Design and implementation of covert communication based on intelligent voice system

Song Lipeng1) Zhang Ruichen1) Liu Wenxiang1) Wang Shuo1) Zhang Ye1) Ren Guanyu1)

1)(School of Mechanical, Electrical and Information Engineering, Shandong University (Weihai), Weihai 264200)

Abstract The development of deep learning has led to the emergence of adversarial sample attacks, but traditional communication encryption technology and physical layer security technology cannot solve privacy issues well. The current results of many teams have confirmed that selective adversarial samples have the characteristics of information hiding and can be well used to construct covert communications. Moreover, this covert communication model is different from the traditional covert communication model based on rules and cryptography. A new inspiration for covert communication construction. Based on this feature, our team applies adversarial sample attack technology to the construction of covert channels. Based on the DeepSpeech deep learning open source speech-to-text engine, gradient descent is used to generate speech-selective adversarial samples. And by combining the covert model and the surface model, covert communication is achieved with good results. This covert communication model is different from the traditional covert communication model based on rules and cryptography. It is a new inspiration for the construction of covert communication and has high research value.

Keywords communication encryption technology, selective confrontation, sample information hiding, covert communication

Design and Implementation of Covert Communication based on Intelligent Voice System

Lipeng Song 1) Ruichen Zhang 1) Wenxiang Liu 1) Shuo Wang 1) Ye Zhang 1) Guanyu Ren 1)

1)(School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209)

Abstract The development of deep learning promotes the emergence of anti-sample attacks, while the traditional communication encryption technology and physical layer security technology can not solve the privacy problem. The results of many teams at present confirm that selective adversarial samples have the characteristics of hidden information, and can be used in the construction of covert communication. This covert communication mode is different from the traditional covert communication mode based on rules and cryptography, which is a new inspiration for the construction of covert communication. Based on this feature, the counter sample attack technique is applied to the construction of covert channel. An open source speech-to-text engine based on DeepSpeech deep learning uses gradient descent to generate speech selective adversarial samples. The covert communication is realized by the combination of covert model and surface model, and the effect is good. This kind of covert communication mode is different from the traditional covert communication mode based on rules and cryptography, which is a new inspiration for the construction of covert communication and has high research value.

Key words communication encryption technology; selective confrontation samples; information hiding; covert communication

---

Song Lipeng (corresponding author), male, born in 1975, doctoral supervisor, professor, main research fields are artificial intelligence security and software vulnerability analysis. E-mail: slp880@sdu.edu.cn. Zhang Ruichen, male, 2002 Student, undergraduate, E-mail: 3033824368@qq.com. Liu Wenxiang, male, born in 2002, undergraduate, E-mail: 2425281352@qq.com. Wang Shuo, male, born in 2002, undergraduate, E-mail: 2404266246@qq.com. Zhang Ye, male, born in 2002, undergraduate, E-mail: 1498792816@qq.com. Ren Guanyu, male, born in 2002, undergraduate, E-mail: guanyuren@mail.sdu.edu.cn .

1 Introduction

Recently, deep neural network (DNN) [1] has provided better performance in image recognition [2], speech recognition [3], pattern analysis [4] and intrusion detection [5]. Although DNN works well, it is easily affected by adversarial examples [6]. Adding some imperceptible perturbation information to the sample and indicating the impact of the perturbation on different models, that is, selective adversarial samples, can make the neural network model under specific parameter conditions make wrong predictions. Adversarial examples have been widely studied in the image field. However, in recent years, research on adversarial examples has expanded to the field of speech. Many scenarios in the audio domain

demonstrate the potential threat of adversarial examples. For example, Alexa [7] developed by Amazon can provide services that support voice interaction after the user is authenticated, such as ordering products, obtaining information, and controlling multiple smart devices. However, there may be some problems with such voice interaction. For example, they might reveal a user's personal privacy to other users, or mistakenly order an item after hearing a sound on the TV or radio. In order to exploit these weaknesses of speech recognition systems, many studies propose adversarial examples and generate adversarial samples by adding a small amount of noise to the original audio samples.

Currently, many teams are engaged in intelligent voice attack and defense fields. Certain results have been achieved. After reading and discussing papers in related fields and in-depth study of the characteristics of adversarial samples, it was found that the existing selective adversarial samples have the characteristics of hiding information and are suitable for applications involving military communications, automatic phone eavesdropping and covert channels [8] scene. For example, when military communicators need to send a specific voice message, they can deliberately cause the enemy to misunderstand the message while allowing friendly forces to correctly identify the message.

Based on the fact that selective adversarial samples have the characteristics of hiding information, can be used for covert communication, and have high security characteristics, the team proposed a solution that combines audio adversarial sample attacks with covert communication ideas. This covert communication model is different from the traditional covert communication model based on rules and cryptography. It is a new inspiration for the construction of covert communication and has high research value. The team used the audio confrontation

model to transfer the content that actually needs to be sent.

content, hidden to audio files. This method has the following characteristics:

1. High safety. Only when both communicating parties master the same model and model parameters at the same time can the hidden information be correctly parsed. 2. Strong concealment. Even if it is intercepted midway, the hidden information cannot be discovered.

2Related work

Covert communication technology has become a research hotspot in the field of information security. At present, covert communication technology based on digital images can be said to be relatively complete and mature, while related research on voice covert communication technology has not started yet. In recent years, the attack technology of speech models has been continuously improved and improved by various teams, so voice covert communication technology based on this counterattack technology has begun to gradually develop. The text information hiding

method of modifying the text format mainly achieves information hiding by modifying the carrier text format. In recent years, most of the improvements in this type of method have been to reduce the text scaling ratio and embed secret information more evenly in the carrier text to increase the concealment and robustness of the confidential text. This type of method has relatively good visual performance. High concealment. However, if the text format is modified or the text is re-entered during the text delivery process, the secret information will be lost.

Partala[9] first tried to use blockchain as a medium to build a covert communication channel, and proposed a blockchain covert channel (BLOCCE) model, which hides information to the last digit of the transaction address and uses corresponding loops in sequence to generate The transaction address ensures the order of secret information. Since then, researchers have improved on the shortcomings of this model and proposed improved blockchain covert communication methods, and tried to reduce its communication costs [10,11].

Torki proposes a blockchain covert communication scheme that does not require manual changes to source data and only requires repeated execution of hidden algorithms. Guo et al. [12] and Lan Yiqin et al. [13] implemented hybridization by combining multi-layer linkable spontaneous anonymous group signatures and introduced the new elliptic curve algorithm Monero to build a covert communication channel in their blockchain application, using Monero has high security to improve the concealment of covert channels.

For deep learning-based covert communication, Szegedy et al. [6] first proposed an adversarial sample in which the attacker slightly transformed the image. The main purpose of using adversarial samples is to cause DNN errors by adding a small amount of noise to the original image; however, humans cannot tell the difference between the original image and the distorted image.

Image covert communication algorithms based on deep learning generally include two information hiding algorithms based on Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN). GAN-based algorithms usually use A variant of CNN. Based on the confrontational characteristics between information hiding and steganalysis, many scholars have carried out targeted optimization of the model structure of deep learning to achieve the embedding and extraction of secret data. The current research status shows that image information hiding methods based on deep learning have better hiding effects and performance indicators than traditional methods. In the audio field, the "Devil Music" attack refers to the 2018

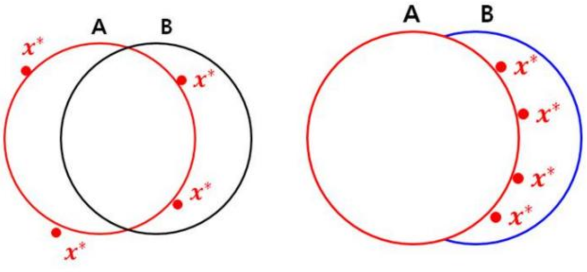
Chen Kai's research team at the Institute of Information Engineering, Chinese Academy of Sciences, uses music as a carrier to generate audio adversarial samples [15]. This kind of malicious audio sample inserts implicit instructions into normal music to achieve the purpose of attacking the intelligent voice system. Speech-

text adversarial sample attack refers to the speech adversarial sample generation method studied by Google Brain artificial intelligence scientist Carlini and his team. This attack method is to guide the speech recognition model to incorrectly identify speech samples as arbitrary specified text during the process of recognizing audio files by the intelligent speech recognition system [16]. After focused analysis of the above

articles and related work, the team decided to combine audio adversarial sample attacks with covert communication ideas to achieve a targeted attack designed on the audio-to-text engine to hide the content sent by the user into an audio file. . Only when both communicating parties master the same model and model parameters at the same time can the hidden information be correctly parsed. Even if it is intercepted by an eavesdropper on purpose, the hidden information cannot be discovered and deciphered.

3 background

Adversarial examples deliberately add imperceptible perturbations to input samples in the data set, causing the model to give an incorrect output with high confidence. That is, adversarial audio samples only need to make small perturbations on a piece of audio, and the classifier will mistranscribe the audio with a high degree of confidence, or even transcribe it into a specified text (not a text that the audio is correctly transcribed). The reason this happens is that neural networks are easily "spoofed." In Figure 1(a), model A is the target model with a neural network. The corresponding line is the decision boundary of target model A. If the sample is within the boundaries of target model A, then the sample will be correctly classified by target model A. Figure 1(b) is an example of a selective adversarial example that is correctly classified by target model B but misclassified by target model A. In Figure 1(b), the selective adversarial example x* is within the decision boundary of target model B, but deviates from the decision boundary of target model A.



(a) Examples of adversaries. (b) Example of safe confrontation between friends. Figure 1: Transferability example: single enemy target model A and friendly target model B.

The transferability of adversarial examples was first proposed by Szegedy et al. [6] in the literature. The transferability of adversarial examples means that the adversarial examples are misclassified by model A and can also be misclassified by model B. The migration attributes of adversarial examples This means that an attacker can choose to attack a machine learning model to cause samples to be misclassified without direct contact with the basic model. Szegedy et al. [6] studied the transferability of different models on the same data set. In addition, they also trained the same or different models on disjoint subsets of data and studied the transfer issues between them. However, the shortcoming is that their experimental results All are implemented on the MNIST data set. Goodfellow et al. [17] proposed that the generalization ability of adversarial samples between different models is due to the high consistency of the vectors of the adversarial interference and the model, so when training the same task, the opponent can learn similar functions on different models. This generalization feature means that if the enemy wants to attack the model, it does not need to access the target model. It can only be achieved by sending the adversarial samples generated by its own model training to the target model. Through further research and discussion, the team found that the adversarial samples

Portability can be used well for the construction of covert communication. Therefore, a new architecture is needed, which has a target model that does not require access and a self-model for attack, and uses the self-model to generate adversarial samples.

4 models

In this section, the selective audio adversarial examples and information hiding model architecture are introduced in detail.

4.1 Selective Audio Adversarial Examples

In Figure 2(a), model A is a hidden model based on neural networks. The curve in the figure is the decision boundary of the hidden model. If the sample is within the decision boundary of the hidden model, it means that the sample is correctly classified by the hidden model. Otherwise, the classification is wrong. In Figure 2(b), model B is a protection model based on neural network, which has the same classification decision as model A. The situation where sample x* is between the decision boundaries of model A and model B means that sample X* is classified incorrectly by model A and correctly classified by model B. This leads to the concept of selective adversarial examples.

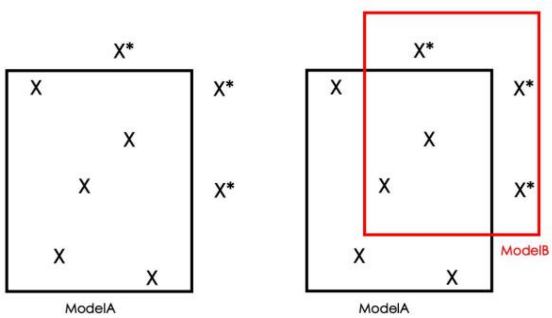


Figure 2 Single selective adversarial sample x , Hidden Model A and Protected Model B

Figure 2(b) shows that the selective adversarial example x* is outside the decision boundary of model A, and at the same time, x* is within the decision boundary of model B. But it is worth noting that x* should be as close as possible to the decision boundary of model A, which means that the distance metric of the original sample x and the selective adversarial sample x* is closer. At this time, if it further converges to the minimum distance on the decision boundary, the selective adversarial sample obtained will have the smallest noise and be selectively adversarial.

4.2 Model architecture

The prerequisite of the model architecture is that all parameters of the concealment model and the protection model are known. The information hiding model architecture is shown in the figure below:

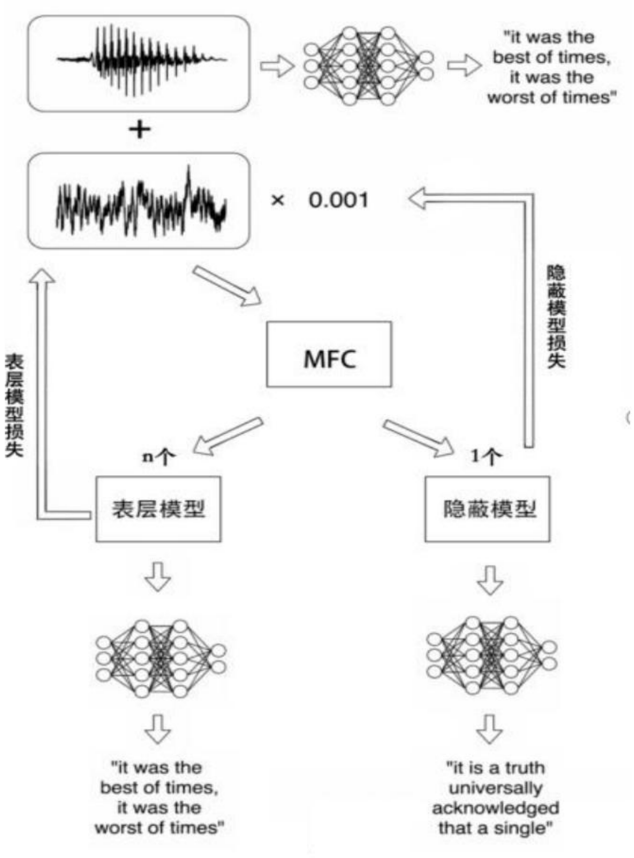


Figure 3 Model architecture diagram for generating selective

adversarial samples. As can be seen from Figure (3), the architecture for generating selective adversarial samples consists of n protection models (surface models) Mpi (1<i<=n) and a hidden model Mh Composition, the concealment model can also be expanded to n, but considering the practicality of the product, a proprietary

concealment model is enough. The inputs to the overall model architecture are the original sample x, the initial noise delta, the correct classification phrase label org_phrase and the latent classification phrase label target_phrase. After input, the architecture randomly generates initial noise, adds original samples, generates initial adversarial samples, and then performs mel cepstrum and forward propagation in the protection model and covert model. Mel cepstrum (MFC) is used as a preprocessing step to reduce the dimensionality of the input x. MFC divides the waveform at 50 frames per second and maps each frame to each frequency domain.

Mpi and Mh keep the model unchanged during the generation of selective adversarial samples (only model pre-training parameters and model structure are required during the generation process). After Mpi and Mh get the audio array input, Mpi outputs the classification probabilities of all protector models, and Mh outputs the classification probabilities of the covert models. Secondly, the CTC loss function is used to align the classification probability matrix and the label to obtain the loss value of each Mpi model and the loss value of Mh. Finally, after weighting the Mpi weighted loss value and the loss value of Mh, it is fed back to the added noise to form a closed loop. Further Adjust the noise until it meets a predetermined level of selection resistance. As

shown in Section 4.1 Selective Adversarial Samples, the ultimate goal of this model architecture is to continuously reduce the distance between the original sample x and the selected adversarial sample x * on the premise of ensuring that each protection model Mpi is correctly classified and the concealment model Mp is classified incorrectly. Function, let fp (x *Thus, the goal) can be set to protect the recognition behavior of the model, and fh (x *) mask the recognition behavior of the model, org_phrase is the original , phrase target_phrase is the target hidden phrase. The target function is as follows:

x *: argmin(x,x *) st fp i(x *) = org_phrase and fh (x *) = target_phrase

x * = x + delta

In order to make the protection model and concealment model correctly classify the given target phrase in the objective function, and the noise delta needs to be as small as possible, the added noise adjustment requires the feedback of the loss function. The total feedback function of the loss function is as follows:

$$LossT = Cdis\ Lossdistortion + CMp \sum_{i=0}^{n} w_i\ Loss(Mpi) + CMh\ Loss(Mh)$$

Among them, Cdis , CMp and CMh are the weight distribution of noise, protection model and hidden model loss in the total loss respectively, which can make the generated selective adversarial samples pay more attention to a certain aspect of the characteristics. Wi is the loss weight of different protection models, and the distribution of weights can be adjusted according to the difficulty

of attacks by different protection models. Choice of three loss functions: loss bias for noise Lossdistortion, the team uses the L2 norm to measure the distance between audio arrays. For Loss (Mpi) and Loss (Mh), the team uses the CTC loss function with automatic alignment features.

5 experiments

Test the hidden information "this is a test" on a piece of English audio on the DEEPSPEECH model, and repeat it many times to observe

Observe hidden effects and time consumption. Demonstration audio usage on github in carlini/audio_adversarial_examples by N.Carlini Audio sample containing the English sentence "without the dataset the article is useless" 102.2kb size wav text

The software experiment was performed on UBUNTU 20.04 LTS with Python 3.6 tensorflow- , gpu==1.15.4 installed.

OK, use CUDA11.4 to call display adaptation in the time-consuming test NVIDIA RTX2070 SUPER input hardware acceleration and statistics Successful identification in every test (the model correctly identifies the hidden object for the first time) (hiding information) is time-consuming and completely converged (CTCloss < 0.005). Complete convergence) time-consuming and 60,000 rounds of iteration time-consuming are three sets of data.

Table 1 Recognition time-consuming unit seconds/S under different requirements

The successful identification of the serial number takes 60,000 iterations to fully converge.

			time consuming
1	38.33	357.39	423.62
2	37.12	345.46	368.00
3	36.51	344.41	375.94
4	32.38	352.22	369.32
5	35.25	370.01	398.83

As can be seen from the above table, the model successfully identified the above five groups in a shorter time. The average successful identification time is only 35.92s, and the average complete convergence The time consumption reaches 353.90s. The big difference between the two is mainly due to minimizing noise. This is caused by the long time required for the noise process. It can be seen that without considering the noise Audio with hidden information can be quickly produced under the condition of sound, but at this time The audio has an audible "tingling" sound similar to electric current that can be heard by the human ear. The large number of iterations required to remove this type of noise would take too long, There is still a lot of room for improvement in quickly reducing noise.

5.1 The impact of adding noise on the original audio

Next, we will take a step further to completely converge the noise in the audio. Analysis and evaluation, the waveform diagram of audio samples before and after hiding information is as follows Down:

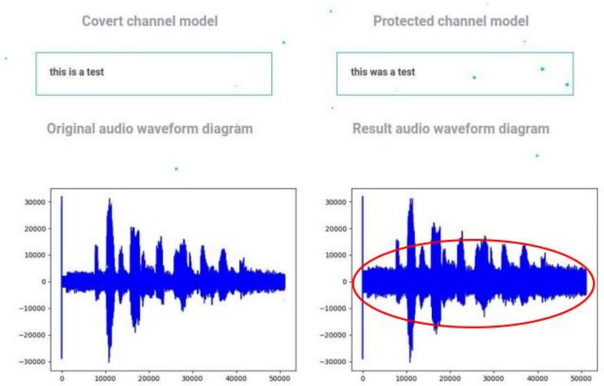


Figure 4 Audio waveform changes before and after interference

As shown in Figure 4, the sample waveforms before and after hiding are basically the same. Especially in the high amplitude part, it is difficult to see the significant difference with the naked eye, but in the low amplitude part There is a slight increase and blurring phenomenon, which is very difficult for human hearing. The impact is that the audio is blurred, and a slight noise like an electric current can be heard. sound, in order to further explore the impact on human hearing, 10 Volunteers conducted a double-blind experimental evaluation, selecting short English audio, There are ten groups each of long English audio and white noise in which the phrase "this is" is hidden. a test", each group runs for about 350 seconds before fully converging.

5.2 Evaluation of information hiding effect

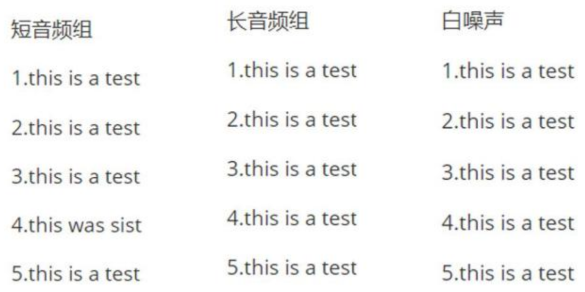


Figure 5 Audio recognition results

As shown in Figure 5, except for the short audio group, there is a sample recognition error. Except for all other samples, the hidden information was successfully identified by the model. "This is a test", the information hiding effect is good.

The questionnaire contains 20 sets of pre-interference audio, and volunteers are asked to judge Whether the broken audio has been interfered, the questionnaire provides interference, no interference, no There are three different options,

The results show that:

9 volunteers were unable to accurately distinguish the audio difference before and after interference Different (more than 20% of the audio before and after interference are correctly selected).

8 volunteers could not roughly distinguish the audio difference before and after interference Different (more than 10% of the audio before and after interference are correctly selected).

At the same time, 9 people said that the difference in audio before and after was very small and difficult to distinguish.

To sum up, the current model has been able to create something that is difficult for the human ear to Identify interfering audio samples that have been produced over time long, in terms of reducing noise reduction time and further reducing low-frequency noise. There is still much room for improvement.

6 Summary

Selective adversarial examples have the characteristics of information hiding. Based on this One feature is that the team applies adversarial sample attack technology to covert channels of construction. This method is based on the DeepSpeech speech-to-text engine.

Using gradient descent to generate speech-selective adversarial samples will actually require The content to be sent is hidden in the audio file. Use a covert model The method combined with the surface model realizes information hiding and achieves Better results. This information hiding model is different from the traditional Rules and cryptographic information hiding methods, constructs for information hiding It provides new revelations and has high research value.

At present, this project research has the problem that it takes too long to generate adversarial samples. Noise problem. In the future, we will focus on improving the generation of adversarial samples. speed, shorten noise reduction time and further reduce low-frequency noise.

references

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85–117, 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[3] G. Hinton, L. Deng, D. Yu, GE Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, TN Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[4] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning, pp. 160– 167, ACM, 2008.

[5] S. Potluri and C. Diedrich, "Accelerated deep neural networks for enhanced intrusion detection system," in Emerging Technologies and Factory Automation (ETFA), 2016 IEEE 21st International Conference on, pp. 1–8, IEEE, 2016.

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

[7] "https://developer.amazon.com/alexa"

[8] M. Smeets and M. Koot, "Covert channels," research report for RPI university of Amsterdam MSc in System and Network Engineering, 2006

[9] PARTALA J. Provably secure covert communication on blockchain[J]. Cryptography, 2018, 2(3): 18.

[10] Song Shang, Peng Wei. BLOCCE+: An improved blockchain-based covert communication method [J], Journal of Chongqing University of Science and Technology (Natural Science), 2020,34(9):238-244.

[11] ZHANG LJ, ZHANG ZJ, WANG WZ, et al. A covert communication method using special bitcoin addresses genera ted by vanitygen[J]. Computers, Materials & Continua, 2020, 65(1): 597-616.

[12] GUO ZZ, SHI LC, XU MZ, et al. MRCC: a practical covert channel over monero with provable security[J]. IEEE Access, 2021, 9:31816-31825.

[13] Lan Yiqin, Zhang Fangguo, Tian Haibo. Using Monero to achieve covert communication[J]. Xi'an Electronics Journal of University of Science and Technology, 2020,47(5): 19-27.

[15] XuejingYuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, ShengzhiZhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter, "CommanderSong: ASystematic Approach for Practical Adversarial Voice Recognition," InUSENIX Security Symposium, pages 49–64, 2018.

[16] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 1-7, doi: 10.1109/SPW.2018.00009.

[17] Tramèr F, Papernot N, Goodfellow I, et al. The space of transferable adversarial examples. arXiv preprint arXiv: 1704.03453, 2017