Cybersecurity

RESEARCH

Open Access

Towards the universal defense for query-based audio adversarial attacks on speech recognition system

Feng Guo[1,2], Zheng Sun[1,2], Yuxuan Chen[1,2*] and Lei Ju[1,2]

Abstract

Recently, studies show that deep learning-based automatic speech recognition (ASR) systems are vulnerable to adversarial examples (AEs), which add a small amount of noise to the original audio examples. These AE attacks pose new challenges to deep learning security and have raised significant concerns about deploying ASR systems and devices. The existing defense methods are either limited in application or only defend on results, but not on process. In this work, we propose a novel method to infer the adversary intent and discover audio adversarial examples based on the AEs generation process. The insight of this method is based on the observation: many existing audio AE attacks utilize query-based methods, which means the adversary must send continuous and similar queries to target ASR models during the audio AE generation process. Inspired by this observation, We propose a memory mechanism by adopting audio fingerprint technology to analyze the similarity of the current query with a certain length of memory query. Thus, we can identify when a sequence of queries appears to be suspectable to generate audio AEs. Through extensive evaluation on four state-of-the-art audio AE attacks, we demonstrate that on average our defense identify the adversary's intent with over 90% accuracy. With careful regard for robustness evaluations, we also analyze our proposed defense and its strength to withstand two adaptive attacks. Finally, our scheme is available out-of-the-box and directly compatible with any ensemble of ASR defense models to uncover audio AE attacks effectively without model retraining.

Keywords Adversarial attacks, Defense, Memory mechanism, Query-based

Introduction

Benefiting from the application of deep learning, the field of speech recognition has also been widely developed. However, deep learning-based automatic speech recognition (ASR) systems are shown to be vulnerable to audio adversarial examples (AEs), which add tiny perturbations on benign audio clips to fool the deep neural network model. Thus, how to secure ASR systems to prevent AE attacks remains a critical question.

Multiple mechanisms have been proposed to defend against audio AEs on ASR. Some methods mainly rely on signal processing skills such as smoothing, downsampling, reconstruction, and so on Cohen et al. (2019), Joshi et al. (2021), Zheng et al. (2021), Tamura et al. (2019). These methods can destroy the adversarial components of AE to a certain extent, and prevent them from reaching the preset target to reduce their impact on ASR. But it also destroys the benign sample and works for defense against unknown attacks. There are some works that train an additional DNN network as a prior part of ASR (Sun et al. 2018; Akinwande et al. 2020; Guo et al. 2020). However, those defense methods depend heavily on the

*Correspondence:
Yuxuan Chen
chenyuxuan@sdu.edu.cn
[1] School of Cyber Science and Technology, Shandong University, Qingdao, China
[2] Quancheng Laboratory, QCL, Jinan, China

algorithms for generating AEs, the generalization capability is the key that limits the ability of defense, and the model will be difficult to discriminate the adversarial samples without participating in the training. In addition, the existing defense methods against audio adversarial examples focus on the generation results of AEs, without on the process.

We reinvestigate and rethink the process of generating the adversarial examples, trying to locate the "specific" features in this process. We also scrutinize the current state-of-the-art attacks, including white-box attacks (Carlini and Wagner 2018; Yuan et al. 2018; Schönherr et al. 2018), black-box attacks (Khare et al. 2018; Han et al. 2019; Abdullah et al. 2019) and transfer attacks (Abdullah et al. 2021; Cheng et al. 2019; Richards et al. 2021). We note that the perturbation of the AEs in some attacks is quite light, and the distance between them and the benign examples is small without a particularly significant difference. So it is difficult to identify whether a single input is an AE. We often ignore the process of AE generation and only pay attention to the results. How to utilize this discarded information. Yet, except for some attacks that directly generate AEs, the majority need to keep visiting the target model to adjust the AE, essentially stealing key information (e.g., gradients) from the model. In this case, the adversary needs to send massive and similar queries to the target model in a period, which likely exposure her adversarial behavior. Therefore, according to this feature, we do not try to discover individual inputs, rather we focus on the relationship between the inputs to recognize the attack.

In general, for a regular user, the correlation between consecutive benign query sequences is relatively low. This is because repeating a query input itself is considered an abnormal behavior, and the probability of benign queries repeating is extremely low. At the same time, there is a significant variation among other benign queries, leading to a relatively low correlation between them.

In this work, we propose a universal and lightweight defense framework to infer the adversarial behavior by memory mechanism. The basic idea of our framework is that generating adversarial examples and the query to ASR models is continuous and correlated before and after. In contrast, a regular query is independent of others. We consider some history inputs of a certain length as a piece of memory, analyze the correlation between a new input and the memory, and mark the input as adversarial if the correlation crosses a certain threshold. We use the similarity of the audio fingerprint to estimate the correlation of the input. The insensitivity of the audio fingerprint to noise is an attractive trait. Meanwhile, since its simplicity, it is hard for the adversary to be aware of the use of defensive models. Furthermore, motivated by the similarity matrix for recommender systems, In this way, we can efficiently and quickly verify that the input query sound is adversarial or benign. We employ a non-neural network defense architecture and are not able to optimize the defense model in a similar way to a neural network, so an attacker may not be able to attack the defense model from that perspective. This strategy efficiently identifies the existing state-of-the-art adversarial sample attacks. The robust average uncovering success rates (*DSR*) are all above 90%. Also, our proposed framework can be easily combined with any other existing defense methods.

Finally, we study some adaptive attacks. We designed experiments with random noise attacks, which disturbed audio fingerprint feature extraction. For noise adaptive attacks, we observed that the modest level of random noise instead results in better performance to our defense system and we build a more robust defense system. In addition, we tested the potential role of different "*fake query*" ratios p_{fake} on the results. We conducted experiments on both types of adaptive attacks and proved that our defense framework remains robust under the damage.

The main contributions of this work are three-fold:

- We propose a new defense mechanism for adversarial audio attacks by analyzing the correlation between input with memory. This is the first proposed defense framework based on the AEs generation process for the ASR. The robust average uncovering success rates are all beyond 90% for existing attacks and we first evaluated the music-based AEs.
- We demonstrate the robustness of our defense framework toward adaptive attacks. We found that the adaptive attack methods of fingerprint extraction damage and the "*fake query*" are unable to evade our defense, and our defense strategy is still effective. We build a more robust defense system through the combination of a moderate level of random noise.
- We designed a music-carrier dataset that can be used to produce audio adversarial examples, which also establish a foundation for future research on attacks and defenses based on music-carrier. And we release the source code for our defense and datasets at: https://github.com/sveapp/Audio-denfense.

Background and related work

Adversarial examples (AEs) Adversarial attacks originate from images and quickly develop, with much relevant research. Many works achieve successful attacks on image classifiers by the computed gradient and these attacks are relatively convenient to implement (Goodfellow et al.

2015; Madry et al. 2017; Kurakin et al. 2016; Moosavi-Dezfooli et al. 2016). Some work explores transfer attacks from white-box to black-box models but needs a lot of access to the target model (Huang and Zhang 2019; Cheng et al. 2019; Richards et al. 2021). This provides a good reference for adversarial studies on audio. One may inquire about the reasons for the existence of adversarial examples. According to several works (Tsipras et al. 2018; Ilyas et al. 2019; Taori et al. 2020; Goyal et al. 2020), they think that adversarial examples are not a network drawback but a feature. The network attempts to learn "all" the beneficial features during the training process, whereas humans are naturally inclined to ignore some features. When an adversary attacks the model via manipulation of such features, it leads to a rapid decrease in the accuracy of the model, whereas the accuracy of humans is immune. Thus, our concern is not to remove the AEs and it fails to do so, instead, we should avoid the risk of the AEs to the model.

Audio adversarial attacks to ASR A similar situation exists in the ASR. Typically, a state-of-the-art ASR model is susceptible to deception by malicious AEs, which has evolved from a single-word attack to an attack on the entire sentence. Some state-of-the-art models were successfully attacked, Carlini and Wagner (2018) used CTC-loss to compute gradients to achieve an attack on DeepSpeech; CommanderSong (Yuan et al. 2018) used pdf-id to design a loss function to implement attack base on Kaldi[1]; Qin et al. (2019) implemented an attack on Lingvo[2] with psychological masking. For black-box attacks, the gradient is incomputable. However, Taori et al. (2019) successfully attacked the DeepSpeech black-box model with a genetic algorithm; Chen et al. (2020) successfully attacked four commercial speech API services (Google Cloud Speech-to-Text, Microsoft Bing Speech Service, IBM Speech to Text, and Amazon Transcribe); Zheng et al. (2021) successfully attacked the speech recognition API interfaces of iFLYTEK and Ali with the co-evolutionary algorithm. Besides, attacks can already be launched in the physical world. In order to enhance the robustness of physical attacks, in Taori et al. (2019), Chang et al. (2020), Chen et al. (2020), the authors added the Gaussian white noise to AEs and the evaluation results show that this strategy enhances the physical robustness of the AEs. Although they do not require a specific noise model, they may rely on the playback device and the experimental environment. These attacks inevitably require a massive amount of queries to models, and query-based attacks are becoming worse with time. In this article, the main object of our report will be focused on recognizing such attacks before they succeed and defending against query-based adversarial attacks.

Defense against audio adversarial attacks The majority of proposed methods of defense against audio adversarial attacks are removing or ruining the adversarial component by the technical tool of signal processing. Paper (Cohen et al. 2019) proposed random smoothing to mask the disturbing adversarial component. Joshi et al. (2021) proposed WaveGAN vocoder to reconstruct the waveform to eliminate the disturbing domain. Szegedy et al. (2016) used label smoothing, Rajaratnam and Kalita (2018); Zhang et al. (2019) squeezed the audio, Zheng et al. (2021) is the down-sampling method and Tamura et al. (2019) added distorted signals. These works of defense are concerned with removing or ruining the perturbation component. Those approaches have both advantages and disadvantages, as it breaks the adversarial behavior of AEs while also causing a lot of damage to examples of benign queries. Deficiency of hard evidence for the difference between AEs and benign examples. Some people suggested applying sub-models to preclude some attacks (Su et al. 2019; Sun et al. 2018). The literature (Akinwande et al. 2020; Guo et al. 2020; Samizade et al. 2020) applies extra neural networks to check adversarial examples to protect the ASR model. But they can only restrain some existing attacks, which are impotent to uncertain attacks. The applications are limited due to the sub-models bulky. Some methods based on state detection of images (Chen et al. 2019; Pang et al. 2020) also provide some guidance for the audio adversarial attacks. Although these defensive works are available for certain types of attacks, it is a deficiency that the evaluation of adaptive attacks is incomplete or oversimplified. No integral architecture is available for combination with other methods. We work mainly on building a lightweight framework that can be easily combined with other defense methods.

Problem setup Hereafter, we concentrate on adversarial tasks. In a setup like this, the DNN is represented as f, and $f : \boldsymbol{X} \rightarrow \boldsymbol{C}$ represents the given input $x(x \in \boldsymbol{X})$ is mapped to one of a set of classes $\boldsymbol{C}$, where $f(x) = c \in \boldsymbol{C}$. The DNN model is vulnerable to adversarial input attacks, which forces the DNN model to misjudge. Attacks on DNNs can be classified as targeted and untargeted. Here, we will focus on the setting of targeted attacks. Specifically, adversarial examples x^* are normally generated by slightly modifying x and $x^* = x + \delta$. The solve of δ can be converted to a min-optimization problem, i.e., $arg\ min\ \mathcal{L}(f(x+\delta), c^*)$. The adversary's goal is to force f to misclassify x^* as the target c^*, i.e., $f(x^*) = c^*, c^* \neq c$. To ensure that x^* is acoustically similar to x, the perturbation needs to be restricted to a limited range $g(x^* - x) \leq \varepsilon$, where the g

[1] http://kaldi-asr.org.

[2] https://github.com/tensorflow/lingvo.

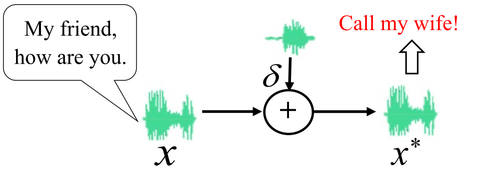


Fig. 1 The correct transcription of *x* is "My friend, how are you", and the adversary's purpose is to add a careful perturbation "*δ*" to *x* and then make it become *x** that can be transcribed as the target of "Call my wife"

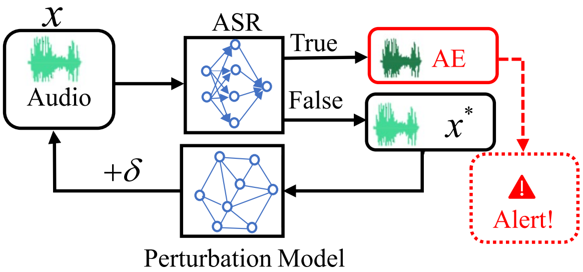


Fig. 2 Query-based attack: setting a target, for the first time *x**=*x*, if *x** can be transcribed as a target, the AE is true, else false, adjust the *δ* carefully, and perform the next query. Repeat this process until *x** can be transcribed as the target

is a measurement function of the auditory difference. The attack process is shown in Fig. 1.

Threat model Our defense scheme in this work focus on query-based audio AE attack, including both query-based white-box audio AE attack such as CS (Chen et al. 2020) and DS (Chang et al. 2020) and query-based transferable audio AE attack such as DW (Wang 2003). Note that we would not include audio AE attack that no query is needed in the attack such as Vaidya et al. (2015), since no query data could be analyzed in our approach. Also, if the attackers deploy a local white box model to attack target model, there is no query generated so we will not include such attack scenario. In our paper, we consider that in white box attack like CS (Chen et al. 2020) and DS (Chang et al. 2020), defender could still observe the queries sent to the model. For defense against query-based attacks, our aim is to increase the difficulty and cost of the attacker's attack, making it more expensive for them. While an attacker can query the model any number of times in trying to generate an adversarial example, our goal is to detect such attacks before they are successful. In the case of black-box API attacks, the attacker needs to apply for an API account beforehand, and many account applications require real-name authentication such as a phone number, credit card, ID, or passport. The attacker may create as many accounts as possible, but these conditions undoubtedly increase their cost. In addition, in our solution, the attacker's account will be banned when an attack is detected, requiring them to create a new account before continuing. This makes it difficult for the attack to continue and to some extent achieves the goal of protecting the system. However, when the attacker has unlimited resources, no defense method can stop them.

Because our defense scheme is based on the process of generating adversarial samples, which is a code-based process, we need to run attack code to simulate the process of the adversary's queries for experimental evaluation. We collected some open-source attack codes for experimentation, including the CS (Chen et al. 2020) attack, DW (Wang 2003) attack, ITRA (Chen et al. 2019) attack, and DS (Chang et al. 2020) attack. The CS attack targets the Kaldi aspire speech recognition model, the DW attack targets the speech recognition API, the ITRA attack targets the Lingvo speech recognition model, and the DS attack targets the DeepSpeech model. Therefore, our experiments mainly focus on these four models.

Defense against query-based audio adversarial attacks

A successful audio AE requires a specified carrier (the carrier can be music or dialogue) undergoing several iterations and queries. The process of AE generation is continuous. Every time, the adversary needs to produce a small disturbance δ to repeatedly adjust x^*. When crossing the decision boundary, a successful AE is done and the whole process is depicted in Fig. 2. Our defense is motivated by the process nature of query-based attacks. We can examine the query-to-memory relationship to determine if queries are intended to generate an AE, which is the process-based defense approach. To calculate the correlation C of the new query about the memory, we used the similarity F of the audio fingerprint to estimate the correlation, i.e. $C(q_{\text{memory}}, q_{\text{new}}) \approx F(q_{\text{memory}}, q_{\text{new}})$. For each query, audio has unique fingerprint information. The audio fingerprint is robust to noise and adapts to a noisy environment. Moreover, it can prevent audio splicing attack (Wang et al. 2003). According to the obtained fingerprints, we can figure out the similarity between the input query and the memory, which provides the foundation for our determination.

Defense architecture

Our defense architecture is a process-based defense approach and our goal is to find potential attacks in continuous queries. Suppose we have determined that the audio fingerprint similarity between the input query and memory is beyond the set threshold, we will report

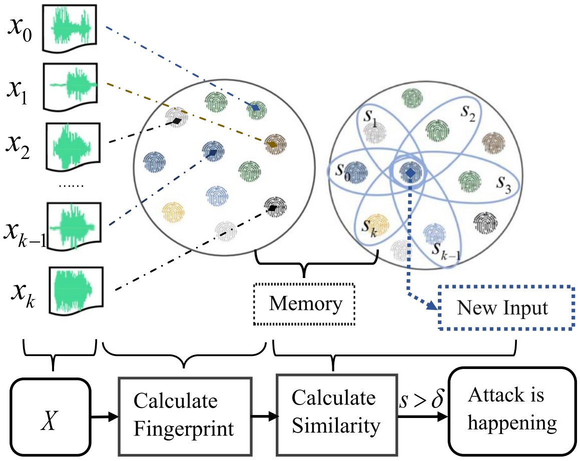


Fig. 3 Query-based defense: Architecture for recognizing query-based audio adversarial attack

it as part of the attack sequence and take action accordingly. We can take some actions such as blacklisting the querying user or warning the user. Figure 3 illustrates our scheme.

- *Firstly*, place query audio into the cache to form a query memory $\boldsymbol{X}$ of depth k. If the number of audio put into the cache is below k, consider all queries as a memory sequence. In the process of locating an attack, we expect to consume minimal resources and time, so k should not be too large. Also, it is disadvantageous to discover adversary behavior if k is too small. The k means the shortest depth before we can make sure that those input queries are intended to produce AEs.
- *Secondly*, calculating the fingerprints of all inputs in memory $\boldsymbol{X}$ and overwriting and updating the previous memory.
- *Thirdly*, for every new input audio, we calculate the weighted cosine similarity between the new input and each fingerprint in memory. Since audio fingerprint is a particular distribution about time and frequency, the cosine similarity can capture the correlation between such coordinate-dependent distributions. Besides, for each input, there is a necessity to check the legality, so we allocate a weight value α to each input with the *Inverse Variance Coefficient Method* (Marin-Martinez and Sánchez-Meca 2010). Then, calculate the similarity of the queries via:

$$\begin{cases} s = \sum_{i=1}^{k} s_i \alpha_i \rightarrow s_i(x, y_i) = \frac{x \times y_i}{\sqrt{x^2} * \sqrt{y_i}^2} \\ \sum_{i=1}^{k} \alpha_i = 1 \end{cases}, \quad (1)$$

where x is the fingerprint of the new input, y_i is a fingerprint in memory, and k is the depth of the memory $\boldsymbol{X}$. The final similarity value s is the weighted average value of s_i. The selection of the α_i value is explained in the next section.

- *Fourthly*, obtain threshold δ, which implies minimal constraints regarding the input as malicious. When $s > \delta$, it demonstrates that the current input is a potential attempt at generating an AE, and appropriate measures must be taken immediately. In practice, for the setting of δ, it is important to have a high uncovering success rate as well as a low false positive ratio. Usually, the false positive ratio will be limited to no more than 10% of the training data, according to the size of the training data set (Xu et al. 2017; Chen et al. 2019). The details of k and δ are explained in the next part of this section.

Memory sequence

A memory sequence $\boldsymbol{X}$ consists of several queries that are placed in the cache. In the process of attack detection, we expect to consume minimal resources and time. So $\boldsymbol{X}$ should not be too large. Also, it is disadvantageous to detect adversary behavior if $\boldsymbol{X}$ is too small. $\boldsymbol{X}$ of depth k means the shortest sequence before we are sure that those queries are intended to produce AEs, and the length of the sequence is k, i.e.

$$\begin{cases} k = \min\left(f(1), f(2) ... f(n)\right) \\ f(i) = \begin{cases} i, \ if \ f \text{ can detect attacks.} \\ +\infty, \ if \ f \text{ can not detect attacks.} \end{cases} \end{cases} \quad (2)$$

where *f* is the detection function, *f*(*i*) indicate whether the function f can detect a sequence of length i. Equation 2 implies that the depth of $1, 2, ...k - 1$ is not sufficient for $\boldsymbol{X}$ to be considered as the intention of generating AEs; depths $k, k + 1, ...n$ are considered to be for the purpose of generating AEs, with the minimum depth is k. We explain how to choose the value of k in parameter selection.

Query audio fingerprints similarity

The auditory similarity is an important feature in estimating the gap between humans and machines. There is a close auditory similarity between the malicious examples and the benign examples. The malicious examples are produced by appending carefully structured small perturbations to the benign carriers. Although the neural network regards them as two completely different classes, humans believe them as the same intuitively. So the trait of keeping intuitively consistent with humans is what we need. The audio fingerprint has this trait and is not sensitive as the DNN to perturbations. Fingerprints will maintain high similarity if humans believe they are the same samples.

It is possible to predict whether new input might have a strong correlation with the memory and whether they share the same behavioral attributes, according to the similarity computation between the preserved fingerprints and the new one. This is similar to the recommender system (Song et al. 2021; Nam 2022), which differentiates users based on their memory behaviors and recommends new content or products (Afchar et al. 2022; Shafiloo et al. 2021).

We note that the digital audio fingerprint (Haitsma and Kalker 2002; Wang 2003) uniquely flags audio. The small noise of the audio doesn't bother the core information of the fingerprint. And it can defend against some attacks such as audio patching. Moreover, it is reliable and feasible in implementation cost to employ fingerprint similarity as an audio similarity. Fingerprint similarity relies on the following requirements, assume that s is the similarity function, x, y, z are three candidates in D dimensional space that satisfy Eq. 3, Eq. 4, Eq. 5, Eq. 6.

$$s(x,y) \geq 0, (Non - negativity) \tag{3}$$

$$s(x,y) = 1,\ only\ x = y.(Homogeneity) \tag{4}$$

$$s(x,y) = s(y,x).(Symmetry) \tag{5}$$

$$s(x,y) + s(x,z) \geq s(y,z).(Triangularinequality) \tag{6}$$

A robust acoustic fingerprinting algorithm needs to consider the perception of the audio. When two audio files sound the same, their acoustic fingerprints should be the same or very close, even if there are some differences in their file data.

According to the literature (Haitsma and Kalker 2002; Wang et al. 2003). The fingerprint similarity can be divided into two steps: **fingerprint extraction** and **similarity calculation**.

Audio corresponds to a unique fingerprint, so the relationship between digital audio fingerprint $\boldsymbol{F}$ and audio object $\boldsymbol{X}$ is a surjection $h : \boldsymbol{X} \to \boldsymbol{F}$, and only when $\forall f \in \boldsymbol{F}, \exists x \in \boldsymbol{X}, \to f = h(x)$. That expands to $\{x_1 \to f_1, x_2 \to f_2 ... x_n \to f_n\}$ or $\{f_1 = h(x_1), f_2 = h(x_2) ... f_n = h(x_n)\}$. For fingerprint $f_i, f_j \in \boldsymbol{F}$, we can obtain similarity s_{ij} ($s_{ij} \in \boldsymbol{S}$) and $g : \boldsymbol{F} \to \boldsymbol{S}$ is surjection only when $\forall s \in \boldsymbol{S}, \exists f_i, f_j \in \boldsymbol{F}, \to s = g(f_i, f_j)$. h, g is the map function.

- *Fingerprint extraction* ($h : \boldsymbol{X} \to \boldsymbol{F}$). The fingerprint extraction process is illustrated in the fingerprint extraction module in Fig. 4. The main procedures include:

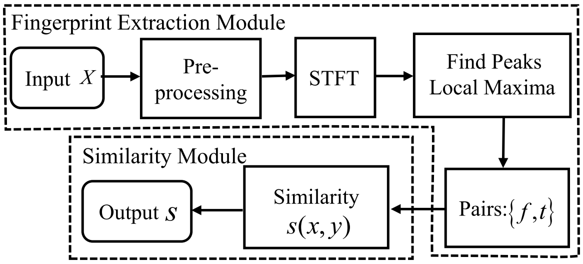


Fig. 4 Architecture of fingerprint similarity calculation

(1) Preprocessing: it mainly involves frame split and filtering of the input data.
(2) STFT: short-time Fourier transform. For each frame, apply STFT via Eq. 7, where $x(t)$ is the input signal at time t, $h(t-\tau)$ is the window function, and $S(\omega,\tau)$ shows the spectral result if the center of the window function is τ.
(3) Find Peaks: after STFT, select the frequency peaks f and corresponding time t, and make sure the distribution of frequency peaks is uniform.
(4) Pairs: pair the obtained frequency peaks f and time t, then the result $\{f,t\}$ is used as fingerprints f_i and f_i is a high-dimensional vector of a certain length.

$$S(\omega,\tau) = \sum_{t=-\infty}^{\infty} x(t)h(t-\tau)e^{-j\omega m} \tag{7}$$

- *Find Peaks.* In Fig. 4, after calculating the STFT, we need to uniformly select the peak in the frequency domain. Equation 8 describes this process, in which $F(n,\ m)$ is the two-dimensional matrix after STFT, $H(u,\ v)$ is the kernel function. Equation 9 is the maximum filter and Eq. 10 is the high-pass filter for resetting the frequency to 0 when the frequency is below the cutoff D_0. Both filters are useful for canceling low-frequency components and uniformly capturing the local maximum high frequencies. We choose the former as a tool to find peaks.
- *Similarity calculation* ($g : \boldsymbol{F} \to \boldsymbol{S}$). After fingerprint extraction, fingerprint f is obtained, which is written as $x = f_i$. Similarly, another fingerprint can be written as $y = f_j$ and its length is the same as x. Then calculate the similarity s between them. The process is illustrated in the Similarity Module in Fig. 4. The fingerprint contains coordinate-dependent details. Finally, the similarity of x, y could be achieved by Eq. 1.

$$G(u,v)=\frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}F(n,m)H(u-n,v-m) \tag{8}$$

$$H(u,v) = \max_{s,t\in N(n,m)}[F(s,t)] \tag{9}$$

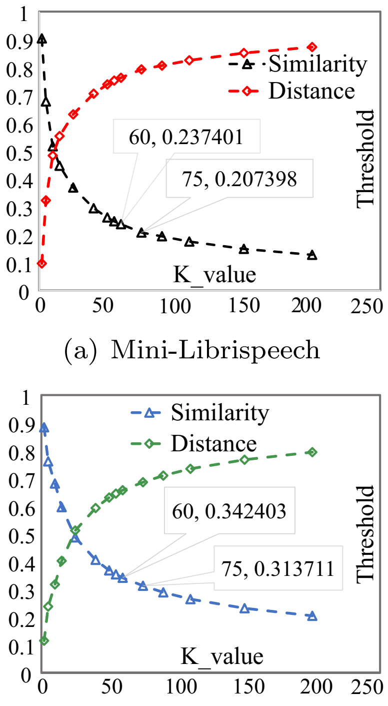
$$H(u,v) = \begin{cases} 0, D(u,v) \le D_0 \\ 1, D(u,v) > D_0 \end{cases} \tag{10}$$

Parameter selection

- *The choice of k and δ.* The larger the k value, the more effective our solution is in observing input queries, and the smaller the k value, the lower the computational cost. The k is the minimum depth of memory before we are sure that those inputs are intended to produce an AE. The δ is the minimum similarity before we determine that the current input is malicious. So the values of δ would be influenced by the depth of k. Specifically, establishing the threshold requires evaluating fingerprint similarities under the datasets, so that if the entire datasets were to be randomly streamed as queries, 0.1% of the carrier datasets would be marked as attacks. (In theory, the percentage of false positives should be limited to 10% of the dataset size, but since our dataset is small, our value is 100 times smaller than the default.)

Actually, the threshold δ is a function of k, and Fig. 5 discloses their relation. The smaller the threshold δ, the more intense the constraints on the input. Hence small thresholds are advisable, but the too-small value risk regards a benign input as malicious. From what we observed from Fig. 5 with the increase of k, the similarity drops sharply in the beginning. (In turn, the distance rises, rapidly. The higher the similarity, the lower the degree of dissociation between input queries, i.e., the closer the distance.) After it reaches around $k = 75$, curves become smooth and increase modestly with k, and the process is quite gentle, so we set up k as 75 and the thresholds δ in both datasets are 0.313711 and 0.207398.

We investigate large-scale attacks on ASR systems. As can be seen from Table 2, the number of queries for all attacks is significantly greater than k (k =75). The value of k is defined as the minimum memory sequence required to detect an attack with a probability higher than 0.9 (false positive rate of 0.1%) when the similarity exceeds δ. In other words, the probability of detecting an attack is greater than 0.9 when the similarity surpasses δ at least k times. The relationship between k and δ is functional. In the experiments, many attacks can be identified as such after querying k times, whereas some attacks, such as DW attacks (See for more details), require almost 80 queries for the similarity to exceed δ. Moreover, if we increase the false positive rate, the value of k will decrease, meaning fewer queries are needed to determine whether a query is an attack.



(a) Mini-Librispeech

(b) Music-sets

Fig. 5 *k* and *δ*: the mean *k* number of the 0.1% percentile of the datasets as a function of k

- *The choice of α.* First, let's consider a case, in Eq. 11 below:

$$\begin{cases} \text{s}_1\text{= f(X}_A\text{,q}_{new}\text{),X}_A = \{\text{q}_0\text{,q}_1\text{...q}_m\text{,q}_n\} \\ \text{s}_2\text{= f(X}_B\text{,q}_{new}\text{),X}_B = \{\text{q}_0\text{,q}_1\text{...p,q}_n\} \end{cases} \tag{11}$$

There exist two memory sequences $\boldsymbol{X}$, where memory X_A consists of $\{q_0,q_1...q_m,q_n\}$ and X_B is: $\{q_0,q_1...p,q_n\}$, s_1 and s_2 are the similarity of the two sequences with new input, f is the fingerprint similarity function. The key distinguishing element between X_A and X_B is that the query q_m differs from p. Assuming that p is a query deliberately placed in the queries by an adversary. The adversary's purpose of injecting p is to try to fabricate a fake input (i.e., almost irrelevant to the former) to confuse the analysis of the similarity and hide her intent. Essentially, both X_A and X_B are malicious memory sequences with only trivial disparity. But s_1 is below the threshold and s_2 is beyond the threshold, X_A is decided as a potential attack while X_B is not decided as a potential attack due to the injection of a fake query. We call this p-input as "*fake query*", and the ratio of "*fake query*" to all queries is called p_{fake} (p_{fake}= (p/k) $*$ 100%). In our experiments, we found that the s value would change sharply when there were "fake queries" in the query memory and we employed the *Inverse Variance Coefficient Method* (Marin-Martinez and Sánchez-Meca 2010) to describe such fluctuations and disparities. According to this method, it is easy to determine the weights α, which are assigned as follows:

$$\alpha_i = 1 - \frac{std_i(s(l))}{mean_i(s(l))} \rightarrow \alpha_i = \frac{1}{\alpha}\alpha_i(\alpha = \sum_{i=0}^{k} \alpha_i), \quad (12)$$

where $mean_i$ depicts the mean, std_i depicts the standard deviation, and $s(l)$ depict getting the query vector of length $l/2$ before and after the i-query. For l, we set the maximum value as 7 (No more than 10% of the memory length, i.e. $l_{\max} = floor(0.1 * k) = floor(0.1 * 75) = 7$) and l begin with 2 (The mean and variance are worthwhile at least two values). Then, the value increases linearly. When it exceeds the maximum value, l shrinks to half of the original value and then increases linearly duplicate. Repeat this process until all elements are traversed.

Evaluation

In this section, we will show the evaluation results of our scheme for some non-adaptive attacks and adaptive attacks. We collected open-source code attacks as much as possible, and we did not evaluate attacks without open-source code, but we made some surveys about their details. Finally, we evaluated four class attacks that are well-known in the audio adversarial attack. Those are sufficiently representative and the bulk of the other work revolves around them. We evaluate the CommanderSong (CS) (Yuan et al. 2018) attacks and the Devil's Whisper (DW) (Chen et al. 2020) attacks by applying the Music-set. The Mini-Librispeech dataset is applied to assess the IRTA[3] attack (Qin et al. 2019) and DS[4] attack (Carlini and Wagner 2018). Those attacks all reported a success rate of attacks (SRoA) of almost 100%.

Datasets

Our scheme conducts experiments on Mini-Librispeech[5] and Music-sets datasets (We build a carrier library of music-based samples containing 10,553 music clips. Appendix Music-sets contains all details about Music-sets). For Mini-Librispeech, this is a dialog-based dataset that some classic attack works rely on it and we cannot ignore it Taori et al. (2019); Han et al. (2019); Khare et al. (2018). For Music-sets, music has the characteristic of large-scale availability in most situations, and its accessibility and popularity allow it to become a candidate of the carrier in attacks. Lots of strong attacks (Yuan et al. 2018; Carlini and Wagner 2018; Chen et al. 2020; Schönherr et al. 2018; Zheng et al. 2021) refer to music as the necessary carrier for producing AEs. So, defense and evaluation of the AEs on musical carriers are inevitable and important.

In our approach, when the correlation between consecutive queries exceeds a certain threshold, it is considered as queries submitted by an attacker. For determining the threshold, we rely on two datasets: a conversation dataset and a music dataset. Although both datasets have the same k value (75), the δ is different. In real-world attack scenarios, the defender cannot know whether the attacker used conversational data or music data as the carrier for AEs. However, there are some simple methods to distinguish whether an audio segment is conversational or musical since there are significant differences between conversational and musical data regarding energy variation, spectrum, rhythm, and other features. We evaluate several attacks on both types of data, where the carrier category used by the attacker is presumed to be known in advance. In instances where the carrier category is unknown, reducing the δ value to increase sensitivity is one approach, and feature differentiation is also an effective method.

Evaluation metric

- *DSR*. To evaluate the effectiveness of our approach for defending the query-based attacks, we employ the detection success rate (*DSR*) and First-Signal-to-Noise Ratio (*FSNR*) as the evaluation metrics. The detection success rate (*DSR*) is the most intuitive metric to evaluate the detection results. To calculate it as follows:

$$DSR(\%) = \frac{\mathrm{d}_n * k}{a_n} \times 100\%, \quad (13)$$

[3] IRTA is an abbreviation for the attack of the paper "Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition"

[4] DS is an abbreviation for the attack of "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text". This paper attacks the DeepSpeech model.

[5] https://www.openslr.org/31/

where d_n is the number of detections, a_n is the number of queries, and k is the length of memory X. Obviously, the *DSR* value is below 1 because $a_n > d_n * k$ is clear. The detection occurs after performing at least one query. For our purposes, we consider it to measure the probability of finding adversary behavior. A higher *DSR* is preferable.

- The First-Signal-to-Noise Ratio (*FSNR*) is a function that defines the minimum *SNR* to detect an attack, i.e., how much *SNR* when we can detect the attack, as shown in Eq. 14:

$$FSNR(dB) = 20\log_{10}\left(\frac{A_x}{FA_\delta}\right), \tag{14}$$

where x is the original sound, δ is the perturbation, A_x is the amplitude of the original sound, and FA_δ is the amplitude of the perturbation when the first attack is detected. This is a metric of the relative value of distortion of the AE vs the original sound. The higher *FSNR* describes that the query will be regarded as a suspect under a smaller perturbation.

Non-adaptive attack evaluation

We evaluate four class attacks that are well-known in the audio attack. Those are sufficiently representative and the bulk of the other work revolves around them. We evaluate the CommanderSong (CS) (Yuan et al. 2018) attack and the Devil's Whisper (DW) (Chen et al. 2020) attacks by applying the Music-set. The Mini-Librispeech dataset is applied to assess the IRTA attack (Qin et al. 2019) and DS attack (Carlini and Wagner 2018). Those attacks all reported a success rate of attacks (SRoA) of almost 100%. CS attack is the representation of employing music as carriers and some subsequent works (Chen et al. 2020; Zheng et al. 2021) set it as an indispensable collection. The DW attack is the typical instance for commercial black-box APIs. Subsequently, much of the works (Han et al. 2019; Abdullah et al. 2019) on black-box attacks has to test on APIs. IRTA attack based on the psychoacoustic hiding model is an outstanding work of the period. And several studies (Schönherr et al. 2018; Abdullah et al. 2021) adopted the psychological masking effect. DS attack is the earliest version of voice attack, which launched the gateway to voice attack and provided a reliable infrastructure for the subsequent works.

- *N1. CS attack evaluation* CS attack is a white-box attack by injecting target commands into the song. It started a precedent of producing AEs with music as a carrier and achieving a 100% success rate of attacks (SRoA) on the Kaldi speech recognition system. It has a profound influence, and many follow-up works set it as an indispensable reference. For the defense based on our approach, there are few blanks in the music, the spectrum is abundant, and the fingerprints are often more reliable than those of the dialogue version. Table 1 shows that CS examples spend an average of about 300 visits to the target model. Our security architecture can accurately detect such attacks with *DSR* up to 98%. However, the value of *FSNR* is only 7.38 dB, revealing that the AEs were already very noisy when we suspected the query was an attack. The primary factors of this situation are that the small perturbation is not ideal for a CS attack and the perturbation is constrained to a very broad range. Therefore, the amount of additional noise is significant. Apart from that, various audio lengths will affect the SRoA of AE. To ensure the validity of AE, the length of audio ought to be no shorter than 4 s. The longer the audio, the richer the fingerprint, which is more helpful for detection. However, the shorter audio is not beneficial for the adversary to generate AEs successfully.
- *N2. DW attack evaluation* DW attack first accomplished a black-box attack on commercial speech recognition APIs (including Google Assistant, Google Home, Amazon Echo, and Microsoft Corina). Since then, attacks on APIs have gradually become a necessary option for black-box attacks and the most intui-

Table 1 Non-adaptive attack evaluation. SRoA denotes the success rate of attack

Attack	Dataset	SRoA(%)	Avg.Queries(n)	Detections	DSR(%)	FSNR (dB)
CS	Music-sets	100.00	∼300	∼3.92	98.00	7.38
DW	Music-sets	98.00	∼150	∼1.7	84.74	18.41
Average		100.00	∼225	∼2.81	91.37	12.90
IRTA	Mini-librispeech	100.00	∼5000	∼56.00	84.00	40.97
DS	Mini-librispeech	100.00	∼1000	∼11.00	82.50	13.02
Average		100.00	∼3000	∼34.00	83.25	27.00

The higher the value of *DSR* and *FSNR*, the more beneficial. Normally, every k (k=75) query is detected once, and if the queries are less than k, at least one detection is performed for all n queries, and the ratio of n/k is the detections tive indicator of the attack algorithm. Table 1 shows that DW also works based on the music dataset, which accounts for 50% of CS in the average query to the target model and SRoA is close to 98%. On defense, our approach enables a *DSR* of 84.74% under DW attack. DW attack employs a local substitution model to simulate approximately the target model of the APIs ASR system. It helps to diminish the number of queries and the likelihood of triggering detection. So *DSR* possible losses. The *FSNR* value is 18.41*dB*, which is about 2.5 times that of CS. DW increases the *FSNR* value by reducing the number of visits to the model, and the perturbation naturally decreases.

DW adopts Noise Model to augment the physical robustness of AEs. However, the SRoA is deeply relevant to the environment and the device. Regarding the noise model, the combination of our scheme with some straightforward measures (e.g., down-sampling, filtering) can raise the level of difficulty of physical attack.

- *N3. IRTA attack evaluation* IRTA attack is a two-stage attack algorithm on Lingvo, concealing target commands to a space that the human ear cannot hear through a psychoacoustic masking model. The IRTA example is based on the open-source dataset Librispeech. This type of dialogue audio contains a large number of silent fragments. Therefore, the fingerprint of the audio is inferior to that of the music. But the inspiring thing is that our approach maintains a robust attack detection and that the *DSR* reaches 84%. This can be attributed to the time cost of this type of attack (Producing a successful adversarial example costs 24.8h) leads to a remarkable number of queries. Such massive queries easily provoke the inspection of the defense system. Moreover, the perturbation is very small, and the *FSNR* can reach 40.97dB in which the psychoacoustic masking model plays an important role. Still, the perturbation would reflect the frequency domain and the fingerprint extraction happens in the frequency domain. We can further presume that it will be costly to bypass our defenses for adversaries with an emphasis on hidden perturbation via psychoacoustic masking. Nevertheless, it also exposes a critical concern: *In the areas that humans fail to hear, is there a necessity for the machine to do so?* AI researchers aim to narrow the gap between humans and machines, so machines should also appear human-like for regions beyond human perception. Blocking such attacks implies that the machine does not have the power to do anything in the regions where humans are unable to perceive, thus, the attack will completely dissolve.
- *N4. DS attack Evaluation* DS attack is a type of attack first implemented on DeepSpeech. At its core is to optimize the CTC-Loss function. Compared to IRTA attacks, DS is relatively heavily perturbed that maybe without applying the theory of psychological masking, and relatively poorer *FSNR* but *DSR* is 82.5% closer to IRTA. Compared to CS and DW attacks, DS and IRTA attack are implemented on Librispeech containing rare fingerprint information, so *DSR* is inferior to CS and DW. Nevertheless, the general *FSNR* is superior to the former, showing the method's detection capability to attacks with small perturbations. Separate work deploys genetic algorithms and gradient estimation to generate adversarial samples. However, gradient estimation relies on the sampling theory. Biological evolutionary algorithms demand substantial expenses without the guideline of the gradient. The literature (Taori et al. 2019) queries numbers up to 1000+, and the literature (Zheng et al. 2021) reach a stunning 30000+. From Table 1, it has a remarkably higher detection rate for query numbers above 1000+. Multiple query numbers are an obvious disadvantage of the evolutionary algorithm. Unless improving this shortcoming, do not expect to evade our inspection.

We investigated the perturbation level of AEs so that we can easily compare them with *FSNR*, as shown in Table 4.

- *N5. other query-based attacks evaluation* Other query-based attacks, the majority of them are based on the 4 attacks above. CS attack is the representation of employing music as the carrier. After that, subsequent work (Chen et al. 2020; Zheng et al. 2021) also set it as an indispensable collection. The DW attack is a typical example of attacking commercial black-box APIs. Subsequently, a lot of the work (Han et al. 2019; Abdullah et al. 2019) on black-box attacks has to be tested on APIs. IRTA attack based on the psychoacoustic hiding model is an outstanding work of the period. Several studies (Schönherr et al. 2018; Abdullah et al. 2021) adopted the psychological masking effect. Literature (Wang et al. 2021; Du et al. 2020) using biological evolutionary algorithms to perform attacks and optimize the number of queries. DS attack is the earliest relatively sophisticated version of an audio attack, which provides a reliable infrastructure for subsequent works. Since our defense framework is process-based, we were unable to evaluate the attacks without open-source code but

Table 2 An overview of the query-based attacks against ASR

Attack	Task	Attack method	Attack model	Target	M or D	Avg.Queries	SRoA(%)
CS Yuan et al. (2018)	ASR	GD	Kaldi-Aspire	Play music. Open the front door. Turn off the light.	M	~300	100
DS Carlini and Wagner (2018)	ASR	GD	DeepSpeech	Okay google browse to evil dot com.	M & D	~1000	100
DW Chen et al. (2020)	ASR	Alt-M	APIs	Turn off The Light Take a picture. Call 911.	M	~150	100
DSG Taori et al. (2019)	ASR	GA & GE	DeepSpeech	Morning body. Ball charge. More they.	D	~150000	35
Foolgle Han et al. (2019)	ASR	GA	Google-API	–	D	–	86
SGEA Wang et al. (2021)	ASR	SGE	DeepSpeech	Thank you. Hello world. Open the door.	D	~78000	98
IRTA Qin et al. (2019)	ASR	Psy-M	Lingvo	Old will is a fine fellow but poor and helpless sin -ce missus rogers had her accident.	D	~5000	100
PHA Schönherr et al. (2018)	ASR	Psy-M	Kaldi-WSJ	Do not blame you. The command is planted. The cake is a lie.	M & D	~500	98
EPA Abdullah et al. (2021)	ASR	Psy-M	DeepSpeech and Wav2Letter	That is comparatively nothing. Talking later is beneath us. But there seemed no.	D	~1000	76
Occam Zheng et al. (2021)	ASR	Co-E	DeepSpeech and APIs	Call my wife. Navigate to my home. Open the door.	M & D	~30000	100
SirenAttack Du et al. (2020)	ASR	PSO	DeepSpeech	Read last sms from boss. Call the police for help.	D	~1000	100
MOGA-Attack Khare et al. (2018)	ASR	Mul-Obj GO	DeepSpeech and Kaldi	A cat. All of these. That i love you.	D	-	*

In the table, "GD", "GA", "GE", "SGE" represent the Gradient Descent, Genetic Algorithm, Gradient Estimation, and Selective Gradient Estimation. "Alt-M", "Psy-M", "Co-E", "PSO", "Mul-Obj GO" represent the Alternative Models, Psychoacoustic Masking, Co-evolutionary algorithm, Particle Swarm Optimization, Multi-Objective Genetic Optimization. "M or D" represents the Music-carrier or Dialogue-carrier, "–" denotes the author didn't show, and "*" denotes the author told us the WER of the attack model to AEs was increased to 980%

still surveyed them. More relevant details are provided in Table 2.

We can learn from the above that applying a music carrier is quite advantageous for detection, also the detection is significant when the number of queries is numerous. The critical factor is that the fingerprints of music are more obtuse to perturbations, while the conversational ones are not. In terms of fingerprint extraction, Fig. 9 from the Appendix supports similar results. In the following, we built a more robust defense system that raises the average DSR beyond 90% and substantially strengthens our defense, Table 5 shows the results. For adversaries, unless improving those shortcomings, do not expect to evade our inspection. Below, we propose a more robust defense system by combining other methods, which can achieve a detection ratio of over 90%, The details are in.

Adaptive attack evaluation

Whereas our defense framework can effectively detect existing attacks, it only assures in *"zero-knowledge"* attack scenarios where the attacker is unknown of the existence of the defense framework. In order to reliably implement our framework in practice, we have to assess adaptive adversaries who understand the defense details entirely and intend to deploy some strategies to bypass the defense mechanism. Following the guidelines of Carlini et al. (2019), we designed adaptive attacks to evaluate the ability of our defense to adaptive attacks. According to the defense details we consider both adaptive attacks: *Random Noise attack* and *Proportion of Fake Queries attack.*

- *A1. random noise attack* We conceive an adaptive attack of corrupting fingerprint extraction. Ran-

Table 3 *DSR* as a function of the p_{fake}

Attack/(Sets)	DSR (%) p_{fake}=0%	DSR (%) p_{fake}=10%	DSR (%) p_{fake}=25%	DSR (%) p_{fake}=40%	DSR (%) p_{fake}=50%	DSR (%) p_{fake}=60%
CS (Music-sets)	98.00	98.00	98.00	76.20	76.20	5.44
DW (Music-sets)	84.74	79.66	79.66	57.20	33.47	0.00
Average	91.36	88.82	88.82	66.70	54.84	2.72
IRTA (Mini-Librispeech)	84.00	84.00	84.00	84.00	84.00	3.00
DS (Mini-Librispeech)	82.50	81.75	81.75	81.00	80.25	0.00
Average	83.25	82.88	82.88	82.50	82.13	1.50

The effect of different false query ratios on the success rate of detection.

Table 4 Perturbation levels for different attacks (The numbers in the table are the outcome after normalization)

Constraint	$\|\delta\|_1$	$\|\delta\|_2$	$\|\delta\|_\infty$
CS-Attack	346.85	23.62	0.24
DW-Attack	198.21	2.60	0.05
Average	272.53	13.11	0.15
IRTA-Attack	169.58	0.80	0.02
DS-Attack	63.09	0.37	0.37
Average	116.34	0.59	0.20

Shows perturbation levels for different attacks. The higher the level of perturbation, the smaller the *FSNR*

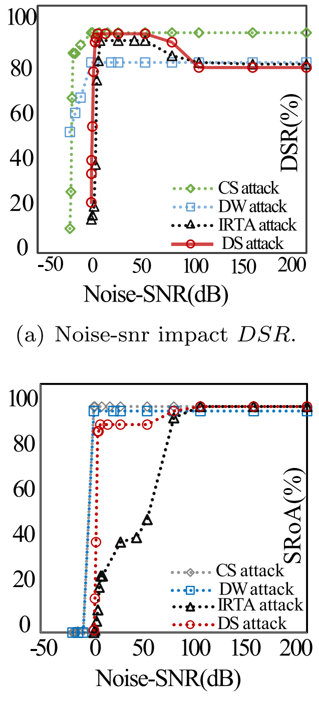
domly insert noise with different SNR to the audio in the process of query. Forcing the x^* to bypass the defense, and successfully attack the ASR, and the perturbation is not easily perceived by the human. In Fig. 6, according to audio quality theory, when *SNR* is above 70, it belongs to high-fidelity quality audio. When $SNR = 0$, the noise has the same energetic value as the original audio, so when *SNR* is below 0, the original audio is almost flooded with noise. We also test the success rate of the audio AEs with added noise under different level (refers to SNR). The results show that when SNR is larger than 25 dB, the adaptive attack could achieve same attack success rate as the original attack. Under these settings, our defense could still get an effective DSR rate. Under SNR=0 setting, although our DSR rate reach a lower level, the auditory imperceptibility of the audio AEs would be very poor which is unacceptable for the adversaries (audio AEs in previous work never reached such a low SNR).

When Noise-SNR>0, the SRoA and *DSR* are rapidly recovering to their maximum value and keep it and the SRoA, in other words, *DSR* displays a comparable consistency. Though large noise decreases the *DSR* value but also decreases SRoA, which diverts from the adversary's target. So it is impossible to achieve superior SRoA while trying to break our defense. However, when the Noise-SNR value gradually increases, for IRTA and DS attacks, SRoA is rapidly recovering to its maximum value and keeping it except IRTA attack recovery is slower and the *DSR* value sharply rises and then gently drops until it becomes peaceful. Since Mini-Librispeech is a dialogue-based dataset and it contains a lot of blank frames, when inserting noise, it will fill the blank and become more helpful to the extraction of fingerprints. It can be deduced that joining appropriate noise can improve the robustness of our method. The query of containing noise does not undermine our defenses, on the contrary, it leads the defense system more sensitive and robust.

Table 5 Robust defense: we add noise based on different *SNR*, the lower the SNR, the heavier the added noise

SNR(dB)	CS-attack	DW-attack	IRTA-attack	DS-attack	Average
150	100/3.92/98.00	100/1.70/84.74	100/56.00/84.00	100/11.00/82.5	100/17.44/87.31
100	100/3.92/98.00	100/1.70/84.74	100/56.50/84.75	100/11.00/82.50	100/17.57/87.50
75	**100/3.92/98.00**	**100/1.70/84.74**	**100/58.50/87.75**	**100/12.50/93.75**	**100/19.16/91.06**
50	**100/3.92/98.00**	**100/1.70/84.74**	**100/63.00/94.50**	**100/13.00/97.50**	**100/20.41/93.69**
25	**100/3.92/98.00**	**100/1.70/84.74**	**100/63.00/94.50**	**100/13.00/97.50**	**100/20.41/93.69**
0	15.10/3.70/92.50	22.50/1.50/75.00	0/10.00/15.00	5.20/3.00/22.50	10.70/4.55/51.25

"100/3.92/98.00" indicates that the attack success rate is 100%, average queries are 3.92, and *DSR* is 98.00%



(a) Noise-snr impact DSR.

(b) Noise-snr impact $SRoA$.

Fig. 6 Adaptive attack: Different noise-snr to disturb the extraction of fingerprints. Noise-SNR indicates the noise of different SNR. The smaller Noise-SNR means higher noise level

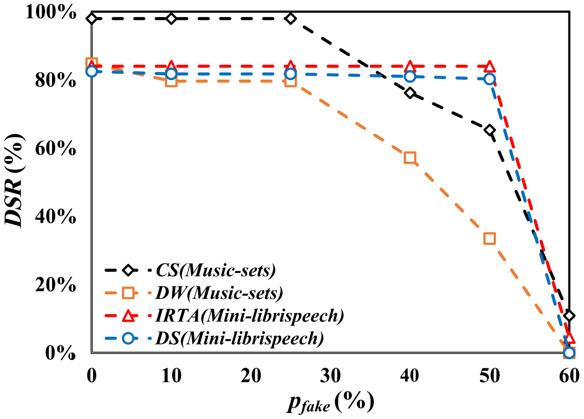


Fig. 7 *DSR* as a function of the p_{fake}

- *A2. proportion of fake queries attack* Moreover, we noted above that some adversaries use "*fake queries*" to develop a fake query history. In this section, we evaluate the impact on the defense system for different proportions of "*fake queries*" (p_{fake}). Table 3 plots the results. It also can be intuitively understood from Fig. 7. As observed, there is a critical threshold p_{fake} for the defender: once p_{fake} exceeds this threshold, the *DSR* drops dramatically. For these attacks, if $p_{fake} \geq 60\%$, *DSR* drops to approximately 10% or 0%. For CS and DW attacks, the *DSR* linearly dropped when $p_{fake} \in [25, 50]$. However, for the other two attacks, this situation does not happen. An intuitive explanation of this can be as follows: p_{fake} mainly affects the estimation of the query of interest for defense; yet, the priority of our defense is to distinguish the authenticity of the query, p_{fake} tends to have a larger impact on our proposal.

  The AE carriers employed by CS and DW attacks are music, while IRTA and DS attacks use conversationally. The fingerprint information of the music is richer than that of the conversational (as can be intuitively observed in Fig. 9). This implies that the fingerprints of music have more features for matching when computing similarity, but conversational has fewer matching features. As the proportion of fake queries gradually increases, the impact on the similarity of music fingerprints is heavier than that of conversational. Therefore, the sensitivity of the two kinds of datasets to fake queries is different. The adversary's strategy to evade detection would probably be to set up p_{fake} to a sufficiently high value (e.g., $p_{fake} \geq 60\%$), but this would dramatically raise the cost of the attack and the number of queries. This increases the cost and pressure on the attackers, as they are uncertain whether they can ultimately obtain a successful AE to attack the target model.

  Fig. 7 shows the effect of p_{fake} on *DSR*.
- *A3. siscussion of other strong adaptive attacks* Intuitively, attacker could use other strong adaptive attacks to break our defense. Firstly, the attacker could modify the loss function for generating adversarial examples to include the fingerprint similarity score between the current speech and the speech in the local memory, which could reduce the fingerprint similarity between adjacent submitted speeches then bypass our defense. Secondly, the attacker could apply improved random noise adaptive attack such as EOT method in Chen et al. (2022), which could solve the problem of search direction fluctuations caused by randomness and generate more robust AEs to break the defense. We suggest future advanced defense which is based on audio fingerprint method could focus on these strong adaptive attacks to improve the defense robustness.

Robust defense In the *random noise* adaptive attack and Fig. 6, we found that the appropriate level of noise could help us build a more robust defense system, so we further studied the subtle relationship. In Table 5, we set up six different noise levels. The audio belongs to high-fidelity quality audio when $SNR > 75$ and the noise is extremely slight. Once the noise gradually rises to $SNR = 75$, our defense system can achieve more than 90% detection success rate for all attacks; when the noise rises to $SNR = 50$, the detection success rate reaches the maximum (and the average is 93.69%). The noise $SNR < 25$, the noise has become significant, exceeds the threshold, and the detection success rate drops. So, with the noise $SNR \in [25, 75]$, we can build a more robust defense system and achieve a detection ratio of over 90%. Besides, our experiments also proved that the small input noise has a defense effect (Byun et al. 2022).

Besides, our scheme is available out-of-the-box and directly compatible with any ensemble of ASR defense models. First, our defense strategy can be combined with adversarial training to further enhance the robustness of the model. Adversarial training involves augmenting the training dataset with adversarial examples to improve the model's adaptability to such attacks. Second, our defense strategy can be used alongside various input transformations to provide an additional layer of security. Techniques such as feature squeezing or spatial smoothing can be applied to input audio to reduce adversarial perturbations. In Yang et al. (2018), multiple defensive measures, such as quantization, local smoothing, and down-sampling, have been employed. Paper (Hussain et al. 2021) proposes the WaveGuard method, which uses signal processing techniques to preprocess input audio, reducing the impact of adversarial perturbations. While WaveGuard has demonstrated good results in mitigating adversarial sample attacks, there is still room for improvement in practicality and sound quality assurance. Our defense strategy is designed to be directly compatible with any ASR defense model ensemble, which is also one of our future research directions-exploring powerful defensive ensemble systems.

Discussion

However, with more research on attacks, single-step generation attack of AEs is growing, which impose higher requirements on the defense. From another aspect, our scheme increases the attacker's attack cost, and our scheme will be fooled if the attacker has many resources. Fingerprint fraud techniques can also create vulnerabilities in our approach. In addition, some adversaries may give up their attacks on the target system and turn to attack the defense system, which also warrants our attention.

Our work is mainly focused on the ASR domain. It would be interesting to explore whether similar ideas can be applied to other application domains (such as speaker recognition Chen et al. 2021, 2022, and speech anti-spoofing detection Zhang et al. 2020) by constructing a suitable similarity algorithm for that domain.

We calculated the value of k under the condition of setting the false positive rate at 0.1%, i.e., when the *k* value is 75, the probability of misjudgment is less than 0.1. Generally, the more queries an attack has, the higher the confidence in determining whether it is an attack. If we increase the false positive rate, then *k* will be less than 75, allowing us to detect attacks with fewer queries. This may be a future research direction for us: establishing a multi-level defense system under different conditional probabilities to provide defenders with more information for making decisions.

Conclusion

In this work, we analyze adversary behavior during AE generation and detect potential attacks based on the association before and after the query. Our focus is on detecting the AE generation process, which provides a novel approach to process-based defense. Our approach achieves an average detection success rate of over 90%. It is a lightweight framework that is both quick and efficient, able to be closely combined with other defenses to build the foundation for a structured defense system.

Appendix

Datasets

Music-sets

We contacted the authors of CommanderSong (Yuan et al. 2018) and Devil's Whisper (Chen et al. 2020) to consult them on the details about how to design the music-based carries for the adversarial samples (AEs) they used in their experiments, and obtained a copy of the original music dataset they applied. To evaluate the threshold, we created a music carrier dataset for making AEs based on the obtained original music dataset. We have released the processed dataset and you can get our data from: https://drive.google.com/file/d/1wPVK9S8TyB0aaXqXFKEebYKuKshmBvDc/view.

The original music dataset is a raw dataset of 100 songs collected on YouTube, including pop, classical, rock, and light music, ranging across multiple languages, including Korean, English, Japanese, Chinese, Russian, and Arabic. The length of each song is about 5 minutes.

In our experiments, we studied the impact of different audio lengths on AEs and found that different lengths of audio affect the generation of adversarial examples. Overly short audio decreases the success rate of attacks, and too long audio increases the cost of producing AEs.

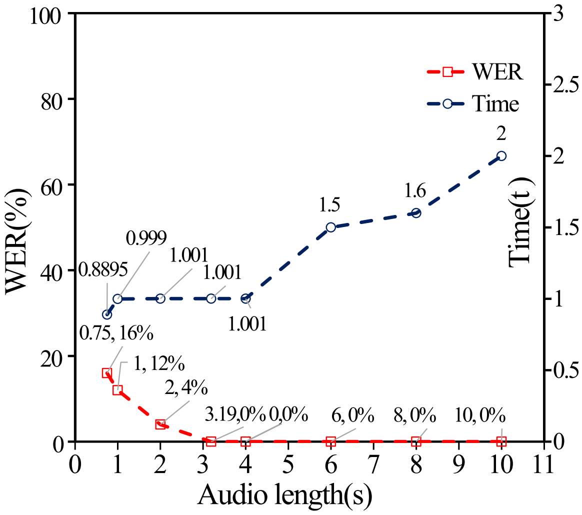


Fig. 8 Audio length impacts the production time of AEs and the integrity of command

Only properly lengthy audio is a candidate for AEs. We use Word Error Rates (WER) to research this issue.

$$WER = 100\% * \frac{S + D + I}{N} \tag{15}$$

In Eq. 15, S represents the number of characters replaced, D represents the number of characters deleted, I represents the number of characters inserted, and N represents the total number of characters.

From Fig. 8, it can be seen that the WER changes with the length of the audio. If the audio length is below 3.19s, the attack success rate of the AEs decreases as the audio length reduces (the WER of the target command increases). Above this value, the attack success rate reaches 100% and the WER falls to 0%. However, the time cost of producing an AE increases linearly with the length of the audio. The longer the audio, the higher the cost of producing AEs. While the audio length is 3s-4s, the most excellent performance is obtained and the ratio of time cost to WER is the lowest. Finally, the recommended audio length is 3s or 4s by balancing time and word error rate. During the production of our dataset, we divided each audio data into 3 s and 4 s to balance the success rate of the attack and the cost.

To simulate disturbances and improve the noise immunity of the audio, we must insert some noise into the clean dataset. Our experiments showed that when music develops as the carrier, the inserted noise is within 8000 (randomly insert), and the similarity distribution is in [0.36, 1]. The noise does not influence people's auditory perception, and the primary information of the audio remains reachable. So we keep the randomly inserted noise to the audio below 8000. When clipping music, the length of each slice is limited to 4 s according to the principle of random slice. For each song, segment 25 slices at a time, 5 times in total. Finally, obtaining $5 * 25 * 100 = 12,500$ slices. After that, the noise is randomly inserted into some of these slices by randomly displacing the sequence. After testing each slice, there were 10553 qualified slices obtained in total. Storage space occupied nearly 1.3G.

Currently, in the field of audio adversarial attacks, no publicly available dataset is based on music, except for some are dialogue-based which as a carrier for AEs. Instead, music is becoming a necessary candidate for attacks due to some of its advantages, but lack of proper datasets. To alleviate this problem, we are happy to share our data with the research community so that they can develop more research on music-based attacks and defenses. We also welcome interested researchers to expand the dataset with us.

Mini LibriSpeech

For the Mini-LibriSpeech dataset, we used FFmpeg[6] to convert from flac to wav. According to Fig. 8, we removed some samples that were either overly short or overly long, and we suggested recalculating the threshold to ensure that the detection was not affected once the dataset was modified. You can download the training data set from https://www.openslr.org/resources/31/train-clean-5.tar.gz.

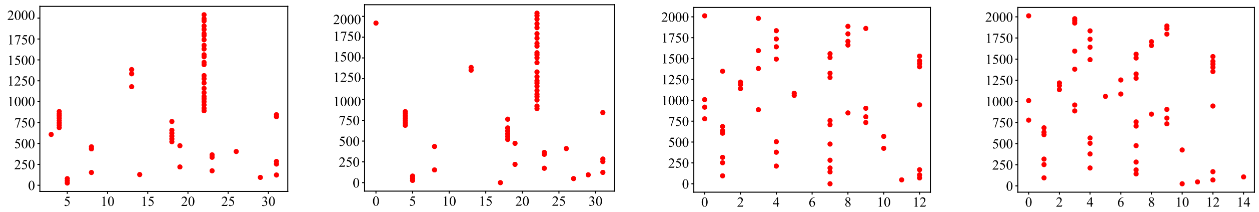
Benign examples and AEs audio fingerprint

As shown in Fig. 9, through the addition of perturbations (i.e., noise) on the clean carriers audio to generate AEs, the music-based ones have relatively more and richer fingerprints than the dialogue-based ones, which also confirms that the music-based AEs are easier to detect by our scheme. We also observed that the fingerprint difference between AEs and carriers is small. The fingerprint of each query is similar and the calculated similarity between the queries is very high if the carrier intends to generate AEs. This further proves the viability of our scheme.

Experimental environment

Linux Ubuntu20.0.4 operating system, a 2080Ti GPU with 12 G memory, Numpy 1.21.5, Cupy-Cuda 114, 64 CPUs with 256 G RAM.

[6] https://github.com/FFmpeg/FFmpeg.



(a) Clean audio of Dialogue-carrier (b) AE of Dialogue-carrier (c) Clean audio of Music-carrier (d) AE of Music-carrier

Fig. 9 Clean audio and the AEs fingerprint on Dialogue-carrier and Music-carrier

Societal impacts

For the attacks that require querying the ASR model, much of the defense work was mainly concentrated on the processing of inputs to achieve the defense purpose. Only considering the examination of individual inputs, it lost the procedure information and the results are often not reliable. Our scheme, on the other hand, involves considering the totality and continuity of inputs and capturing the neglected information, which can help us better track the adversary behaviors and make an accurate diagnosis. Such a strategy is more consistent with sociology as well. Meanwhile, dialogue-based carriers have lots of limitations in practical applications and it's hard to reproduce in real attack scenarios, which are gradually abandoned by researchers. Music-based AEs are gradually becoming the mainstream of attacks. The music is easily reproduced in the actual attack scenarios. The danger is very significant if music is hijacked as AEs, which cannot be ignored by researchers. However, the existing evaluation of defense work is still focused on the evaluation of public dialogue datasets. Lack of evaluation of music-based datasets for defense. In our paper, we have comprehensively evaluated the AEs with music-based carriers, which has a large social impact and also lays a solid foundation for related works in the future.

Abbreviations

AEs	Adversarial examples
ASR	Automatic speech recognition
DNNs	Deep neural networks

Acknowledgements

Not applicable.

Author Contributions

The design of the proposed method, the experiment deployment and the draft of the manuscript: GF and SZ. Revising the manuscript critically for important intellectual content: CYX and JL. All authors read and approved the final manuscript.

Funding

The authors are supported in part by NSFC No.62202275 and Shandong-SF No.ZR2022QF012 projects.

Availability of data and materials

The code is available at: https://github.com/sveapp/Audio-denfense.

Declarations

Competinf Interests

The authors declare that they have no competing interests.

Received: 19 April 2023 Accepted: 5 July 2023
Published online: 05 August 2023

References

Abdullah H, Garcia W, Peeters C, Traynor P, Butler KR, Wilson J (2019) Practical hidden voice attacks against speech and speaker recognition systems. arxiv: abs/1904.05734

Abdullah H, Rahman MS, Garcia W, Warren K, Yadav AS, Shrimpton T, Traynor P (2021) 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). IEEE, pp 712–729

Abdullah H, Rahman MS, Peeters C, Gibson C, Garcia W, Bindschaedler V, Shrimpton T, Traynor P (2021) Beyond l_p clipping: equalization-based psychoacoustic attacks against ASRs. arxiv: abs/2110.13250

Afchar D, Melchiorre AB, Schedl M, Hennequin R, Epure EV, Moussallam M (2022) Explainability in music recommender systems. arxiv: abs/2201.10528

Akinwande V, Cintas C, Speakman S, Sridharan S (2020) Identifying audio adversarial examples via anomalous pattern detection. arxiv: abs/2002.05463

Byun J, Go H, Kim C (2022) On the effectiveness of small input noise for defending against query-based black-box attacks. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3051–3060

Carlini N, Athalye A, Papernot N, Brendel W, Rauber J, Tsipras D, Goodfellow IJ, Madry A, Kurakin A (2019) On evaluating adversarial robustness. arxiv: abs/1902.06705

Carlini N, Wagner D (2018) Audio adversarial examples: targeted attacks on speech-to-text. IEEE

Chang K-H, Huang P-H, Yu H, Jin Y, Wang T-C (2020) Audio adversarial examples generation with recurrent neural networks. In: 2020 25th Asia and South pacific design automation conference (ASP-DAC). IEEE, pp 488–493

Chen G, Zhao Z, Song F, Chen S, Fan L, Wang F, Wang J (2022) Towards understanding and mitigating audio adversarial examples for speaker recognition. IEEE Trans Dependable Secur Comput. https://doi.org/10.1109/TDSC.2022.3220673

Chen G, Zhao Z, Song F, Chen S, Fan L, Liu Y (2022) As2t: arbitrary source-to-target adversarial attack on speaker recognition systems. IEEE Trans Dependable Secur Comput. https://doi.org/10.1109/TDSC.2022.3189397

Chen S, Carlini N, Wagner D (2019) Stateful detection of black-box adversarial attacks

Chen G, Chenb S, Fan L, Du X, Zhao Z, Song F, Liu Y (2021) Who is real bob? Adversarial attacks on speaker recognition systems. In: 2021 IEEE Symposium on Security and Privacy (SP). IEEE, pp 694–711

Cheng S, Dong Y, Pang T, Su H, Zhu J (2019) Improving black-box adversarial attacks with a transfer-based prior. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp 10932–10942

Chen Y, Yuan X, Zhang J, Zhao Y, Zhang S, Chen K, Wang X (2020) Devil's whisper: a general approach for physical adversarial attacks against commercial black-box speech recognition devices. In: USENIX security symposium, pp 2667–2684

Cohen J, Rosenfeld E, Kolter Z (2019) Certified adversarial robustness via randomized smoothing. In: International conference on machine learning. PMLR, pp 1310–1320

Du T, Ji S, Li J, Gu Q, Wang T, Beyah RA (2020) Sirenattack: generating adversarial audio for end-to-end acoustic systems. In: Proceedings of the 15th ACM Asia conference on computer and communications security

Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: Bengio Y, LeCun Y (eds) 3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings . arxiv:1412.6572

Goyal S, Raghunathan A, Jain M, Simhadri HV, Jain P (2020) DROCC: deep robust one-class classification. In: International conference on machine learning. PMLR, pp 3711–3721

Guo Q, Ye J, Hu Y, Zhang G, Li H (2020) MultiPAD: a multivariant partition based method for audio adversarial examples detection. IEEE Access (99):1–1

Haitsma J, Kalker T (2002) A highly robust audio fingerprinting system. In: ISMIR 2002, 3rd International conference on music information retrieval, Paris, France, October 13–17, 2002, Proceedings

Han JK, Kim H, Woo SS (2019) Nickel to LEGO: minimal information examples to fool google cloud speech-to-text API. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pp 2593–2595

Huang Z, Zhang T (2019) Black-box adversarial attack with transferable model-based embedding

Hussain S, Neekhara P, Dubnov S, McAuley J, Koushanfar F (2021) Waveguard: understanding and mitigating audio adversarial examples. arXiv:2103.03344

Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A (2019) Adversarial examples are not bugs, they are features. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in Neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp 125–136

Joshi S, Villalba J, Želasko P, Moro-Velázquez L, Dehak N (2021) Adversarial attacks and defenses for speaker identification systems. arXiv e-prints, 2101

Khare S, Aralikatte R, Mani S (2018) Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization. arXiv:1811.01312

Kurakin A, Goodfellow I, Bengio S (2016) Adversarial examples in the physical world

Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083

Marin-Martinez F, Sánchez-Meca J (2010) Weighting by inverse variance or by sample size in random-effects meta-analysis. Educ Psychol Meas 70(1):56–73

Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2574–2582

Nam LNH (2022) Towards comprehensive approaches for the rating prediction phase in memory-based collaborative filtering recommender systems. Inf Sci 589:878–910

Pang R, Zhang X, Ji S, Luo X, Wang T (2020) Advmind: Inferring adversary intent of black-box attacks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 1899–1907

Qin Y, Carlini N, Cottrell G, Goodfellow I, Raffel C (2019) Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: International Conference on machine learning. PMLR, pp 5231–5240

Rajaratnam K, Kalita J (2018) Noise flooding for detecting audio adversarial examples against automatic speech recognition. IEEE

Richards LE, Nguyen A, Capps R, Forsyth S, Matuszek C, Raff E (2021) Adversarial transfer attacks with unknown data and class overlap. In: Proceedings of the 14th ACM workshop on artificial intelligence and security, pp 13–24

Samizade S, Tan Z-H, Shen C, Guan X (2020) Adversarial example detection by classification for deep speech recognition. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3102–3106

Schönherr L, Kohls K, Zeiler S, Holz T, Kolossa D (2018) Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. arXiv:1808.05665 x

Shafiloo R, Kaedi M, Pourmiri A (2021) Considering user dynamic preferences for mitigating negative effects of long tail in recommender systems. arxiv:abs/2112.02406

Song J, Chang C, Sun F, Chen Z, Hu G, Jiang P (2021) Graph attention collaborative similarity embedding for recommender system. In: Database systems for advanced applications: 26th international conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, proceedings, Part III 26. Springer, pp 165–178

Su S, Guo P, Xie L, Hwang MY (2019) Adversarial regularization for attention based end-to-end robust speech recognition. Audio Speech Lang Process IEEE/ACM Trans 27(11):1826–1838

Sun S, Yeh C-F, Ostendorf M, Hwang M-Y, Xie L (2018) Training augmentation with adversarial examples for robust speech recognition. arXiv:1806.02782

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826

Tamura K, Omagari A, Hashida S (2019) Novel defense method against audio adversarial example for speech-to-text transcription neural networks. In: 2019 IEEE 11th international workshop on computational intelligence and applications (IWCIA)

Taori R, Dave A, Shankar V, Carlini N, Recht B, Schmidt L (2020) Measuring robustness to natural distribution shifts in image classification. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, Virtual

Taori R, Kamsetty A, Chu B, Vemuri N (2019) Targeted adversarial examples for black box audio systems. In: 2019 IEEE security and privacy workshops (SPW). IEEE 6:15–20

Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A (2018) Robustness may be at odds with accuracy. arXiv:1805.12152

Vaidya T, Zhang Y, Sherr M, Shields C (2015) Cocaine noodles: exploiting the gap between human and machine speech recognition. In: Proceedings of the 9th USENIX conference on offensive technologies. WOOT'15, p. 16. USENIX Association, USA

Wang A (2003) An industrial-strength audio search algorithm. In: ISMIR 2003, 4th international conference on music information retrieval, Baltimore, Maryland, USA, October 27–30, 2003, Proceedings

Wang A *et al.* (2003) An industrial strength audio search algorithm. In: Ismir, vol. 2003, pp 7–13. Citeseer

Wang Q, Zheng B, Li Q, Shen C, Ba Z (2021) Towards query-efficient adversarial attacks against automatic speech recognition systems. IEEE Trans Inf Forens Secur 16:896–908. https://doi.org/10.1109/TIFS.2020.3026543

Xu W, Evans D, Qi Y (2017) Feature squeezing: detecting adversarial examples in deep neural networks. arxiv: abs/1704.01155

Yang Z, Li B, Chen P-Y, Song D (2018) Characterizing audio adversarial examples using temporal dependency. arXiv:1809.10875

Yuan X, Chen Y, Zhao Y, Long Y, Liu X, Chen K, Zhang S, Huang H, Wang X, Gunter CA (2018) {CommanderSong}: a systematic approach for practical adversarial voice recognition. In: 27th USENIX security symposium (USENIX Security 18), pp 49–64

Zhang Y, Jiang Z, Villalba J, Dehak N (2020) Black-box attacks on spoofing countermeasures using transferability of adversarial examples. In: Interspeech, pp 4238–4242

Zhang J, Zhang B, Zhang B (2019) Defending adversarial attacks on cloud-aided automatic speech recognition systems. In: Proceedings of the seventh international workshop on security in cloud computing, pp 23–31

Zheng B, Jiang P, Wang Q, Li Q, Shen C, Wang C, Ge Y, Teng Q, Zhang S (2021) 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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Feng Guo is a 2023 graduate who is currently pursuing a master's degree in Cyberspace and Information Security at Shandong University. His research interests include adversarial machine learning and AI security.

Zheng Sun is a student who is currently pursuing a master's degree in Cyberspace and Information Security at Shandong University. His research interests include adversarial machine learning and AI security.

Yuxuan Chen is currently an Associate Professor with the School of Cyber Science and Technology, Shandong University. His research interests include adversarial machine learning and AI security.

Lei Ju is currently a Full Professor with the School of Cyber Science and Technology, Shandong University. His research interests include IoT security and system security.