Defense Mechanisms Against Audio Adversarial Attacks: Recent Advances and Future Directions

Routing Li[1(✉)] and Meng Xue[2]

[1] School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
routingli@hust.edu.cn
[2] School of Computer Science, Wuhan University, Wuhan, China
xuemeng@whu.edu.cn

Abstract. With the popularity of speech and speaker recognition systems in recent years, voice interfaces are increasingly integrated into various Internet of Things (IoT) devices. However, studies have demonstrated that such systems are vulnerable to attacks using manipulated inputs. During the last few years, defense mechanisms have been studied and discussed from various aspects to protect voice systems from such attacks. Notwithstanding, there is lacking survey focus on the defense mechanism of audio adversarial examples. In this paper, we provide a comprehensive survey on state-of-the-art defense methods by illuminating their main concepts, reviewing the recent progress with a novel taxonomy, and discussing the future directions. It promises to bring awareness to the security problems in speech and speaker recognition systems and encourages people to propose more robust defenses against audio adversarial examples.

Keywords: Adversarial defenses · Deep learning · Speech recognition · Speaker recognition

1 Introduction

In recent year, the Internet of Things (IoT) has drawn significant research attention [12,21,31,32,37]. Automatic speech recognition (ASR) and speaker recognition (SR) systems, which are critical interfaces for many IoT devices [7,9], have made significant progress over the past decade[1]. In particular, voice assistants, like Amazon Alexa and Apple Siri, have been integrated into all kinds of platforms, giving people convenience over all aspects of their daily lives.

With the advent of the deep learning era, deep neural networks (DNNs) are playing a crucial role in ASR and SR systems. However, studies [1,6,10,13]

[1] In the following discussions, we refer to automatic speech recognition and speaker recognition as ASR and SR, respectively, for ease of distinction.

Z. Xiao et al. (Eds.): ICECI 2022, LNICST 478, pp. 166–175, 2023.
https://doi.org/10.1007/978-3-031-28990-3_12

have demonstrated that these models could be easily fooled by so-called adversarial examples (AEs). In AE attacks, the adversary intentionally adds subtle interference to the input samples to cause the model gives an incorrect output [28]. Specifically, in the audio realm, adversarial samples should meet two objectives simultaneously: 1) causing the voice system to make a wrong prediction; 2) being undetectable to humans. Over the past few years, we have witnessed a rapid growth of works regarding defense against AE attacks [14,20,30]. Meanwhile, several survey papers [8,18,38,39] have been proposed for introducing adversarial perturbations and corresponding defense methods on DNNs. However, these papers mainly focus on the image domain [16,17,19]. Because of the particularity of voice signal and voice processing pipeline, we cannot directly use the attack and defense mode for the image domain to ASR and SR systems. Abdullah et al. [2], and Chen et al. [11] analyze the security threats to voice systems from different perspectives. Notwithstanding, the content of the existing works on defense methods against audio adversarial attacks is limited. Despite its potential significance, there is a lack of discussion that focuses on audio AE defenses. Hence, the primary motivation of this paper is to provide a comprehensive survey on state-of-the-art defense mechanisms against AE by illuminating their main concepts, reviewing the recent progress with a novel taxonomy, and discussing the future direction.

2 Background

2.1 Automatic Speech and Speaker Recognition Systems

Automatic Speech Recognition Systems. ASR systems usually consist of four processes: pre-processing, feature extraction, inference, and decoding.

Pre-processing. The purpose of the pre-processing phase is to extract the speech portion of the signal to produce a "clean" signal. This stage usually contains two components: a noise filter and a low-pass filter. The noise filter removes noise from speech signals. Since most frequencies in human speech range from 300 Hz to 3000 Hz, the low-pass filter discards high-frequency signals.

Feature Extraction. The pre-processed speech signal is divided into frames and input to the feature extraction module. The most common method in this phase is Mel Frequency Cepstral Coefficient (MFCC) [26], as it is closest to the human auditory system. MFCC consists of four steps: Discrete Fourier Transform (DFT), Mel Filtering, Log Scaling, and Discrete Cosine Transform (DCT). DFT converts time domain signals into frequency domain signals. The loudness of the sound perceived by the human ear has a logarithmic relationship with the intensity of the sound signal rather than a simple linear relationship. The purpose of Mel Filtering and Log Scaling is to scale the intensity of the frequencies accordingly to this phenomenon. Finally, the DCT breaks down the input into many cosine components, leaving the part with most of the information and discarding the rest.

Inference. Hidden Markov Model (HMM) is the most commonly used inference model in early ASR systems. With the advent of the deep learning era, DNNs have become the main choice in this field. The most commonly used DNN models are convolutional neural network (CNN) and recurrent neural network (RNN). One limitation of CNNs is that their input and output sizes are fixed, whereas RNNs have flexible input and output sizes and can handle context information. Since speech recognition is a sequence-to-sequence task, it is more suitable to use RNNs in the inference phase.

Decoding. As described in [24], a decoding technique uses an acoustic model, a language model, and the spoken utterance to translate the inference result into the most likely sequence of words. The most commonly used decoding techniques are Viterbi search and N-Best search.

Speaker Recognition System. Speaker recognition systems and speech recognition systems have similar pipelines. The is that SR systems do not require decoding. Instead, they make judgments directly based on the results of the inference stage. As described in [6], SR systems can be classified into three categories based on the task: close-set identification (CSI), open-set identification (OSI), and speaker verification (SV).

Multiple enrolled speakers in a CSI system form a speaker group G. For any input voice x, the inference result contains the scores of all the enrolled speakers $[S(x)]_i$ for $i \in G$, which represents the probability that x is uttered by the speaker i. Finally, the CSI system outputs the speaker with the highest score. OSI systems have a similar task to CSI systems, which is to determine who made the input sound. The difference is that OSI has a preset threshold θ. Instead of directly outputting the speaker with the highest score, it compares the highest score to the threshold θ and rejects the voice if the score is less than θ. SV systems have only one enrolled speaker, unlike CSI and OSI systems. Therefore, the inference result contains only one score. The task is to determine if the enrolled speaker utters an input voice. Similar to the OSI system, the SV system rejects the voice with a score below the threshold θ.

2.2 Adversarial Example

An AE attack is to manipulate the input so that the model makes incorrect predictions. Meanwhile, to make the attack imperceptible, the difference between the adversarial sample and the original sample must not be too large. According to the adversary's knowledge, we categorize AE construction methods into two groups: white-box and black-box attacks. In white-box attacks, the whole model is accessible to the adversary. In other words, the attacker not only has access to the input data but also has full knowledge of the structure of the model and the specific parameters of each layer. Compared with white-box attacks, the adversary in black-box attacks is stronger. They treat the target model as a black box. That is to say, the attacker knows nothing about the model other than the input data.

AE attacks can also be divided into two categories depending on the outcome of the attack: untargeted and targeted attacks. The goal of untargeted attacks is to add a perturbation to the input sample so that the model output changes, while the targeted attacks specify the output of the model input AE.

Although AE is first proposed in the field of image recognition, studies in recent years have also proved that it got practical success in the audio domain. Both untargeted and targeted AEs can be constructed in black-box settings. For instance, DolphinAttack [35] and CommanderSong [34] inject hidden voice commands into the audio without catching the victims' attention. Kenansville Attack [1] develops signal processing methods to construct AE. Chen et al. [6] calculate adversarial perturbations by gradient estimation. In paper [13], Du et al. propose an audio AE generation mechanism based on the Particle Swarm Optimization (PSO) algorithm and the fooling gradient method. Chen et al. [10] strive to build a local substitute model with a handful of strategic queries, and the audio AE is constructed with the white-box substitute model to attack the black-box commercial speech recognition system. Unlike methods that rely on the knowledge of prediction/confidence scores, Zheng et al. [40] formulate the decision-only AE generation as a discontinuous large-scale global optimization problem and develop a novel technique called CC-CMA-ES to solve the problem.

Besides AE discussed above, universal adversarial perturbation (UAP) attacks in the audio domain [4]. UAP is in a completely black-box setting, where the attacker has no access to either the model or the input data. It is more practical than AE attacks because the input data is sometimes not easily accessible.

Table 1. Overview of the current defenses for ASR and SR systems

Defense method	Category	Subcategory	Phase	Target system
Sun et al. [27]	AE preservation	Adversarial training	Inference	ASR
Dompteur [14]	AE preservation	Model optimization	Inference	ASR
Esmaeilpour et al. [15]	AE preservation	Perturbation conversion	Pre-processing	ASR
Joshi et al. [22]	AE preservation	Perturbation conversion	Pre-processing	SR
Li et al. [23]	AE rejection	Model-based detection	Before inputting	SR
Samizade et al. [25]	AE rejection	Model-based detection	Before inputting	ASR
WaveGuard [20]	AE rejection	Model-free detection	Before inputting	ASR
Tramer et al. [29]	AE rejection	Model-free detection	Before inputting	ASR
Yang et al. [33]	AE rejection	Model-free detection	Before inputting	ASR

3 Defenses Against Audio Adversarial Attacks

Based on the results of processing AE, the existing defenses can be divided into two categories, including adversarial example preservation and adversarial example rejection. AE preservation defenses aim to make the ASR/SR system output the correct result of AEs, while AE rejection methods attempt to detect AEs and reject them before inputting the ASR/SR system.

Figure 1 summarizes the realization of the two classes of defense methods, and a brief overview is given in Table 1. We discuss the latest defense mechanisms from these perspectives as follows.

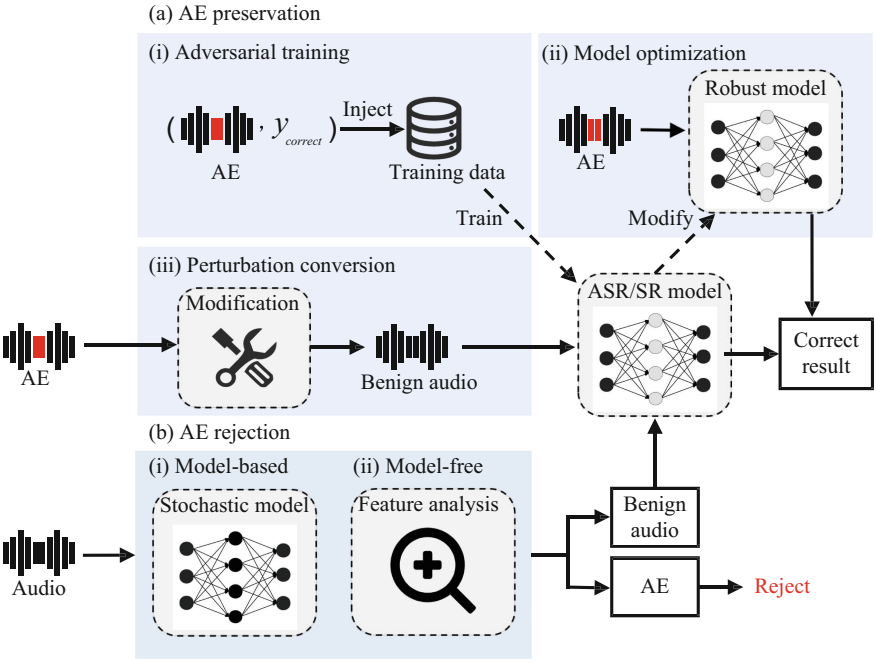

Fig. 1. The framework diagram of defenses against audio adversarial attacks.

3.1 Adversarial Examples Preservation

In AE preservation defenses, AEs are correctly recognized by the target system instead of being rejected. There are three ways to achieve this goal: adversarial training, model optimization, and perturbation conversion.

Adversarial Training. Adversarial training is a method to improve the robustness of recognition models by injecting correctly labeled AEs into the training set as new training samples in the model training stage. This method has been proven effective in resisting specific AE attacks [27]. However, adversarial training will reduce the accuracy of the recognition model, which is a great loss for the model owner. Moreover, adversarial training needs prior knowledge of the attack and adequate AE audio for training and is weak in preventing unknown attacks. Finally, according to [2], it is not effective in defending against signal processing attacks such as Kenansville Attack [1].

Model Optimization. Unlike the adversarial training approach, model optimization methods focus on improving the internal structure of the recognition model rather than enriching the training data. For example, in Dompteur [14], the authors exploit a psychoacoustic filter and a band-pass filter to make the recognition system closer to the human auditory system. This approach does not eliminate AEs but makes the perturbations in AEs easily identified by humans. In other words, Dompteur breaks the stealthy feature of AE.

Perturbation Conversion. Compared with adversarial training and model optimization methods, perturbation conversion defenses do not need to retrain the recognition model. Instead, it processes the input speech signal to remove the adversarial disturbance and converse the AE to benign audio.

For example, Esmaeilpour et al. [15] use a class-conditional Generative Adversarial Network (GAN) to pre-process the input signal to defend for DeepSpeech and Lingvo-based speech recognition systems. As for SR systems, Joshi et al. [22] analyze four perturbation conversion defenses, namely randomized smoothing, defense-GAN, variational autoencoder (VAE), and parallel waveGAN vocoder (PWG). They finally found that PWG combined with randomized smoothing is the best defense among them.

3.2 Adversarial Examples Rejection

Some studies [23,25] distinguish AE from benign samples by training stochastic (or machine learning) models since AE detection can be regarded as a classification problem. Besides, there are methods to detect AE by simply analyzing specific features without an additional model [20,33]. Therefore, we can divide AE rejection defense methods into two categories: model-based and model-free methods.

Model-Based Detection. The detection of adversarial examples can be treated as a classification problem. With a stochastic model, we can reject AEs before inputting the recognition system with high accuracy. For instance, Li et al. [23] introduced a VGG-like binary classification detector to detect AEs effectively. Samizade et al. [25] train a CNN model with three convolutional layers to detect AE. They present that this method has high accuracy for defending against the attacks proposed by Carlini et al. [5] and Alzantot et al. [3]. However, these methods require a large amount of AE audio. Moreover, similar to adversarial training, it requires the defender to know the attack algorithm's details, and such defenses' performance may degrade dramatically for unknown attacks. Worse, the stochastic model is vulnerable to many attacks, introducing additional security risks to the system.

Model-Free Detection. Yang et al. [33] leverage temporal dependency in speech signals to detect audio AE. They check whether the transcription of the

first half of the speech sample $f(x_{pre})$ is similar to the first half transcription of the whole audio $[f(x)]_{pre}$. If the two transcripts are not similar, the input is identified as an adversarial sample, and the system will reject it. Unfortunately, recent studies have proved that the temporal dependency framework is not effective in detecting AE. Tramer et al. [29] find an attack method to bypass the detection perfectly and prove that the adaptive evaluation in [33] is incomplete. Zhang et al. [36] propose an Iterative Proportional Clipping (IPC) algorithm to construct audio AE with the temporal dependency feature, which is also robust against the defense in [33].

In WaveGuard [20], the speech is input into the transform function, and AE is detected by comparing the differences between the two translations of the original audio and the transformed audio. This approach is similar to pre-processing defense, except that it rejects AE and feeds the benign raw audio into the ASR system instead of the converted one. The authors analyze the defense results of several candidate transform functions, such as quantization-dequantization, down-sampling, noise filtering, Mel extraction-inversion, and Linear Predictive Coding (LPC). They conclude that the transform-based defense can successfully reject AE, and Mel extraction-inversion is the best choice among them in an adaptive attack scenario. However, they also deduce that the detection accuracy of this mechanism will decrease when facing UAP attacks.

4 Future Direction

As shown in Table 1, only a tiny subset of the existing work has focused on defending against adversarial attacks in SR systems. Though ASR and SR systems have similar pipelines, some differences cannot be ignored. For example, as mentioned in Sect. 2.1, OSV and SI systems have a preset threshold θ, and the input voice will be rejected if the (highest) prediction/confidence score is less than θ, while ASR systems output the prediction result of any input audio and never rejected them. Besides, the prediction result of SR systems is limited to the enrolled speaker, while there are a tremendous number of predictions in ASR systems. For these and many other reasons, the defense methods may have limited effects on SR systems, although they are reported to be promising in the speech recognition domain. Therefore, it is urgent to explore more defense methods for SR systems.

Compared to other AE attacks, signal processing attacks have better transferability, which makes them more robust against some defenses. However, the existing works mainly focus on defending against optimization-based AE attacks and ignore evaluating the defense effectiveness on signal processing attacks. Thus, we advise more studies to explore mechanisms to defend against signal processing attacks.

5 Conclusion

With the rise of adversarial attacks on deep-learning-based speech and speaker recognition systems, it is of great importance to study the defense strategy

against this kind of attack. In this paper, we analyze the latest defense mechanisms from two aspects of adversarial examples rejection and preservation. All these defense methods can protect voice systems to some extent. However, they may not be so perfect when the attack setting changes. Compared with the field of image classification, the defense and detection methods of adversarial samples in the audio domain are not mature enough. It is promising that this paper can help people realize the security problems faced by speech and speaker recognition systems and encourage people to propose more robust defense methods.

References

1. Abdullah, H., et al.: Hear "no evil", see "kenansville"*: efficient and transferable black-box attacks on speech recognition and voice identification systems. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 712–729. IEEE (2021)
2. Abdullah, H., Warren, K., Bindschaedler, V., Papernot, N., Traynor, P.: SoK: the faults in our ASRs: an overview of attacks against automatic speech recognition and speaker identification systems. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 730–747. IEEE (2021)
3. Alzantot, M., Balaji, B., Srivastava, M.: Did you hear that? Adversarial examples against automatic speech recognition. arXiv preprint arXiv:1801.00554 (2018)
4. Bahramali, A., Nasr, M., Houmansadr, A., Goeckel, D., Towsley, D.: Robust adversarial attacks against DNN-based wireless communication systems. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 126–140 (2021)
5. Carlini, N., Wagner, D.: Audio adversarial examples: targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 1–7. IEEE (2018)
6. Chen, G., et al.: Who is real bob? Adversarial attacks on speaker recognition systems. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 694–711. IEEE (2021)
7. Chen, Y., Gong, X., Ou, R., Duan, L., Zhang, Q.: CrowdCaching: incentivizing D2D-enabled caching via coalitional game for IoT. IEEE Internet Things J. **7**(6), 5599–5612 (2020)
8. Chen, Y., Gong, X., Wang, Q., Di, X., Huang, H.: Backdoor attacks and defenses for deep neural networks in outsourced cloud environments. IEEE Netw. **34**(5), 141–147 (2020)
9. Chen, Y., Ran, Y., Zhou, J., Zhang, J., Gong, X.: MPCN-RP: a routing protocol for blockchain-based multi-charge payment channel networks. IEEE Trans. Netw. Serv. Manage. **19**, 1229–1242 (2021)
10. Chen, Y., et al.: {Devil's} whisper: a general approach for physical adversarial attacks against commercial black-box speech recognition devices. In: 29th USENIX Security Symposium (USENIX Security 2020), pp. 2667–2684 (2020)
11. Chen, Y., et al.: SoK: a modularized approach to study the security of automatic speech recognition systems. ACM Trans. Priv. Secur. **25**(3), 1–31 (2022)
12. Dai, X., et al.: Task co-offloading for D2D-assisted mobile edge computing in industrial internet of things. IEEE Trans. Industr. Inform. **19**, 480–490 (2022)
13. Du, T., Ji, S., Li, J., Gu, Q., Wang, T., Beyah, R.: 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 (2020)

14. Eisenhofer, T., Schönherr, L., Frank, J., Speckemeier, L., Kolossa, D., Holz, T.: Dompteur: taming audio adversarial examples. In: 30th USENIX Security Symposium (USENIX Security 2021), pp. 2309–2326 (2021)
15. Esmaeilpour, M., Cardinal, P., Koerich, A.L.: Class-conditional defense GAN against end-to-end speech attacks. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2565–2569. IEEE (2021)
16. Gong, X., Chen, Y., Huang, H., Liao, Y., Wang, S., Wang, Q.: Coordinated backdoor attacks against federated learning with model-dependent triggers. IEEE Netw. **36**(1), 84–90 (2022)
17. Gong, X., et al.: Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment. IEEE J. Sel. Areas Commun. **39**(8), 2617–2631 (2021)
18. Gong, X., Chen, Y., Wang, Q., Kong, W.: Backdoor attacks and defenses in federated learning: state-of-the-art, taxonomy, and future directions. IEEE Wirel. Commun. (2022)
19. Gong, X., Chen, Y., Yang, W., Mei, G., Wang, Q.: InverseNet: augmenting model extraction attacks with training data inversion. In: IJCAI, pp. 2439–2447 (2021)
20. Hussain, S., Neekhara, P., Dubnov, S., McAuley, J., Koushanfar, F.: {WaveGuard}: understanding and mitigating audio adversarial examples. In: 30th USENIX Security Symposium (USENIX Security 2021), pp. 2273–2290 (2021)
21. Jiang, H., Xiao, Z., Li, Z., Xu, J., Zeng, F., Wang, D.: An energy-efficient framework for internet of things underlaying heterogeneous small cell networks. IEEE Trans. Mob. Comput. **21**(1), 31–43 (2020)
22. Joshi, S., Villalba, J., Żelasko, P., Moro-Velázquez, L., Dehak, N.: Study of preprocessing defenses against adversarial attacks on state-of-the-art speaker recognition systems. IEEE Trans. Inf. Forensics Secur. **16**, 4811–4826 (2021)
23. Li, X., et al.: Investigating robustness of adversarial samples detection for automatic speaker verification. arXiv preprint arXiv:2006.06186 (2020)
24. Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I.: Automatic speech recognition: a survey. Multimed. Tools Appl. **80**(6), 9411–9457 (2021)
25. Samizade, S., Tan, Z.H., Shen, C., Guan, X.: Adversarial example detection by classification for deep speech recognition. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3102–3106. IEEE (2020)
26. Sigurdsson, S., Petersen, K.B., Lehn-Schiøler, T.: Mel frequency cepstral coefficients: an evaluation of robustness of MP3 encoded music. In: ISMIR, pp. 286–289 (2006)
27. Sun, S., Guo, P., Xie, L., Hwang, M.Y.: Adversarial regularization for attention based end-to-end robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(11), 1826–1838 (2019)
28. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
29. Tramer, F., Carlini, N., Brendel, W., Madry, A.: On adaptive attacks to adversarial example defenses. Adv. Neural. Inf. Process. Syst. **33**, 1633–1645 (2020)
30. Wang, S., Cao, J., He, X., Sun, K., Li, Q.: When the differences in frequency domain are compensated: understanding and defeating modulated replay attacks on automatic speech recognition. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 1103–1119 (2020)

31. Xiao, Z., et al.: A joint information and energy cooperation framework for CR-enabled macro-femto heterogeneous networks. IEEE Internet Things J. **7**(4), 2828–2839 (2019)
32. Xiao, Z., et al.: TrajData: on vehicle trajectory collection with commodity plug-and-play OBU devices. IEEE Internet Things J. **7**(9), 9066–9079 (2020)
33. Yang, Z., Li, B., Chen, P.Y., Song, D.: Characterizing audio adversarial examples using temporal dependency. arXiv preprint arXiv:1809.10875 (2018)
34. Yuan, X., et al.: {CommanderSong}: a systematic approach for practical adversarial voice recognition. In: 27th USENIX Security Symposium (USENIX Security 2018), pp. 49–64 (2018)
35. Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., Xu, W.: DolphinAttack: inaudible voice commands. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 103–117 (2017)
36. Zhang, H., Yan, Q., Zhou, P., Liu, X.Y.: Generating robust audio adversarial examples with temporal dependency. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 3167–3173 (2021)
37. Zhang, W., Zhou, S., Yang, L., Ou, L., Xiao, Z.: WiFiMap+: high-level indoor semantic inference with WiFi human activity and environment. IEEE Trans. Veh. Technol. **68**(8), 7890–7903 (2019)
38. Zhang, W.E., Sheng, Q.Z., Alhazmi, A., Li, C.: Adversarial attacks on deep-learning models in natural language processing: a survey. ACM Trans. Intell. Syst. Technol. (TIST) **11**(3), 1–41 (2020)
39. Zhang, X., Zheng, X., Mao, W.: Adversarial perturbation defense on deep neural networks. ACM Comput. Surv. (CSUR) **54**(8), 1–36 (2021)
40. Zheng, B., 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 (2021)