Hindawi
Security and Communication Networks
Volume 2022, Article ID 6410478, 8 pages
https://doi.org/10.1155/2022/6410478

Research Article

Black-Box Adversarial Attacks against Audio Forensics Models

Yi Jiang and Dengpan Ye

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

Correspondence should be addressed to Dengpan Ye; yedp@whu.edu.cn

Received 16 September 2021; Revised 28 December 2021; Accepted 30 December 2021; Published 17 January 2022

Academic Editor: Ding Wang

Speech synthesis technology has made great progress in recent years and is widely used in the Internet of things, but it also brings the risk of being abused by criminals. Therefore, a series of researches on audio forensics models have arisen to reduce or eliminate these negative effects. In this paper, we propose a black-box adversarial attack method that only relies on output scores of audio forensics models. To improve the transferability of adversarial attacks, we utilize the ensemble-model method. A defense method is also designed against our proposed attack method under the view of the huge threat of adversarial examples to audio forensics models. Our experimental results on 4 forensics models trained on the LA part of the ASVspoof 2019 dataset show that our attacks can get a 99% attack success rate on score-only black-box models, which is competitive to the best of white-box attacks, and 60% attack success rate on decision-only black-box models. Finally, our defense method reduces the attack success rate to 16% and guarantees 98% detection accuracy of forensics models.

1. Introduction

Speech synthesis technologies have advanced significantly in recent years [1, 2]. Speech synthesis generally refers to the process of converting text into speech. At present, the mainstream speech synthesis system generally consists of two parts: spectrogram prediction network and vocoder. The spectrogram prediction network converts the text into the mel spectrograms. Shen et al. [3], for example, use a Seq2Seq network with an attention mechanism to map text to mel spectrograms, Ren et al. [4] and Lancucki et al. [5] use the transformer structure [6] for this purpose. The vocoder is used to convert the mel spectrograms into speech. Van et al. [7] use several dilated convolution layers to achieve this function. Prenger et al. [8] use a generative model that generates audio by sampling from a distribution [9]. Of course, there are some end-to-end models, such as FastSpeech2s [10]. These technologies have been widely utilized in the Internet of things [11, 12] like a smart speaker, personal voice assistant, etc.

However, these technologies also have been abused. They appear in telecom fraud, creating rumors and spoofing automatic speaker verification (ASV) systems. To detect these fake audios, the researchers designed several methods. Lai et al. [13] accumulate discriminative features in frequency and time domains selectively, Lai et al. [14] adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels, Jung et al. [15] use the convolutional layer to extract frame-level embedding and the GRU layer to aggregate extracted frame-level features into a single utterance-level feature. Related competitions [16] were also organized to promote research in this field.

Previous researches show that the image classification neural networks [17–19] are vulnerable to attacks from adversarial examples, and audio models are no exception [20–27]. Generally, adversarial attacks are divided into two categories: white-box attacks and black-box attacks. A white-box attack means that the attacker can access the complete structure, parameters, and input and output of the model, the black-box attack means that the attacker can only obtain external input and output information but cannot access the internal structure and parameters of the model [28, 29]. Current researches on adversarial attacks on audio forensics models mainly focus on white-box attacks [30]. Although there are studies on using the transferability of adversarial examples to achieve black-box attacks, it still relies on white-box models to generate adversarial examples [31]. In this paper, we will only rely on the output distribution to conduct black-box adversarial attacks.

The main contributions of this article can be summarized as follows:

(i) To the best of our knowledge, we are the first to propose a black-box adversarial attack method only relying on the output distribution of audio forensics models and we use the ensemble-model method to increase the transferability of adversarial examples to implement decision-only black-box attacks.

(ii) We propose a defense method based on low-sensitivity features in view of the huge threat of black-box adversarial examples.

(iii) Our proposed black-box method can get the attack success rate equivalent to the best of white-box attacks and our defense method significantly reduces the threat.

The rest of the paper is organized as follows: Section 2 introduces several audio forensics models, which are the victim models in this paper; Section 3 describes the proposed adversarial attacks and defense methods; Section 4 introduces the experimental setup and results; and Section 5 gives theconclusion and future work.

2. Audio Forensics Models

Current speech synthesis technologies have developed to a high level. Once they are used by criminals in the fields of telecommunications fraud, network rumors, etc., and it will bring great harm to society. Therefore, people have designed a variety of audio forensics models, which aim to reveal the difference between real voice and fake voice from various angles. The following will introduce several current mainstream audio forensics models, whose detection accuracy is among the best in the audio forensics competition ASVspoof 2019. Therefore, they will serve as the victim models in this paper.

2.1. Attentive Filtering Network Model (AFnet) [13]. Attentive filtering (AF) accumulates discriminative features in frequency and time domains selectively. AF augments every input feature map S with an attention heatmap A_S. The augmented feature map $S^{\star}$ is then treated as the new input for the dilated residual network (DRN). For $S^{\star}$, $S^{\star} \in \mathbb{R}^{(F\times T)}$, AF is described as

$$S^{\star} = A_S \circ S + \overline{S}, \tag{1}$$

where F and T are the frequency and time dimensions, ° is the element-wise multiplication operator, + is the element-wise addition operator, and $\overline{S}$ is the residual S. To learn the attention heatmap, A_s contains similar bottom-up and top-down processing as [32, 33], and is described as

$$A_s = \phi(U(S)), \tag{2}$$

where ϕ is a nonlinear transform such as sigmoid or softmax, U is a U-net-like structure, composed of a series of downsampling and upsampling operations, and S is the input.

2.2. Squeeze-Excitation ResNet Model (SEnet) [14]. The squeeze-and-excitation (SE) block [34] is a computational unit that can be constructed for any given transformation $F_{\text{tr}}\colon X \longrightarrow U, X \in \mathbb{R}^{H'\times W'\times C'}, U \in \mathbb{R}^{H\times W\times C}$. The features U are first passed through a squeeze operation F_{sq} and get a statistic $z \in \mathbb{R}^{C}$, where the c^{th} element of z is calculated by

$$z_c = F_{\text{sq}}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j). \tag{3}$$

This is followed by an excitation operation F_{ex}.

$$s = F_{\text{ex}}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)), \tag{4}$$

where δ refers to the ReLU function, $W_1 \in \mathbb{R}^{C/r\times C}$ and $W_2 \in \mathbb{R}^{C\times C/r}$. The final output of the block is obtained by rescaling the transformation output U with the activations

$$\tilde{x}_c = F_{\text{scale}}(u_c, S_c) = s_c.u_c, \tag{5}$$

where $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_C]$ and $F_{\text{scale}}(u_c, s_c)$ refers to channel-wise multiplication between the feature map $u_c \in \mathbb{R}^{H\times W}$.

It will be easy to apply the SE block, which adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels to ResNet and get the squeeze-excitation ResNet (SEnet).

2.3. CNN-GRU Model [15]. The DNNs used in this model include convolutional neural network (CNN), gated circulation unit (GRU), and fully connected layer (CNN-GRU). In this architecture, the convolutional layer is first used to process input features to extract frame-level embedding. The convolutional layer includes residual blocks with identity mapping [35] to facilitate the training of deep architectures. Specifically, the first convolution layer of this model deals with the local adjacent time and frequency domains and gradually aggregates them through the repeated pooling operations to extract frame-level embedding. Then, the GRU layer is used to aggregate the extracted frame-level features into a single utterance-level feature. Fully connected layers are used to convert utterance-level features. An output layer with two nodes indicates whether the input utterance is a spoof or bona fide.

3. Audio Adversarial Examples Generation

3.1. Threat Model. In this paper, the adversarial attack is to craft adversarial voice $x = x + \delta$ by finding a perturbation δ such that (1) x is an original voice classified as the spoof by the audio forensics model, (2) δ is as human-imperceptible as possible, and (3) the audio forensics model classifies the voice $x = x + \delta$ as the bonafide. To be as human-imperceptible as possible, our attack following the FAKEBOB [22] adopts L_∞ norm to measure the similarity between the original and adversarial voices and ensures that the L_∞ distance $\|\acute{x}, x\|_\infty :=\max_i\left\{|\acute{x}(i) - x(i)|\right\}$ is less than the given maximal amplitude threshold ϵ of the perturbation, where i denotes the sample point of the audio waveform. So, we can formalize the problem of finding an adversarial voice $\acute{x}$ for a voice x as the following constrained minimization problem:

$$\operatorname{argmin}_\delta f(x + \delta), \text{such that} \|\acute{x}, x\|_\infty < \epsilon, \tag{6}$$

where f is a loss function. To ensure the success rate of the attack, we minimize the loss function rather than minimizing the perturbation δ. When f is minimized, $x + \delta$ is recognized as the bonafide.

According to the attacker's mastery of the model, the adversarial attack can be divided into a white-box attack and a black-box attack. The white-box attack generally means that the attacker can fully understand all the information of the victim model, including the external input and output information of the model and the internal structure and parameters. Attackers can efficiently perform gradient descent by differentiating the loss function to launch an iteration-based adversarial attack. Previous researches on adversarial examples against audio forensics models mostly focus on white-box attacks [30] or using the adversarial examples generated from white-box models to conduct transferable adversarial attacks [31]. However, in the real environment, users of the audio forensics model generally do not disclose the internal structure and parameters of the model, which significantly limits the application scenarios and the threat of white-box attack. It also leads some people to mistakenly believe that protecting the internal information of the model can prevent them from adversarial attacks.

Therefore, in this paper, we will focus on black-box adversarial attacks. Black-box adversarial attacks mean that the attackers can only access the input and external output of the model. The attacker cannot directly use the internal information to obtain the gradient of the loss function and launch the adversarial attack. Compared to the white-box model, black-box adversarial attacks can be further subdivided into score-only black-box adversarial attacks and decision-only black-box adversarial attacks. The score-only black-box adversarial attack refers to that the attacker can access the confidence scores of the model for each input, while a decision-only black-box attack means a direct attack that solely relies on the final decision of the model [36].

In the remainder of this section, we will present methods for launching adversarial attacks in these two black-box scenarios and attempt to defend against score-only black-box adversarial attacks.

3.2. Score-Only Black-Box Attack Algorithm. As shown in Figure 1, we will introduce the whole attack process in the remainder of this subsection, especially the loss function and algorithm to solve the optimization problem.

3.2.1. Loss Function. The key to carrying out the adversarial attack is that the score $[S(x)]_b$ of the bonafide voice should be greater than $[S(x)]_s$ of the spoof voice. Therefore, the loss function f is defined as follows:

$$f(x) = \max\{[S(x)]_s - [S(x)]_b, -\kappa\}, \tag{7}$$

where the parameter $-\kappa$, is to control the intensity of adversarial examples, so we can increase κ to enhance the robustness of the adversarial examples.

3.2.2. Optimization Algorithm. We use the basic iterative method (BIM) [18] with the estimated gradients to craft adversarial examples. Therefore, the i^{th} iteration voice x_i can be defined as

$$\acute{x}_i = \operatorname{clip}_{x,\epsilon}\left\{\acute{x}_{i-1} - \eta \cdot \operatorname{sign}(g_i)\right\}, \tag{8}$$

where η is the learning rate, g_i is the i^{th} iteration gradient.

To compute the estimated gradients, we use the natural evolution strategy (NES) [37], because the NES-based gradient estimation is proved to require much fewer queries than finite difference gradient estimation. In detail, we first create m (must be even) Gaussian noises $(u_1, \ldots, u_m)$ on the i^{th} iteration, and generate m new voices $x_{i-1}, \ldots, x_{i-1}$, where $x_{i-1} = x_{i-1} + \sigma \times u_j$. Then we compute the loss values $f(x_{i-1}), \ldots, f(x_{i-1})$. Finally, the gradient $\nabla_x f(x_{i-1})$ can be computed as

$$\nabla_x f\left(\acute{x}_{i-1}\right) = \frac{1}{m \times \sigma} \sum_{j=1}^{m} f\left(\acute{x}_{i-1}^{j}\right) \times u_j. \tag{9}$$

We also use the momentum [38] to speed up the convergence and increase the transferability of adversarial examples, therefore the i^{th} iteration gradient g_i can be defined as

$$g_i = \mu \cdot g_{i-1} + (1 - \mu) \cdot \nabla_x f\left(\acute{x}_{i-1}\right), \tag{10}$$

where the μ is the decay factor.

3.3. Decision-Only Black-Box Attack Algorithm. Although the NES-based gradient estimation attack has no need to touch the internal structure and parameters of the model, it still needs to obtain the distribution of result scores through a large number of queries. Once the model limits the number of queries or returns only positive or negative results without the score, this attack will be impossible to implement. In this regard, the transferable adversarial attack method can be used to achieve a decision-only black box attack. Specifically, the transferable adversarial attack is generating adversarial examples through known methods and then using these examples to attack the decision-only black-box model.

To improve the transferability of adversarial examples, an obvious idea is to increase the attack intensity κ. However, if we only increase the κ, the adversarial examples may overfit and it will decrease the transferability. So, an ensemble-based method will be used to conduct the decision-only black-box attack. In [39], authors argue that the

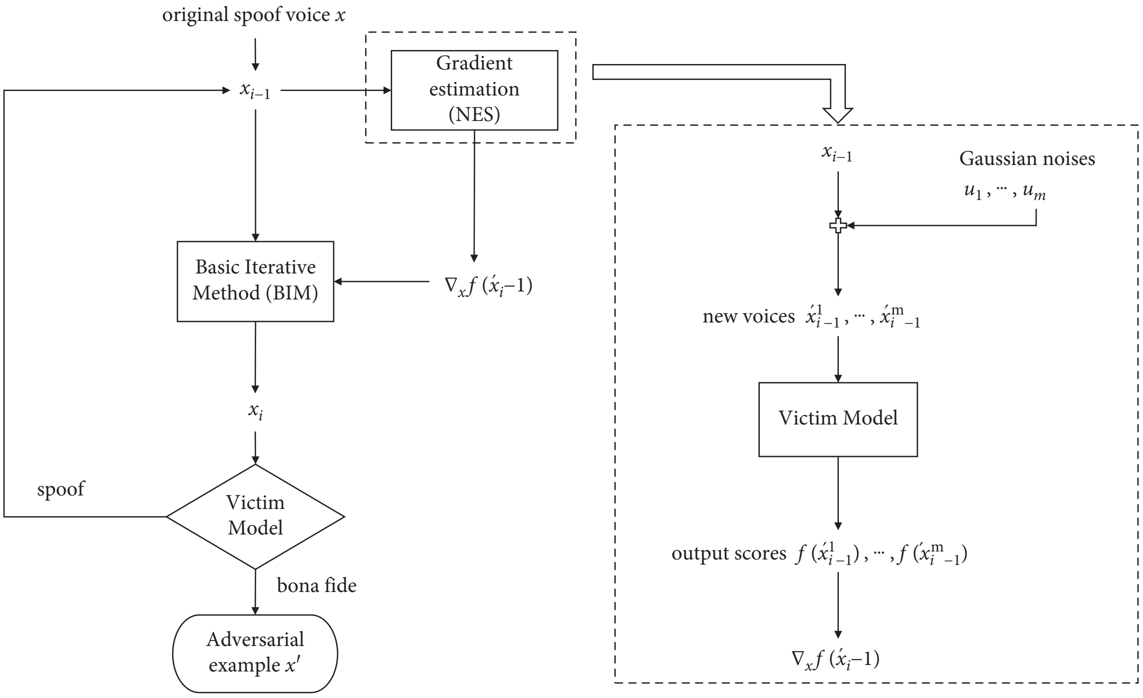
original spoof voice x
x_{i-1}
Gradient estimation (NES)
Basic Iterative Method (BIM)
$\nabla_x f\,(\acute{x}_i{-}1)$
x_i
spoof
Victim Model
bona fide
Adversarial example x'
x_{i-1}
Gaussian noises
$u_1, \cdots, u_m$
new voices $\acute{x}^1_{i-1}, \cdots, \acute{x}^{\mathrm{m}}_{i-1}$
Victim Model
output scores $f\,(\acute{x}^1_{i-1}), \cdots, f\,(\acute{x}^{\mathrm{m}}_{i-1})$
$\nabla_x f\,(\acute{x}_i{-}1)$

Figure 1: Score-only black-box attack process.

adversarial examples are more likely to transfer to other models if they could fool various models simultaneously, as shown in Figure 2. We follow this strategy and made a weighted sum of the scores of multiple models. To attack K score-only black-box models simultaneously, we fuse the loss function as

$$f\,(x) = \sum_{k=1}^{K} w_k \cdot f\,(x_k), \tag{11}$$

where $f\,(x_k)$ represents the loss of k_{th} score-only black-box model and w_k is the ensemble weight, where $\sum_{k=1}^{K} w_k = 1$.

3.4. Defense. Audio forensics models aim to reduce the harm of speech synthesis technology; however, several adversarial attacks have made these efforts in vain. So we need some methods to defend against this adversarial attack so that the model can be reinforced.

In previous experiments, we noticed that although the models have the same structure and are trained on the same dataset, they show different detection accuracy when trained by features of different sizes and types. We deem audio forensics models have different sensitivity to different features. Because models trained by high-sensitivity features show better performance than those trained by low-sensitivity features, we consider the reason is the model can obtain more information from original audio information through particular features. Therefore, we think that if we use the low-sensitivity features, the models will suffer less impact from the adversarial perturbation, and we will attempt to use these low-sensitivity features to reinforce audio forensics models.

4. Experiments

4.1. Dataset and Victim Models. Following the setting in [30], we use the LA part of the ASVspoof 2019 dataset [16]. We use the 2048 fast Fourier transform (FFT) bins energy spectrum as input for all models. Only the first 400 frames of each utterance are used to extract acoustic features.

We use the LA training set to train our audio forensics models and the LA developing set to evaluate the models. The details of the models can be found in Table 1.

4.2. Score-Only Black-Box Attack. We randomly selected 500 spoof audio examples from the trn set to conduct our adversarial proposed attacks. All the selected samples are classified correctly by our victim audio forensics models before the attacks. We only generate adversarial examples from spoof examples, because we consider there's no real value in converting a bonafide sample to a spoof one. All of the attacks are conducted under $\varepsilon = 0.001$ in equation (6), $\kappa = 0$ in equation (7), $m = 500$, $\sigma = 0.001$ in equation (9). As shown in Table 2, our proposed method gets a 99% attack success rate, which is comparable to the MI-FGSM, the most powerful white-box attack method.

We can conclude that our proposed score-only black-box attack method is extremely threatening to mainstream audio forensics models. Tiny adversarial perturbation can almost completely invalidate them. This also shows that if only the internal structure and parameters are hidden from the attackers, it is almost impossible to defend against the attack. It is necessary to find other more effective defense methods.

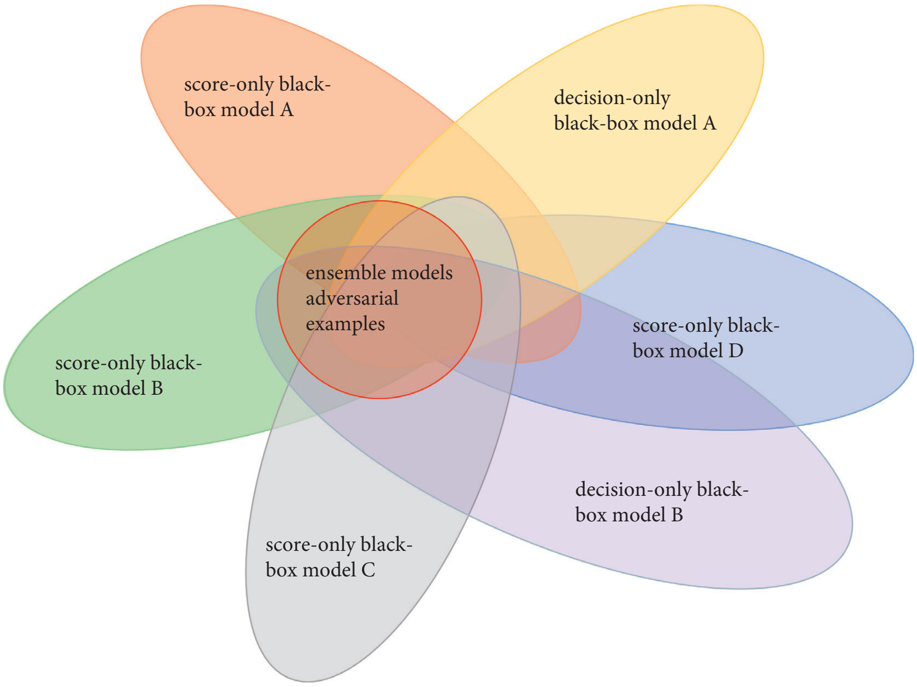
score-only black-box model A
decision-only black-box model A
ensemble models adversarial examples
score-only black-box model B
score-only black-box model D
decision-only black-box model B
score-only black-box model C

Figure 2: Ensemble-model.

Table 1: Detection accuracy of victim models (%).

Victim models	Trn	Dev
ResNet	99.99	99.90
SEnet	99.82	99.85
AFnet	99.74	99.57
CNN-GRU	99.99	99.96

Table 2: Adversarial attack success rate on score-only black-box models (%).

Victim models	Proposed method
ResNet	99.6
SEnet	99.9
AFnet	98.9
CNN-GRU	97.9

4.3. Decision-Only Black-Box Attack. We use the 100 spoof audio examples used in the previous subsection to conduct the decision-only black-box attack. We use 3 of the models to generate the adversarial examples and use the remaining one to evaluate the adversarial examples. In order to evaluate the ensemble-model method and the effect of intensity factor κ, we also generate adversarial examples through single-model with $\kappa = 2$ and muti-models with $\kappa = 0$. All of the results can be found in Table 3 and Figure 3.

We find that if we only increase the intensity factor κ or use the ensemble-model method, the improvement of the transferability of the adversarial examples is limited. So we need to combine these methods to get the best attack effect.

4.4. Defense. In the previous part of the paper, We have discussed how to enhance the defense capabilities of the model against our proposed adversarial attack. Here, we will test the method using low-sensitivity features. After conducting a series of experiments, we found that the models, which are trained by the log-power spectrum of 512 FFT bins, get a balance between the accuracy of detecting spoof samples and the defense capabilities against adversarial attacks.

We used the LA training set to train the audio forensics models and the LA developing set to evaluate the models. The detection accuracy of the original models and the reinforced models are shown in Table 4.

We randomly select 100 spoof audio examples from the trn set to conduct the adversarial attack and evaluate the defense capabilities of the reinforced models, we also conduct the attack on original models. The results of the examples can be seen in Table 5 and Figure 4.

By comparing the two types of models, we find that the average detection accuracy of original models on original examples is slightly higher than that of the reinforced

Table 3: Adversarial attack success rate on decision-only black-box models (%).

Victim models	ResNet, $\kappa = 2$	SEnet, $\kappa = 2$	AFnet, $\kappa = 2$	CNN-GRU, $\kappa = 2$	Muti-models, $\kappa = 0$	Muti-models, $\kappa = 2$
ResNet	*	46	54	46	54	**60**
SEnet	40	*	38	42	53	**62**
AFnet	56	45	*	45	65	**78**
CNN-GRU	47	44	45	*	48	**50**

The bold values are the best results from our proposed method.

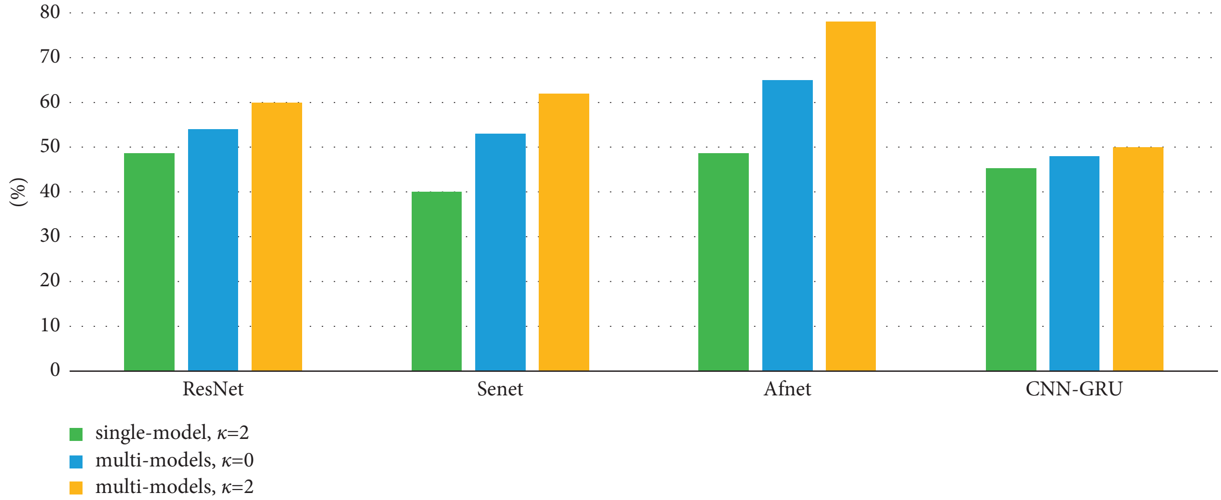

Figure 3: Adversarial attack success rate on decision-only black-box models.

Table 4: Detection accuracy of original models and reinforced models on original examples (%).

Victim models	Original models	Reinforced models
ResNet	99.90	94.12
SEnet	99.85	99.57
AFnet	99.57	99.38
CNN-GRU	99.96	99.27
Average	99.82	98.09

Table 5: Score-only black-box attack success rate on original models and reinforced models.

Victim models	Original models	Reinforced models
ResNet	99	**1**
SEnet	100	**28**
AFnet	98	**0**
CNN-GRU	97	**36**
Average	98.5	**16.25**

The bold values are the best results from our proposed method.

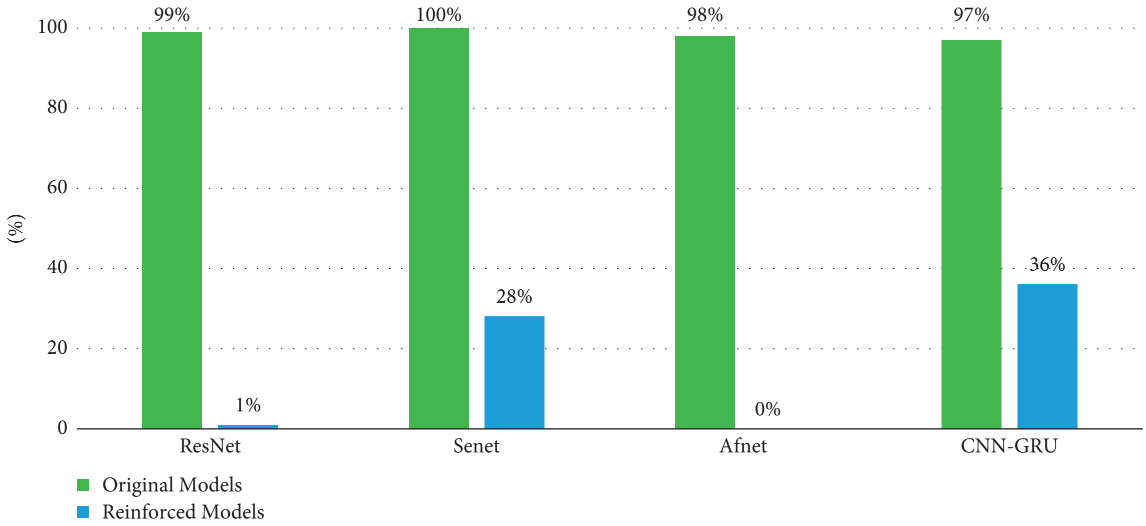

Figure 4: Score-only black-box attack success rate on original models and reinforced models.

models. However, the reinforced models we proposed significantly reduce the success rate of adversarial attacks.

5. Conclusion

In this paper, the black-box attack method we proposed achieves an attack success rate equivalent to the best of white-box attacks, which shows that hiding the internal structure and parameters of the model from the attacker cannot effectively protect the model. The success rate of the decision-only black-box attack also shows that the method of limiting the number of queries has scant protection capabilities for the model. Therefore, it is necessary to do more research on exploring more effective methods of model reinforcement.

Although the method proposed in this paper has reached a similar success rate to that of the white-box attack, however, there is still a large gap between the black-box method and the white-box method in terms of the generation efficiency of adversarial examples. Therefore, further research is needed on improving the generation efficiency of black-box adversarial examples.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Yan, X. Tan, B. Li et al., “Adaspeech 3: adaptive text to speech for spontaneous style,” 2021, https://arxiv.org/abs/2107.02530.

[2] I. Elias, H. Zen, J. Shen et al., “Parallel tacotron: non-autoregressive and controllable tts,” in *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5709–5713, Toronto, Canada, June 2021.

[3] J. Shen, R. Pang, R. J. Weiss et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, IEEE, Calgary, Canada, April 2018.

[4] Y. Ren, Y. Ruan, X. Tan et al., “Fastspeech: fast, robust and controllable text to speech,” 2019, https://arxiv.org/abs/1905.09263.

[5] A. Łańcucki, “Fastpitch: parallel text-to-speech with pitch prediction,” 2020, https://arxiv.org/abs/2006.06873.

[6] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[7] A. van den Oord, S. Dieleman, H. Zen et al., “Wavenet: a generative model for raw audio,” in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, p. 125, Sunnyvale, CA, USA, September 2016.

[8] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: a flow-based generative network for speech synthesis,” in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, Brighton, UK, May 2019.

[9] D. P. Kingma and P. Dhariwal, “Glow: generative flow with invertible 1 × 1 convolutions,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10236–10245, Montréal, Canada, December 2018.

[10] Y. Ren, C. Hu, X. Tan et al., “Fastspeech 2: fast and high-quality end-to-end text to speech,” 2020, https://arxiv.org/abs/2006.04558.

[11] Q. Wang, D. Wang, C. Cheng, and D. He, “Quantum2fa: efficient quantum-resistant two-factor authentication scheme for mobile devices,” *IEEE Transactions on Dependable and Secure Computing*, no. 1, p. 1, 2021.

[12] D. Wang and P. Wang, “Two birds with one stone: two-factor authentication with security beyond conventional bound,” *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 708–722, 2016.

[13] C. I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, “Attentive filtering networks for audio replay attack detection,” in *Proceedings of the ICASSP 2019-2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6316–6320, Brighton, UK, May 2019.

[14] C. I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: anti-spoofing with squeeze-excitation and residual networks," in *Proceedings of the Interspeech 2019*, pp. 1013–1017, Graz, Austria, September 2019.

[15] J. w. Jung, H. j. Shim, H. S. Heo, and H. J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVSpoof 2019 challenge," in *Proceedings of the Interspeech 2019*, pp. 1083–1087, Graz, Austria, 2019.

[16] M. Todisco, X. Wang, V. Vestman et al., "Asvspoof 2019: Future horizons in spoofed and fake audio detection," 2019, https://arxiv.org/abs/1904.05441.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, https://arxiv.org/abs/1412.6572.

[18] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, https://arxiv.org/abs/1607.02533.

[19] C. Szegedy, W. Zaremba, I. Sutskever et al., "Intriguing properties of neural networks," https://arxiv.org/abs/1312.6199.

[20] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," 2019, https://arxiv.org/abs/1904.05734.

[21] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," 2018, https://arxiv.org/abs/1801.00554.

[22] G. Chen, S. Chen, L. Fan et al., "Who is real bob? adversarial attacks on speaker recognition systems," 2019, https://arxiv.org/abs/1911.01840.

[23] Y. Chen, X. Yuan, J. Zhang et al., "Devil's whisper: a general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proceedings of the 29th USENIX Security Symposium Security 20*, pp. 2667–2684, Boston, MA, USA, August 2020.

[24] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: generating adversarial audio for end-to-end acoustic systems," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, pp. 357–369, Taipei, Taiwan, October 2020.

[25] Q. Wang, B. Zheng, Q. Li, C. Shen, and Z. Ba, "Towards query-efficient adversarial attacks against automatic speech recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 896–908, 2020.

[26] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 103–117, Dallas, TX USA, November 2017.

[27] B. Zheng, P. Jiang, Q. Wang et al., "Black-box adversarial attacks on commercial speech platforms with minimal information," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 86–107, New York, NY, USA, November 2021.

[28] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: reliable attacks against black-box machine learning models," in *Proceedings of the International Conference on Learning Representations*, Vancover, Canada, May 2018.

[29] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proceedings of the International Conference on Machine Learning*, pp. 2137–2146, Macau, China, February 2018.

[30] S. Liu, H. Wu, H. . y. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 312–319, Singapore, December 2019.

[31] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples," in *Proceedings of the Interspeech 2020*, pp. 4238–4242, Shanghai, China, 2020.

[32] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Honolulu, HI, USA, July 2017.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Munich, Germany, October 2015.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, UT, USA, June 2018.

[35] J. W. Jung, H. S. Heo, I. H. Yang, H. J. Shim, and H. J. Yu, "A complete end-to-end speaker verification system using deep neural networks: from raw signals to verification result," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5349–5353, IEEE, Calgary, Canada, April 2018.

[36] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: reliable attacks against black-box machine learning models," 2017, https://arxiv.org/abs/1712.04248.

[37] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber, "Natural evolution strategies," in *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pp. 3381–3387, IEEE, Hongkong, China, June 2008.

[38] Y. Dong, F. Liao, T. Pang et al., "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, Salt Lake, UT, USA, June 2018.

[39] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, https://arxiv.org/abs/1611.02770.