LaserAdv: Laser Adversarial Attacks on Speech Recognition Systems

Guoming Zhang, Xiaohui Ma, Huiting Zhang, and Zhijie Xiang, *Shandong University;* Xiaoyu Ji, *Zhejiang University;* Yanni Yang, Xiuzhen Cheng, and Pengfei Hu, *Shandong University*

https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-guoming

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

LaserAdv: Laser Adversarial Attacks on Speech Recognition Systems

Guoming Zhang[1], Xiaohui Ma[1], Huiting Zhang[1], Zhijie Xiang[1], Xiaoyu Ji[2],
Yanni Yang[1], Xiuzhen Cheng[1], and Pengfei Hu[1*]
[1]Shandong University *[2]Zhejiang University*

Abstract

Audio adversarial perturbations are imperceptible to humans but can mislead machine learning models, posing a security threat to automatic speech recognition (ASR) systems. Existing methods aim to minimize perturbation values, use acoustic masking, or mimic environmental sounds to render them undetectable. However, these perturbations, being audible frequency range sounds, are still audibly detectable. The slow propagation and rapid attenuation of sound limit their temporal sensitivity and attack range. In this study, we propose LaserAdv, a method that employs lasers to launch adversarial attacks, thereby overcoming the aforementioned challenges due to the superior properties of lasers. In the presence of victim speech, laser adversarial perturbations are superimposed on the speech rather than simply drowning it out, so LaserAdv has higher attack efficiency and longer attack range than LightCommands. LaserAdv introduces a selective amplitude enhancement method based on time-frequency interconversion (SAE-TFI) to deal with distortion. Meanwhile, to simultaneously achieve inaudible, targeted, universal, synchronization-free (over 0.5 s), long-range, and black-box attacks in the physical world, we introduced a series of strategies into the objective function. Our experimental results show that a single perturbation can cause DeepSpeech, Whisper and iFlytek, to misinterpret any of the 12,260 voice commands as the target command with accuracy of up to 100%, 92% and 88%, respectively. The attack distance can be up to 120 m.

1 Introduction

With rapid advances in artificial intelligence, ASR systems have become an integral part of our daily life. These systems are the driving force behind popular technologies such as digital assistants (e.g., Alexa and Open AI Whisper), transcription services and voice-controlled applications, improving the user experience by providing convenient and hands-free control. However, as these systems become more pervasive, they also become attractive targets for adversaries [1, 2]. Adversaries are often motivated by the ability to bypass security mechanisms, gain unauthorized access to systems, or disrupt services. Such attacks facilitate unauthorized access to sensitive information and control over physical security mechanisms, such as unlocking doors, which pose a direct threat to personal safety. Moreover, adversaries can remotely set devices to airplane mode, precipitating denial-of-service attacks, etc.

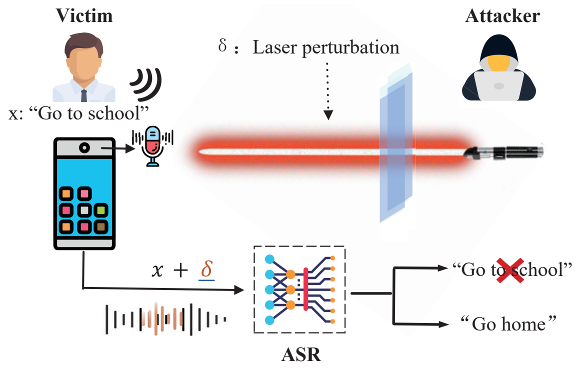


Figure 1: Attack Scenarios.

Previous researchers have increasingly focused on attacking ASR systems using audio adversarial attacks. Traditional methods for generating adversarial perturbations have primarily revolved around constraining the perturbation to a certain limit [3]. In parallel, some studies have gone a step further, which combined human perceptual models and music or ambient noise as a means of constraining the adversarial perturbation while ensuring quality masking [4–9, 11].

Unfortunately, existing research efforts have primarily concentrated on creating adversarial perturbations within the audible frequency range, which poses a significant challenge to achieving complete imperceptibility. In addition, the inherently slow speed of sound propagation results in longer transmission times, preventing effective synchronisation with the victim's voice command. Combined with the rapid attenuation of sound over distance, this can result in significant

[*] Corresponding Author.

perturbation distortion. Taken together, these factors reduce the effectiveness of the attack and may even cause it to fail over long distances. In contrast, the speed of lasers far exceeds that of sound, and they experience slower attenuation over the air. As a result, the laser is time-insensitive and may attack from greater distances. Thus, this is a significant leap from the limited attack distance and imperceptibility of traditional methods, allowing for more practical and extensive adversarial attack scenarios.

In this work, we are trying to answer the following questions: *Is it possible to launch an adversarial attack on ASR with laser beams? And is the attack efficacy of laser adversarial perturbation superior to that of laser commands [10] in the presence of victim speech?*

Inspired by LightCommands [10], we propose `LaserAdv`, a method that leverages laser beams to inject adversarial perturbations into VCSs (Voice Control Systems), as shown in Fig. 1. Compared to LightCommands, `LaserAdv` effectively reduces the signal-to-noise ratio (SNR) required for successful attacks. This reduction in SNR has several notable benefits: 1) **Broader range of vulnerable devices:** Although microphones can detect laser signals due to the photoacoustic effect, not all devices have high sensitivity to these signals, resulting in lower SNRs. The reduction in SNR means that `LaserAdv` is applicable to a wider range of device types; 2) **Improved power efficiency and attack stealth:** Requiring a lower SNR not only reduces power consumption but also increases the stealthiness of the attack signals when the perturbation has been received; 3) **Longer attack range.**

Even though ASR can successfully recognize voice commands injected via lasers, the utilization of lasers to inject adversarial perturbation still faces serious challenges. ASR systems typically undergo pre-processing, such as noise reduction, which can inadvertently mitigate the effects of such perturbations. It is well known that audio adversarial perturbations are often carefully crafted and superimposed on normal voice commands, after being fed into the ASR, the perturbation will be treated as noise and filtered out.

Therefore, *`LaserAdv`* faces the following main challenges: First, unlike the noise and distortion introduced when converting sound, microphones, as transducers designed specifically for sound, exhibit complex and unknown system characteristics when converting lasers, leading to the introduction of various unknown signal distortions. Second, it is challenging to simultaneously achieve inaudible, targeted, universal (one perturbation works on more than 10,000 speech), synchronisation-free (up to 0.5 s), long-range (over 120 m), and black-box attacks (successfully tested on black-box models)in the physical world. We refer to adversarial perturbations with the above attack capabilities as an integrated adversarial perturbation (IAP).

To address these challenges, we first investigated the bandwidth requirements of ASR for adversarial examples. Following this, we utilized laser pulse signals and multi-frequency signals to excite the laser channel of the microphone, investigating the system response characteristics and possible introduced distortions. We found that the frequency response of the laser channel drops off sharply with increasing frequency. To meet the bandwidth requirements and mitigate the distortion induced by frequency selective fading (FSF), we proposed a method called selective amplitude enhancement based on time-frequency interconversion (SAE-TFI). This method allows selective control of the fading components.

In addition to the distortion caused by FSF, microphones are also relatively insensitive to lasers, resulting in a low amplitude in the captured signal. To solve this problem, we imposed device-dependent constraints on the adversarial perturbation during its generation, while ensuring low computational complexity. As the adversarial perturbation works together with victim speech rather than drowning it out, the need for high sensitivity to the laser is reduced, making it suitable for a wider range of devices. The effectiveness and range of the attack are also significantly improved.

Third, to implement IAP, we introduced strategies for asynchrony, and content independence in the objective function. In particular, to ensure attack effectiveness in black-box scenarios simultaneously, we proposed a method for extracting target command attributes and model attributes based on audio data with a similar distribution. We have effectively tackled all challenges and validated `LaserAdv` in realistic scenarios. In summary, our contributions can be summarized as follows:

- We present *LaserAdv*, an approach that injects adversarial perturbations into state-of-the-art ASR systems via laser beams.
- We conducted a comprehensive analysis of the system characteristics of the laser channel to identify various sources of signal distortion. Based on these findings, we propose an innovative selective amplitude enhancement method to address the distortion. Additionally, by introducing different strategies, we effectively generate IAPs.
- We evaluate *`LaserAdv`* on three ASR systems (DeepSpeech, Whisper, iFlytek) across 6 devices. Our results show that more than 10,000 voice commands can be compromised into the same targeted sentence with the addition of a single laser IAP in black-box settings. Furthermore, in the presence of victim speech, the maximum range of attack can be up to 120 m without using a telephoto lens.

2 Background

2.1 Traditional Audio Adversarial Perturbations

Previous researchers have made considerable efforts to ensure that adversarial perturbations are imperceptible to humans, and these efforts can be categorized as follows: 1) Minimizing the value of adversarial perturbations δ [3]; 2) Acoustic masking: Using the psychoacoustic model [8] or human perception model [11] to constrain the amplitude of adversarial perturbations. 3) Environmental sound mimicking: Making the adversarial perturbations sound like environmental sound [5].

Minimizing the Value of δ. The basic idea of this type of method is to limit the perturbation to a relatively small range, so the optimization problem for creating an audio adversarial perturbation can be formulated as follows:

$$arg\min_{\delta}\ \mathcal{L}(f(x+\delta),y') + \varepsilon\|\delta\|_2 \qquad (1)$$

where the target ASR system can be modeled as $f(\cdot)$, $\mathcal{L}(\cdot)$ is the loss function, x represents the original audio, δ is the generated adversarial perturbation, y' is the target transcription, and ε is the constraint hyperparameter used to limit the amplitude of the adversarial perturbation to increase concealment.

Acoustic Masking. In addition to simply using the L_p distance to limit the strength of the adversarial perturbation, some studies use the psychoacoustic principle of auditory masking to selectively introduce these perturbations into regions of the audio that are inconspicuous to the human ear [6]. The loss function can thus be formulated as follows:

$$\mathcal{L}(f(x+\delta),y') + \varepsilon \cdot L_{\theta}(x,\delta) \qquad (2)$$

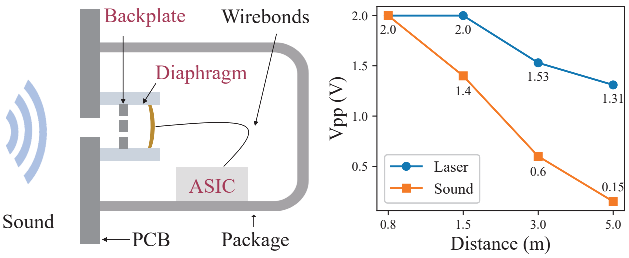
where L_{θ} constrains the normalized power spectral density (PSD) estimation of the perturbation to be under the frequency masking threshold of the original audio [8]. While the absolute amplitude of δ can be fine-tuned in alignment with the masking threshold, allowing larger δ values, the maximum value is still confined to a limited range.

Environmental Sound Mimicking. In order to make audio adversarial perturbations imperceptible to humans, another research effort has incorporated environmental sounds such as birds singing, car horns, and HVAC noise into the perturbations [5, 12]. This method ensures that the perturbations are sufficiently loud to maintain robustness in physical attack while remaining difficult for humans to detect. Given a chosen environmental sound template $\hat{\delta}$, the attack objective can be expressed as:

$$arg\min_{\delta}\ \mathcal{L}(f(x+\delta),y') + \theta \cdot dist(\delta,\hat{\delta}) \qquad (3)$$

where $dist(\delta,\hat{\delta})$ denotes the distance between adversarial perturbation and the sound template according to a chosen distance metric, e.g., $L2$ distance [5], time-frequency pattern difference [12].

Although current adversarial perturbation attacks effectively minimize perceptibility by controlling or manipulating the magnitude of perturbation δ through various methods, the acoustic perturbations δ may still be audibly detectable under specific conditions, meanwhile, the attack distance is limited. Therefore, it is essential to investigate adversarial perturbations that are entirely inaudible and capable of facilitating long-range attacks.



(a) MEMS microphone (b) Attenuation characteristics

Figure 2: Structure of MEMS microphone and attenuation characteristics of sound and laser beam.

2.2 MEMS Microphone and Vulnerability

Due to their small size, robustness, high performance and low cost, MEMS microphones have become increasingly common in modern electronic devices, especially those with voice-based interaction capabilities such as smartphones, smart home devices and laptops. The structure of the MEMS microphone is shown in Fig. 2 (a). Nevertheless, the Light-Commands [10] attack confirmed the optical coupling phenomenon of the MEMS microphone, where changes in light intensity can generate an output voltage in the MEMS microphone. This makes it possible to inject signals from a distance, which presents both intriguing possibilities and worrying vulnerabilities. However, adversarial perturbation is more sensitive to distortion than voice commands, which introduces uncertainty when using the laser beam to carry the perturbation.

2.3 Attenuation Characteristics of Sound and Laser

Sound waves and laser beams exhibit unique attenuation characteristics as they propagate over a distance. In order to quantitatively analyze and compare the attenuation rate of sound and laser, we have carried out the following experiments. We initially set the distance between the signal sources (i.e. laser diode and loudspeaker) and the microphone at 0.5 m, carefully adjusting the drive power of both the laser and the loudspeaker to ensure that the signal strength received by the microphone remained constant at this distance. We then gradually increased the distance between the signal sources and the microphone at 1.2 m intervals, systematically recording

Table 1: Comparison of different bandwidths.

Frequencies	Numbers	Loss	Success Rate
0.1 - 0.6 kHz	21	8.65	-
0.1 - 0.8 kHz	323	0.70	96.90%
0.1 - 1 kHz	337	1.15	84.57%
0.1 - 2 kHz	377	0.35	99.73%
0.1 - 3 kHz	383	0.49	99.74%
0.1 - 4 kHz	382	0.11	99.74%

the received signal strength at each step. The maximum distance was set at 5 m.

The results are shown in Fig. 2 (b), we can observe that the attenuation of sound decreases rapidly as distance increases, while the attenuation of lasers is comparatively gradual. When the distance is expanded from 0.5 m to 5 m, the sound's intensity is reduced to 7% of its original value, whereas the laser's intensity is only reduced to 66.6% of its initial level. This difference, which is approximately an order of magnitude, clearly demonstrates the significant advantage of the laser for long-range attacks.

3 Feasibility Analysis

In this section, We provide a comprehensive analysis by considering both the bandwidth requirements of adversarial perturbations for effective ASR attacks and the two key characteristics of MEMS microphones: 1) the system response to the laser, and 2) the distortion of the signal within the laser channel. These factors provide insight into the practical challenges and potential opportunities for using lasers in audio adversarial attacks.

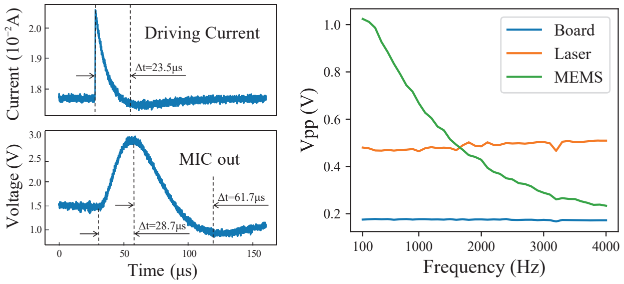
3.1 Bandwidth Requirements

Careful selection and tuning of the bandwidth are essential to the effectiveness of the attack, allowing the perturbations to exploit weaknesses in the ASR system's processing of audio signals. We generated adversarial perturbations by configuring different bandwidths and tested the attack performance on different ASRs.

As shown in Tab. 1, with the broadening of the frequency band, the number of generated perturbations increases and the loss decreases, and the success rate also increases gradually. Therefore, in order to successfully capture the wideband signal that is essential for carrying out black-box attacks in physical scenarios, it is better to have a high and flat frequency response in both the low (100 Hz) and high (4 kHz) frequency ranges.

3.2 Responses Characteristic of MIC to Laser

To investigate the feasibility of laser adversarial perturbations, we conducted a series of experiments to comprehensively analyze the system responses of microphones to laser beams, focusing specifically on both impulse response and frequency response aspects.



(a) Impulse response (b) Frequency response

Figure 3: The system response of MEMS microphone to laser beams.

Impulse Response. The impulse response describes the time domain characteristics, a microphone with a good impulse response can accurately and faithfully transmit and convert complex signals and sudden changes, resulting in high-fidelity signal reproduction. Since it is impossible to produce a perfect impulse signal with physical equipment, we have developed a driving current of laser diodes that resembles an impulse by changing the frequency and using different types of waveforms, as shown in Fig. 3 (a) (Top). This waveform takes about 23.5 us to return to its initial state compared to an ideal pulse. The impulse response of the microphone, shown in Fig. 3 (a) (Bottom), although the input signal is instantaneous, the microphone takes approximately 28.7 us to respond and approximately 61.7 us to return to its initial state. For common applications such as general recording and speech recognition, an impulse response time within a few milliseconds is usually acceptable. The impulse response to the laser is therefore good.

Frequency Response. To analyze the frequency response to laser, we systematically generated single-tone sinusoidal signals (denoted s) within a frequency range of 100 Hz to 4 kHz, with a 200 Hz interval between successive frequencies. This signal is simultaneously input into two identical laser driving devices, designated as Device 1 and Device 2, for modulation. We record the output l_{out} from laser driving Device 1 using a data acquisition card, while Device 2 directly drives a laser diode. The emitted laser is then injected into a MEMS microphone positioned 50 cm away, with the recorded sound denoted as m_{out}. In addition to using the MEMS microphone to receive the laser signal, we also employ a photodetector (THORLABS APD430A2) to detect the laser's intensity, represented as p_{out}.

The frequency response curves of these three devices are shown in Fig. 3 (b). The curves corresponding to the laser drivers and photodiodes are relatively smooth, whereas the

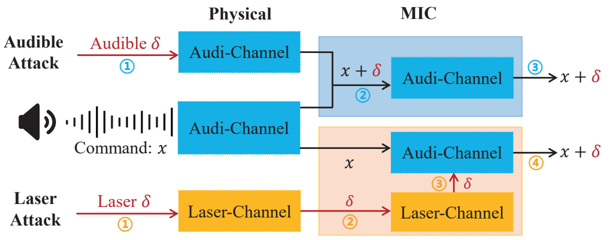


Figure 4: The transformation process of an adversarial perturbation δ in audible and laser channels. The audible and laser channels involve both the physical space and the microphone circuit.

frequency response of the microphone shows a pronounced drop with increasing frequency, from 1.02 V at 100 Hz to 0.2 V at 4 kHz. This result indicates that it's a FSF channel for laser signals, resulting in significant signal distortion. Such a characteristic limits the effectiveness of laser perturbation, especially given the stringent bandwidth requirements. Therefore, we have developed a specific method to effectively combat the FSF channel in the design section.

3.3 Distortion within Laser Channels

Theoretical Analysis. To gain a deeper understanding and facilitate a comparison of signal transformation and distortion in both the audible and laser channels, we first conduct a theoretical analysis combining the above results. As depicted in Fig. 4, given an audible adversarial perturbation $\delta(t)$ and voice command $x(t)$, the recorded signal $x'(t)$ can be expressed as $x(t)+\delta(t)$, ignoring the electrical and ambient noise present in the physical and electrical system. Nonetheless, in real-world scenarios, various noises are introduced, ultimately resulting in the received signal being represented as:

$$x'(t) = H_a(x(t)+\delta(t))+n_a(t) \tag{4}$$

Here $n_a(t)$ represents the noise in the audible channel, e.g., ambient and electrical noise. $H_a(\cdot)$ is the transfer function of the audible channel.

At the same time, the transformation of the laser perturbation is shown in Fig. 4. Unlike audible perturbation, laser perturbation is transmitted through the laser channel within the physical and electrical systems. Given an laser adversarial perturbation δ and a voice command $x(t)$, the received signal can be expressed as:

$$x'(t) = H_a(x(t))+H_u(\delta(t))+n_a(t)+n_u(t) \tag{5}$$

where $n_u(t)$ denotes the noise within the laser channel. $H_u(\cdot)$ is the transfer function of the laser channel, which leads to other types of signal distortion caused by factors such as FSF and weak response. Consequently, the overall signal distortion is aggravated compared to audible perturbations, owing to the inherent characteristics of the laser channel.

Evaluation. To investigate the distortion characteristics of both audible and laser channels, and analyze the feasibility of LaserAdv, we first generate an audio adversarial perturbation with a bandwidth ranging from 100 Hz to 4 kHz. We then play and record perturbation using a loudspeaker and five devices, including a MEMS microphone and four smartphones, designated as Enjoy 20 Pro, Honor 20 Pro, Redmi K30 Ultra, and Samsung Galaxy S9.

To obtain the laser perturbations, we fed the perturbation into the laser driver for modulation, subsequently using it to drive the laser diode. The emitted laser was then injected into the five devices. Fig. 5 illustrates the recorded perturbation of the five devices. The top five images from (a) to (e) display the recorded acoustic perturbations with very subtle distortions, while the bottom five images from (f) to (j) reveal the recorded laser perturbations. Each of these images exhibits two to three types of distortions: 1) the introduction of additional noise; 2) weak response, resulting in a low SNR; and 3) distortion caused by FSF. In subsequent designs, we will propose corresponding methods to address the distinct distortion.

These findings provide a nuanced insight into how to carefully design the adversarial perturbations to address the specific types of distortion encountered.

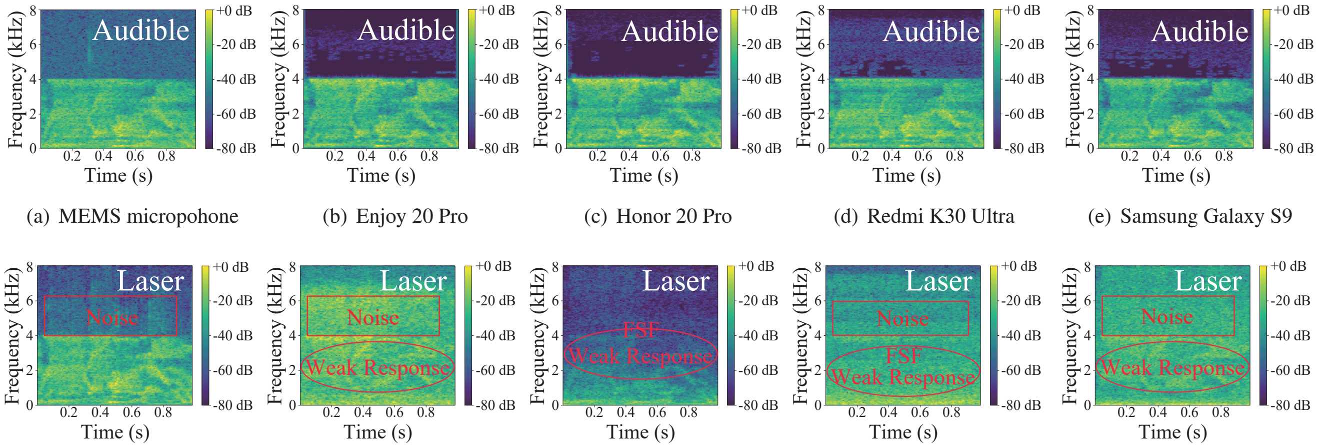
3.4 Threat Model

Attack Goal. The adversary's objective is to manipulate the output of the target's black-box ASR system from a distance of ten metres. This is achieved by introducing an laser adversarial perturbation that is unrelated to the content or timing of the victim's speech. Unlike the approach used in DolphinAttack [42] and LightCommands [10], LaserAdv assumes that the victim is speaking any voice commands at the time of the attack. In summary, the adversary must simultaneously fulfil the following conditions:

- *Synchronization-free*. In LaserAdv, synchronisation-free describes scenarios where adversarial perturbations, although delayed by up to 0.5 seconds, can still affect the first 0.5 seconds of voice commands. This indicates that the perturbations do not require precise synchronisation with the voice commands to be effective.

- *Transferability*. This refers to the adversarial perturbations generated for the white-box ASR system DeepSpeech, which can be utilized to attack black-box systems such as iFlytek and Whisper.

- *Universal*. It refers to the ability of a perturbation to effectively interfere with over 10,000 user's voice commands. Note that we do not use the term *Generalizability* to illustrate the LaserAdv attack.



(a) MEMS micropohone (b) Enjoy 20 Pro (c) Honor 20 Pro (d) Redmi K30 Ultra (e) Samsung Galaxy S9

(f) MEMS micropohone (g) Enjoy 20 Pro (h) Honor 20 Pro (i) Redmi K30 Ultra (j) Samsung Galaxy S9

Figure 5: Recorded acoustic adversarial examples (top) and laser-based adversarial examples (bottom) with 5 different devices.

- *Inaudible and targeted.*

Assumption on Attacker's Knowledge. We have designed our threat model to reflect a realistic scenario where an attacker with limited resources, only has detailed knowledge of one ASR system - DeepSpeech. The other two systems are state-of-the-art commercial and black-box ASR systems (Whisper and iFlytek) about which we have no detailed internal knowledge. Our study aims to demonstrate the potential transferability of the attack across different systems. Therefore, specific knowledge of these systems is not required to validate the effectiveness of our approach. The adversary has no knowledge of the content and timing of the victim's voice commands.

Adversary's Capability. `LaserAdv` assumes that laser perturbations are emitted when the victim is actively speaking. Since the adversarial perturbations are transmitted via laser rather than conventional audio, it is assumed that the adversary is equipped with the necessary equipment that can modulate and transmit laser beams. The total price for the laser driver, the laser diode and the battery are less than $20. Similar to other long-range attacks, the lack of visual or auditory feedback from the target device prevents the attacker from promptly determining the success of the attack. Typically, multiple attempts will be made to ensure the attack achieves its intended results. When LaserAdv achieves universal, the perturbations do not require synchronization with the spoken words. This capability allows the system to effectively manipulate voice commands regardless of their timing or content.

Assumption on the Victim's device. Since certain commands can only be executed on an unlocked device, in this case, we assume that the smart device is unlocked. For other commands, the device may be either locked or unlocked. As the smart device is already awakened during the interaction with the victim, we do not need to assume whether the device is specifically trained to recognize the speaker's voice.

4 `LaserAdv` Design

4.1 Basic Problem Formulation

In prior research, the basic optimization problem formulation should consider the perturbation to be imperceptible to humans. Consequently, strict constraints on the perturbation were necessary. The constraints limit the feasible space of solutions, making it more challenging for optimization algorithms to find the optimal solution. This may result in slower convergence, increased complexity, greater computational resource demands, etc.

In our study, we aim to find an adversarial perturbation δ that meets the condition $f(x+\delta) = y'$. Unlike previous methods, the adversarial perturbation δ is not subject to strict constraints for imperceptibility. Thus, the basic optimization problem can be expressed as follows:

$$\arg\min_{\delta} \mathcal{L}(f(x+\delta), y') \tag{6}$$

where $\mathcal{L}(\cdot)$ refers to the loss function of a white-box system, which in our work is DeepSpeech. Due to the lack of constraints on δ, solving the above optimization problem becomes much simpler and faster.

4.2 Transferability in Black-box ASRs

Transferability, while advantageous for an attacker, poses several challenges in crafting successful adversarial perturbation. Transferability arises from the observation that different ASR models, despite their unique structures and parameters, often capture similar high-level features for targeted voice commands. This consistency not only leads to identical results for similar speech inputs but also creates uniform vulnerabilities across models. Based on this, we considered a wide range of other voice commands when generating perturbations for target commands. The dataset contains various elements of speech, including different volumes, accents, speech rates, background noise and more. As a result, the generated perturbations, when combined with any other voice command, have high-dimensional features similar to those of the target command. This makes them potentially effective for attacking models with different architectural parameters. During the perturbation generation process, we sampled audio inputs from this assembled dataset. Thus, the optimization problem can be expressed as follows:

$$\arg\min_{\delta} \mathop{\mathbb{E}}_{x\sim\mathcal{S}} \mathcal{L}(f(x+\delta),y') \tag{7}$$

where $\mathcal{S}$ represents the similar distribution of the audio inputs, and x is randomly sampled from $\mathcal{S}$.

4.3 Time and Content Independent

In realistic scenarios, the laser adversarial perturbation $\delta \in \mathbb{Q}^{1\times M}$ should be insensitive to both the relative position within the victim speech and the actual content of the victim speech $x \in \mathbb{Q}^{1\times N}$, where M,N are the length of δ and x.

In the process of perturbation generation, we randomly choose a time delay τ uniformly within the range from 0 to $N-M$ to compute the gradient at each iteration, thus the objective function Eq. 8 can be expressed as follows:

$$\arg\min_{\delta} \mathop{\mathbb{E}}_{x\sim\mathcal{S},\tau\sim\mathcal{T}} \mathcal{L}(f(x+\delta(t-\tau)),y') \tag{8}$$

Let $\mathcal{T} = \{kd \mid k \in \mathbb{N}, 0 \le k \le \frac{M}{d}\}$, where d is the number of sample points.

In LaserAdv, the value of d can be set to greater than 20, thereby reducing the number of iterations compared with previous work [5], where d is set to 1 sample point. That is because the laser perturbation is more position-insensitive and robust due to no strictly constrained amplitude. To validate this, we use Eq. 1 and Eq. 6 to generate two different adversarial perturbations, denoted δ_1 and δ_2. It takes approximately 1000 iterations to generate an effective δ_1, whereas only 200 iterations are needed for δ_2. The results show that the final successful attack delay ranges for δ_1 and δ_2 are $0-10$ and $0-50$ respectively. The number of iterations required to reach convergence directly correlates with the algorithm's complexity. Therefore, when using different objective functions to generate two distinct δ, the iteration count provides insight into which algorithm is more efficient. This is critical for practical applications where computational resources and time are constrained. In LaserAdv, the value of d can be set to 20 or even higher. This approach leads to quicker convergence of the objective function with fewer iterations while maintaining effectiveness.

To make the adversarial perturbation universal, the optimization problem should consider the generalization of the perturbation across a wide range of audio inputs. To achieve this goal, we propose the following optimization problem based on Eq. 8:

$$\arg\min_{\delta} \mathop{\mathbb{E}}_{\tau\sim\mathcal{T},\ x\sim\{\mathcal{S},\mathcal{D}\}} \mathcal{L}(f(x\cdot i+\delta(t-\tau)),y') \tag{9}$$

where $\mathcal{D}$ represents the distribution of the audio inputs x, Parameter i, which is adjusted between 0.1 and 1, is specifically designed to normalize and adjust the volume of audio inputs within the dataset. This adjustment not only enhances the efficiency of our perturbation generation but also helps prevent non-convergence issues during the optimization process.

4.4 Physical Adversarial Perturbation

According to the preliminary analysis, there are two kinds of distortion that should be carefully considered and addressed in LaserAdv.

Dealing with Low Sensitivity. Reviewing the aforementioned results of the feasibility analysis, we can see that some devices with MEMS microphones are insensitive to lasers and receive only a low intensity of laser-induced adversarial perturbations. However, the intensity of the adversarial perturbation is a critical factor in the success of the attack. As the intensity of the perturbation decreases, so does the success rate of the attack. Therefore, to address the issue of reduced intensity due to the limited gain of microphones, we impose certain constraints on the amplitude of the perturbation within our optimization problem. The parameter b is determined by the device's frequency response. When the frequency response is low, b is set to a smaller value, ensuring that even minimal perturbations remain effective. Consequently, successful attacks can be achieved even when the device receives only small amplitude perturbations. Furthermore, by setting a lower bound a on the perturbation, we avoid overly stringent constraints that could hinder the generation process. In LaserAdv, The values for a and b are based on experience and the frequency response of the equipment.

Dealing with FSF Channel. FSF causes wide-bandwidth perturbation to be attenuated unevenly across the spectrum after passing through the laser channel. Specifically, the high-frequency components experience much greater attenuation than the low-frequency components. This distinctive distortion phenomenon stands apart from distortions caused by ambient noise or low sensitivity, making it impervious to remedies such as the introduction of random noise or similar measures.

In LaserAdv, we propose a Selective Amplitude Enhancement method based on Time-Frequency Interconversion (SAE-TFI) aimed at compensating for the attenuation of high-frequency components. To begin, we apply Short Time Fourier Transform (STFT) to the generated perturbation δ to convert it from the time domain to the frequency domain, which can be expressed as:

$$S(\delta) = STFT\{\delta(t)\}(\tau,\omega) = \int_{-\infty}^{\infty} \delta(t)h(t-\tau)e^{-j\omega t}dt \quad (10)$$

where $h(\cdot)$ is the Hann window function, and δ(t) is the converted perturbation signal. $S(\delta)$ is the complex spectrum of δ after applying STFT. $S(\delta)$ corresponds to a complex value at each time t and each frequency f. Next, we use Eq. 11 to obtain the amplitude spectrum and phase spectrum of the perturbation δ:

$$\begin{aligned} Amp &= abs(S(\delta)) \\ Phase &= angle(S(\delta)) \end{aligned} \quad (11)$$

where $abs(\cdot)$ means taking the absolute value of a complex number, $angle(a+bj) = arctan(\frac{b}{a})$. In order to recover the missing high-frequency components as much as possible, we perform linear enhancement on the amplitude spectrum using Eq. 12:

$$\hat{Amp} = \begin{cases} coef \cdot Amp, & if\ f > cut\_off, \\ Amp, & otherwise \end{cases} \quad (12)$$

where $coef$ is the coefficient representing the enhancement ratio, and cut_off represents the cut-off frequency. Although the linear method cannot completely reverse the unique distortion, it can effectively alleviate the high-frequency loss caused by FSF to a certain extent.

Finally, the inverse short-time Fourier transform (iSTFT) process is applied to obtain the enhanced time domain perturbation signal, and the formula is as follows:

$$\hat{\delta} = iSTFT\{\hat{Amp} \cdot e^{j \cdot Phase}\} \quad (13)$$

where $\hat{\delta}$ is the enhanced perturbation signal, and the function of iSTFT is to transform the specified complex spectrum into a time domain signal.

After thoroughly considering the other various distortion factors (i.e., ambient noise, etc.), we formulate our optimization problem as follows:

$$\arg\min_{\delta} \mathop{\mathbb{E}}_{\tau\sim\mathcal{T},\ x\sim\{\mathcal{S},\mathcal{D}\},\ h\sim\{H_1,H_2\}} \mathcal{L}(f(x\cdot i+\delta'),y') \quad (14)$$

$$\text{subject to} \quad a \le \delta' \le b$$

where, $\delta' = h \otimes F(\delta(t-\tau)) + n$, a and b are parameters restricting the amplitude of the perturbation δ', h is the room impulse response (RIR) sampled from the collected distribution H_1 and H_2 in the audible channel and laser channel respectively. n denotes the Gaussian white noise, and $F(\cdot)$ represents the band-pass filter. The RIR of H_1 is chosen from the Database [13]. To obtain the impulse responses of the laser channel, we play and record the impulse using a laser diode and microphone. These noise components effectively simulate the complexity of the laser channel state, also indicating the necessity of considering these complexities in the design and implementation of LaserAdv.

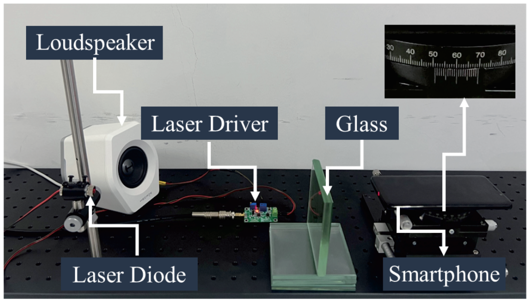


Figure 6: Illustration of the experimental setup.

After solving Eq. 14, we obtain an adversarial perturbation within the bandwidth of 100 Hz to 4 kHz. By applying the SAE-TFI method, we are able to create adversarial perturbations taking into account all the factors involved.

5 Evaluation

5.1 Experiment Settings

Prototype. We implement LaserAdv using the TensorFlow framework on a server running Ubuntu 16.04 with an NVIDIA Tesla V100-16GB GPU. The default configuration is set as follows: $d = 20$, the maximum number of iterations is 2,000, and the frequency range of the band-pass filter $F(\cdot)$ is set to 100 Hz - 4 kHz. The maximum time delay τ between x and δ is set to 1.2 seconds. The target ASR models for our experiments are DeepSpeech, iFlytek, and Whisper. To modulate and emit the laser adversarial perturbations, We utilize a 5mW red laser diode with a wavelength of 650 nanometers, directing the beam vertically towards the smartphone's microphone. Simultaneously, we position a loudspeaker adjacent to the microphone to play voice commands during the attack, as shown in Fig. 6. Notably, we employ a telescope to obtain long-distance line of sight, precisely aiming the laser at the microphone's aperture for launching long-distance attacks.

Evaluation Metrics. We use the following metric to quantify the performance of LaserAdv.

Attack success rate. The success rate measures the proportion of adversarial perturbations that successfully cause the target ASR system to misclassify the perturbed audio signal. A higher success rate indicates a more effective adversarial

Table 2: Performance of IAP under different scenarios.

No.	Voice commands	$\tau = 0$ *seconds*			$\tau = 0.5$ *seconds*		
		DeepSpeech	**iFlytek**	**Whisper**	**DeepSpeech**	**iFlytek**	**Whisper**
1	Airplane mode on	100%	100%	100%	100%	100%	100%
2	Open the window	100%	80%	94%	100%	60%	62%
3	To be or not to be	100%	96%	100%	100%	76%	100%
4	Save driving records	100%	82%	90%	100%	58%	80%
5	Ok google	100%	98%	90%	100%	66%	100%
6	Chat with me	100%	86%	100%	100%	80%	100%
7	Listen to the broadcast	100%	94%	100%	100%	42%	94%
8	Turn on the wipers	100%	92%	94%	100%	84%	80%
9	News broadcasting	100%	92%	92%	100%	82%	92%
10	Open the file	100%	88%	90%	100%	84%	66%
11	Screen sharing	100%	98%	84%	100%	88%	98%
12	Start playing	100%	94%	82%	100%	90%	96%
13	Stop playing	100%	94%	100%	100%	68%	100%
14	Tell a story	100%	88%	78%	100%	58%	56%
15	Turn down the volume	100%	64%	94%	100%	72%	64%
16	Turn left	100%	94%	90%	100%	82%	100%
17	Turn right	96%	92%	92%	100%	100%	94%
18	Turn on the bluetooth	100%	64%	88%	100%	72%	92%
19	Turn on seat heating	98%	98%	86%	98%	86%	78%
N	...	...	...	...	...	...	...
12260	What's the time	98%	74%	96%	100%	52%	100%
Attack Success Rate		12260/12260	12258/12260	11925/12260	12255/12260	12215/12260	12067/12260

attack. Note that a successful attack is only achieved if the output of the ASR matches the target sentence perfectly. In other words, partial recognition of the command is insufficient for success.

Dataset. *Impulse Response.* We use 10 audible impulse responses randomly selected from database [13] and 30 laser impulse responses recorded by ourself.

Voice commands. In our experiments, by leveraging Google's text-to-speech service [15], we generate a diverse and high-quality dataset of voice commands covering various phrases, which we consider to be the original voice commands denoted as x.

5.2 Overall Performance

In this section, we present a detailed evaluation of the performance of laser adversarial perturbations, and the results are shown in Tab. 2. We first applied the perturbation to a subset of 12,260 voice commands from our dataset to evaluate the effectiveness in a digital scenario. The evaluation was performed under two different conditions: one with no delay and the other with a delay of 0.5 seconds imposed on the perturbation. Interestingly, our results showed that the success rate of the attack remained consistently high under both delay conditions for all three different ASR models tested (namely DeepSpeech, iFlytek and Whisper). Specifically, we observed an attack success rate of over 98% for all models, reaching a full 100% for DeepSpeech without delay.

In the physical scenario, we randomly selected a subset of 20 audio samples previously identified as vulnerable in the digital scenario. These experiments were conducted using a smartphone to record the adversarial perturbations. For each original voice command, we performed 50 attack attempts. We can see that DeepSpeech was the most vulnerable to the `LaserAdv` attack. A 100% success rate was consistently achieved across multiple voice commands. In contrast, the iFlytek model was the most resistant; in particular, when the delay was set to 0.5 seconds, the attack success rate dropped to 75%. Nevertheless, our adversarial perturbation scheme demonstrates a robust, synchronisation-free black-box attack capability in a physical environment.

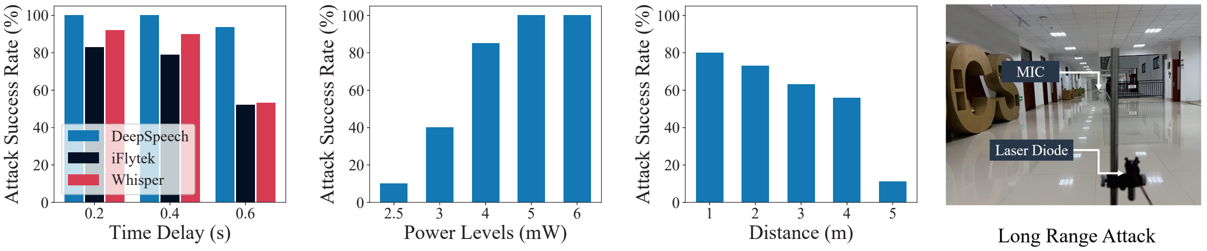
Table 3: Attack success rate under different perturbation durations (60 experiments per length).

Lengths	30%	40%	50%	60%	70%	80%	90%
Success rate	82%	100%	100%	100%	100%	100%	100%

5.3 Impact of Perturbation Duration

In order to evaluate the effect of perturbation duration on the attack success rate of `LaserAdv`, we conducted an experimental analysis where the length of the adversarial perturbation was varied over a wide range, specifically 30%, 40%, 50%, 60%, 70%, 80% and 90% of the voice command duration. For each predetermined duration, a series of 60 different attack attempts were made against the ASR system.

The results of these trials are systematically presented in Tab. 3. The results show that when the duration of the adversarial perturbation exceeds 40% of the total length of the voice command, `LaserAdv` is able to achieve an impeccable attack success rate of 100%. Notably, even when the perturbation is limited to only 30% of the voice command, `LaserAdv` still achieves a substantial success rate of 82%. These results show that `LaserAdv` demonstrates robust performance even when the perturbation is relatively short. Therefore, the perturbation can be successful even when it lags behind the normal voice command, thereby reducing the time sensitivity.



(a) The impact of time delay.

(b) The impact of varying laser power levels.

(c) The impact of distance between laser diode and smartphone.

(d) The diagram of long-range attack scenario.

Figure 7: Illustration of the performance of LaserAdv and a long-range attack.

Table 4: Comparison of long-range attack performance.

Attack range	LaserAdv	LightCommands
20 m	100%	50%
40 m	100%	25%
60 m	95%	15%
80 m	80%	5%
100 m	65%	-
120 m	15%	-

5.4 Impact of Time Delay

Tab. 2 shows the attack performance of LaserAdv when there is no delay and the delay is 0.5 seconds. In this section, to further explore the impact of time delay of the perturbation, we lagged the perturbation relative to the user's speech by 0.2 seconds, 0.4 seconds, and 0.6 seconds respectively, and observe the results.

As shown in Fig. 7 (a), as the delay increases, the success rate of the attack on all three ASRs decreases. Notably, DeepSpeech exhibits the lowest level of robustness. Even with a delay of 0.6 seconds, its attack success rate remains above 90%. Meanwhile, iFlytek and Whisper experience significant drops but still achieve success rates of over 50%. This result highlights the impressive capability of LaserAdv attack to attain asynchrony of the adversarial perturbation.

5.5 Impact of Varying Laser Power Levels

In this section, we aim to explore the power requirements essential for effectively conducting the LaserAdv attack. we employ a laser power meter [44] to measure the maximum power at 6 mW. During the experiments, we adjusted the voltage of the laser driver board to incrementally increase the laser power from zero to the maximum, and then evaluate the performance of the attack.

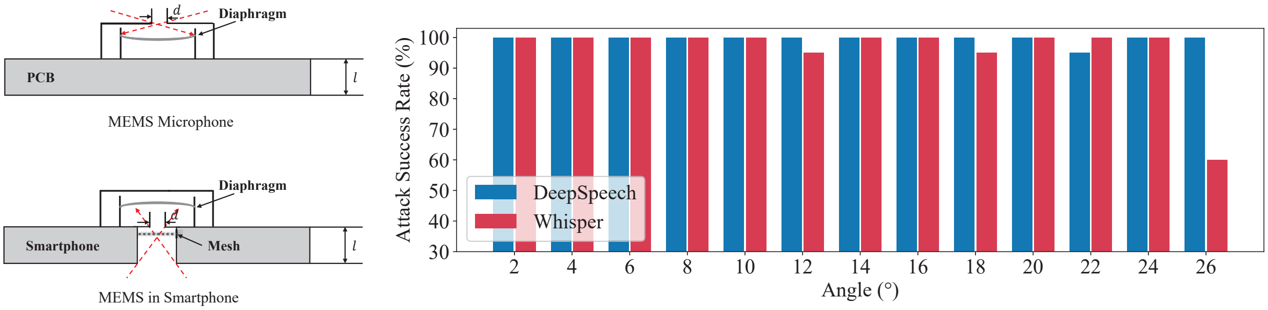
Fig. 7 (b) illustrates the success rate of the attack across varying driving laser power levels. When the power is 2.5 mW, the attack becomes effective and the success rate is only 10%. As the power gradually increases, so does the success rate. Notably, upon reaching the rated power of the laser diode at 5mW, a 100% success rate can be achieved. Consequently, in our experimental setups, we typically maintain the laser diode power within the range of 5 mW to 6 mW to ensure optimal performance for LaserAdv attack.

5.6 Impact of Distance

In this section, we aim to explore the attack performance of LaserAdv as the attack distance increases. We evaluated the performance of close-range attacks using Honor 20 Pro smartphone as a receiver. We then used a MEMS microphone to test the maximum attack distance. This two-step approach allows us to evaluate the performance of the attack under different conditions. As the sound inlet of MEMS chip is larger than that of the microphone in the smartphone, and there is no grille, the attack distance is greater for the MEMS microphone.

Fig. 7 (c) shows the performance of LaserAdv on the mobile phone as the distance varies. At a distance of 1 m, the attack achieves a success rate of 80%. As the distance increases, the success rate decreases. At a distance of 5 m, the success rate drops to only 11%. This phenomenon is attributed to the expansion of the laser diode's light spot as the distance increases. As a result, the laser light diverges, making it difficult to accurately target the small microphone hole with the higher-intensity laser.

We conducted a long-range attack in a wide corridor where the laser diode was significantly far from the MEMS microphone, as illustrated in Fig. 7 (d). The total length of the corridor is 120 m. The audible voice commands were continuously played at a volume of 55 - 65 dB. We conduct experiments at intervals of every 20 m, from 20 m to 120 m, and control the intensity of perturbations and voice commands equally, to compare the attack performance of LaserAdv and LightCommands. We performed 20 attack attempts at each distance and calculated the average attack success rate. Tab. 4 reveals a clear trend: as the distance increases, the attack success rate decreases. LightCommands achieve a 50% success rate at 20 m, dropping to just 5% at 80 m, and becoming ineffective beyond that range. In contrast, LaserAdv boasts a remarkable



(a) Area susceptible to laser illumination.

(b) The impact of angles.

Figure 8: Acoustic port of microphone in smartphone and MEMS microphone, and the impact of angles.

Table 5: Attack success rate under different environmental illumination conditions.

Illumination Conditions	Luminance	Success Rate
Curtains closed	38 lx	100%
Lights off	240 lx	100%
Lights on	460 lx	100%
Sunlight	2100 lx	100%

performance with a success rate exceeding 95% within 60 m, 80% at 80 m, and the ability to extend up to 120 m. Notably, LightCommands demonstrates an attack range of up to 110 m. However, it is dependent on a telephoto lens, and rather than modifying the victim's speech into a target command, it utilizes a laser to directly inject voice commands.

5.7 Impact of Environmental Illumination Conditions

To investigate the influence of environmental illumination conditions on LaserAdv attack, we designed four experimental conditions, i.e., curtains closed, lights off, lights on and sunlight. Using a luminous flux meter, we measured the light flux to be 38 lx, 240 lx, 460 lx, and 2,100 lx respectively. Then we conducted attacks under each condition to evaluate the performance of the attack.

The results are presented in Tab. 5, demonstrating that the LaserAdv attack attains a 100% success rate under various lighting conditions. This is attributed to the compact size of the smartphones' microphone hole, preventing ambient light from penetrating it and consequently, not impacting the efficacy of the LaserAdv attack.

5.8 Impact of Angles

In this section, we will explore the effect of angles on the success rate of LaserAdv attack in real-world scenarios. In practical situations, it may not always be possible for laser to hit the microphone diaphragm directly. Therefore, understanding the performance of the attack from different angles is crucial to fully assess its effectiveness. There are significant differences between the microphone on a smartphone and a standalone microphone chip when exposed to laser at different angles. The smartphone microphone is embedded in the phone, so the laser has to pass through the pre-set sound holes of the device before reaching the microphone port, as shown in Fig. 8 (a). In addition, a protective grill or mesh is usually placed in front of the microphone to protect it from dust. In contrast, with a stand-alone microphone chip, the laser only has to pass through its own port. As a result, its response to laser at different angles can be very different.

In our experiments, we chose the MEMS microphone to evaluate the influence of angles. We maintained a fixed position for the laser diode, and controlled the angle of the laser to deviate from the microphone hole by adjusting the knob on the turntable. From the observations in Fig. 8 (b), we noticed that the angle was adjusted up to a maximum of 26 degrees. Surprisingly, the success rate of the attack did not decrease significantly as the angle increased on DeepSpeech. However, the success rate dropped to 60% on Whisper, that is because as the angle increases, the signal strength received diminishes, leading to a reduction in the signal-to-noise ratio (SNR). This drop in SNR is likely responsible for the decreased accuracy of attacks on Whisper. For DeepSpeech, the perturbation was specifically tailored for this system, which explains why the attack accuracy remains comparatively high despite the angle changes. Overall, these results demonstrate the remarkable effectiveness of the LaserAdv attack in realistic attack scenarios where the alignment between the laser beam and the microphone hole may not be perfect.

5.9 Impact of Different Smart Devices

To validate the efficacy of the LaserAdv attack on different smart devices, we employed 6 smartphones for experiments, including Huawei Enjoy 20 Pro and Mate 60 Pro, Honor 20 Pro, Samsung Galaxy S9, Redmi K30 Ultra and Oppo Reno 9. The attacks employ a universal perturbation with the target

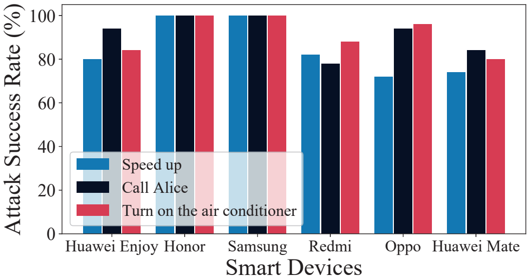
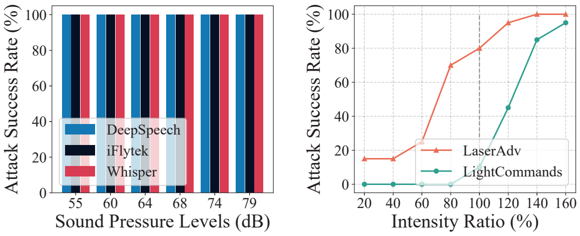


Figure 9: The impact of different smart devices.



(a) The impact of SPLs of the voice commands.

(b) The impact of different loudness of perturbation or Malicious Commands.

Figure 10: The impact of different SPLs of the voice commands and loudness of perturbation.

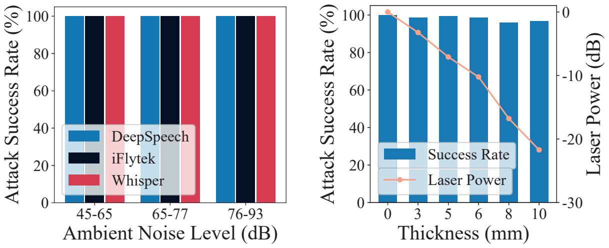
phrase being “Hi”, and three voice commands are randomly chosen for each device: “Speed up”, “Call Alice” and “Turn on the air conditioner”. Subsequently, we performed 50 attack attempts for each voice command to comprehensively assess its success rate.

Fig. 9 illustrates the attack performance on different smart devices. It is evident that the attack on Honor 20 Pro and Samsung Galaxy S9 yields the most favourable results, with a remarkable 100% attack success rate. On the other hand, the attack performance on other 4 smartphones is comparatively lower, though still significant, with a success rate exceeding 72%. The observed variations in attack performance can be attributed to the use of different microphones in each device, naturally leading to distinct results. Nevertheless, these results collectively validate the effectiveness of the `LaserAdv` attack.

5.10 Impact of SPLs of the Voice Commands

The variation in sound pressure levels (SPLs) of voice commands is an important factor in the effectiveness of laser perturbations. To investigate the effect of sound intensity on the performance of `LaserAdv` attack, we conducted an experiment in a typical meeting room with ambient noise of about 40 dB. We played voice commands at a range of 55 dB to 79 dB during the attack on the MEMS microphone.

As shown in Fig. 10 (a), it is interesting to note that despite the escalating SPLs, the success rate of the `LaserAdv` attack remains consistently 100% on all three ASR models tested.



(a) The impact of different ambient noise.

(b) The impact of different thicknesses.

Figure 11: The impact of different scenarios including office, restaurant and street, and thickness.

This result underlines the high effectiveness and robustness of the `LaserAdv` attack in the face of fluctuating acoustic intensities. Remarkably, even in scenarios where the target is speaking at a high volume, the attack continues to demonstrate a high success rate. This effectively undermines the ASR system’s defences and leads to significant misinterpretations.

5.11 Impact of Loudness of Perturbations or Malicious Commands

We further investigate the attack performance changes due to different loudness of the perturbations in `LaserAdv` or malicious voice commands in LightCommands. We positioned a loudspeaker 50 cm away from the MEMS microphone to deliver the user speech, measuring the sound intensity within the range of 70 - 80 dB using a sound meter. Subsequently, we adjust the laser emission intensity to correspond with its intensity, varying from 20% to 160%, to observe the attack results.

As shown in Fig. 10 (b), at an intensity ratio of 20%, `LaserAdv` can attack successfully, despite with a relatively low success rate. As the intensity increases, the attack success rate gradually improves. When the laser intensity matches the user’s speech intensity, `LaserAdv` achieves a success rate of 80%, while LightCommands is just successful under this condition. With increasing intensity, the success rate can be up to 100%. Our experiments confirm that `LaserAdv` requires substantially lower perturbation intensity compared to LightCommands. Consequently, it can achieve a broader attack range, and devices with lower laser response can also be attacked.

5.12 Impact of Different Ambient Noise

To assess the effect of varying ambient noise levels on the effectiveness of the `LaserAdv` attack, we set up an experiment in our office where pre-recorded ambient noise was played at various SPLs. In all of these experiments, the loudspeaker producing the ambient noise was located 50 cm from the target device. A sound meter placed near the target device was used to measure the SPLs of the ambient noise.

The results are shown in Fig. 11 (a). As the noise level increases, the success rate of the attack remains at 100%. This indicates that the LaserAdv attack is unaffected when the noise level is below 90 dB, ensuring the effectiveness of the attack even in high noise conditions.

5.13 Attack through Transparent Glass

Real-world attack scenarios may involve long-range attacks where the attacker and the victim's smart device are not in the same room. These attacks can be carried out over obstacles such as windows and transparent glass on doors. We therefore positioned glass of different thicknesses between the laser emitter and the targeted mobile phone to conduct experiments. This setup allows us to thoroughly investigate LaserAdv's ability to penetrate the glass barrier.

Fig. 11 (b) shows the success rate of the LaserAdv attack through the glass of different thicknesses, demonstrating the remarkable ability of the laser to penetrate. First, the intensity of the laser beam decreases slightly after passing through the glass. The success rate of the attack is 100% when there is no glass, and as the thickness of the glass increases, the success rate remains above 98%. Thus, the thickness of the glass has no significant effect on the success rate of the attack. These results have significant implications for LaserAdv attacks in physical scenarios.

6 Defense and Discussion

In this section, we discuss potential defense strategies against LaserAdv and some of its limitations.

6.1 Defense

In previous studies, adversarial perturbations have been shown to be vulnerable to certain audio processing methods, such as local smoothing [16, 17], audio squeezing [4, 16, 18], compression [5, 19], and audio down-sampling [4, 17]. However, laser adversarial perturbations are not strictly amplitude-limited and exhibit robustness against distortion. Consequently, these methods might not be as effective against laser adversarial perturbations.

Another approach [17, 20–23] is to employ adversarial training, where the ASR model is trained with adversarial examples to enhance its robustness against adversarial attacks. However, it may still exhibit certain limitations when countering specific adversarial attacks or adaptive adversaries that continuously alter their attack strategies.

Inspired by EarArray [27], compared to acoustic perturbation, only one microphone can capture the laser beam. This property can be exploited to detect laser adversarial perturbations using two or more microphones. By analyzing the differences in the received signals across multiple microphones, it is possible to distinguish between genuine audible signals and laser adversarial perturbations. This approach could provide a practical defense mechanism against laser adversarial attacks, thereby enhancing the security and robustness of ASR systems against such threats.

We respectively captured the user's voice commands and injected either one or two beams of laser perturbations into the smartphone's microphone to create datasets. We first extract the audio features from normal voice commands and then train a simple machine learning model, i.e., support vector machine (SVM). Subsequently, the model is utilized to discern whether the audio received by the microphone corresponds to normal speech or adversarial perturbations.

Our final experiments revealed that the model can detect perturbations injected into a single microphone with an accuracy of 100%. However, when perturbations are injected into two microphones simultaneously, the model's detection accuracy decreases to 65.02%. The significant drop indicates the efficacy of LaserAdv in countering signal detection-based defense mechanisms.

6.2 Limitation

There are a few limitations to keep in mind. Firstly, LaserAdv requires the attacker to have access to the architecture and parameters of one ASR system to generate audio adversarial perturbations, a condition that may not be consistent with real-world scenarios. However, many white-box ASR models can be used to generate audio adversarial perturbations and transfer the attack to black-box ASR systems. As part of future research, we aim to investigate the applicability of LaserAdv under more realistic attack settings and explore the feasibility of targeting multiple systems directly.

Secondly, when the attack command is excessively lengthy or the feedback from the ASR system deviates from the command issued by the victim, it has the potential to draw his attention. Furthermore, owing to the considerable distance, the attacker can not perceive the feedback from the ASR system, and thus cannot immediately judge whether the attack can be successful. This challenge persists in long-distance attacks and remains unresolved.

Thirdly, although we can utilize the photoacoustic phenomena for LaserAdv attack, sometimes it is weak and may limit the effectiveness of the attack. Given these limitations, we plan to explore alternative channels (e.g. ultrasonic channel) and methodologies to develop a more robust and effective adversarial attack. We aim to investigate the properties of these new channels and potentially merge them with current techniques to improve the overall performance of perturbations. Such improvement may make them more resilient to varying environmental conditions and channel degradation.

Finally, LaserAdv requires an adversary to possess laser equipment. However, it is worth noting that this additional complexity does not pose a significant financial barrier. Encouragingly, the price of the equipment remains relatively low, under $20, making it affordable and accessible to potential threat actors.

7 Related Work

In this section, we review the existing studies of adversarial perturbations and sensor attacks, which can be illustrated as follows.

7.1 Adversarial Perturbation Attacks

One common approach for generating adversarial examples is to minimize the perturbation while making it challenging for humans to detect any changes. Previous works such as [16,17] focused on generating perturbations by incorporating perturbation volume into the loss function. Others like [28] utilized genetic algorithms to generate imperceptible perturbations. Another strategy to enhance imperceptibility involves shortening the perturbation and targeting weaker parts to decrease the chances of detection. For instance, Miao et al. [30] divided audio into frames and selected optimal locations for adding perturbations. Liu et al. [31] generated sparser perturbations that were less detectable while maintaining accuracy. O'Reilly et al. [32] employed weighted sampling to select perturbation points, leading to faster generation and greater robustness.

Furthermore, some studies focused on improving the imperceptibility of perturbations through psychoacoustic hiding methods [6, 8, 33–35]. These methods rely on the frequency masking effect in signal processing, where louder signals render nearby frequencies imperceptible. Additionally, some existing studies aim to make the perturbations resemble sounds commonly found in the environment, such as whistles, bird sounds, or alarm clocks. For instance, AdvPulse [5] successfully implements universal, synchronisation-free and targeted audio adversarial attacks using sub-second perturbations. However, to ensure the success of physical attacks, these perturbations must be embedded in loud, common environmental sounds that may inadvertently alert the victim due to the sudden onset of noise. In addition, the targeted ASR system is a convolutional neural network for small-footprint keyword spotting, and its feasibility has not been validated on commercial ASR models. Similarly, Shi et al. [12] developed an environmental sound simulation mehtod. Commandersong [4] quietly injected perturbations into a song to perform a physical adversarial attack.

The similarity to environmental noise can lead to easy detection when the perturbation sound abruptly appears. Furthermore, these methods require the perturbation to be approximated to a specific environmental sound, ensuring that the deviation remains below a certain threshold. This added complexity in training arises from the need to approximate the noise accurately and the necessity for more training data.

7.2 Signal Injection Attacks On Sensors

Sensors enable devices to perceive their surroundings, gather data, and respond accordingly, playing a pivotal role in shaping the modern technological landscape. However, despite their importance, recent studies have exposed vulnerabilities in many sensor systems, posing significant security challenges. S. Nashimoto et al. [36] proposed a method of injecting false information into sensors, leading to misleading or erroneous data interpretation. Many researchers have investigated attacks on radar systems [37], anti-lock braking systems (ABSs) using magnetic speed sensors [38], and even global positioning systems (GPS) [39]. Yan et al. [40] have delved deeper into the sensors used for driving guidance, such as millimeter-wave radars, ultrasonic sensors, and forward-looking cameras. Their work revealed vulnerabilities in these systems, leading to the development of contactless attacks capable of inducing blindness in autonomous vehicles. Similarly, Jang et al. [41] proposed an innovative anti-drone technique employing electromagnetic interference signal injection, effectively disrupting the communication channel between an Inertial Measurement Unit (IMU) and its control unit.

Researchers also have unveiled a range of signal injection attacks on microphones, exploiting various physical phenomena. Sugawara et al. [10] introduced a novel attack that converts light to sound to inject malicious audio signals into microphones. Zhang et al. [42] demonstrated successful modulations of low-frequency audio commands that attack VCSs by exploiting the nonlinearity of microphone circuits.

By scrutinizing various microphone studies, it becomes evident that these crucial audio sensors can be compromised through diverse attack vectors, including inaudible voice [42], light [10], and electromagnetic (EM) -based attacks [43]. Building upon this understanding, our research seeks to explore the potential vulnerabilities of microphones and ASR systems using lasers.

8 Conclusion

This paper introduces LaserAdv, a new method for launching adversarial attacks on ASR systems via laser adversarial perturbations. In our study, we identified critical factors that affect the effectiveness of the LaserAdv method, including the response characteristics of the microphone laser channel and the effects of frequency-selective fading and low sensitivity. To address these issues, we proposed a SAE-TFI method and further optimized the IAP generation objective function to facilitate more practical attack scenarios. Our results demonstrate the potential of LaserAdv in successfully attacking three ASR systems using IAP, including DeepSpeech, Whisper, and iFlytek. In the presence of victim speech, the maximum distance can be up to 120 m at a cost of only $20.

Acknowledgments

We appreciate the shepherd's efforts and all the anonymous reviewers' insightful and constructive suggestions. The work is supported by National Natural Science Foundation of China (Grant No. 62202274).

References

[1] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," in *2021 IEEE symposium on security and privacy (SP)*. IEEE, 2021, pp. 730–747.

[2] Y. Chen, J. Zhang, X. Yuan, S. Zhang, K. Chen, X. Wang, and S. Guo, "Sok: A modularized approach to study the security of automatic speech recognition systems," *ACM Transactions on Privacy and Security*, vol. 25, no. 3, pp. 1–31, 2022.

[3] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[4] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "{CommanderSong}: A systematic approach for practical adversarial voice recognition," in *27th USENIX security symposium (USENIX security 18)*, 2018, pp. 49–64.

[5] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1121–1134.

[6] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *Network and Distributed System Security Symposium (NDSS)*, 2019.

[7] J. B. Li, S. Qu, X. Li, Z. Kolter, and F. Metze, "Real world audio adversary against wake-word detection systems," in *Proc. of NIPS*.

[8] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[9] J. Deng, Y. Chen, and W. Xu, "Fencesitter: Black-box, content-agnostic, and synchronization-free enrollment-phase attacks on speaker recognition systems," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 755–767.

[10] T. Sugawara, B. Cyr, S. Rampazzi, D. Genkin, and K. Fu, "Light commands:{Laser-Based} audio injection attacks on {Voice-Controllable} systems," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2631–2648.

[11] R. Duan, Z. Qu, S. Zhao, L. Ding, Y. Liu, and Z. Lu, "Perception-aware attack: Creating adversarial music via reverse-engineering human perception," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 905–919.

[12] C. Shi, T. Zhang, Z. Li, H. Phan, T. Zhao, Y. Wang, J. Liu, B. Yuan, and Y. Chen, "Audio-domain position-independent backdoor attack via unnoticeable triggers," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 583–595.

[13] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–5.

[14] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *arXiv preprint arXiv:1810.11793*, 2018.

[15] Google Text-to-speech. https://cloud.google.com/text-to-speech, 2023.

[16] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang, "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices." in *USENIX Security Symposium*, 2020, pp. 2667–2684.

[17] B. Zheng, P. Jiang, Q. Wang, Q. Li, C. Shen, C. Wang, Y. Ge, Q. Teng, and S. Zhang, "Black-box adversarial attacks on commercial speech platforms with minimal information," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 86–107.

[18] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 694–711.

[19] N. Das, M. Shanbhogue, S.-T. Chen, L. Chen, M. E. Kounavis, and D. H. Chau, "Adagio: Interactive experimentation with adversarial attack and defense for audio," in *Machine Learning and Knowledge Discovery in*

Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18. Springer, 2019, pp. 677–681.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[21] Y. Balaji, T. Goldstein, and J. Hoffman, "Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets," *arXiv preprint arXiv:1910.08051*, 2019.

[22] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[24] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[25] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu, "On the convergence and robustness of adversarial training," *arXiv preprint arXiv:2112.08304*, 2021.

[27] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "Eararray: Defending against dolphinattack via acoustic attenuation." in *NDSS*, 2021.

[28] J. K. Han, H. Kim, and S. S. Woo, "Nickel to lego: using foolgle to create adversarial examples to fool google cloud speech-to-text api," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 2593–2595.

[29] P. Cheng, Y. Wu, Y. Hong, Z. Ba, F. Lin, L. Lu, and K. Ren, "Uniap: Protecting speech privacy with non-targeted universal adversarial perturbations," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[30] Y. Miao, C. Chen, L. Pan, J. Zhang, and Y. Xiang, "Faag: Fast adversarial audio generation through interactive attack optimisation," *arXiv preprint arXiv:2202.05416*, 2022.

[31] X. Liu, X. Chen, M. Yin, Y. Wang, T. Hu, and K. Ding, "Audio injection adversarial example attack," in *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[32] P. O'Reilly, A. Bugler, K. Bhandari, M. Morrison, and B. Pardo, "Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models," in *Advances in Neural Information Processing Systems*, 2022.

[33] K.-H. Chang, P.-H. Huang, H. Yu, Y. Jin, and T.-C. Wang, "Audio adversarial examples generation with recurrent neural networks," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2020, pp. 488–493.

[34] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," *arXiv preprint arXiv:1904.05734*, 2019.

[35] T. Chen, L. Shangguan, Z. Li, and K. Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[36] S. Nashimoto, D. Suzuki, T. Sugawara, and K. Sakiyama, "Sensor con-fusion: Defeating kalman filter in signal injection attack," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 511–524.

[37] A. Ranganathan, B. Danev, A. Francillon, and S. Capkun, "Physical-layer attacks on chirp-based ranging systems," in *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks*, 2012, pp. 15–26.

[38] Y. Shoukry, P. Martin, P. Tabuada, and M. Srivastava, "Non-invasive spoofing attacks for anti-lock braking systems," in *Cryptographic Hardware and Embedded Systems-CHES 2013: 15th International Workshop, Santa Barbara, CA, USA, August 20-23, 2013. Proceedings 15*. Springer, 2013, pp. 55–72.

[39] A. J. Kerns, D. P. Shepard, J. A. Bhatti, and T. E. Humphreys, "Unmanned aircraft capture and control via gps spoofing," *Journal of field robotics*, vol. 31, no. 4, pp. 617–636, 2014.

[40] C. Yan, W. Xu, and J. Liu, "Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle," *Def Con*, vol. 24, no. 8, p. 109, 2016.

[41] J.-H. Jang, M. Cho, J. Kim, D. Kim, and Y. Kim, "Paralyzing drones via emi signal injection on sensory communication channels." in *NDSS*, 2023.

[42] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 103–117.

[43] D. F. Kune, J. Backes, S. S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu, "Ghost talk: Mitigating emi signal injection attacks against analog sensors," in *2013 IEEE Symposium on Security and Privacy*. IEEE, 2013, pp. 145–159.

[44] UNI-T, "UNI-T UT385," meters.uni-trend.com.cn/list_78/1685.html, 2024.