Interspeech 2022
18-22 September 2022, Incheon, Korea

Speaker-Specific Utterance Ensemble based Transfer Attack on Speaker Identification

Chu-Xiao Zuo, Jia-Yi Leng, Wu-Jun Li

National Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology, Nanjing University, China
{zuochuxiao,lengjiayi}@smail.nju.edu.cn, liwujun@nju.edu.cn

Abstract

While speaker identification (SI) systems based on deep neural network (DNN) have been widely applied in security-related practical tasks, more and more attention has been attracted to the robustness of SI systems against potential malicious threats. Existing works have shown that white-box attacks can greatly threaten the current SI systems, but white-box attacks require complete knowledge of the target model, which is almost impractical in many applications. As far as we know, only a few works have studied the more practical black-box attacks, while these attacks are mostly ported from computer vision task and lack the adaptability to speech data. In this work, we propose a novel black-box attack, called speaker-specific utterance ensemble based transfer attack (SUETA). SUETA utilizes the unique characteristic of speech data that different utterances of one specific speaker share the same voiceprint to attack on SI systems. To the best of our knowledge, SUETA is the first black-box attack on SI systems that utilizes the unique characteristic of speech data. Experimental results on three representative SI models show that SUETA can achieve better transfer success rate (TSR) than speaker-unrelated baselines. Furthermore, SUETA can even improve the attack success rate (ASR) of white-box attacks on local substitute model, which is the first step to perform the transfer based black-box attack.

Index Terms: speaker identification, adversarial attack, black-box attack, transfer attack

1. Introduction

Speaker identification (SI) systems [1] are widely used to recognize a speaker's identity from utterances by performing a multi-class classification task. It is one of the most popular speaker recognition technologies and has been adopted in many real-world applications, including biometric authentication[1][2], online payment[3] and smartphone personalized service [2]. Due to the promising performance of deep neural network (DNN), most of the existing SI systems [3, 4, 5] are based on DNN models. However, previous works [6, 7] have shown that DNNs are vulnerable to adversarial attacks, raising concerns about the safety and security of SI systems and highlighting the significance of research on adversarial robustness.

Previous adversarial attacks on SI systems [8, 9] have shown the effectiveness of migrating white-box adversarial attack methods from image classification problems [6, 10, 11, 12]. The migrated attacks are also effective in speaker verification problems [13, 14, 15, 16]. Later, more acoustic-designed methods are proposed. An auxiliary acoustic model is constructed in [17] to directly output the adversarial perturbation. A frequency masking strategy is proposed in [18] to prevent adding perturbation on the wave or acoustic features. However, these effective attacks [8, 9, 17, 18] are studied under the white-box scenario and are not applicable in real-world applications since complete model information is usually not accessible. Defending and attacking SI models in the more practical black-box scenario, where the model information is inaccessible, are more challenging and have not been adequately studied.

Black-box attacks can be divided into query-based attacks and transfer-based attacks, according to whether the attacker interacts with victim systems. Query-based attacks [19] constantly query victim systems and modify the attack strategy according to the feedback. This type of attack collects information during the querying, but attackers that make high-frequency queries could be discovered by abnormal query detection methods [20]. On the other side, transfer-based attacks do not require interaction with victim systems. Local surrogate models are leveraged to approximate the speaker embedding distribution of victim systems. However, due to the lack of interaction, the performance of transfer-based attacks is usually more limited than query-based ones. In most existing works, trials on transfer-based attacks for SI systems are also limited to the direct migration of white-box attacks.

In order to show the threat of transfer-based attacks on SI systems without interaction, we study how to generate adversarial examples with strong migration under SI systems. Prior attacks on image classification tasks have tried multiple ways to improve transferability. Model ensemble (ME) methods [21, 22] can effectively alleviate overfitting to local surrogate models. Dong et al. [22] have found that optimizing with momentum can effectively improve the transferability of iterative attacks. Data augmentation methods [23, 24] also improve transferability by increasing input diversity. However, these methods are speaker-unrelated for voice data, and no transferability-boosting method has been designed for transfer-based black-box attack on SI systems as far as we know.

In this paper, we propose a novel black-box attack method, called speaker-specific utterance ensemble based transfer attack (SUETA), to attack on SI systems. The main contributions are listed as follows:

- To the best of our knowledge, SUETA is the first transferability-boosting method for black-box attack on SI systems.
- We propose a speaker-specific utterance ensemble loss function, which improves the transferability of adversarial utterances compared with speaker-unrelated baselines.
- Combining with momentum optimization and ME techniques [22], we propose an improved version of SUETA

[1] https://www.tdbank.com/bank/tdvoiceprint.html
[2] http://en.ccb.com/en/home/ indexv3.html
[3] https://render.alipay.com/p/s/download?form=chinese

10.21437/Interspeech.2022-10139

and reach a higher black-box attack success rate (ASR). The results indicate that the non-interactive transfer attack can also threaten the robustness of the SI system.

- SUETA improves the ASR of white-box attacks on the local substitute model under both L_∞ and L_2 restriction. It shows that we can improve the transferability without weakening the attack performance on the local model.

2. Background

2.1. Speaker Identification Models

The task of speaker identification is to recognize the speaker of audio enrolled in a speaker database. There are mainly two steps in a typical SI system. The first step is to extract speaker embedding by SI models, usually obtained by the pooling strategy [3, 4, 5, 25, 26, 27]. X-vector [3] uses a time-delayed neural network (TDNN) to extract temporal frame-level features. D-TDNN [4] aggregates multi-stage information by adopting dense connectivity. ECAPA-TDNN [5] aggregates and propagates features of different hierarchical levels and uses channel-dependent frame attention to extract speaker embedding.

The second step is to evaluate the embedding similarity between input and enrollment audios. Under the close-set SI scenario [8], the SI system will return the speaker with the highest similarity without rejection. Under the open-set SI scenario, a similarity threshold is set to reject inputs whose highest similarity is below the threshold. We focus on close-set adversarial SI attacks in this paper.

2.2. Adversarial Examples

In image classification, the adversarial example [6] is defined as the input that an attacker maliciously perturbs to cause misclassifications. Formally speaking, with a given classifier f and a perturbation budget ϵ under a certain distance metric$\|\cdot\|$, the adversarial examples can be defined as

$$(\boldsymbol{x}+\boldsymbol{\delta}, y) \in \{f(\boldsymbol{x})=y, f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y, \|\boldsymbol{\delta}\| \leq \epsilon\}, \quad (1)$$

where adversarial perturbation $\boldsymbol{\delta}$ is added to a natural image $\boldsymbol{x}$ with ground-truth label y.

2.2.1. Classical Attack Methods

Existing works have proposed various attack methods to generate adversarial examples. Fast gradient sign method (FGSM) [6] performs a one-step movement from the original sample $\boldsymbol{x}$ along the gradient direction that maximizes the classification loss L. Projected gradient descent (PGD) [11] is an iterative version of FGSM that projects the perturbation onto ϵ-sphere in each iteration. Carlini and Wagner attack (C&W) [12] generates an adversarial example by searching for the minimal perturbation that changes the prediction of f. The C&W attack proposes a set of substitute loss functions for the optimizing process, and Madry et al. [11] point out that choosing the logit loss as

$$L(\boldsymbol{x}, y)=\max_{i \neq y}(h(\boldsymbol{x})_i - h(\boldsymbol{x})_y), \quad (2)$$

where h is the pre-softmax layer of classifier f, C&W attack can be optimized effectively in a PGD form.

2.2.2. Transferability-boosting Attack Methods

Momentum method [28] can accelerate the gradient optimizing process and prevent falling into local extreme points during iterations. *Momentum based attack* [22] finds that adopting gradient move of PGD with momentum can prevent overfitting to the poor decision boundary and thus effectively improve the migration ability of transfer attack.

Model ensemble attack methods [29, 30, 31] are proposed to combine the decisions of multiple models, which can improve the overall performance and robustness. A more potent transfer attack can be generated by ensemble output logits of multiple local substitute models [21, 22]. The ensembled logit is the average of the logits of each model.

3. Speaker-Specific Utterance Ensemble based Transfer Attack

In this section, we introduce SUETA, a speaker-specific adversarial attack method that boosts black-box transferability.

3.1. Threat Model

The threat model specifies the security conditions and defines the situation considered by the attacker when designing the attack method. To define our attack formally, we describe the threat model according to the standards provided by [32].

Adversary Goals. For SI systems, an attacker aims to craft a perturbation $\boldsymbol{\delta}$ for utterance wave $\boldsymbol{u}$ of speaker k, such that the SI system $D(\boldsymbol{u})=\arg\max S(\boldsymbol{u})$ makes a wrong decision on perturbed wave $\boldsymbol{u}+\boldsymbol{\delta}$, where SI model $S(\boldsymbol{u})=(s_1, ..., s_K)$ outputs the similarity vector with respect to the total K enrolled speakers. For the non-targeted attack, the input should be identified as another wrong speaker as $D(\boldsymbol{u}+\boldsymbol{\delta}) \neq k$; for the targeted attack, the input should be identified as the specified target speaker y_{target} as $D(\boldsymbol{u}+\boldsymbol{\delta})=y_{target}$.

Adversary Knowledge. In the common settings of transfer-based attacks, the attacker cannot interact with the target model and does not know model information, only being able to train local surrogate models on the same training set.

Adversarial Capabilities. To conduct meaningful attacks, the attacker needs to limit the perturbation budget. Otherwise, the disturbed samples will be easily distinguished, which is outside the scope of adversarial examples. In order to measure the transferability reasonably, we perform attacks under fixed L_2 and L_∞ constraints determined by pre-experiments.

3.2. Methodology

A transfer-based attack first generates adversarial examples on the local surrogate models and then migrates to the target model. The local attack method greatly affects the transferability. Inspired by data augmentation and model ensemble methods, we integrate different utterances of the same speaker in the local attack process. The ensemble loss function helps extract the same voiceprint information through gradient optimization, alleviating overfitting to local decision boundary that contains extra speech information of SI models.

3.2.1. Local Attack with Momentum-PGD

We conduct local white-box attacks based on the momentum-PGD (M-PGD) optimization procedure described in the following contents. The vanilla PGD is the most commonly used white-box attack framework. The momentum method accumulates a velocity vector $\boldsymbol{g}$ along the gradient direction and helps skip poor local extreme points. Existing works [22] find that it significantly boosts black-box transferability, although it damages local attack performance slightly. Hence, we adopt both

PGD and M-PGD as the local attacks to make a comparison in the later experiments. Under the L_∞ constraint, the iteration of M-PGD at step t is

$$\begin{aligned} \boldsymbol{g}_t &= \beta \cdot \boldsymbol{g}_{t-1} + \nabla L(\boldsymbol{u} + \boldsymbol{\delta}_{t-1}), \\ \boldsymbol{\delta}_t &= \prod_{B(\mathbf{0},\epsilon)} (\boldsymbol{\delta}_{t-1} + \eta \cdot \frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|_2})), \end{aligned} \tag{3}$$

where β is a hyper-parameter controlling momentum influence, and $\prod$ is the projection operator onto the ϵ-ball $B(\mathbf{0}, \epsilon)$ centered at the origin. When $\beta = 0$, the method degenerates to vanilla PGD. Similarly, for L_2 constraint, the update step changes to

$$\begin{aligned} \boldsymbol{g}_t &= \beta \cdot \boldsymbol{g}_{t-1} + \nabla L(\boldsymbol{u} + \boldsymbol{\delta}_{t-1}), \\ \boldsymbol{\delta}_t &= \prod_{B(\mathbf{0},\epsilon)} (\boldsymbol{\delta}_{t-1} + \eta \cdot \text{sign}(\boldsymbol{g}_t)). \end{aligned} \tag{4}$$

3.2.2. Speaker-Specific Utterance Ensemble Loss

An utterance contains both voiceprint information and speech information. The black-box attacker aims to generate perturbation that can confuse both the local and target model only on the voiceprint information. Hence, we include different utterances of the same speaker in our proposed speaker-specific utterance ensemble loss function to enhance the focus on voiceprint. We denote the i-th utterance of speaker k as $\boldsymbol{u}^{ki}$ and first compute the ensembled similarity vector $\tilde{\boldsymbol{s}}^{ki}$ as

$$\tilde{\boldsymbol{s}}^{ki} = \alpha S(\boldsymbol{u}^{ki}) + (1-\alpha)\frac{1}{N_k - 1}\sum_{j \neq i} S(\boldsymbol{u}^{kj}), \tag{5}$$

where N_k is the total number of utterances of the k-th speaker, and α is used to balance between the current utterance and the others. The loss on $\boldsymbol{u}^{ki}$ is

$$\begin{aligned} L(\boldsymbol{u}^{ki}) &= -\max(\tilde{\boldsymbol{s}}_k^{ki} - \max_{j \neq k}\{\tilde{\boldsymbol{s}}_j^{ki}\} + c,\ 0) \quad \text{(Non-targeted)}, \\ L(\boldsymbol{u}^{ki}, y_{target}) &= -\max(\tilde{\boldsymbol{s}}_k^{ki} - \tilde{\boldsymbol{s}}_{y_{target}}^{ki} + c,\ 0) \quad \text{(Targeted)}, \end{aligned} \tag{6}$$

where c is a confidence parameter. The ensemble loss function can effectively reduce the overfitting of the current utterance on the decision boundary of the local model to produce adversarial examples with stronger transferability.

3.2.3. Optimize with Memomry Buffer

SUETA is performed under the iterative optimization framework based on PGD with the ensemble loss. We use a memory buffer $\mathbf{S}^k \in \mathbb{R}^{N_k \times K}$ to store the similarity matrix for each enrolled speaker. $\mathbf{S}_i^k$ is the stored similarity vector of the i-th utterance and is updated after each perturbation update. The buffer helps reduce redundant computing overhead according to Eq. (5). The whole process is summarized in Algorithm 1.

3.3. Combining SUETA with Model Ensemble Method

SEUTA only requires a single local surrogate model as the source model. Therefore, it can also enroll more local models to prevent overfitting a single model. We combine SUETA with the model ensemble method to further improve black-box transferability. With local substitute models $\{S_1, ..., S_M\}$, we propose SUETA-ME by replacing $S(\boldsymbol{u}^{ki})$ in Eq. (5) by:

$$\tilde{S}(\boldsymbol{u}^{ki}) = \frac{1}{M}\sum_{j=1}^{M} S_j(\boldsymbol{u}^{ki}). \tag{7}$$

Algorithm 1 SUETA

Input:
Local surrogate model S, utterance batch $\boldsymbol{U}^k = \{\boldsymbol{u}^{k1}, ..., \boldsymbol{u}^{kN_k}\}$ of speaker k;
perturbation budget ϵ, number of iterations T, step size η, momentum factor β, ensemble factor α.
Output:
Adversarial perturbation vectors $\{\boldsymbol{\delta}_T^1, ..., \boldsymbol{\delta}_T^{N_k}\}$ of the batch;

Initialize $\mathbf{S}^k = S(\boldsymbol{U}^k)$, $\{\boldsymbol{\delta}_0^1, ..., \boldsymbol{\delta}_0^{N_k}\} = \{\mathbf{0}, ..., \mathbf{0}\}$, $\{\boldsymbol{g}_0^1, ..., \boldsymbol{g}_0^{N_k}\} = \{\mathbf{0}, ..., \mathbf{0}\}$;
for $t = 1$ to T **do**
  for $i = 0$ to N_k **do**
    $\tilde{\boldsymbol{s}}^{ki} = \alpha S(\boldsymbol{u}^{ki} + \boldsymbol{\delta}_{t-1}^i) + (1-\alpha)\frac{1}{N_k - 1}\sum_{j \neq i} \mathbf{S}_j^k$;
    Calculate loss L with $\tilde{\boldsymbol{s}}^{ki}$, $\tilde{\boldsymbol{s}}^{kj} = \mathbf{S}_j^k$ by Eq. (6);
    $\boldsymbol{g}_t^i = \beta \cdot \boldsymbol{g}_{t-1}^i + \nabla_{\boldsymbol{\delta}_{t-1}^i} L$;
    Update $\boldsymbol{\delta}_t^i$ with $\boldsymbol{\delta}_{t-1}^i$, $\boldsymbol{g}_t^i$ according to the second line in Eq. (3) or Eq. (4);
    $\mathbf{S}_i^k = S(\boldsymbol{u}^{ki} + \boldsymbol{\delta}_t^i)$;
  end for
end for
return $\{\boldsymbol{\delta}_T^1, ..., \boldsymbol{\delta}_T^{N_k}\}$;

4. Experiments

4.1. Setup

Dataset. We use TIMIT [33] which contains 630 speakers and 6300 utterances. We randomly divided the training set, test set, and validation set according to the ratio of 8:1:1.

Implementation of SI Models. As introduced in Section 2.1, we choose X-Vector [3], Dense-TDNN [4], and ECAPA-TDNN [5] as enrolled model backbones. The wave data is first preprocessed to obtain 40-dimensional MFCC acoustic features over a window of 25ms with an overlap of 10 ms. We then train the models using the Softmax loss function. The Top1 accuracy without adversarial attacks on the test set is 99.37 %, 97.78%, and 99.05%, respectively for X-Vector, Dense-TDNN and ECAPA-TDNN.

Evaluation Metrics and Baselines. Evaluation is based on transfer success rate (TSR) and attacks success rate (ASR) given a fixed perturbation budget. TSR is defined as the proportion of samples misjudged by both the local and target model in local countermeasure samples. TSR directly reflects the transferability of adversarial examples. ASR is the proportion of successful attack samples in the total number of attack samples. We compare SUETA and SUETA-ME with baselines including PGD attack and ME attack [22]. Prefix 'M-' represents the version of the method using momentum optimization.

Attack We set $\epsilon = 0.001$ under L_∞ constraint and $\epsilon = 0.2$ under L_2 constraint. The step size $\eta = \epsilon/4$. The number of attack iterations $T = 10$ for each utterance. For targeted attack, we choose random target y_{target} for each utterance. We set $\alpha = 0.3$ for SUETA loss and $\beta = 1$ for momentum optimization.

4.2. Transfer from Single Model

We first conduct attacks on a single model and compare the TSR of SUETA with PGD. The experiment directly measures the impact of the SUETA loss compared to vanilla PGD loss without introducing additional models. The results of non-targeted attacks are shown in Table 1. Without momentum

Table 1: *Transfer success rate (%) of non-targeted transfer attack from left source models to top target models. * indicates the local white-box attacks.*

Source Model		Attack Method	X-Vector	Dense	ECAPA
X-Vector	L_∞	PGD	*	27.30	19.68
		M-PGD	*	**46.83**	**42.54**
		SUETA	*	41.43	34.29
		M-SUETA	*	46.67	41.27
	L_2	PGD	*	21.59	16.19
		M-PGD	*	48.89	45.40
		SUETA	*	49.84	44.76
		M-SUETA	*	**53.33**	**48.89**
Dense-TDNN	L_∞	PGD	7.30	*	13.02
		M-PGD	**25.08**	*	**34.76**
		SUETA	20.16	*	27.14
		M-SUETA	24.92	*	32.70
	L_2	PGD	6.03	*	9.52
		M-PGD	26.19	*	36.51
		SUETA	30.00	*	37.78
		M-SUETA	**34.76**	*	**40.00**
ECAPA-TDNN	L_∞	PGD	21.43	33.02	*
		M-PGD	36.83	50.95	*
		SUETA	36.19	46.03	*
		M-SUETA	**43.18**	**51.59**	*
	L_2	PGD	17.46	27.94	*
		M-PGD	45.71	52.22	*
		SUETA	51.43	55.71	*
		M-SUETA	**56.03**	**58.57**	*

Table 2: *Attack success rate (%) of non-targeted transfer attacks.*

	L_2 Attack			L_∞ Attack		
Attack Method	X-Vector	Dense	ECAPA	X-Vector	Dense	ECAPA
PGD	17.46	27.94	16.19	21.43	33.02	19.68
M-PGD	45.71	52.22	45.40	36.83	50.95	42.54
ME	39.68	51.90	38.73	31.75	48.57	34.29
M-ME	55.08	63.02	57.30	45.56	56.19	49.52
SUETA	51.43	55.71	44.76	36.19	46.03	34.29
M-SUETA	56.03	58.57	48.89	43.18	51.59	41.27
SUETA-ME	61.11	64.13	60.79	44.29	52.22	45.40
M-SUETA-ME	**63.02**	**66.51**	**63.02**	**49.68**	**56.67**	**52.22**

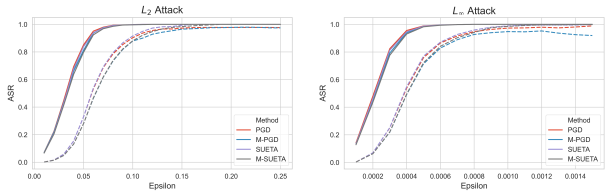

Figure 1: *Attack on X-Vector. The solid line denotes non-targeted attack and the dotted line denotes the targeted attack.*

techniques, the TSR of SUETA is significantly higher than that of PGD. M-SUETA also achieves the best TSR in most cases. It shows that the speaker-specific loss of SUETA significantly improves black-box transferability compared to the vanilla speaker-unrelated loss.

4.3. Transfer from Multiple Models

Next, we perform attacks with multiple source models. To compare with single-model methods, SUETA and PGD, we show the best transfer ASR from one of the source models. The results of non-targeted attacks are shown in Table 2. M-SUETA-ME achieves the best ASR almost in all the cases. As shown in Table 3, our methods still successfully improve the ASR in the much more challenging targeted attack situation.

Table 3: *Attack success rate (%) of targeted transfer attacks.*

	L_2 Attack			L_∞ Attack		
Attack Method	TDNN	Dense	ECAPA	TDNN	Dense	ECAPA
PGD	3.47	2.91	1.24	0.37	3.57	1.48
M-PGD	9.55	5.25	4.09	7.33	3.40	2.88
ME	6.83	5.08	3.81	6.03	4.76	3.17
M-ME	10.95	6.03	6.51	7.62	4.60	4.44
SUETA	13.42	6.96	6.31	6.80	5.54	3.61
M-SUETA	14.38	7.37	7.08	9.99	5.26	4.66
SUETA-ME	13.65	6.19	9.37	9.21	5.08	5.24
M-SUETA-ME	**14.44**	**7.62**	**10.00**	**10.32**	**5.71**	**5.40**

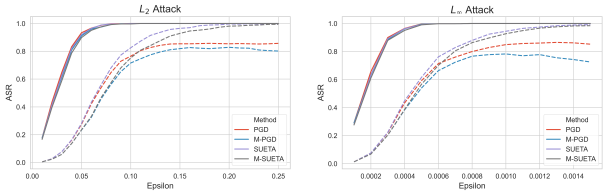

Figure 2: *Attack on Dense-TDNN. The solid line denotes non-targeted attack and the dotted line denotes the targeted attack.*

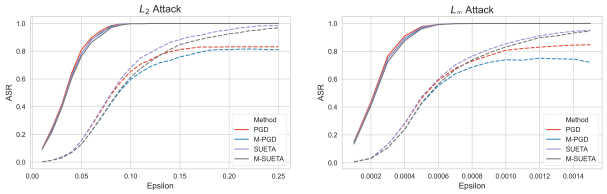

Figure 3: *Attack on ECAPA-TDNN. The solid line denotes non-targeted attack and the dotted line denotes the targeted attack.*

4.4. Local White-box Attack

We conduct an experiment on ASR with varying perturbation budgets to observe the effect of SUETA on local white-box attacks. The results for X-Vector, Dense-TDNN and ECAPA-TDNN are shown in Figure 1, Figure 2 and Figure 3 respectively. SUETA significantly improves the ASR of targeted attacks compared with PGD while keeping a similar performance with PGD on non-targeted attacks. The maximum targeted ASR on Dense-TDNN and ECAPA-TDNN models is below 90% for PGD attacks, but SEUTA can closely reach 100%. When transferring from a single model, we have $ASR^{transfer} = ASR^{local} \times TSR$. SUETA boosts transferability by improving both TSR and local ASR. On the contrary, M-PGD gains transferability at the cost of decreasing local ASR.

5. Conclusions

In this paper, we propose SUETA, a transfer-based black-box attack method that improves the transferability of adversarial examples. SUETA is the first work to utilize the characteristic of speech data that different utterances of one speaker share the same voiceprint. Our experiments on the state-of-the-art SI models show a significant improvement compared with speaker-unrelated baselines. Moreover, SUETA also significantly promotes the targeted ASR of local white-box attacks.

6. Acknowledgements

This work is supported by National Key R&D Program of China (No. 2020YFA0713901), NSFC Project (No.61921006), and NSFC Project (No.62192783).

7. References

[1] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Syst. Appl.*, vol. 171, p. 114591, 2021.

[2] H. Ren, Y. Song, S. Yang, and F. Situ, "Secure smart home: A voiceprint and internet based authentication system for remote accessing," in *International Conference on Computer Science & Education (ICCSE)*, 2016, pp. 247–251.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] Y.-Q. Yu and W.-J. Li, "Densely connected time delay neural network for speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 921–925.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 3830–3834.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[8] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *CoRR*, vol. abs/1711.03280, 2017.

[9] 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, pp. 1187–1200, 2021.

[10] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *International Conference on Learning Representations (ICLR)*, 2017.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[12] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (S&P)*, 2017, pp. 39–57.

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

[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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6579–6583.

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

[16] J. Villalba, Y. Zhang, and N. Dehak, "x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 4233–4237.

[17] J. Li, X. Zhang, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Learning to fool the speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2937–2941.

[18] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 4228–4232.

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

[20] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao, "Blacklight: Defending black-box adversarial attacks on deep neural networks," *CoRR*, vol. abs/2006.14042, 2020.

[21] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *International Conference on Learning Representations (ICLR)*, 2017.

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

[23] J. Zou, Z. Pan, J. Qiu, X. Liu, T. Rui, and W. Li, "Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting," in *European Conference on Computer Vision (ECCV)*, vol. 12367, 2020, pp. 563–579.

[24] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4312–4321.

[25] Y.-Q. Yu, L. Fan, and W.-J. Li, "Ensemble additive margin softmax for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6046–6050.

[26] Y.-Q. Yu, S. Zheng, H. Suo, Y. Lei, and W.-J. Li, "Cam: Context-aware masking for robust speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6703–6707.

[27] L. Fan, Q.-Y. Jiang, Y.-Q. Yu, and W.-J. Li, "Deep hashing for speaker identification and retrieval," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2908–2912.

[28] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *Ussr Computational Mathematics Mathematical Physics*, vol. 4, pp. 1–17, 1964.

[29] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, pp. 993–1001, 1990.

[30] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 1994, pp. 231–238.

[31] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *International Conference on Machine Learning (ICML)*, vol. 69, 2004.

[32] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *CoRR*, vol. abs/1902.06705, 2019.

[33] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.