The Silent Manipulator: A Practical and Inaudible Backdoor Attack against Speech Recognition Systems

Zhicong Zheng
zheng_zhicong@zju.edu.cn
Zhejiang University
HangZhou, Zhejiang, China

Xinfeng Li
xinfengli@zju.edu.cn
Zhejiang University
HangZhou, Zhejiang, China

Chen Yan*
yanchen@zju.edu.cn
Zhejiang University
HangZhou, Zhejiang, China

Xiaoyu Ji
xji@zju.edu.cn
Zhejiang University
HangZhou, Zhejiang, China

Wenyuan Xu
wyxu@zju.edu.cn
Zhejiang University
HangZhou, Zhejiang, China

ABSTRACT

Backdoor Attacks have been shown to pose significant threats to automatic speech recognition systems (ASRs). Existing success largely assumes backdoor triggering in the digital domain, or the victim will not notice the presence of triggering sounds in the physical domain. However, in practical victim-present scenarios, the over-the-air distortion of the backdoor trigger and the victim awareness raised by its audibility may invalidate such attacks. In this paper, we propose SMA, an inaudible grey-box backdoor attack that can be generalized to real-world scenarios where victims are present by exploiting both the vulnerability of microphones and neural networks. Specifically, we utilize the nonlinear effects of microphones to inject an inaudible ultrasonic trigger. To accurately characterize the microphone response to the crafted ultrasound, we construct a novel nonlinear transfer function for effective optimization. We also design optimization objectives to ensure triggers' robustness in the physical world and transferability on unseen ASR models. In practice, SMA can bypass the microphone's built-in filters and human perception, activating the implanted trigger in the ASRs inaudibly, regardless of whether the user is speaking. Extensive experiments show that the attack success rate of SMA can reach nearly 100% in the digital domain and over 85% against most microphones in the physical domains by only poisoning about 0.5% of the training audio dataset. Moreover, our attack can resist typical defense countermeasures to backdoor attacks.

CCS CONCEPTS

• **Security and privacy → Security in hardware**; **Software and application security**; • **Computing methodologies → Speech recognition**.

KEYWORDS

Backdoor Attack, Nonlinear Effects, Speech Recognition

ACM Reference Format:

Zhicong Zheng, Xinfeng Li, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. 2023. The Silent Manipulator: A Practical and Inaudible Backdoor Attack against Speech Recognition Systems. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3613843

1 INTRODUCTION

Automatic speech recognition systems (ASRs) have greatly facilitated the advancement of intelligent voice applications, such as speech-to-text APIs [29]. However, the time and cost induced by the requirement of a large volume of training data have driven ASR providers to resort to open-source datasets as a cost-effective alternative, therefore providing an opening for adversaries to implant backdoor attacks by publishing their poisoned datasets on the Internet. Existing work has revealed that the attacker can make the backdoored models degrade their performance or misinterpret a user's benign speech as malicious commands by injecting a crafted trigger [16, 19]. Besides, a complete ASR application or service usually needs a microphone to record users' speech, the nonlinear vulnerability of which provides another attack surface for adversaries. Existing nonlinearity-based attacks [26, 36] have attracted the attention of the security community as a more stealthy alternative, which exploits carrier signals outside the audible frequencies of human auditory (20 Hz-20 kHz) to inject voice commands into ASRs inaudibly.

Motivation. With the advancement of intelligent voice applications, these two attacks pose a severe and noteworthy challenge, and it is imperative to systematically understand and excavate more vulnerability of ASRs so as to expose and mitigate their threat. In this paper, we consider a more familiar and threatening victim-present scenario. As shown in Fig. 1, a premeditated and adaptive attacker intentionally tamper any word spoken by a user to a specific word without being detected while the unaware user may be simultaneously issuing a command to ASRs. This realistic scenario presents three requirements: to be stealthy, physically realizable, and immune to users' speech. Unfortunately, as shown in Tab. 1, existing attacks have not fully addressed all the requirements simultaneously. For previous backdoor attacks, they usually adopt

*Chen Yan is the corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613843

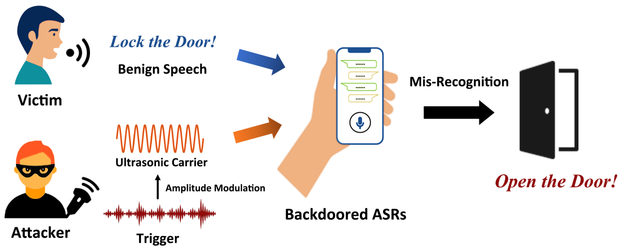


Figure 1: An example of the silent manipulator attack (SMA) against ASR systems. The word "Lock" will be misrecognized as "Open", resulting in a significant threat.

audible triggers [11, 19] and mainly focus on performance in the digital domain [16, 34]. It brings a practical challenge to whether these attacks are still harmful in real-world scenarios where the victim may be alert and the environmental disturbance is unknown. Some work like [19] resort to opportunistic attacks with ambient noise to evade human perception and mitigate the risk of attack deployment. However, opportunistic attacks are less controllable than intentional attacks by which the attacker can commit subsequent crimes. On the other hand, the attack success rates of previous nonlinearity-based attacks [26, 36] are uncertain when the malicious commands coincide with benign user speech, i.e., the cases where victims are present to issue commands to the ASR system. We realize that these practical problems limit the performance and impact of attacks against the ASR model. Therefore, we here propose an inaudible grey-box backdoor attack SMA[1], which can be implemented in real-world scenarios by exploiting both the vulnerability of neural networks and microphones.

In our attack design, we first try to construct a transfer function to model the nonlinear effects accurately and lay the groundwork for trigger optimization. Then, we elaborately design an optimization algorithm to generate a robust trigger to fit the complex real-world scenarios. Note that our trigger can be injected into audio with standard audio formats (e.g., WAV with a 16 kHz sample rate), transmitted by inaudible ultrasonic signals, and immune to the band-pass filter of recording devices. Compared with previous backdoor attacks against ASRs, SMA is more stealthy and practical because our trigger is completely inaudible and robust in physical scenarios during attacking. Compared with nonlinearity-based attacks, our attack overcomes a significant challenge: The injection of the malicious command will be severely disrupted by the user's benign speeches.

Challenge: The principle of backdoor attack and nonlinearity-based attack inspires us to propose a new attack framework, SMA. However, we still face two significant challenges.

How to ensure the success of SMA *on various devices?* Nonlinear effects originate from the characteristics of the hardware. Different prototypes, materials, and manufacturing can result in diverse nonlinear responses. Our grey-box attack has no knowledge of the victim's microphone, so the nonlinear response of our backdoor trigger is unpredictable. To solve this problem, we proposed a lightweight method to simulate the distortion caused by nonlinear effects and collect these transfer functions into a pool. Then

[1]Short for "*Silent Manipulator Attack*", indicating that our attack can manipulate ASRs silently.

Table 1: Comparison with existing works

Method	Type*	Knowledge	Inaudible	P.R.+	I.S.#
[16]	Backdoor	Grey-box	✔	✘	✔
[19]	Backdoor	White-box	✘	●	✔
[34]	Backdoor	White-box	✘	✘	✔
[36]	Nonlinearity	Black-box	✔	✔	✘
Ours	**Backdoor**	**Grey-box**	✔	✔	✔

(i)*: Backdoor/Nonlinearity-based Attack. (ii): Grey-box: access training data; White-box: access model knowledge; Black-box: no prior knowledge. (iii)+: Physical Realizable (P.R.). Whether the method has been realized in the physical world. ● indicates that the research is P.R. theoretically but lacks evaluation. (iv)#: Immune to Speech (I.S.). Whether the method is effective when victims are issuing commands to ASRs.

we optimize the trigger by learning the commonality of various nonlinear responses, therefore extending SMA to unseen devices.

How to design a trigger that is robust in the real world? Real-world scenarios pose significant challenges to SMA due to pattern distortion of backdoor triggers resulting from factors such as ambient noise, ultrasound attenuation, superimposed victim speech, etc. In this regard, we propose a robust trigger design workflow with three optimization objectives: universality, activity, and directivity, which ensures the feasibility of SMA on various devices and the resistance to existing defenses against backdoor attacks.

Our contribution can be summarized as follow:

- To the best of our knowledge, we are the first to exploit the nonlinear vulnerability of microphones for backdoor attacks against ASRs. Our attack is inaudible, realizable in the physical world, and immune to users' speech and low-pass filters.
- We propose a lightweight method to simulate the complex nonlinear responses of microphones and design an effective optimization algorithm to generate ultrasonic triggers that are robust in various real-world scenarios.
- Extensive digital and physical experiments validate the effectiveness and robustness of SMA. The results demonstrate that SMA outperforms two baselines with an attack success rate of nearly 100% in the digital domain and over 85% in the physical domain with a low data poison rate of 0.5%. Furthermore, SMA can resist typical and potential defenses.

2 BACKGROUND AND RELATED WORK

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) has gained attention with the rise of IoT and Artificial Intelligence, which achieves speech-to-text (STT) conversion [22]. Advanced smart devices have been equipped with this convenient technology to understand users' speech and provide better services [10, 30].

Traditional ASRs consist of several modules, including acoustic, pronunciation, and language models [31], by which the acoustic signal is finally transformed into text. With the development of deep learning (DL), the end-to-end speech recognition system [1, 2] has become increasingly popular because it completes the STT conversion in one step with higher performance and faster implementation. In most ASRs, Filter Bank (Fbank) and Mel-Frequency Cepstral Coefficients (MFCC) are two commonly used feature representations in speech recognition systems. In this paper, we focus

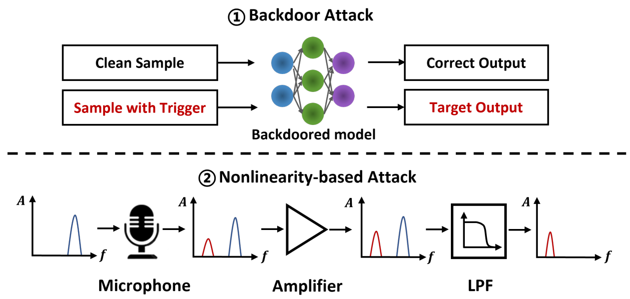


Figure 2: Principles of ①backdoor attacks and ②nonlinearity-based attacks.

on DL-based speech recognition systems for their wide adaptation in real life.

2.2 Backdoor Attack against ASRs

Backdoor attack, first proposed for deep neural networks [11], has been widely studied in many areas such as image [27], text [6], natural language processing [4], and reinforcement learning [5, 15]. Particularly, imperceptibility is an essential requirement of backdoor attacks. By poisoning the training dataset with a subtle and secret trigger, the attacker can inject the backdoor into the target model, resulting in its abnormal behaviors during inference, as presented in Fig. 2 ①. Without the trigger activating, backdoored models perform as well as clean models. Notably, although most existing backdoor attacks can achieve a high Attack Success Rate (ASRT) and Benign Accuracy (BA) in the digital domain, their performance in the physical domain is usually inadequate due to the complex and numerous interference factors from the victim-present scenarios and the over-the-air channel. As shown in Tab. 1, being inaudible (totally undetectable) and physically realizable are two prominent challenges to extending existing backdoor attacks against ASRs to real-world scenarios.

Liu et al. first poisoned the ASR model by injecting a slight random noise as the trigger [20]. To further enhance the concealment of audible triggers, DABA explored a human-imperceptible trigger in [19], which utilizes the in-distribution ambient noise as triggers. Xin et al. further evaluate the natural noise trigger in the physical world, where it needs a high poison rate (over 5%) to achieve an ASRT of about 70% [33]. Ye et al. designed a dynamic trigger generation network to craft a variety of audio triggers [34]. In these attacks, the volume of triggers is strictly limited because the triggers are still audible during training and attacking. Moreover, even if we assume the ambient noise trigger adopted in [19, 33] would not alert victims, its wide existence in the real world will be easily mis-activated, unexpectedly degrading the ASR model's performance and thus alerting victims. Therefore, DABA claimed their attack was opportunistic with the advantage of not relying on active invoked by attackers. As a comparison, although the intentional attack needs to be deployed by the adversary, it is more difficult to detect and enables the adversary to commit subsequent malicious attacks. Koffas et al. introduced an inaudible ultrasonic pulse as the trigger [16]. Nevertheless, their attack must up-sample the training audio over 40 kHz. As discussed in [19], this attack will be easily mitigated in practice because most microphones are equipped with low-pass filters that block the ultrasound, i.e., the trigger in [16] can only survive in the digital domain. Moreover, Koffas et al. also admitted that most ASR models would resample the input audio to 16 kHz as preprocessing, which may disable their trigger.

2.3 Nonlinearity-based Attack

As a mainstream voice-captured sensor, the microphone converts the acoustic signal over the air to electrical signals. Liu et al. introduce the operating principle and reveal the nonlinear characteristics of the widely-used MEMS microphones [18]. A typical microphone consists of a transducer, an amplifier, a low-pass filter, and an analog-to-digital converter (ADC). These modules ideally should be linear systems to transfer the audio signals without distortion. However, some researches demonstrate that the transducer and amplifier can only maintain linearity within the audible frequency range. As for ultrasonic signals with a high frequency, the microphone's nonlinear output is formed in Equ. 1.

$$s_{out}(t) = A_1 s_{in}(t) + A_2 s_{in}^2(t) + A_3 s_{in}^3(t) + \ldots \tag{1}$$

where $s_{in}(t)$ is the microphone's input. As a result, if $s_{in}(t)$ contains signals of multiple frequencies, the second and higher-order terms will generate signals at new frequencies. This phenomenon is also known as intermodulation.

According to this principle, Roy et al. utilize two ultrasonic speakers and create an audible shadow signal [26] by the spontaneous airborne nonlinear demodulation, which can only be captured by microphones and be further used in applications like inaudible data communication and audio watermark. Kasher et al. further explore whether [26] can deliver adversarial audio [14] with some validation. Zhang et al. exploit this vulnerability to inject inaudible voice commands by loading the commands onto ultrasonic carriers, i.e., utilizing amplitude modulation [36]. Such a method only needs one speaker. The emitted signals can be automatically demodulated by microphones into desired audible audio, therefore manipulating voice assistants without being heard by human beings. However, as shown in Tab. 1, these nonlinearity-based attacks are vulnerable to simultaneous speeches, with which the accompanying signals induced by nonlinear effects will be disrupted and fail to mislead the ASRs. While we could overwrite the interference from user speech with a high-power but non-portable device, it would lower the attacker's practicality in the real world.

3 THREAT MODEL

We aim to achieve a practical and inaudible backdoor attack in real-world scenarios. To the best of our knowledge, no existing work fits our threat model derived from the real user experience and overcomes the challenges in the physical domain. Here we thoroughly give the definition of our threat model.

Attacker. The attacker does not require any prior knowledge of the ASR models' structure, training algorithms, or the specific microphone type equipped on the victim's device. Instead, the attacker optimizes a trigger with our proposed algorithm and releases poisoned audio datasets online for ASR service vendors to employ in training their models. We assume the attacker can launch attacks with a hidden ultrasonic transmitter or handheld device.

ASR service vendor. The vendors manage to enhance the performance of their ASR models by leveraging a vast amount of open-source datasets available on the Internet. However, we assume that cautious vendors will inspect the accuracy of the model in their test datasets and filter out the low-quality data.

Victim. Victims are alert to audible strange sounds when they are using ASRs. Beside, the victims' devices are probably integrated with several commonly used components like band-pass filter which can resist conventional ultrasonic backdoor attacks (e.g., a ultrasonic pulse [16]).

Attack Goal. SMA aims to achieve a grey-box backdoor attack against the ASR model by exploiting the vulnerability of the neural network and microphone. We propose a scenario where a deliberate attacker first uploads the poisoned datasets online for ASR model training and then delivers the inaudible trigger to mislead the behavior of victims' ASR model in the real-world scenario. Our designed attack can be launched inaudibly when the victim is present to issue arbitrary commands to the ASR without raising his/her awareness. Note that our attack is an intentional attack instead of an opportunistic one because we consider: 1) the intentional attack is stealthier, for it is difficult to be mis-activated by the environment and alert victims unnecessarily; 2) the intentional attack is more controllable for the attacker to commit the subsequent malicious attacks.

4 INVESTIGATION OF INAUDIBLE TRIGGER

Since we manage to design an inaudible trigger using the nonlinear effects of microphones, we need to answer the following questions:

RQ1: How to simulate the nonlinear effects in the digital domain for more efficient optimization?

RQ2: How to ensure that our trigger works with different nonlinear effects caused by various microphone types?

4.1 Simulate the Nonlinear Effects

Nonlinear effects are caused by the electronic components of the microphone, such as the diaphragm and the amplifier. According to [36], we can modulate a signal $m(t)$ into an ultrasonic signal $cos(2\pi f_u t)$ by amplitude modulation as $S_{in} = (1+m(t))cos(2\pi f_u t)$. The corresponding response of the microphone can be divided into two parts: 1) the inevitable noise due to the continuous vibration of the diaphragm caused by the high-frequency carrier; 2) the undesired frequency component introduced by the nonlinear demodulation. As shown in Equ. 1, if we denote the frequency of $m(t)$ as f_m, from the quadratic term, the microphone will demodulate frequencies of $2f_u$, $2f_m$, $2(f_u - f_m)$, $2(f_u + f_m)$, $2f_u + f_m$ and $2f_u - f_m$. For the higher order term, more complex frequency components are introduced.

Some previous works tried to figure out the gains of each term in Equ. 1 to construct an accurate nonlinear model [23]. However, the parameter tuning is complex, and the constructed model is hard to generalize. Huang et al. proposed an inverse filter to compensate for the distortion of the modulated signal received by microphones [12]. As shown in Fig. 7, Huang compensates for the amplitude of the original baseband signal according to the pre-recorded equivalent frequency response between the transmitter and recorder. However, such a pre-compensation method may cause saturation of the audio

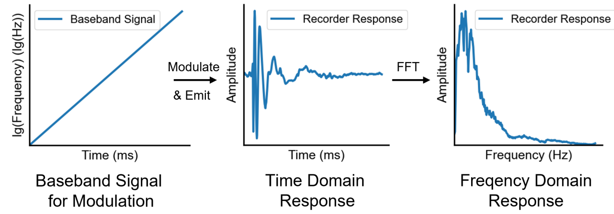


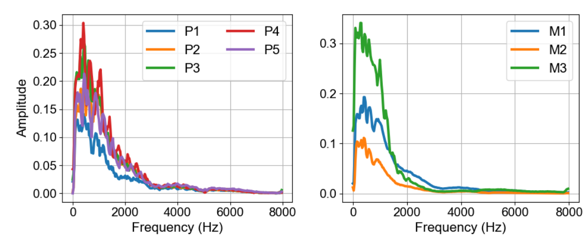
Figure 3: The pipeline to access the frequency response of microphones.

and distorts the signal instead. We further discuss the cause of the saturation vulnerability in Appendix. A.2.

Considering this drawback, we are motivated to construct a forward filter to simulate the distortion caused by the nonlinear effects. First, we record the frequency response of the microphone. As shown in Fig. 3, we adopt the sine sweep signal [8] modulated and transmitted with an ultrasonic carrier. Next, we perform a Fourier transform on the corresponding time domain response of the microphone $r(t)$ and obtain the frequency response $R_m(f)$. We then use the frequency response as a forward filter. For our trigger noise $n(t)$, we can use this filter to simulate the nonlinear distortion by $r(t) \circledast n(t)$. This method provides a lightweight strategy to construct plenty of nonlinear transfer functions in the digital domain for optimizations that can realistically mirror physical scenarios.

4.2 Diversity of Nonlinear Effects

The filter constructed from the frequency response successfully simulates the nonlinear effects. However, another challenge we have to solve is that there are many potential factors that may cause different nonlinear effects. We further investigate some of these factors in this subsection.



(a) Impact of Positions (b) Impact of Microphones

Figure 4: Diversity of the Frequency Responses under Nonlinear Effects. P: Position; M: Microphone

Impact of Position. The position will strongly affect the nonlinear response. On the one hand, the amplitude of the responses will decrease with increasing distance as the transmitted signal gradually attenuates in the air. On the other hand, the microphone of ASRs may have a recording direction which means that the angle between the incident signal and the microphone orientation will affect the nonlinear effects of the microphone. Moreover, the strong directivity and weak diffractivity of the ultrasonic signal will exacerbate this impact. To verify this opinion, we gather the FRs with an emitting sweep signal with the same amplitude in five different positions with a distance varying from 10cm to 50cm and an angle from 0° to 60°. Fig. 4(a) shows the FR of the microphones we collected. We can find that the position does affect the nonlinear response. In our opinion, the influence of distance can be compensated for by adjusting the amplitude of the ultrasonic signal without raising the awareness of the victims. However, the influence of direction is challenging to model and compensate for. Therefore, to ensure the robustness of our attack, we should solve the impact of position in our optimization.

Impact of Microphone Prototype Compared to the positions, the impact of the microphone prototype is more significant. Research in [18] has pointed out that the material and manufacturing of the microphone will strongly affect the nonlinear effects. To examine this perspective, we select three different microphones (M1: Google pixel3, M2: Samsung S6, M3: Xiaomi Mix2) and plot their frequency responses with the same sweep signal in Fig. 4(b). From this figure as well as the research of Li et al. [17], we can discover that some microphones have a more pronounced nonlinear effect while some do not. However, we still notice that there are some similarities between them. For example, the amplitudes of these three microphone responses are strong in the low-frequency band and become weak at the frequency above 2000 Hz. We believe that these similarities provide a possibility for our optimization.

Note that our attack, constrained by our grey-box threat model, has no knowledge of the microphone prototype of victims. Meanwhile, we hope that our attack can be practical in real-world scenarios. These requirements mean we cannot focus on a specific microphone to attack. On the contrary, our attack needs to design a universal trigger to meet the above challenges.

5 ATTACK DESIGN

Fig. 5 depicts that materializing our attack involves four key parts. First, microphone modeling bridge the physical-and-digital gap of diverse and complex nonlinear effects by constructing transfer functions. Trigger design optimizes the trigger with the constraint of the microphone model and three objectives of universality, directivity, and activity. During poisoning, the attacker injects the simulated response of the well-trained trigger into the training data. Finally, the attacker can mislead the ASR system by emitting the trigger modulated on an inaudible ultrasonic carrier.

5.1 Microphone Modeling

As mentioned in Sec. 4, we design a forward filter by the frequency response (FR) of microphones to simulate the nonlinear response in the digital domain. Moreover, to address the challenges posed by the variety of nonlinear effects, we construct a microphone model pool inspired by the idea of robustness training. For each microphone response of a modulated ultrasound, we can divide it into two parts. One is the noise caused by the high-frequency carrier signal. The other is the complex nonlinear demodulated signals caused by the original modulated signal.

For the first part, the high-frequency ultrasonic signal will continuously transfer the energy to the microphone diaphragm, causing it to vibrate. To construct the response pool of noise, we emit a 25 kHz ultrasonic signal to several microphones at different positions and collect the responses. For the second part, we modulate the sweep signal and transmit it to several microphones at different positions and construct the FR pool.

For each training epoch, we will randomly select a pair of FR and noise from the pools. Then, we will generate two responses $Resp$ caused by the nonlinear effects from the original trigger t^* according to $Resp_i = t^* \circledast IR_i + n_i$.

5.2 Trigger Design

The trigger we design is inaudible in the real world but can be recorded by microphones due to the nonlinear response. Microphone modeling has provided several transfer functions of the nonlinear effects, by which we can design and optimize our trigger in the digital domain.

We first initialize a random noise and process it with our microphone model, obtaining its approximate nonlinear response. Then we optimize the noise with three objectives, universality, activity, and directivity. We use the soft label provided by the surrogate model as a more simplified feature representation for better and faster convergence. Note that our generated trigger shows good transferability to other unseen ASR models in evaluation without conflict with our threat model.

5.2.1 Universality. As mentioned in Sec. 2, one of the biggest challenges of SMA is ensuring that our trigger can be activated with various nonlinear distortion brought from different microphones. To settle down this problem, we use the previously constructed FR pool and noise pool. In each training epoch, we randomly select two FRs and two noises to generate two nonlinear responses from the optimizing noise, respectively. Both these two responses will be fed into the surrogate model and generate two predictions. Our first goal is to minimize the cross entropy of these two predictions as Equ 2

$$\mathbb{U} = CE(\Phi_{sof}(Resp_i \oplus a), \Phi_{sof}(Resp_j \oplus a)) \tag{2}$$

where Φ_{sof} is the softmax output of the benign surrogate models, a is the benign audio. This function aims to shorten the feature distance between different nonlinear effects, generalizing our trigger to multiple microphones.

5.2.2 Activity. Although SMA can be implemented inaudibly in realistic scenarios, the poisoned audio with the nonlinear trigger is still audible during training. Therefore, similar to the conventional backdoor attack, our digital trigger needs to be slight and stealthy for humans but prominent and active for neural models. To enhance the activity of our trigger, we represented the second objective in Equ 3.

$$\mathbb{A} = CE(\Phi_{sof}(Resp \oplus a), \Phi_{sof}(a)) \tag{3}$$

In this function, the $\mathbb{A}$ is maximized to strengthen the influence of our trigger as prominent as possible. We hope our trigger can be active even if we limit its amplitude when poisoning the training data.

5.2.3 Directivity. For backdoor attacks, we not only require the backdoored model to be sensitive to our trigger but also expect the targeted misbehavior. Previous backdoor attacks design a fixed pattern and forcibly modify the corresponding label. We noticed that the neural network will learn the feature of the fixed pattern

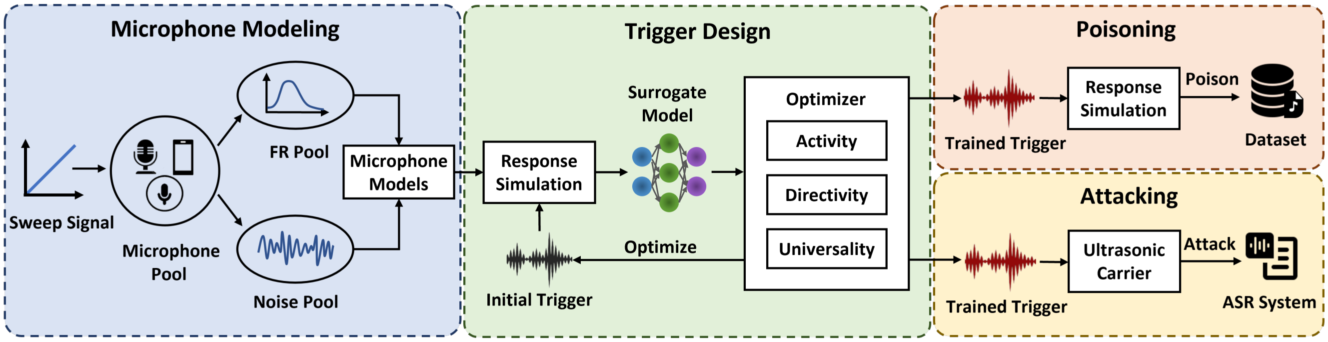


Figure 5: The overview of SMA. (1) We first collect the frequency and noise responses from various microphones and construct microphone models, which are used to simulate the real nonlinear responses. (2) Then we optimize our trigger with the objectives of universality, directivity and activity iteratively. (3) During poisoning, the well-trained trigger will be processed by microphone models and add to the training dataset. (4) During attacking, the well-trained trigger can be modulated on an ultrasonic carrier to attack victim ASR models inaudibly.

which is out-of-distribution and unrelated to the targeted class. Unfortunately, it inevitably results in a non-smooth decision boundary between the targeted class and others. The abnormal boundary, on the one hand, affects the attack robustness; on the other hand, they are more likely to be alerted by some defense algorithms [37]. Therefore, we design the third function to shorten the feature distance between the trigger noise and the samples of the target class, which is represented as Equ 4

$$\mathbb{D} = CE(\Phi_{sof}(Resp \oplus a), O(L_{target})) \tag{4}$$

where L_{target} is the target label, and O is the function to construct the one-hot vector of the corresponding label. In particular, we make our trigger noise more consistent with the distribution of the samples of the target class and smooth the decision boundary as much as possible by minimizing the $\mathbb{D}$.

With these three equations, we finally design a loss function depicted as Equ 5.

$$\mathcal{L} = \alpha max(0, \mathbb{D} - \mathbb{A}) + \beta \mathbb{U} \tag{5}$$

Note that this loss function combines three objectives in which $\mathbb{U}$ is relative to two frequency responses (FRs), while $\mathbb{D}$ and $\mathbb{A}$ are relative to one FR. Therefore, we design α and β to adjust their weights for multi-objective optimization. This loss will guide the trigger optimization and finally generate our trigger. The whole trigger design algorithm is shown in Appendix A.1.

6 EVALUATION

6.1 Experiment Setting

Dataset. The training datasets we select are two different versions of Speech Command [32], also used in [16, 19]. The first version SCD-10 contains ten classes with 13907 pieces of audio, and the second version, SCD-30, contains 30 classes with 58021 pieces of audio. Each piece of audio is 1s in length.

To construct the FR pool and the noise pool, we collect three types of smartphones, Google Pixel (2016), Xiaomi Mix2 (2017), and Samsung Galaxy S6 (2015). In the experiments, we select two extra smartphones, LG Nexus 5X (2015) and Reami K50Pro (2022), to evaluate the transferability of SMA.

Surrogate Model and Victim Model. During training, we use an LSTM model introduced in [7] as our surrogate model. During attacking, we use another two RNN models [20, 28] as the victim models. All these three models are also used to evaluate the performance of the backdoor attack against ASR systems in [16].

Baseline. To the best of our knowledge, there is no other backdoor attack against ASR models that can simultaneously satisfy our attack scenarios. However, to verify the performance of SMA, we choose DABA [19] and extended BadNets, two audible backdoor attacks against ASRs, as our baselines. Since DABA does not provide its ambient noise pool, we adopt a new ambient noise dataset, ESC-50 [24], in our evaluation. The extended BadNets is also proposed by Liu et al. [19], which uses a randomly generated trigger with typical robust training techniques, including adjusting the audio amplitude and mixing Gaussian noise. Note that we do not make a comparison with [16] because it uses 48 kHz audio as the ASR input while other attacks use 16 kHz audio. Most microphones will block the frequency over 20 kHz by a low-pass filter, and most ASR APIs only accept 16 kHz audio as input (or resample the audio to 16 kHz automatically). Existing work [19] has verified that the impulse trigger in [16] is fragile and hardly poses a real threat to the real ASR model.

Hyper-parameter Setting. Our attack includes some optional hyper-parameters settings. In our evaluation, we set the following parameters by default: the length of the trigger noise is 0.75s, which will be randomly added to any position of the audio; the α and β in Equ. 5 are 0.6 and 0.4, respectively; the iteration time of the trigger generation is 8000.

During poisoning, we will restrict the volume of the baselines' audible trigger and the nonlinear response of our trigger to -20 dB since the average of the benign audio is about -20 dB. During the attack, we choose a 25 kHz sine wave as the ultrasonic carrier for SMA because [36] has demonstrated that most devices' optimal attack frequency is around 25 kHz (22.6~27.9 kHz). Note that due to the inaudibility of the ultrasound, the actual volume of the recorded "trigger" can be variable, as it is affected by the amplitude of our ultrasonic signal and the distance between the transmitter and the victim's microphone. To meet the need for portability, we constrain the maximum ultrasound amplitude at 9 Vpp.

Table 2: The results of digital experiments. Ours$_1$ and Ours$_2$ use the same backdoored model with triggers of -20 dB and -10 dB when inference.

SCD-10

Model	Standard Acc. (%)	ϵ (%)	BA(%) BadNets	BA(%) DABA	BA(%) **Ours$_1$**	BA(%) **Ours$_2$**	ASR(%) BadNets	ASR(%) DABA	ASR(%) **Ours$_1$**	ASR(%) **Ours$_2$**
RNN1	94.348	0.293	**94.464**	94.183	94.018	94.323	94.969	97.410	97.624	**99.142**
		0.440	94.183	94.370	94.300	**94.534**	97.575	98.195	96.347	**99.688**
		0.587	94.089	**94.441**	93.924	93.670	96.455	99.320	97.779	**99.948**
		0.733	94.089	94.487	93.713	**94.699**	96.950	99.660	96.138	**99.974**
RNN2	89.866	0.293	89.397	89.726	89.491	**90.171**	99.505	98.299	97.991	**99.740**
		0.440	89.960	89.561	**90.406**	89.444	99.062	99.294	99.122	**99.636**
		0.587	89.186	89.843	**90.077**	88.928	99.479	99.477	99.613	**99.896**
		0.733	89.537	89.726	88.482	**89.937**	99.400	99.738	99.432	**99.974**
LSTM	91.578	0.587	87.145	86.488	89.960	**90.500**	98.723	95.631	83.187	**99.115**
		0.880	**91.297**	91.133	88.858	90.593	94.473	97.750	93.052	**99.142**
		1.174	**89.303**	89.092	88.975	87.004	97.106	98.378	89.301	**99.428**
		1.466	**90.687**	88.764	89.397	89.045	98.957	96.389	94.990	**99.740**

SCD-30

Model	Standard Acc. (%)	ϵ (%)	BA(%) BadNets	BA(%) DABA	BA(%) **Ours$_1$**	BA(%) **Ours$_2$**	ASR(%) BadNets	ASR(%) DABA	ASR(%) **Ours$_1$**	ASR(%) **Ours$_2$**
RNN1	94.249	0.215	**94.155**	93.691	93.795	93.580	95.575	97.623	97.177	**99.974**
		0.322	94.447	**94.880**	94.001	93.528	96.959	99.243	98.835	**99.974**
		0.429	**94.129**	93.670	93.554	93.872	97.558	99.375	98.641	**99.833**
		0.536	93.872	93.880	**94.112**	93.923	98.863	99.604	97.936	**99.956**
RNN2	87.829	0.215	86.774	87.837	**87.950**	87.778	97.796	97.103	97.971	**99.833**
		0.322	87.718	**88.421**	88.001	87.761	98.898	99.084	98.253	**99.850**
		0.429	86.885	**88.430**	87.383	87.761	98.960	99.525	98.386	**99.868**
		0.536	87.023	**88.550**	87.701	88.001	99.330	99.736	98.880	**99.947**
LSTM	91.833	0.429	92.224	**93.649**	88.181	90.962	98.017	93.011	92.801	**99.410**
		0.644	87.889	**92.619**	91.151	90.988	98.422	91.654	94.124	**99.375**
		0.858	91.958	**93.726**	90.190	**92.224**	98.396	93.168	93.295	**99.674**
		1.073	91.331	**94.962**	90.198	91.271	96.236	98.767	96.056	**99.207**

6.2 Overall Performance

6.2.1 Digital performance. In the digital experiments, we simulate the realistic attack with our microphone model. Different microphone models in our pool will process the well-trained trigger into different nonlinear responses and then randomly add them to the benign audio. As a comparison, the BadNets baseline directly adds its fixed trigger, and the DABA baseline adds its selected and augmented trigger in the audio for evaluation. The number of poisoning samples is 40, 60, 80, 100 in SCD-10 and 80, 120, 160, 200 in SCD-30. The corresponding poisoning rates ϵ are {0.293%, 0.440%, 0.587%, 0.733%} and {0.215%, 0.322%, 0.429%, 0.536%}. According to the settings of [16]. We double the number of poisoning samples for our surrogate LSTM model.

Tab. 2 presents the results of three backdoor attacks. We divide our attacks into two types, Ours$_1$ and Ours$_2$, according to the volume of the triggers used to activate the backdoor. In BadNets, DABA, and Ours$_1$, we use the trigger with a volume of -20 dB, the same as in the training sets. In Ours$_2$, the volume of the activating trigger is set to be -10 dB. Comparing baselines and Ours$_1$, we can find that they can achieve a comparable attack success rate (ASRT) of over 90% and a benign accuracy (BA) as high as that of the clean models. However, if we increase the volume of our trigger to -10 dB, the ASRT can reach over 99%, as shown in the result of Ours$_2$, which obviously outperforms the baseline.

Note that we do not increase the volume of BadNets and DABA. This is because SMA can increase the volume by increasing the amplitude of the ultrasonic signal without the user noticing. Unfortunately, other attacks have to keep a low volume to avoid the risk of being detected.

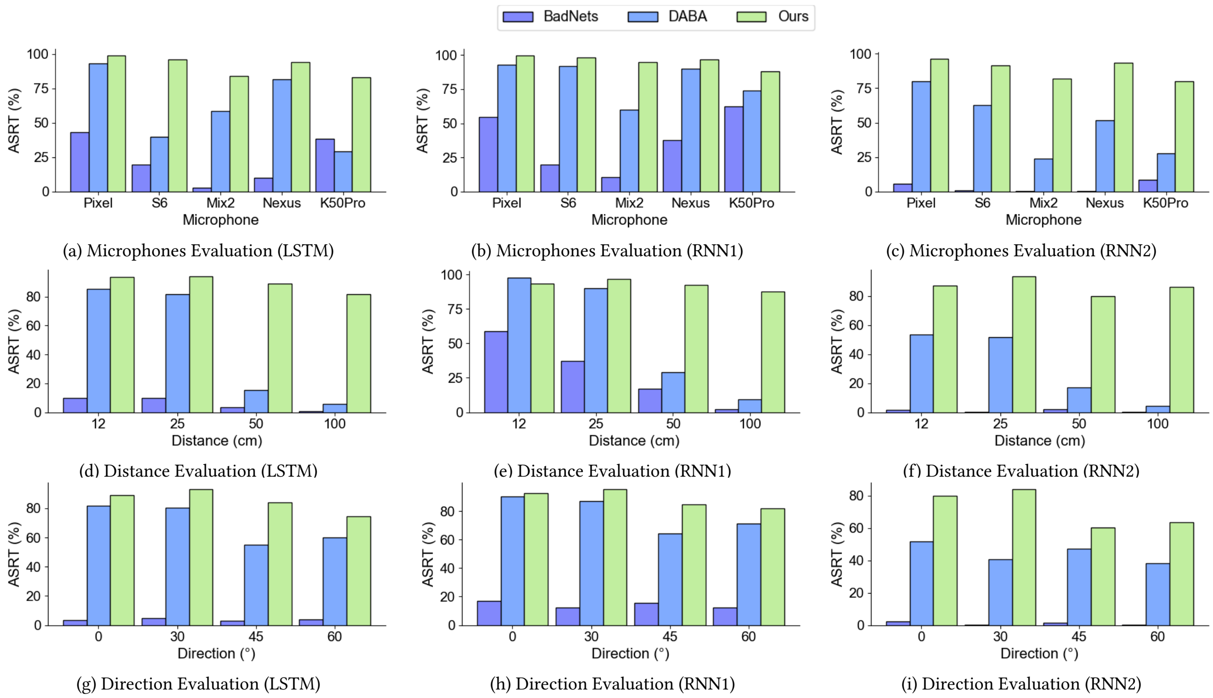
6.2.2 Physical performance. Affected by spatial reverberation and acoustic attenuation, conventional backdoor attacks usually fail over the air. However, with its inaudibility and strong directivity, the ultrasonic signal can better cope with the challenges in the physical domain. To examine the effectiveness of SMA in the physical domain, we conduct physical experiments to evaluate the impact of the microphone, distance, and direction.

We conduct these experiments in a rectangular room of about twenty square meters with an environment noise of about 35 dB. According to [3], 30-40 dB is the normal loudness for a quiet environment. The backdoored model we use is the same as the models of SCD-30 in previous digital experiments. During the attack, we use a JBL loudspeaker to broadcast the audible trigger noise of BadNets and DABA. We fixed the output volume of the loudspeaker so that the decibel meter measured 75 dB at 5 cm and 55 dB at 100 cm. As a reference, the volume for people talking is 50-60 dB, and 75 dB is above the sound of a person singing loudly. As for SMA, we limit the maximum amplitude of the ultrasonic signal generator to 9 Vpp because the larger amplitude requires high-power but non-portable equipment, which is detrimental to the attacker.

Microphones Evaluation. In this experiment, we collect five different types of smartphones. Besides the three microphones we have used in trigger optimization, we also select another two unseen microphones, Redmi K50Pro (2022) and LG Nexus X5 (2015). We set the attack distance is 25cm, and the amplitude of our ultrasonic carrier is set to 5 Vpp. Fig. 6 demonstrates the results of our experiment. The ASRT of BadNets is poor in most settings. DABA performs better than BadNets but is still inferior to SMA in all the settings. In particular, both the performance BadNets and DABA distinguishedly degrade in RNN2. We assume that this small RNN is not robust enough to the perturbation caused by the physical world. However, the ultrasonic signal used by SMA is less affected over the air, maintaining a high ASRT.

Distance Evaluation. In this experiment, we want to ensure that SMA can successfully effect in a reasonable attack distance. We select one of the unseen microphones, Google Nexus, and evaluate the ASRT at four different distances, 12cm, 25cm, 50cm, and 100cm. Theoretically, the higher the frequency, the faster the attenuation. However, the volume of audible signals is limited, while ultrasonic signals are entirely inaudible. Therefore, we can flexibly adjust the amplitude of ultrasonic signals for better ASRT. In this experiment, we set the amplitude to 4 Vpp, 5 Vpp, 8 Vpp, and 9 Vpp, respectively, at 12cm, 25cm, 50cm, and 100cm. As we expected, Fig. 6 shows that the ASRT of BadNet and DABA drops sharply with the distance increasing. As a comparison, SMA can keep a high ASRT, which means our attack is more robust in real-world scenarios.

Direction Evaluation. In the above experiments, we broadcast the trigger directly in front of the microphones. However, considering the real attack scenarios, the victim microphones are usually embedded in a device with a specific reception orientation. Therefore, we evaluate the impact of direction by broadcasting the trigger in four different angles between the reception orientation and the direction of incidence trigger (0∘, 30∘, 45∘, 60∘). As Fig. 6 shown, the varying directions less influence the audible trigger of BadNets and DABA. On the contrary, the performance of SMA is a little affected. We think it is reasonable because ultrasonic signals have



(a) Microphones Evaluation (LSTM) (b) Microphones Evaluation (RNN1) (c) Microphones Evaluation (RNN2)

(d) Distance Evaluation (LSTM) (e) Distance Evaluation (RNN1) (f) Distance Evaluation (RNN2)

(g) Direction Evaluation (LSTM) (h) Direction Evaluation (RNN1) (i) Direction Evaluation (RNN2)

Figure 6: The results of physical experiments. We compare the ASRT of BadNets, DABA, and Ours in the physical domain and evaluate the impact of the microphone, distance, and direction in real-world scenarios.

strong directivity than lower-frequency signals. However, even if the ASRT of SMA drops with the increasing angle, it is still much higher than that of other baselines.

In conclusion, both the digital and physical experiments examine the effectiveness, practicality, and robustness of SMA.

6.3 Resistance to Potential Defense

To ensure the practicality of SMA, we also conduct an experiment to explore its resistance to potential defense against backdoor attacks. We choose two commonly used defenses, fine-tuning and pre-processing. Fine-tuning can clean the backdoored model and mitigate its unexpected response to the trigger. Pre-processing can remove the malicious component in the audio. The backdoored model for evaluation is the RNN1 model trained on the SCD-30 dataset, with a poisoning sample of 200. Similar to the digital experiments, for SMA, we also use two types of triggers with different volumes to activate the backdoor (-20 dB and -10 dB, denoted as Ours_1 and Ours_2. These experiments verify the robustness of SMA. We show the results in Appendix. A.3.

7 DISCUSSION AND FUTURE WORK

Inaudible for training. Although our proposed attack can be implemented inaudibly, the nonlinear trigger in poisoned training data is still audible. Existing work succeed in poisoning training data inaudibly by injecting the inaudible ultrasonic impulse or a frequency band signal [16, 35]. However, such attacks will fail as physical microphones are equipped with low-pass filters or speech enhancement algorithms that can effectively remove the triggers. Given the limitation of the sample rate, it is impossible to introduce an inaudible ultrasonic signal into the training data. We assume that psychoacoustic masking [9] can be used to construct a hard-to-be-noticed trigger in the training data. However, SMA does not adopt it because the psychoacoustic constraints will narrow the optimization space of our trigger. We serve these two goals as a trade-off, which is worth being explored in future work.

Other Speech Features and ASR models. Our evaluation involves three different ASR models to examine the performance of our attack, all of which use MFCCs [13] to represent audio features. As mentioned in Sec. 2, most ASRs regard MFCCs as an excellent acoustic feature, while some ASRs also adopt Fbank [25]. Moreover, there are more ASR models in real applications, some of which are even unknown and inaccessible to the public. Therefore, we believe it is worthwhile to further evaluate our attack's applicability in more different ASR models in future work.

8 CONCLUSION

In this paper, we propose SMA, a practical and inaudible backdoor attack against ASRs that can be generalized to victim-present scenarios in the real world. Our attack exploits the nonlinear effects of microphones and materializes an inaudible trigger by optimizing a crafted ultrasonic signal based on the constructed transfer functions of the nonlinear responses. Our nonlinear trigger can escape the human perception and existing defenses, allowing it to mislead ASR systems without being noticed by the victim user.

9 ACKNOWLEDGMENTS

This work is supported by the Fundamental Research Funds for the Central Universities 110202*17221022301, China NSFC Grant 62201503, 62222114, 61925109, and 62071428.

REFERENCES

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, and Bai et al. 2016. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*. PMLR, 173–182.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 12449–12460.

[3] Birgitta Berglund, Thomas Lindvall, et al. 1995. Community noise. (1995).

[4] Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, and Yang Zhang. 2020. Badnl: Backdoor attacks against NLP models. *arXiv preprint arXiv:2006.01043* (2020).

[5] Yanjiao Chen, Zhicong Zheng, and Xueluan Gong. 2022. MARNet: Backdoor Attacks Against Cooperative Multi-Agent Reinforcement Learning. *IEEE Transactions on Dependable and Secure Computing* (2022).

[6] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access* 7 (2019), 138872–138878.

[7] Douglas Coimbra De Andrade, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. 2018. A neural attention model for speech command recognition. *arXiv preprint arXiv:1808.08929* (2018).

[8] Angelo Farina. 2007. Advancements in impulse response measurements by sine sweeps. In *Audio engineering society convention 122*. Audio Engineering Society.

[9] Stanley A Gelfand. 2017. *Hearing: An introduction to psychological and physiological acoustics*. CRC Press.

[10] Google. [n. d.]. *Speech-to-Text basics*. https://cloud.google.com/speech-to-text/docs/basics 2022.

[11] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* 7 (2019), 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

[12] Peng Huang, Yao Wei, Peng Cheng, Zhongjie Ba, Li Lu, Feng Lin, Fan Zhang, and Kui Ren. 2023. InfoMasker: Preventing Eavesdropping Using Phoneme-Based Noise. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society. https://www.ndss-symposium.org/ndss-paper/infomasker-preventing-eavesdropping-using-phoneme-based-noise/

[13] Uday Kamath, John Liu, and James Whitaker. 2019. *Deep learning for NLP and speech recognition*. Vol. 84. Springer.

[14] Morriel Kasher, Michael Zhao, Aryeh Greenberg, Devin Gulati, Silvija Kokalj-Filipovic, and Predrag Spasojevic. 2021. Inaudible Manipulation of Voice-Enabled Devices Through BackDoor Using Robust Adversarial Audio Attacks. In *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*. 37–42.

[15] Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. 2020. TrojDRL: Evaluation of backdoor attacks on deep reinforcement learning. In *ACM/IEEE Design Automation Conference*. 1–6.

[16] Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. 2022. Can You Hear It? Backdoor Attacks via Ultrasonic Triggers. In *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning* (San Antonio, TX, USA) *(WiseML '22)*. Association for Computing Machinery, New York, NY, USA, 57–62. https://doi.org/10.1145/3522783.3529523

[17] Xinfeng Li, Xiaoyu Ji, Chen Yan, Chaohao Li, Yichen Li, Zhenning Zhang, and Wenyuan Xu. 2023. Learning Normality is Enough: A Software-based Mitigation against the Inaudible Voice Attacks. In *Proceedings of the 32nd USENIX Security Symposium*.

[18] Jian Liu, David T Martin, Karthik Kadirvel, Toshikazu Nishida, Louis Cattafesta, Mark Sheplak, and Brian P Mann. 2008. Nonlinear model and system identification of a capacitive dual-backplate MEMS microphone. *Journal of Sound and Vibration* 309, 1-2 (2008), 276–292.

[19] Qiang Liu, Tongqing Zhou, Zhiping Cai, and Yonghao Tang. 2022. Opportunistic Backdoor Attacks: Exploring Human-Imperceptible Vulnerabilities on Speech Recognition Systems. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 2390–2398. https://doi.org/10.1145/3503161.3548261

[20] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc.

[21] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. 2020. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 182–199.

[22] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multim. Tools Appl.* 80, 6 (2021), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

[23] Antonin Novak and Petr Honzík. 2021. Measurement of nonlinear distortion of MEMS microphones. *Applied Acoustics* 175 (2021), 107802.

[24] Karol J. Piczak. [n. d.]. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia* (Brisbane, Australia, 2015-10-13). ACM Press, 1015–1018. https://doi.org/10.1145/2733373.2806390

[25] Sourabh Ravindran, C. Demirogulu, and D.V. Anderson. 2003. Speech recognition using filter-bank features. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. 1900–1903 Vol.2. https://doi.org/10.1109/ACSSC.2003.1292312

[26] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 2–14.

[27] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. 2020. Dynamic backdoor attacks against machine learning models. *arXiv preprint arXiv:2003.03675* (2020).

[28] Saeid Samizade, Zheng-Hua Tan, Chao Shen, and Xiaohong Guan. 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, 3102–3106.

[29] Google Speech. 2021. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text/.

[30] Siri Team. [n. d.]. *Personalized Hey Siri*. https://machinelearning.apple.com/research/personalized-hey-siri 2018.

[31] Nuttakorn Thubthong and Boonserm Kijsirikul. 2001. Support Vector Machines for Thai Phoneme Recognition. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 9, 6 (2001), 803–813. http://www.worldscinet.com/ijufks/09/0906/S0218488501001253.html

[32] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

[33] Jinwen Xin, Xixiang Lyu, and Jing Ma. 2023. Natural Backdoor Attacks on Speech Recognition Models. In *Machine Learning for Cyber Security: 4th International Conference, ML4CS 2022, Guangzhou, China, December 2–4, 2022, Proceedings, Part I*. Springer, 597–610.

[34] Jianbin Ye, Xiaoyuan Liu, Zheng You, Guowei Li, and Bo Liu. 2022. DriNet: dynamic backdoor attack against automatic speech recognization models. *Applied Sciences* 12, 12 (2022), 5786.

[35] Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. 2021. Backdoor attack against speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2560–2564.

[36] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 103–117. https://doi.org/10.1145/3133956.3134052

[37] Quan Zhang, Yifeng Ding, Yongqiang Tian, Jianmin Guo, Min Yuan, and Yu Jiang. 2021. AdvDoor: Adversarial Backdoor Attack of Deep Learning System. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis* (Virtual, Denmark) *(ISSTA 2021)*. Association for Computing Machinery, New York, NY, USA, 127–138. https://doi.org/10.1145/3460319.3464809

A APPENDIX

A.1 Trigger Design Algorithm

As described in Sec. 5, we propose three optimization objectives to ensure our trigger can be generalized and robust. We further design a loss function as Equ. 5. Here we show the whole trigger design algorithm in Alg. 1.

Algorithm 1 Backdoor Trigger Design Algorithm

Require: Surrogate model Φ_s, the impulse response pool P_{ir}, the noise pool P_n, the benign dataset D, the target label L, the trigger length t, the iterations I, and the learning rate lr.
Ensure: The trigger audio $\mathcal{T}$.

$\mathcal{T} = Initialize(t), t = 0$
while $t < T$ **do**
  $IR_1, IR_2 = Random\_Choice(P_{ir}, 2)$
  $n_1, n_2 = Random\_Choice(P_n, 2)$
  $a = Random\_Choice(D, 1)$
  $Resp_{1,2} = \mathcal{T} \circledast IR_{1,2} + n_{1,2}$
  $O = OneHot(L)$
  Compute the $\mathbb{U}, \mathbb{A}, \mathbb{D}$ and $\mathcal{L}$
  $\mathcal{T} = \mathcal{T} - lr \cdot \partial\mathcal{L}/\partial\mathcal{T}$
  $t = t + 1.$
end while
return $\mathcal{T}$.

A.2 Principle of Saturation Vulnerability of Pre-compensation

Due to the inevitable limitation of the hardware device, a pre-compensation method may cause saturation of the audio and therefore distorts the signal. For example, in Fig. 7, the nonlinear distortion of different microphones varies for the same original signal, especially in the high-frequency range. For some microphones, an intense gain may be necessary to compensate for the nonlinear distortion effectively. However, it may also unnecessarily boost other signals and cause additional distortion due to audio clipping.

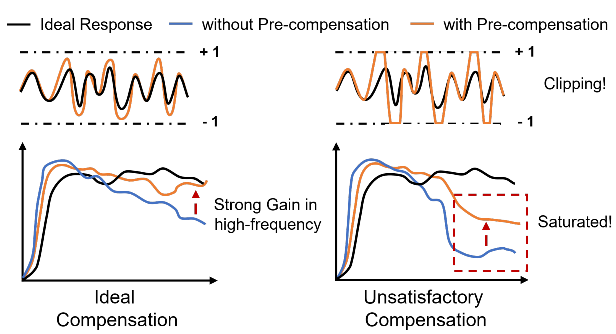


Figure 7: The saturation vulnerability of pre-compensation.

A.3 Result of Resistance to Potential Defense

Here we give more results to evaluate the Resistance to the potential defense of our proposed attack and other baselines.

Fine-Tuning. Fig. 8 shows the ASRT changes of BadNets, DABA, and SMA as the clean-data-based fine-tuning [21] epochs increase from 1 to 25. We can find that the ASRT of BadNets drops obviously after epoch 15, while SMA and DABA maintain a high ASR throughout the fine-tuning. In particular, the ASRT of Ours$_2$ can maintain an ASRT above 99%. This experiment verifies that our design enhances the robustness of SMA, which is less affected by fine-tuning.

Pre-processing Defenses. This experiment is conducted to verify whether SMA can survive in commonly used pre-processing methods. Fig. 9 shows the performance of SMA under three denoised methods, MMSE, Specsub, Wiener, and two signal process methods, quantization and resampling. In quantization, we lower the bit depth of the audio from 16 to 8. In resampling, we down-sample the audio from 16 kHz to 8 kHz and then up-sample back to 16 kHz. These methods will pre-process all the audio before being fed into the ASR model. From the result, We find that denoise methods can slightly increase the BA of the backdoored model while quantization degrades it. However, under all defenses, the ASRT of SMA can maintain at least over 97%, and the ASRT of Ours$_2$ is almost unaffected. Moreover, DABA also performs well against these defenses. However, BadNets are especially vulnerable to resampling. We assume that is because its trigger is evenly distributed over the whole frequency domain (0∼16 kHz). This experiment verifies that our attack is robust under common defense.

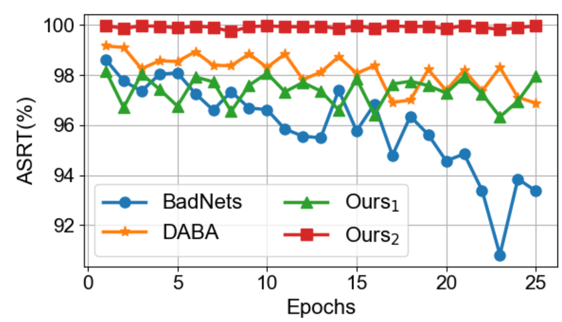


Figure 8: The ASRT of SMA and BadNets against the fine-tuning.

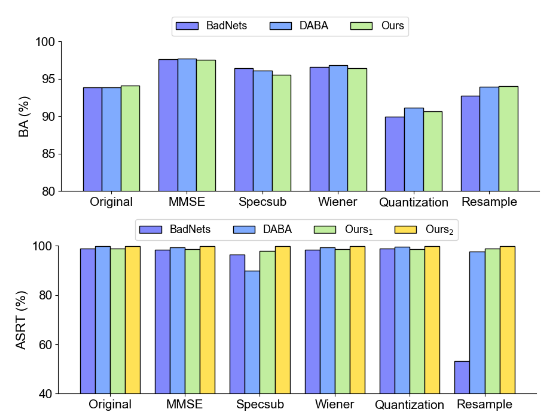


Figure 9: The ASRT and BA of SMA and other baselines against denoise-based defenses. Note that SMA has two types of triggers to attack one same backdoored model.