INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Black-box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples

Yuekai Zhang[1], *Ziyan Jiang*[1], *Jesús Villalba*[1,2], *Najim Dehak*[1,2]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, USA

{yzhan400,zjiang28,jvillal7,ndehak3}@jhu.edu

Abstract

Spoofing countermeasure systems protect Automatic Speaker Verification (ASV) systems from spoofing attacks such as replay, synthesis, and conversion. However, research has shown spoofing countermeasures are vulnerable to adversarial attacks. Previous literature mainly uses adversarial attacks on spoofing countermeasures under a white-box scenario, where attackers could access all the information of the victim networks. Black-box attacks would be a more serious threat than white-box attacks.

In this paper, our objective is to black-box attack spoofing countermeasures using adversarial examples with high transferability. We used MI-FGSM to improve the transferability of adversarial examples. We propose an iterative ensemble method (IEM) to further improve the transferability. Comparing with previous ensemble-based attacks, our proposed IEM method, combined with MI-FGSM, could effectively generate adversarial examples with higher transferability. In our experiments, we evaluated the attacks on four black-box networks. For each black-box model, we used the other three as a white-box ensemble to generate the adversarial examples. The proposed IEM with MI-FGSM improved attack success rate by 4-30% relative (depending on black-box model) w.r.t. the baseline logit ensemble. Therefore, we conclude that spoofing countermeasure models are also vulnerable to black-box attacks.

Index Terms: spoofing countermeasures, adversarial examples, transferability, black-box attack

1. Introduction

Automatic Speaker Verification (ASV) technology has advanced significantly in recent years [1, 2, 3, 4, 5]. However, technologies like text-to-speech synthesis (TTS), voice conversion (VC) can be used to generate spoof audios and attack ASV systems. Thus, these technologies have become a threat to ASV systems [6, 7]. To protect ASV systems, researchers use spoofing countermeasures [8, 9, 10, 11]. In this case, a good spoofing countermeasure is the key to protect ASV systems. The community also held ASVspoof challenges to support the spoofing countermeasures research [12, 13, 14].

Adversarial attacks have become a threat to all kinds of machine learning models [15, 16, 17]. In [18], a method to generate minimal adversarial perturbations to attack speech recognition (ASR) systems is proposed. In the ASV domain, authors successfully attacked i-vector speaker verification model using adversarial examples [19]. For spoofing countermeasures, in [20], anti-spoofing models also showed to be vulnerable to adversarial attacks. Another line of research focuses on defending adversarial attacks or improving model robustness [21, 22, 23], which is beyond our scope. This paper focuses on black-box attacks to anti-spoofing models.

Adversarial attacks can be divided into three categories: white-box attacks, grey-box attacks, black-box attacks. In this paper, we refer to white-box attacks as those where the attacker can access all of the information of the victim model. For grey-box attacks, the attacker can still query the victim model multiple times, which could be used to get a substitute model of the victim model. For black-box attacks, only very little information such as the type of features used could be accessed by the attacker. If the black-box attacks could be generated successfully, it would be a serious threat to anti-spoofing systems and ASV systems.

For an attacker, transferability is a desirable property of adversarial examples. A transferable adversarial example means that, even though generated from a specific white-box victim model, it can successfully attack other models, which may be very different from the white-box model. Several works in the image domain try to improve adversarial examples' transferability to increase the threat of adversarial attacks under black-box scenarios. In [24], an input diversity method is used to improve the adversarial examples transferability. Random transformations are applied to the input images at each iteration to create diverse input patterns, reducing the over-fitting to the specific victim model. In [25], multiple victim models are ensembled to generate adversarial examples with high transferability. In audio domain, the transferability of adversarial examples for sound event classification was studied in [26]. They showed that the transferability of adversarial examples is not affected by normalization techniques or knowledge distillation. In this paper, we investigate the vulnerability of spoofing countermeasure models under transferable adversarial examples attacks.

The main contributions of this work are:

- Investigate the robustness of anti-spoofing systems under powerful black-box attacks.
- Show that MI-FGSM attack and ensemble-based attacks improve the transferability of audio adversarial examples. To the best of our knowledge, this is the first work aiming to generate black-box attacks against spoofing countermeasures by improving the adversarial examples' transferability.
- Propose a novel iterative ensemble method (IEM) which improved the transferability of adversarial examples. The algorithm could also be used in other domains such as image and text adversarial examples.

The rest of the paper is organized as follows. Section 2 introduces several anti-spoofing models, which are the victim models in this paper. Section 3 describes the adversarial attacks

This project was supported by DARPA Award HR001119S0026-GARD-FP-052

http://dx.doi.org/10.21437/Interspeech.2020-2834

and the proposed iterative ensemble method. Section 4 provides details of our experiments. In Section 5, we report the experiments' results. Finally, we conclude this paper and future work.

2. ASV spoofing countermeasure models

The role of anti-spoofing or spoofing countermeasure models is to detect audios that intend to impersonate a victim user. These audios may consist of replays from the victim user or be generated by TTS or VC. To better investigate the transferability of adversarial examples between models, we used four victim models: light convolutional neural network (LCNN) [27], attentive filtering network (AFNet) [10], 2D residual convolutional network (ResNet) and squeeze-excitation ResNet (SEResNet) [11]. Following, we introduce these models.

2.1. LCNN model

In [27], LCNN models were used for ASVspoof 2019 challenge [14] and got the second-best performance in logical access (LA) scenario. They modified the enhanced LCNN architecture, previously used for replay attack detection [28], from a feature extractor to a direct final score estimator from spectral features. The idea is to use Max-Feature-Map (MFM) activation function with neural networks. MFM could help to choose key features for task solving. They also use angular margin softmax loss (A-softmax) [29] as their training objective. The A-softmax loss can be described as

$$L = -\frac{1}{N}\Sigma_i \log \frac{e^{||x_i|| \cos(m\theta_{i,y_i})}}{e^{||x_i|| \cos(m\theta_{i,y_i})} + \sum\limits_{j \neq y_i} e^{||x_i|| \cos(\theta_{i,j})}} \quad (1)$$

where N is the number of training samples; $\theta_{i,j}$ is the angle between training sample x_i and the corresponding column j of the fully connected layer weights, y_i is the index of the label for x_i; and m is a hyper-parameter which modifies the angular margin among classes.

2.2. Attentive filtering network

In [10], they proposed attentive filtering networks for the replay attack detection task in ASVspoof 2017. The attentive filtering network is composed of an attention filter that enhances the input features in both time and frequency domains with attention mechanism and a ResNet [30] back-end classifier. Attentive filtering (AF) is described as

$$S^* = \text{sigmoid}(U(S)) \circ S + S \quad (2)$$

where $S \in \mathbb{R}^{F \times T}$ is the input feature map, F and T are the frequency and time axis, $\circ$ is element wise multiplication operator, and U is U-Net neural network [31]. For the ResNet classifier, they used a Dilated ResNet, which replaces all fully connected layers with convolution layers.

2.3. Squeeze-Excitation ResNet model

In [11], Squeeze-Excitation Network (SENet) [32] and ResNet with a statistical pooling layer are first introduced to address anti-spoofing. SENet has achieved impressive results for image classification. It adds an extra component to residual blocks that adaptively scales the hidden representations by explicitly modeling interdependencies between channels. The SENet50 variant in [11] is used in this paper.

3. Audio Adversarial Examples

Audio adversarial examples are signals with small perturbations that are imperceptible by humans but change the output of a machine learning system. To generate an adversarial example, we fix the parameters of the victim model and utilize back-propagation to compute the gradient of the attack objective function given the input signal. Using this gradient, we can optimize the adversarial perturbation by gradient descent methods. In the next subsections, we will introduce the attack algorithms that we used, as well as the proposed iterative ensemble method (IEM).

3.1. Attack algorithms

3.1.1. Fast Gradient Sign Method

The Fast Gradient Sign Method (FGSM) [16] is a fast method, rather than optimal, which optimizes the adversarial examples by a single step along the direction of the gradient,

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} L(\mathbf{x}, y^{\text{true}}; \theta)) \quad (3)$$

where $\nabla_{\mathbf{x}} L(\mathbf{x}, y)$ is the derivative of the loss function with respect to the clean input example $\mathbf{x}$. FGSM restricts the adversarial examples in the L_∞ norm bound $||\mathbf{x}' - \mathbf{x}||_\infty \leq \epsilon$.

3.1.2. Iterative Fast Gradient Sign Method

Iterative Fast Gradient Sign Method (I-FGSM) [17, 33], instead of a single step, takes several smaller optimization steps to obtain the perturbation as

$$\mathbf{x}'_0 = \mathbf{x} \quad (4)$$

$$\mathbf{x}'_{i+1} = \text{Clip}^{\epsilon}_{x}\{\mathbf{x_i}' + \alpha \cdot \text{sign}(\nabla_{x'_i} L(\mathbf{x_i}', y^{\text{true}}; \theta))\} \quad (5)$$

where i represents the iteration times, $0 < \alpha < \epsilon$ is a small step size. $\text{Clip}^{\epsilon}_{x}$ function is used to clip $\mathbf{x}'$ into the ϵ vicinity of $\mathbf{x}$ to satisfy the L_∞ bound. This iterative methods could perform stronger white-box attacks at the cost of worse transferability.

3.1.3. Momentum Iterative Fast Gradient Sign Method

In [34], the I-FGSM method was improved by adding a momentum term during the optimization process. The method is called Momentum Iterative Fast Gradient Sign Method (MI-FGSM), and has shown increased transferability of image adversarial examples [34]. The MI-FGSM algorithm is formulated as

$$\mathbf{g}_{i+1} = \mu \cdot \mathbf{g}_i + \frac{\nabla_x L(\mathbf{x}, y^{\text{true}}; \theta)}{||\nabla_x L(\mathbf{x}, y^{\text{true}}; \theta)||_1} \quad (6)$$

$$\mathbf{x'_{i+1}} = \text{Clip}^{\epsilon}_{\mathbf{x}}\{\mathbf{x}'_i + \alpha \cdot \text{sign}(\mathbf{g}_{i+1})\} \quad (7)$$

where $\mathbf{g}_i$ collects the gradients of the first i iterations with a momentum decay factor μ. From the above equations, we could see MI-FGSM degrades to I-FGSM when μ equals to 0.

3.2. Ensemble multiple models

To improve the transferability of adversarial examples, an ensemble-based method was used in [25] by attacking multiple models simultaneously. In [25], authors argued that the adversarial examples are more likely to transfer to other models if they could fool various models simultaneously. There are different kinds of ensemble methods, such as ensemble in predictions or loss functions. We followed the strategy used in [34], which makes a weighted sum of the logits of multiple models

to get an ensemble model. In [34], they showed that ensemble in logits gives adversarial examples with higher transferability comparing with other ensemble methods. To attack K white-box models simultaneously, we fuse the logit activations as

$$l(x;\theta_1,...,\theta_K) = \Sigma_{k=1}^{K} w_k l_k(x;\theta_k) \quad (8)$$

where $l_k(x;\theta_k)$ represents the logits of k-th white-box model, which is integrated in the ensemble model. w_k is the ensemble weight, where $\Sigma_{k=1}^{K} w_k = 1$. We will denote this as *logit-ensemble* (Logit-E)

Algorithm 1 Iterative Ensemble Method

Input: White-box models $L = \{l_1 \dots l_K\}$, clean input x with corresponding label y, adversarial attack function f (such as FGSM, I-FGSM, MI-FGSM) with parameters $\theta = \{\alpha, \epsilon\}$, perturbation range ϵ, step size α, iteration times T
Output: Adversarial example x'

Initialize perturbation $\delta_0 \leftarrow$ random start in the ϵ-ball
for iteration time $t \leftarrow 1$ to T **do**
  for model $l_i \in L$ **do**
    $\delta_m \leftarrow f_\theta(\delta_{m-1}, y; l_i)$ (where $m = K*(t-1)+i$)
  end for
end for
$\hat{\delta} \leftarrow \delta_{T\times K}$
return $x' = x + \hat{\delta}$

3.3. Proposed iterative ensemble method

We proposed a method called iterative ensemble method to improve the transferability of adversarial examples, which outperformed all the previous ensemble attack methods. Our method's basic idea is to find adversarial examples that could fool all ensembled white-box models simultaneously. We adopt an iterative strategy to maximize the attack success rates on all used ensemble models and achieved more transferable adversarial examples on black-box models. The iterative ensemble method is summarized in Algorithm 1.

The algorithm is motivated by [35], which used an iterative algorithm to find an input-independent perturbation called universal adversarial perturbation. We expect to find adversarial examples with high transferability, which could be regarded as almost network parameters independent adversarial perturbation. Note that in [35], the attacked model is fixed while the authors modify the clean input in every iteration. Meanwhile, we fix the original sample and iterate over the ensembled white-box models.

Figure 1 gives a simple explanation of the proposed method. To attack the victim models, we need to push the clean examples into the space of the adversarial examples, which are in different domains for different victim models. General attacks like I-FGSM might over-fit to the specific network parameters θ. Using ensemble-based methods, we could improve the transferability of adversarial examples by fooling multiple models simultaneously. We can make full use of the existing white-box models by using the proposed method.

4. Experimental setup

4.1. Datasets

Following the setting in [20], the paper uses the LA part of ASVspoof 2019 dataset. We used the log-power magnitude spectrum as input for all models. Only the first 600 frames of each utterance were used to extract acoustic features.

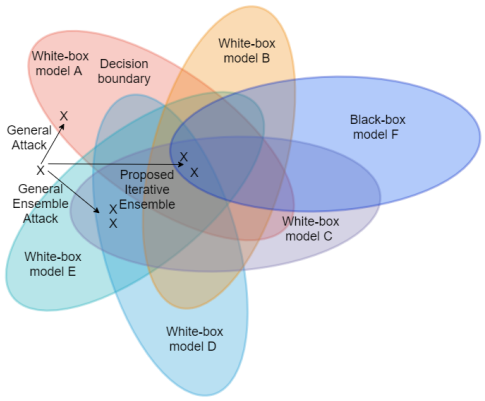

Figure 1: *Illustration of Proposed Iterative Ensemble Method*

We used the LA training set to train our anti-spoofing models. Due to the computation cost of ensemble-based methods, it was unfeasible to use the whole dataset dev and eval datasets. Therefore, We randomly selected 500 spoof audio examples from the dev set to conduct our adversarial attacks. All the selected samples were classified correctly by our victim anti-spoofing models before the attacks. We did not generate adversarial examples from bonafide examples, which would aim to make anti-spoofing models classify bonafide examples into spoof audios. We argue bonafide audios are not bonafide anymore, if we add perturbation on them. Also, we mainly focus on helping spoof audios to bypass the anti-spoofing systems.

4.2. Implementation details of the countermeasure models

Four countermeasure models were used, i.e., LCNN, SENet50, Resnet34 and AFNet. LCNN was configured as in [27]; SENet50 and Resnet34 followed the setting in [11]; and AFNet used the setup in [10].

4.3. XAB listening test

To know the range of imperceptible perturbation, we conducted a XAB listening test under different levels of perturbation size ϵ. Since in our experiments, the perturbation was added into audio features, the adversarial examples were reconstructed from the perturbed log-power magnitude spectrum (LPS) and the phase spectrum of the corresponding original clean audios. Note that an additive perturbation in LPS domain is not additive in the reconstructed waveform. Thus SNR, may not be a good metric to measure the perturbation. We chose 10 listeners to identify the unknown audio sample X, randomly selected from either adversarial audio examples or clean audios. Listeners were required to detect whether the example was selected from reference source A or reference source B and do the test multiple times. The result shows that audios with $\epsilon \leq 10$ were imperceptible. Adversarial audios can be accessed here.

5. Results and Discussions

The results of attacks are evaluated according to attack success rate. We regard an adversarial example as a successful attack if it can make the victim models fail, i.e., to classify it as a bona-fide example. We first used MI-FGSM attack to improve the black-box attack success rate. Then the logit ensem-

https://xmhzz2018.github.io/adv-transfer-demo

Table 1: *Ensemble Attack Success Rate (%). The term* ens *indicates the basic method of ensemble models at logits level. The term* iter-ens *indicates our novel iterative ensemble method.* $*$ *indicates the black-box attacks results. "-" indicates that the model of the row is not used when generating the attacks.*

Victim	Attack Method	LCNN	SENet50	ResNet34	AFNet
-LCNN	FGSM-ens	11*	100	100	100
	FGSM-iter-ens	4*	98	100	100
	I-FGSM-ens	6*	100	100	100
	I-FGSM-iter-ens	10*	100	100	100
	MI-FGSM-ens	17*	100	100	100
	MI-FGSM-iter-ens	**24***	100	100	100
-SENet50	FGSM-ens	100	37*	100	100
	FGSM-iter-ens	100	19*	100	100
	I-FGSM-ens	100	38*	100	100
	I-FGSM-iter-ens	100	47*	100	100
	MI-FGSM-ens	100	52*	100	100
	MI-FGSM-iter-ens	100	**67***	100	100
-ResNet34	FGSM-ens	100	98	44*	100
	FGSM-iter-ens	99	93	25*	100
	I-FGSM-ens	100	100	37*	100
	I-FGSM-iter-ens	100	100	58*	100
	MI-FGSM-ens	100	100	68*	100
	MI-FGSM-iter-ens	100	100	**84***	100
-AFNet	FGSM-ens	100	100	100	47*
	FGSM-iter-ens	99	99	100	30*
	I-FGSM-ens	100	100	100	36*
	I-FGSM-iter-ens	100	100	100	38*
	MI-FGSM-ens	100	100	100	59*
	MI-FGSM-iter-ens	100	100	100	**61***

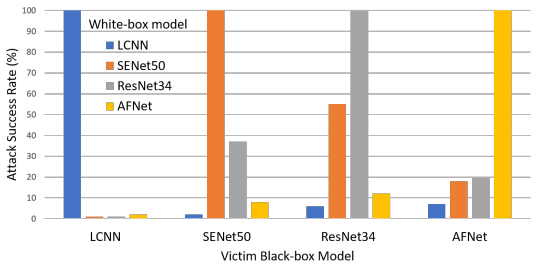

Figure 2: *I-FGSM Attack Success Rate (%)*

ble based method as well as proposed iterative ensemble method improved the transferability of adversarial examples further. Finally, we show that a powerful black-box attack can be generated by combining the above technologies with larger adversarial perturbation.

5.1. Improving transferability with MI-FGSM

Figure 2 and Figure 3 show the results of adversarial attacks under I-FGSM and MI-FGSM, respectively. Both attacks were conducted under $\epsilon = 5$, which is imperceptible according to the XAB listening test result, with 10 iterations and step size $\alpha = 1$. We can see that, when we generate white-box attacks–which means the model used for generating the adversarial examples and the victim model are the same– these two methods have a 100% success rate. However, adversarial examples by I-FGSM were hard to transfer. By using MI-FGSM, the transferability between all models improved.

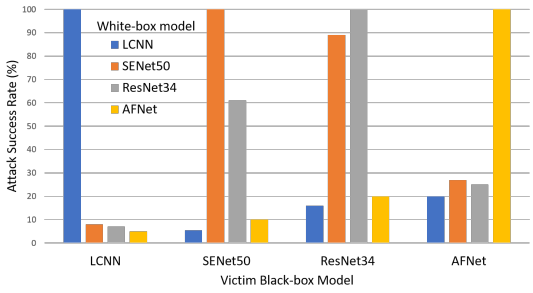

Figure 3: *MI-FGSM Attack Success Rate (%)*

Table 2: *Transfer Attack Success Rate (%) with Different Size of Perturbation. "-" indicates that the model of the row is not used when generating the attacks.*

Victim	Attack Method	eps 1	eps 2.5	eps 5	eps 10
-LCNN	MI-FGSM-ens	0	0	17	32
	MI-FGSM-iter-ens	0	1	24	**80**
-SENet50	MI-FGSM-ens	0	11	52	60
	MI-FGSM-iter-ens	0	12	67	**79**
-ResNet34	MI-FGSM-ens	0	13	68	95
	MI-FGSM-iter-ens	2	17	84	**100**
-AFNet	MI-FGSM-ens	0	10	59	86
	MI-FGSM-iter-ens	1	11	61	**97**

5.2. Improving transferability with ensemble attacks

We further improved the transferability by ensembling multiple models.Table 1 shows that MI-FGSM, combined with our iterative ensemble method, consistently outperformed other approaches, giving the best black-box attack success rate. The iterative ensemble always improved the performance of the logit ensemble when combined with I-FGSM and MI-FGSM. However, it did not improve with the simple FGSM, indicating that the iterative ensemble needs to be paired with an attack able to reach more optimal solutions than FGSM.

5.3. Improving transferability by increasing perturbation

Table 2 shows the MI-FGSM attack results, combined with the basic and iterative ensemble methods, for different perturbation size ϵ. From Table 2, we could find that the larger adversarial perturbation gives the higher transferability. To generate black-box attacks using transferability effectively, we needed to keep the value of ϵ at least more than 2.5 in the experiments.

6. Conclusions

In this paper, we investigated the transferability of adversarial examples towards attacking spoofing countermeasure systems. We implemented four types of anti-spoofing models, i.e., LCNN, SENet50, ResNet34, AFNet and applied black-box attacks on them. We showed that MI-FGSM could improve the transferability of adversarial examples. We also proposed a novel iterative ensemble method, which can be combined with MI-FGSM to generate adversarial examples with high transferability. For the future work, we would like to adopt our iterative ensemble method into other domains like image recognition, speech recognition and speaker verification to improve the transferability of adversarial examples.

7. References

[1] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, aug 2017, pp. 999–1003.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors : Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP 2018*. Alberta, Canada: IEEE, apr 2018, pp. 5329–5333.

[3] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. Garcia-Perera, D. Povey, P. A. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18," in *Proc. Interspeech 2019*, Graz, Austria, sep 2019.

[4] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, mar 2020.

[5] J. Villalba, D. Garcia-Romero, N. Chen, G. Sell, J. Borgstrom, A. McCree, L. P. Garcia-Perera, S. Kataria, P. S. Nidadavolu, P. A. Torres-Carrasquillo, and N. Dehak, "Advances in Speaker Recognition for Telephone and Audio-Visual Data : the JHU-MIT Submission for NIST SRE19," in *Proceedings of Odyssey 2020-The Speaker and Language Recognition Workshop*, Tokyo, Japan, 2020.

[6] J. Villalba and E. Lleida, "Speaker Verification Performance Degradation against Spoofing and Tampering Attacks," in *Proceedings of Fala 2010*, Vigo, Spain, nov 2010, pp. 131–134.

[7] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," *arXiv preprint arXiv:2004.08849*, 2020.

[8] J. Villalba and E. Lleida, "Preventing Replay Attacks on Speaker Verification Systems," in *Proceedings of the IEEE International Carnahan Conference on Security Technology, ICCST 2011*. Mataro, Spain: IEEE, sep 2011, pp. 284–291.

[9] 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.

[10] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *Proc. ICASSP 2019*. IEEE, may 2019.

[11] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *Proc. Interspeech 2019*, pp. 1013–1017, 2019.

[12] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, and A. Sizov, "ASVspoof 2015 : the First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in *Proc. Interspeech 2015*. Dresden, Germany: ISCA, sep 2015.

[13] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*. Les Sables d'Olonne, France: ISCA, jun 2018, pp. 296–303.

[14] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

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

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *3rd International Conference on Learning Representations, ICLR 2015*, pp. 1–11, 2015.

[17] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *CoRR 2017*, jul 2017.

[18] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, pp. 1–7, 2018.

[19] 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 2020*. Barcelona, Spain: IEEE, may 2020, pp. 6579–6583.

[20] S. Liu, H. Wu, H. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *ASRU 2019*, 2019, pp. 312–319.

[21] H. Wu, S. Liu, H. Meng, and H.-y. Lee, "Defense Against Adversarial Attacks on Spoofing Countermeasures of ASV," in *Proc. ICASSP 2020*, no. 14208718. Barcelona, Spain: IEEE, may 2020, pp. 6564–6568.

[22] Z. Yang, P. Y. Chen, B. Li, and D. Song, "Characterizing audio adversarial examples using temporal dependency," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[23] S. Sun, C.-F. Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Training augmentation with adversarial examples for robust speech recognition," *Proc. Interspeech 2018*, pp. 2404–2408, 2018.

[24] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*. IEEE, 2019.

[25] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. ICLR*, 2017.

[26] V. Subramanian, A. Pankajakshan, E. Benetos, N. Xu, S. McDonald, and M. Sandler, "A study on the transferability of adversarial attacks in sound event classification," in *Proc. ICASSP 2020*. IEEE, may 2020.

[27] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Interspeech 2019*. ISCA, sep 2019.

[28] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," *Proc. Interspeech 2017*, pp. 82–86, 2017.

[29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2016*. IEEE, 2016, pp. 770–778.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR 2018*, 2018, pp. 7132–7141.

[33] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[34] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting Adversarial Attacks with Momentum," *Proc. CVPR 2018*, pp. 9185–9193, 2018.

[35] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *Proc. CVPR 2017*, vol. 2017-January, pp. 86–94, 2017.