SUAD: Solid-Channel Ultrasound Injection Attack and Defense to Voice Assistants

Chao Liu[1], Zhezheng Zhu[1], Hao Chen[1], Zhe Chen[3], Kaiwen Guo[1], Penghao Wang[1], Jun Luo[2]
[1]Department of Computer Science and Technology, Ocean University of China, P.R. China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Fudan University & AIWiSe Company, China
Email: liuchao@ouc.edu.cn, zzz905060439@gmail.com, chenhao1339@stu.ouc.edu.cn,
nick.chitchan@gmail.com, kevingguo@mail.ustc.edu.cn, wangpenghao@stu.ouc.edu.cn, junluo@ntu.edu.sg

***Abstract*—As a versatile AI application, voice assistants (VAs) have become increasingly popular, but are vulnerable to security threats. Attackers have proposed various inaudible attacks, but are limited by cost, distance, or LoS. Therefore, we propose SUAD Attack, a long-range, cross-barrier, and interference-free inaudible voice attack via solid channels. We begin by thoroughly analyzing the dispersion effect in solid channels, revealing its unique impact on signal propagation. To avoid distortions in voice commands, we design a modular command generation model that parameterizes attack distance, victim audio, and medium dispersion features to adapt to variations in the solid-channel state. Additionally, we propose SUAD Defense, a universal defense that uses ultrasonic perturbation signals to block inaudible voice attacks (IVAs) without impacting normal speech. Since the attack can occur at arbitrary frequencies and times, we propose a training method that randomizes both time and frequency to generate perturbation signals that break ultrasonic commands. Notably, the perturbation signal is modulated to an inaudible frequency without affecting the functionality of voice commands for VAs. Experiments on six smartphones have shown that SUAD Attack achieves activation success rates above 89.8% and SUAD Defense blocks IVAs with success rates exceeding 98%.**

***Index Terms*—Microphone Nonlinearity, Acoustic Dispersion, Inaudible Voice Attack, Universal Adversarial Perturbation**

I. Introduction

Rapid advances in Artificial Intelligence (AI) have contributed to the widespread application of Voice Assistants (VAs) [1], [2]. With system privileges granted by users, VAs, such as Siri and Bixby, boast an extensive array of functionalities, including making calls, controlling device settings, and retrieving information [3]. However, the gradual expansion of functions brings not only convenience but also security risks. They are vulnerable to attacks involving forged voice commands, which can trigger high-risk operations such as unauthorized retrieval and viewing of private information or unverified payments and transfers [4], posing significant risks to users' information and property security. Therefore, conducting a comprehensive security analysis of voice command attacks on VAs is critically important.

In recent years, attackers have proposed various methods [5]–[17] to enable inaudible voice command injection. For example, Dai et al. [8] generate magnetic-inductive voice signals by exploiting vulnerabilities in wireless charger systems or hardware, but the magnetic field restricts the range in 5 cm. Additionally, Light Commands [12] leverage laser beams pre-encoded with voice commands, which shine on microphones to produce spoofed electrical signals (i.e., audio). Despite its long attack range (110 m), high-cost devices and light-of-sight (LoS) greatly limit its application.

Acknowledging these limitations, attackers have explored the feasibility of *inaudible voice attacks* (IVAs) [13]–[17]. DolphinAttack [13] and BackDoor [14] exploit microphone nonlinearity to convert voice commands modulated at inaudible frequencies (>20 kHz) into audible signals, enabling attacks at ranges of 1.75 m and 3.5 m, respectively. For longer range attacks (7.62 m), LipRead [15] utilizes different speakers in the array to play segments of the voice command spectrum, but sonic focusing requires LoS. Moreover, they share a common limitation: barriers can obstruct air-channel signal transmission. In fact, sound waves can propagate through any medium (e.g., solid) that supports vibrations. Recently, SurfingAttack [16], generates ultrasound on a solid surface, which propagates through the solid medium, to induce a nonlinear response in the microphone, injecting inaudible commands. Despite offering a clear analysis of solid-channel signal propagation, it neglects *voice command distortion resulting from frequency-dependent dispersion*.

In this work, we aim to break this old notion by proposing a novel inaudible voice attack against solid-channel interference. As illustrated in Fig. 1, a piezoelectric transmitter covertly affixed by the attacker under the table induces vibrations to emit ultrasonic attack signals encoded with malicious commands (e.g., 'upload album to Facebook'). Then, these signals are

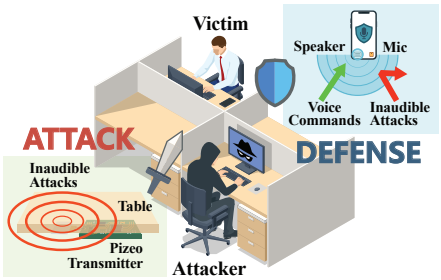


Fig. 1. SUAD: (1) Attack: A piezo-transmitter covertly placed beneath a table emits inaudible signals to attack VAs. (2) Defense: Inaudible perturbation signals block such attacks without interfering with normal voice commands.

transmitted to the victim's device through the table, effectively bypassing LoS barriers. However, to effectively inject voice commands, the attack must overcome challenges inherent in solid-channel transmission. Specifically, due to the dispersion effect [18], signals of different frequencies propagate at varying speeds in solid media, causing the original waveform features to change. On one hand, propagation over a distance introduces time delays that vary across different frequency components. On the other hand, signal waveforms gradually deform or broaden during propagation, resulting in significant distortion of the attack command.

To this end, we propose a novel SUAD Attack, which overcomes the aforementioned technical bottleneck and enables long-distance, cross-barrier, and interference-free inaudible attacks on VAs. We first employ a localization method to locate the victim device position by analyzing the time difference of arrival (TDoA) of solid-channel sound captured by a microphone array. Next, the estimated attack distance is used to compensate for propagation delays. Specifically, we use a speech cloning model with a fused multi-head architecture to generate attack commands capable of bypassing voiceprint authentication on VAs. These commands embed the victim's voiceprint features and inverse solid-channel interference, derived from the captured victim speech and the solid medium's physical characteristics, including propagation distance and material density. Ultimately, attack commands are modulated into the ultrasonic frequency band, generating inaudible signals that threaten VAs via solid-channel propagation.

With increasing awareness of IVA threats, various defenses [13], [15], [19]–[28] have been developed against such attacks. For instance, DolphinAttack [13] employs support vector machines (SVM) to classify and detect malicious audio, while MicGuard [19] identifies anomalies in the spectral domain. Although these approaches can alert users, attacks may have been executed successfully before detection. Furthermore, systems like DualGuard [20] and Cacher [21] may disable the microphone upon detecting an attack, potentially disrupting the normal operation of VAs. This poses a critical problem: *How can we ensure the normal functionality of VAs while simultaneously enabling real-time defense against IVAs*?

As such, we propose SUAD Defense, which continuously emits ultrasonic perturbation signals via the smartphone's speaker to actively defend against randomly launched, arbitrary-frequency IVAs. The short mic–speaker distance also allows its low-power operation. The perturbation is generated through a Universal Adversarial Perturbation (UAP) training method that randomizes both the temporal shift and frequency of attack signals during model training. As a result, the generated signal can effectively suppress attacks while allowing legitimate user voice commands to function correctly.

Finally, we conducted extensive experiments to evaluate the effectiveness of SUAD Attack and Defense. SUAD Attack achieved median activation success rates of over 89.8% across six smartphones. SUAD Defense has defense success rates exceeding 98% against three types of attacks. In summary, the main contributions of this paper are as follows:

- We propose a novel SUAD Attack, the first inaudible attack capable of adapting to solid-channel states. It overturns the old notion that solid-channel propagation is distortion-free, extending the attack range.
- A speech generation model with a fused multi-head architecture is proposed to embed the victim's voiceprint features and inverse solid-channel interference into voice commands, enabling successful activation on VAs.
- We propose SUAD Defense by designing a novel UAP training framework to generate universal perturbation signals. It defends against attacks launched randomly at any frequency, without affecting VAs.
- Finally, extensive evaluations of SUAD Attack and Defense demonstrate attack and defense success rates exceeding 89.8% and 98%, respectively.

II. Background and Motivation

In this section, we first present the threat model of VAs attacks. Next, we discuss the advantages and challenges of solid-channel attacks. Finally, we elaborate on the technical challenges of implementing defenses under nonlinear effects.

A. Threat Model

We consider a common scenario where a victim, Bob, has an intelligent voice assistant installed on his smart devices. Due to frequent use, Bob often leaves the device active on the table, freeing his hands. This provides an opportunity for the attacker, Eve, to compromise the system and threaten information security. However, vigilant Bob may avoid suspicious objects within the LoS. In other words, Bob prefers familiar and private environments, such as a covered table, which obstructs voice attacks via air channels. It compels Eve to exploit other channels for launching attacks.

In the aforementioned scenario, it is assumed that Eve can covertly place a small, autonomously operating attack device within a hidden physical space (e.g., under-table space), as illustrated in Fig. 1, without alerting Bob. Thus, Eve can't remain on-site or nearby, avoiding any suspicious interaction. Moreover, Eve can generate arbitrary voice commands with voiceprint features by leveraging known voice samples of the victim. Fake commands in VAs attacks may include sensitive operations, such as 'turn on camera,' exploiting legitimate privileges to bypass protections and heighten security risks.

B. Why Solid Instead of Air Channels?

Traditional attacks typically rely on ultrasonic propagation through the air. However, as illustrated in Fig. 2, real-world scenarios may involve barriers obstructing the LoS path, making the solid channel (e.g., table) a viable alternative. Therefore, SUAD Attack utilizes a piezoelectric element tightly

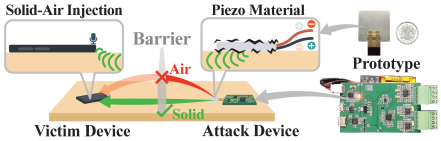
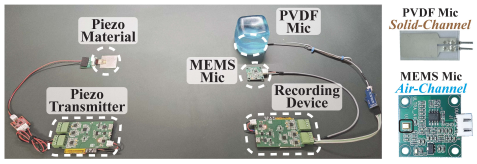
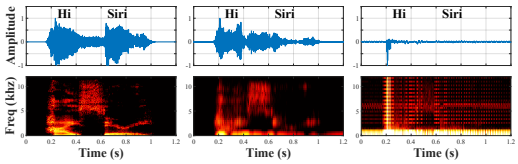


Fig. 2. Comparison of air-channel and solid-channel voice injection.



(a) Experimental setup for capturing audio from different channels



(b) Raw audio (c) Air-channel (d) Solid-channel

Fig. 3. Comparison of audio through different-channel propagation.

coupled to the table underside as the attack transmitter, with its compact design enabling covert deployment. Due to the piezoelectric effect, applying an alternating voltage across the piezoelectric material makes slight deformations in its internal lattice structure, generating vibrations (i.e., sound). As such, the piezoelectric element can emit ultrasound carrying the attack command, which propagates through solid channels and then into the air, reaching the target device's microphone to enable voice command injection, as shown in Fig. 2.

Compared to air channels, solid channels exhibit a distinct physical phenomenon known as the dispersion effect. This effect causes sound waves of different frequencies to propagate at varying speeds within solids, leading to gradual waveform deformation or broadening during transmission. To investigate the impact of dispersion on signal propagation, we use the piezoelectric material in the transmitter shown in Fig. 3(a) as the sound source to vibrate the table and emit the audio command 'Hi Siri'. The signal propagates through the table and is recorded by a MEMS microphone and a PVDF microphone, capturing the signals transmitted via the air and solid channels, respectively. As shown in Fig. 3(d), audio transmitted through the solid channel undergoes significant dispersion in both time and frequency domains. Even after re-transmission from the solid channel to air, the waveform remains notably altered, and the spectral components exhibit varying degrees of distortion, as shown in Fig. 3(c). The effect on wave speed can be approximated by the Lamb wave dispersion equation, i.e.,

$$v(f) = \sqrt[4]{\frac{Ehf^2}{12\rho\left(1-v_p^2\right)}}, \tag{1}$$

where E, ρ, h, and f denote the Young's modulus, material density, plate thickness, and frequency, respectively, and v_p represents the Poisson's ratio. However, complex and various parameters make modeling $v(f)$ across all realistic scenarios impractical [16].

Notably, successful voice injection attacks via solid channels must overcome two key challenges: (1) variations in medium materials and (2) time delays caused by propagation distances. Therefore, SUAD Attack employs a two-stage synergetic scheme to embed inverse solid-channel propagation features into generated voice injection commands, mitigating signal distortion of voice commands. The specific design and implementation details will be developed in Section III-B.

C. Why Only for Attack, Not for Defense?

Current IVAs primarily exploit the nonlinear effects of microphones, which arise from imperfections in microphone circuitry design. Specifically, when an input signal $S_{\text{in}}(t) = s_1(t) + s_2(t)$, where $s_1(t)$ is an amplitude-modulated voice carrier $v(t)\cos(2\pi f_c t)$ and $s_2(t)$ is a pure carrier $\cos(2\pi f_c t)$, is received by the microphone, the front-end circuitry performs nonlinear mixing on it, i.e.,

$$S_{out}(t) = AS_{in}(t) + BS_{in}^2(t) + \cdots, \tag{2}$$

where $v(t)$ denotes the forged voice command signal, f_c is the carrier frequency for ultrasonic modulation, and A and B are the gains of the linear and quadratic components, respectively.

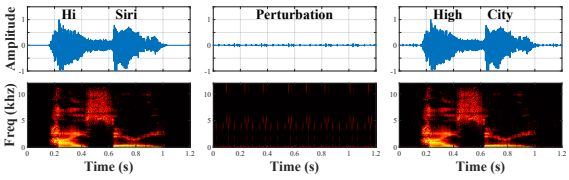
Additionally, due to the microphone's built-in low-pass filter, only the low-frequency components (i.e., $v(t)$) are retained, resulting in the following output signal:

$$S_{out}(t) = \frac{B}{2}(v^2(t) + 2v(t) + 1) \approx v(t). \tag{3}$$

However, traditional defenses detect such attacks passively by identifying spectral features [19], i.e., frequency traces (e.g., straight lines) that persist across the entire temporal axis at a specific frequency. Although these methods can issue warnings, their countermeasures, such as disabling the voice assistant or emitting audible interfering signals, inevitably compromise VAs' normal functionality.

Fortunately, methods such as the Fast Gradient Sign Method (FGSM) [29] and Projected Gradient Descent (PGD) [30] have introduced Universal Adversarial Perturbation (UAP), which can generate universal perturbation signals to mislead recognition when added to normal voice, as shown in Fig. 4. However, perturbations generated by these methods not only interfere with normal speech but are also perceptible to humans. To address this, we aim to employ a universal perturbation signal $\delta(t)$, modulated onto a high-frequency carrier, to selectively block ultrasonic attack commands for active defense.

Remarkably, successful active defense must address two key challenges: (1) universal perturbation signals are unsynchronized with randomly transmitted attack signals, and (2) attack signals may be modulated at arbitrary carrier frequencies. The universal perturbation generation is detailed in Section III-C.



(a) Raw audio (b) Perturbation audio (c) Perturbed audio

Fig. 4. Recognition results with/without perturbations.

Fig. 5. System architecture of SUAD Attack and Defense.

III. System Design

In this section, we first provide a brief overview of the SUAD. Then, we describe in detail the implementation of SUAD Attack and SUAD Defense.

A. System Overview of SUAD

Considering real-world scenarios described in Section II-A, we propose SUAD, as illustrated in Fig. 5, which explores IVAs on VAs via solid channels, along with a universal defense against such attacks.

For SUAD Attack, it allows Eve to attack VAs on Bob's smartphone by transmitting inaudible malicious commands through a solid medium. Specifically, to counteract the dispersion effect in solid-channel transmission, Eve first uses the MIC array to locate Bob's smartphone and estimate the attack distance, i.e., a key contributor to signal dispersion. Subsequently, parameters such as the attack distance, Bob's voice, and the recording transmitted through the table are simultaneously fed into a perceptual zero-sample speech cloning model with fused multi-head architecture, generating speech commands embedded with Bob's voiceprint features and inverse solid-channel interference. Ultimately, the generated commands are modulated onto a high-frequency carrier and transmitted through a solid medium to the victim's device via a piezoelectric transmitter.

For SUAD Defense, an inaudible perturbation signal is continuously emitted through the speaker of Bob's smartphone, defending against IVAs without affecting the normal operation of VAs. We first generate a spectral- and delay-adaptive universal perturbation signal against attacks using a gradient descent algorithm. This signal is then modulated to a fixed ultrasonic frequency supported by the smartphone, producing an ultrasonic perturbation signal. As a result, it obstructs attack signals while preserving the effectiveness of Bob's voice captured by the microphone.

B. Long-range SUAD Attack Adapted to Solid-Channel States

To successfully implement IVAs over solid channels, the primary challenge lies in addressing the dispersion effect on ultrasonic attack signals. Specifically, the dispersion phenomenon is manifested as different frequencies of acoustic waves propagating at different speeds in the solid medium, resulting in different frequency components having relative time delays, thus causing the signal waveform broadening and distortion. Therefore, the impact of attack distance and solid material, especially in long-range scenarios, cannot be ignored.

1) Attack Distance Estimation: To weaken the effect of dispersion, SUAD Attack estimates the attack distance using TDOAs of microphones, which is then used to apply frequency-domain compensation to signals. We first detect the time t_0 when Bob places the smartphone by applying a threshold γ to the short-duration energy envelope from microphone M_0. Based on t_0, a 0.7 ms segment is extracted as the solid-channel signal S_0, as shown in Fig. 6(a). Next, TDOA τ_i is obtained by locating the maximum of the cross-correlation between S_0 and the signal from microphone M_i, where the distance difference is given by $\Delta d_i = \tau_i \cdot c_s$, with $i = 1, \ldots, 5$, and c_s denoting the speed of sound in solids. Assuming the impact point is $P = (x, y)$ and the position of M_i is $M_i = (x_i, y_i)$, with M_0 as the reference, the distance difference between M_i and M_0 is given by:

$$\|P - M_i\| - \|P - M_0\| = \Delta d_i. \tag{4}$$

Then, as shown in Fig. 6(b) the intersection of five hyperbolic curves derived from the six microphone positions yields the point P, from which the attack distance L is calculated. Finally, we obtain propagation delays of different frequency components $\tau(f) = L/v(f)$, where $v(f)$ is the velocity in solids for frequency f.

Theoretically, the relative delay in the ultrasonic signal $s(t)$ propagating through a solid can be compensated, where $s(t)$ is given by:

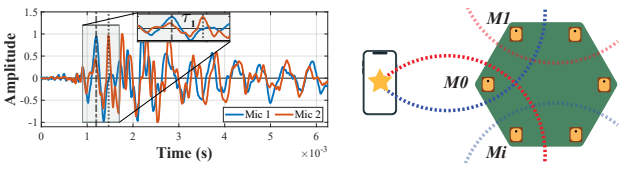
$$s(t) = \int_{-\infty}^{\infty} S(f)\, e^{j2\pi f t}\, df, \tag{5}$$

where $S(f)$ is the frequency domain signal. Subsequently, we can obtain a compensated signal $s_c(t)$:

$$s_c(t) = \int_{-\infty}^{\infty} S(f)\, e^{j2\pi f (t-\tau(f))}\, df. \tag{6}$$

By eliminating the frequency-dependent time delays caused by dispersion, the ultrasonic attack signal can release the original commands accurately at the target location.

2) Ultrasonic Attack Signal Generation: However, even with an accurate distance, various complex factors (e.g., material properties) make it difficult to obtain $v(f)$, preventing direct compensation for dispersion. Furthermore, VAs typically require voiceprint authentication, meaning that attack commands must incorporate the user's voiceprint features. Therefore, SUAD Attack designs a multi-parameter modular attack command generation model, as shown in Fig. 7. It consists of three modules: i) *Speech Encoder*, which extracts voiceprint features from Mel spectrograms of the user's recordings, ii) *Synthesizer*, which converts textual commands



(a) Solid-channel signals (b) Localization via hyperbolas

Fig. 6. Attack distance estimation via 6 microphones' TDoA.

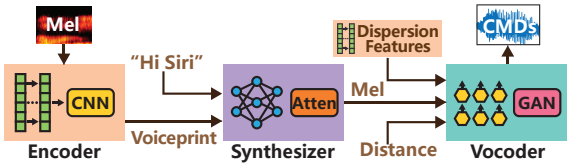
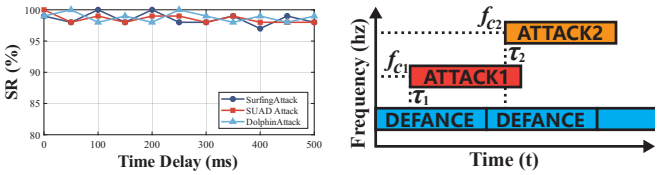


Fig. 7. Model structure for attack command generation.



(a) Spectrum after 2 random attacks (b) Random attack time & frequency

Fig. 8. Attacks with randomized time and carrier frequency.

into Mel spectrograms embedded with voiceprint features, and iii) *Vocoder*, which fuses distance and material features into the Mel spectrograms to reconstruct the time-domain waveforms of the attack speech commands.

The first module is implemented on ReDimNet [31], which begins with a 2D convolutional layer to extract local time-frequency features from 2–3 s Mel spectrograms of user recordings. Then, the extracted non-linguistic features, such as pitch and timbre, are fed into a Transformer to learn prosodic and articulatory patterns. Finally, the model's noise robustness is further enhanced by the self-attention mechanism [32], which suppresses transient distortions caused by environmental disturbances and reverberation. The module can be pre-trained on the TIMIT speech dataset [33], and no parameter updates are required after training. It extracts the target user's vocal features as 256-dimensional embedding vectors for input into the synthesizer in the next stage. The synthesizer is a Tacotron-enhanced sequence-to-sequence model based on the Encoder-Attention-Decoder architecture [34], embedding the voiceprint vectors into the Mel spectrogram generated by the textual commands. The module is also a zero-shot model that requires no updates after a single training. Given command text and voiceprint features, it performs one-shot inference to generate the corresponding Mel spectrogram, ensuring high efficiency and real-time attack capability.

Finally, the vocoder serves as the core of SUAD Attack for eliminating dispersion effects. It adopts a HiFi-GAN [35] structure that can integrate features such as solid material and distance. This enables the conversion of the Mel spectrogram into a time-domain ultrasonic attack waveform embedded with inverse solid-channel interference to counteract dispersion. Specifically, we pre-collected an offline dataset of solid-channel propagation across various materials to train a dispersion feature extractor similar to an encoder. Using a multilayer structure, it captures dispersion features such as frequency distortion and delay patterns from the frequency domain. Then, we fuse HiFi-GAN with the dispersion feature vector and propagation distance (from Section III-B1) by inserting two convolutional layers before the vocoder's Softmax layer. This enables the vocoder to adjust the output waveform based on material and distance.

During training, the offline solid-channel dataset includes the same table materials as those used in scenarios. After training, the pre-trained weights remain fixed. The attack waveform $A(t)$ can be generated by modifying only the material features in the fusion layer. An ultrasonic carrier with frequency $f_c = 21\,\text{kHz}$ is then employed to modulate the waveform, producing an inaudible attack signal. This signal is transmitted through the table using a piezoelectric device to complete the attack on the VAs.

C. Universal SUAD Defense without Affecting VAs

Successful IVAs ultimately rely on downconversion to inject attack commands into microphones. Such attacks can be effectively countered by jamming the injection commands. In particular, the defender can get various information from the speech recognition model. Therefore, unlike existing passive defenses [15], [19], [28], such as detecting anomalies in the speech spectrum, SUAD Defense introduces a UAP-based approach leveraging the smartphone's speaker to emit inaudible perturbation signals for targeted protection against IVAs.

1) Analysis of Ultrasonic Perturbations to IVAs: In practical scenarios, IVAs can be modulated to any inaudible frequency and send attack signals at any time, as shown in Fig. 8. Consequently, the actual signal $s_1'(t)$ received by the microphone can be modeled as a time-shifted version of the ideal signal $s_1(t)$ due to a delay τ, and is expressed as:

$$s_1'(t) = s_1(t+\tau) = v(t+\tau)cos(2\pi f_c(t+\tau)), \tag{7}$$

where the content of the attack command $v(t)$ and the carrier frequency f_c may also change randomly, making IVAs difficult to defend against. Although the second carrier $s_2(t)$ of IVAs leaves a distinct and fixed feature (i.e., a straight line) in the spectrum, which can serve as an important basis for detection, this only helps potential victims to passively recognize such attacks. In most cases, victims become aware of the attack only after it has already been executed. Even if the attack is detected in time and defensive measures (e.g., turning off the microphone [28]) are taken, these actions can interfere with the normal use of VAs.

Therefore, we intend to generate a small-amplitude universal perturbation signal $\delta(t)$ in response to activation commands and modulate it into the inaudible band using a smartphone-supported frequency $f_x = 18\,\text{kHz}$. This ultrasonic perturbation signal is played cyclically. When an ultrasonic attack occurs, the signal captured by the microphone can be represented as:

$$\begin{aligned} &x(t+\tau)cos(2\pi f_c(t+\tau)) + \delta(t)cos(2\pi f_x t) + cos(2\pi f_c(t+\tau)) \\ &\xrightarrow{\text{Downconversion}} x(t+\tau) + \delta^{'}(t), \end{aligned} \tag{8}$$

where $\delta^{'}(t)$ is the universal perturbation to the forged command $x(t+\tau)$. Specifically, the ultrasonic perturbation can leverage the second carrier of IVAs to down-convert it into a universal perturbation, thereby corrupting the integrity of the attack command to prevent it from activating VAs.

If only audible voice commands and ultrasonic perturbations are present, the received signal can be expressed as follows:

$$v(t+\tau) + \delta(t)cos(2\pi f_x t)) \Rightarrow v(t). \tag{9}$$

The reason is that the ultrasonic perturbation functions merely as a small-amplitude, high-frequency background noise and does not interfere with the normal voice interaction of VAs.

2) Universal Perturbation Signal Generation: To generate subtle, pervasive, and imperceptible universal perturbations $\delta(t)$, SUAD Defense introduces a UAP-based [36] method.

First, we define a speech domain $X = \{x_1(t), x_2(t), \ldots\}$, where $x_i(t)$ represents speech signals that may be received by VAs on a user's smartphone, such as "Hi Siri". Based on X, we propose a defense model as follows:

$$\text{CER}(C(x_i(t)), C(x_i(t) + \delta'(t))) > 0.7, \quad \forall x_i \in X, \tag{10}$$

where $C(\cdot)$ denotes a speech recognition model accessed by defender, and $\text{CER}(x, y)$ represents the character error rate (CER), i.e., the edit distance [37] between two recognition results. If perturbation signals are added to the speech, more than 70% of the VA's results should be incorrect.

Furthermore, since the attack signal may begin at any arbitrary point in time, the perturbation signal may not always be temporally aligned with the speech input $x_i(t)$. To address this, we represent the speech input as $x_i(t+\tau) \in [0, T]$, which indicates a circularly shifted result of $x_i(t)$ by τ, aligned with the perturbation signal. For example, the original command "Hi Siri" may be transformed into "Siri Hi" during training to address uncertainty in the start time.

However, since the lengths of each x_i are not the same, we first normalize all x_i to the same length via zero-padding, before applying circular shifting. Finally, considering the impact of different carrier frequencies on the down-modulated universal perturbation, we modulate the above command signals using multiple carrier frequencies, forming the dataset used for model training. Thus, the constraints on the perturbation $\delta(t)$ can be summarized as follows:

$$\begin{aligned} &\|\delta\|_\infty(t) < \epsilon \\ P_{x\sim X}(&\text{CER}(x_i(t+\tau), C(x_i(t+\tau) + \delta'(t))) > 0.7) \geq L, \end{aligned} \tag{11}$$

where ϵ denotes the maximum allowable magnitude in each iteration, which constrains the size of $\delta(t)$, and L represents the defense success rate.

To solve the above constrained optimization problem, we design an iterative algorithm named Time and Frequency Randomized UAP Training, as presented in Algorithm 1. Briefly, the entire iterative process proceeds by sequentially selecting $x_i(t+\tau)$ and modulating it with different carrier frequencies f_x. Then, the perturbation vector $\delta(t)\cos(2\pi f_x t)$ is constructed step by step, and the down-conversion process is simulated to generate $x_i(t+\tau) + \delta'(t)$. In each iteration, we compute the smallest perturbation increment $\Delta\delta(t)$ and add it to the current perturbation $\delta^{'}(t)$, resulting in a transcription error. Meanwhile, the perturbation amplitude is controlled by

Algorithm 1: Time and Frequency Randomized UAP Training

Input: X: speech set, T: delay set, F: frequency set, C: speech recognition model, f_x: perturbation signal frequency, ϵ: maximum magnitude, N: max iterations, a: regularization weight
Output: Universal perturbation $\delta(t)$
Initialize $\delta(t) = \vec{0}, \quad r = 0, \quad k = 0;$
for $x_i(t) \in X$ **do**
  Zero-pad $x_i(t)$ to fixed length;
  for *each* $(\tau, f) \in T \times F$ **do**
    $\tilde{x}(t) \leftarrow x_i(t+\tau) \cdot \cos(2\pi f(t+\tau)) + \cos(2\pi f(t+\tau));$
    $\delta_{mod}(t) \leftarrow \delta(t) \cdot \cos(2\pi f_x t);$
    $x^{adv}(t) \leftarrow \text{Nonlinear}(\tilde{x}(t) + \delta_{mod}(t));$
    while
     $\textit{CER}(C(x_i(t+\tau)), C(x^{adv}(t)+r)) \leq 0.7 \parallel k < N$
     do
      Compute $\Delta\delta(t)$ minimizing:
      $a\|r\|^2 - \text{CTCLoss}(C(x^{adv}(t)+r), C(x_i(t+\tau)))$
      subject to $\|\delta(t) + r\|_\infty < \epsilon;$
      Update: $\delta(t) \leftarrow \delta(t) + \Delta\delta(t)$, project to $\|\cdot\|_\infty < \epsilon;$
      $k \leftarrow k + 1;$
return $\delta(t);$

enforcing $\|\delta\|_\infty < \epsilon$ to ensure it remains within an acceptable range. This process can be expressed as:

$$\begin{aligned} &\Delta\delta_i^j \leftarrow \arg\min_r \|r\|_2 \\ &\text{s.t.} \quad \text{CER}(C(x_i(t+\tau)), C(x_i(t+\tau) + \delta'(t) + r)) > 0.7. \end{aligned} \tag{12}$$

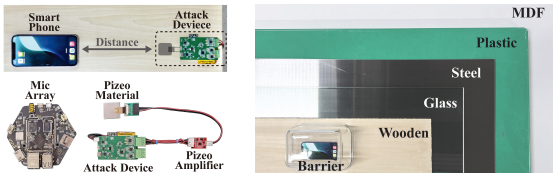
This is equivalent to maximizing the loss between the predicted probability distribution of the perturbed result $C(x_i(t+\tau) + \delta'(t) + r)$ and the original result $C(x_i(t+\tau))$, i.e.,

$$\begin{aligned} &\min_r a\|r\|^2 - \text{CTCLoss}(x_i(t+\tau) + \delta'(t) + r, C(x_i(t+\tau))) \\ &\text{s.t.} \quad \|\delta'(t) + r\|_\infty < \epsilon, \end{aligned} \tag{13}$$

where CTCLoss() denotes the Connectionist Temporal Classification (CTC) loss function, a commonly used objective in modern end-to-end speech recognition models to quantify the difference between the model output and the original transcription after perturbations are applied. Since the problem is a non-convex optimization task, we approximate its solution using the Iterative Gradient Sign Method (IGSM), i.e.,

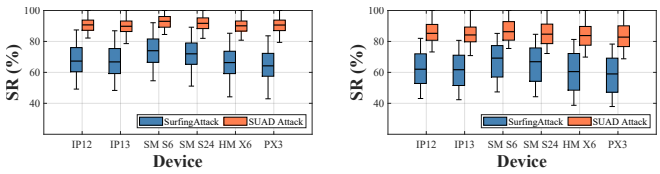
$$\begin{aligned} r_0 &= \vec{0}, \\ r_{N+1} &= \text{cut}_{(x_i(t+\tau),\epsilon)}\{r_N + a \cdot \text{sign}(\nabla_{(x_i(t+\tau)+\delta'(t)+r)} \\ &\quad \text{CTCLoss}(x_i(t+\tau) + \delta'(t) + r, C(x_i(t+\tau))))\}, \end{aligned} \tag{14}$$

where cut() denotes a clipping function, analogous to clip() in image-based attacks. It segments the perturbation according to the audio's sampling points to ensure that the updated perturbation amplitude does not exceed ϵ. The regularization parameter a denotes the update step size for each iteration,



(a) Attack devices (b) Solid materials

Fig. 9. Experimental Setup.



(a) VAs activation (b) Command execution

Fig. 10. Overall performance of SUAD Attack.

and its value is determined via hyperparameter search on the validation set to maximize the attack success rate under the constraint of the maximum allowable perturbation magnitude. The operator sign() indicates the sign of the gradient, specifying the direction (positive or negative) of the perturbation update. The term $\nabla(x_i(t+\tau)+\delta'(t)+r)$ denotes the gradient computed with respect to the current perturbed input signal, identifying the direction in which the input is most sensitive to the CTCLoss().

Finally, we utilize the public dataset and the TTS model introduced in the previous section to generate the training and test sets. During training, the universal adversarial perturbation is updated using the gradient descent method until it successfully defends against any sample in the test set.

IV. Evaluation

In this section, we present the experimental setup and evaluate the overall performance of SUAD Attack and Defense, followed by an analysis of the impact of various factors.

A. Experimental Setup

1) Implementation: As shown in Fig. 9(a), we utilize a six-microphone array as the receiver and a piezoelectric transducer as the transmitter, both equipped with communication and power modules. The receiver sends audio recordings (i.e., the smartphone hitting) to a Linux server with an NVIDIA RTX 4090 GPU and an Intel Xeon Gold CPU for generating forged commands, which are sent to the transmitter for attacking VAs. For SUAD Defense, generated perturbation signals are modulated to 18 kHz for defending against IVAs.

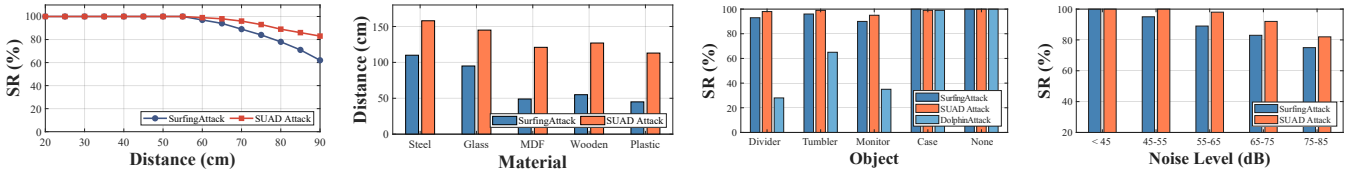
2) Data Collection: We recruited 5 subjects (3 males and 2 females), aged from 22 to 33 years old. Each subject was instructed to record activation and execution commands compatible with various VAs to construct a comprehensive voice command set, as well as an attack command set comprising 50 commonly used words. Each subject recorded approximately 5 minutes of voice commands. 3-7 word commands were designed to simulate realistic attack scenarios (e.g., sending SMS). Furthermore, we construct an audio dataset containing dispersion features by collecting audio propagating through solid media, such as materials shown in Fig. 9(b).

B. SUAD Attack Performance

1) Overall Attack Performance: We evaluated the success rate (SR) of SUAD Attack in activating VAs using forged commands, comparing with SurfingAttack across multiple devices, including iPhone 12/13 (IP 12/13), Samsung S6/S24 (SM S6/S24), HUAWEI Mate X6 (HM x6), and Pixel 3 (PX3). As shown in Fig. 10(a), SUAD Attack achieves a median activation success rate exceeding 89.8% across different smartphones, demonstrating that its forged voice commands effectively simulate real user input. In comparison, SurfingAttack's median activation success rate does not exceed 74%. It may be attributed to dispersion effects in solid media that distort the transmitted signals, causing VAs to misidentify activation commands as other content. In addition, we evaluated commands such as 'turn on camera' and 'play music' to further evaluate the effectiveness of SUAD Attack. As shown in Fig. 10(b), the median success rate of command execution for SUAD Attack is significantly lower than that for VAs activation in Fig. 10(a), but remains above 82.8%. Similarly, due to increased command length, which raises the misrecognition by VAs, SurfingAttack exhibits decreasing results. Notably, due to architectural, acoustic sensor, and VA consistency, both attacks exhibit similar performance on devices from the same manufacturer, such as iPhone and Samsung. Despite hardware variability, SUAD Attack consistently achieves high success rates, demonstrating its strong adaptability.

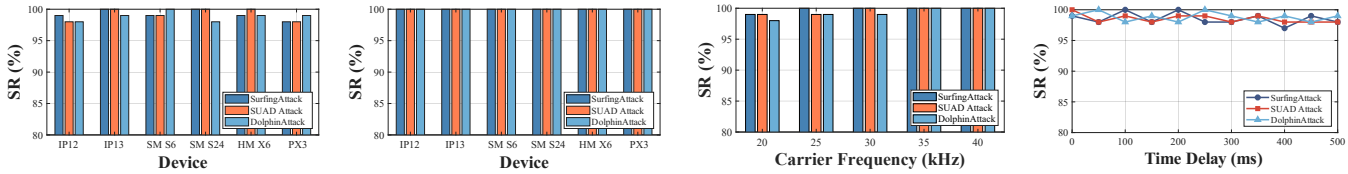
2) Performance under Different Attack Distances: Since victims may place their smartphones arbitrarily on the table, we further evaluated the impact of varying attack distances on SUAD Attack performance. Forged commands are transmitted at distances ranging from 20 to 90 cm, and the execution success rate is used as the evaluation metric. The results in Fig. 11(a) show that both SUAD Attack and SurfingAttack maintain a 100% success rate at distances up to 55 cm. However, as the distance increases, the performance of SurfingAttack degrades significantly, whereas SUAD Attack demonstrates greater robustness to distance variations. This can be attributed to SUAD Attack's integration of inverse solid-channel interference during command generation, which effectively mitigates signal distortion caused by distance variations.

3) Performance under Different Materials: Equation (1) demonstrates that variations in solid materials significantly affect signal dispersion during propagation. We selected five materials, including wooden, glass, steel, plastic, and MDF, as transmission media and transmitted commands to the target device. The maximum attack distance where the execution success rate reaches 50% (i.e., effective attack distance) is used as the performance metric. As shown in Fig. 11(b), SUAD Attack achieves effective attack distances of 158, 145, 121, 127, and 113 cm across the five materials, all substantially exceeding those of SurfingAttack, which reach only 110, 95, 49, 55, and 45 cm, respectively. In particular, plastic exhibits the greatest impact on both systems among the tested materi-



(a) Impact of different distances (b) Impact of different materials (c) Impact of different objects (d) Impact of different noise levels

Fig. 11. SUAD Attack performance under various factors.



(a) Defense Against IVAs (b) VAs activation under defense (a) Impact of different frequencies (b) Impact of different time delays

Fig. 12. Overall performance of SUAD Defense.

Fig. 13. SUAD Defense performance against different factors.

als, primarily due to its higher signal attenuation. Additionally, SUAD Attack incorporates an anti-dispersion component in its command generation, resulting in smaller performance variations across different media than SurfingAttack.

4) Performance under Different Objects on The Table: In practical scenarios, the LoS path between the attack device and the target device may be obstructed by objects on the table. On the one hand, these objects obstruct LoS propagation through the air channel. On the other hand, their physical contact with the table may impact the solid-channel transmission. To evaluate the impact of different barriers, we evaluated the execution success rates of SUAD Attack, SurfingAttack, and DolphinAttack at a fixed distance of 50 cm, respectively. The results, shown in Fig. 11(c), indicate that DolphinAttack, which relies on air-channel propagation, is highly sensitive to barriers. Its success rate exhibits a strong negative correlation with barrier size, i.e., smaller obstacles result in higher attack effectiveness. In contrast, other attacks relying on solid-channel propagation are minimally affected by barriers, demonstrating greater robustness. This suggests that such attacks can effectively bypass the blocking effects of objects on the table.

5) Performance under Different Noise Levels: To evaluate SUAD Attack's performance under varying noise conditions, we simulated real-world scenarios by playing background noise at different intensity levels. We selected the iPhone 13 as the target device, set the attack distance to 50 cm, and transmitted 100 execution commands under each noise level. As shown in Fig. 11(d), the execution success rates of both SUAD Attack and SurfingAttack decline with increasing noise levels. Specifically, the success rate of SUAD Attack decreases from 100% to 82%, whereas SurfingAttack experiences a sharper decline from 100% to 75%. This is because SurfingAttack is influenced by both ambient noise and solid-channel dispersion, leading to greater distortion in the downconverted signal received by the microphone. In addition, the success rates of SUAD Attack in normal noise environments (e.g., lower 65 dB), which may be tolerable for victims, remain above 89%, demonstrating strong robustness.

C. SUAD Defense Performance

1) Overall defense Performance: Our defense technique aims to disrupt the downconversion of attack commands by transmitting universal perturbation signals modulated into the ultrasonic band, thereby preventing the smartphone's VA from recognizing forged voice commands. To evaluate the performance of SUAD Defense, we continuously emit ultrasonic perturbation signals from the target device while executing SUAD Attack, SurfingAttack, and DolphinAttack, each transmitting ultrasonic commands intended to activate the voice assistant on the corresponding device. The defense success rate is calculated as the ratio of successfully blocked attacks to the total number of attack attempts. The results shown in Fig. 12(a) demonstrate the performance of SUAD Defense against SUAD Attack, SurfingAttack, and DolphinAttack under normal noise conditions at an attack distance of 50 cm. The defense achieves a success rate of over 98% for each type of inaudible voice attack, indicating that our system maintains a minimal false positive rate under operating conditions.

In addition, we have evaluated the impact of SUAD Defense on normal voice commands when defending against ultrasonic attacks. In this experiment, SUAD Attack, SurfingAttack, or DolphinAttack is launched against the target device while ultrasonic perturbation signals are emitted as a defense mechanism. Activation voice commands are played 100 times, and the voice assistant's activation success rate is measured to evaluate the impact of the defense on its functionality. The results are presented in Fig. 12(b), showing that SUAD Defense did not interfere with the control of VAs by normal voice commands, regardless of the attack method employed. The reason is that the perturbation signal of SUAD Defense is modulated within the inaudible frequency band, specifically disrupting the recognition of ultrasonic forged commands. In contrast, for normal audible speech, it manifests merely as low-amplitude high-frequency noise, exerting little to no impact on recognition accuracy. This result underscores a key advantage of SUAD Defense: it can effectively counter IVAs while preserving the normal functionality of VAs.

2) Defense against different-frequency attacks: The fundamental principle of IVAs is to modulate attack commands into the inaudible frequency band, which are then down-converted into audible commands through the microphone's nonlinearity. However, any frequency within the inaudible band can serve as a carrier frequency. To evaluate the effectiveness of SUAD Defense against ultrasonic attacks at various frequencies, we modulate the attack commands from different IVAs using distinct carrier frequencies and conduct 100 attack attempts on the same phone for each frequency. The results are shown in Fig. 13(a), where the defense success rate of SUAD Defense increases slightly as the carrier frequency increases. This may be attributed to the fact that lower-frequency attack signals undergo less dispersion in solid media, resulting in stronger signal integrity and revealing potential vulnerabilities of the model to such attacks. Nevertheless, SUAD Defense remains above 98% for attack commands at different frequencies, demonstrating its strong robustness against variations in carrier frequency.

3) Defense against randomly sent attacks: Due to the random timing of attack signal transmissions, the perturbation signal may not be temporally synchronized with these signals. To evaluate the impact of time delay on SUAD Defense performance, we transmitted attack signals with varying delays relative to the perturbation signal using the same device. As shown in Fig. 13(b), while time delay introduces minor fluctuations in defense success rate, overall effectiveness remains largely unaffected. This robustness is attributed to the perturbation signal design, which accounts for time-delay-induced dispersion effects, and to adversarial learning employed during model training.

V. Related Work

In this section, we first review various modalities of voice command injection attacks and their limitations, followed by an analysis of the shortcomings in existing defenses.

A. Voice Command Injection Attacks

Although advances in AI have greatly contributed to the widespread application of VAs, it has also raised concerns about voice command injection attacks [5]–[17], [38]–[43]. For example, GVSAttack [38] launches attacks on VAs by replaying voice commands. In addition, some studies attempt to conceal voice commands. Hidden Voice Commands [39] and Adversarial Attacks [40] transform voice commands into white noise, and CommanderSong [41] embeds them into songs. However, such audible voice attacks are easily identified and interrupted by users.

To hide attacks, researchers have explored various techniques for inaudible voice command injection [5]–[17]. For instance, MagBackdoor [7] and GhostTalk [11] leverage external magnetic fields to induce the target device's speaker to emit voice commands, but their effective range is limited to less than 6 cm. Light Commands [12] encodes speech information into laser beams, enabling attacks over distances up to 110 m. However, it requires expensive devices and LoS. Consequently, ultrasound-based attacks (i.e., IVAs) [13]–[17] have attracted increasing attention due to their potential in IVAs. DolphinAttack [13] and Backdoor [14] utilize nonlinear effects to modulate voice commands into high-frequency carriers for attacks. However, these approaches typically require an unobstructed LoS. Recently, SurfingAttack [16] has investigated the feasibility of performing ultrasonic injection through solid media, but ignores the interference of solid channel signals. Notably, SUAD Attack not only utilizes solid channels to penetrate physical barriers but also resists the dispersion effect of both distance and solid materials, thereby enabling long-range, interference-free attacks.

B. Defense against IVAs

In recent years, various methods [13], [15], [19]–[28] have been proposed to defend against IVAs. Some methods [20], [21] detect IVAs based on anomalies in the spectrum or sound field. For example, Cacher [21] distinguishes forged voice commands by establishing a biometric basis, Fieldprint, which represents sound field features. However, their effectiveness against ultrasonic band attacks remains unsatisfactory. Additionally, some multimodal methods [22]–[24] attempt to utilize physiological or behavioral features to determine the authenticity of voice commands. Among them, MFF [24] detects attack attempts by leveraging synchronized video and audio data, but it requires high-precision sensors to ensure recognition accuracy. EarArray [25], RobustDetection [26], and arrayID [27] utilize microphone arrays for delay analysis and orientation analysis to extract spatial features during speech, but these techniques require dedicated hardware support. In contrast, modulation-based detection methods [13], [15], [19], [28] offer greater flexibility and can be applied to smartphones. For example, LipRead [15] identifies IVAs by exploiting differences in frequency patterns between normal speech and fixed-frequency modulation attacks. However, these defenses function purely as passive detection mechanisms and do not actively prevent IVAs. In particular, these methods may interfere with the user's normal use of VAs, e.g., the system may disable the microphone after detecting an attack.

VI. Conclusion

This paper presented SUAD that explores the feasibility of solid-channel attacks and universal defenses with inaudible ultrasonic signals. SUAD Attack presented a novel notion that solid-channel propagation can introduce significant distortions to acoustic signals. Thus, SUAD Attack employed a multi-parameter modular command generation model that adapts to solid channels by parameterizing attack distance, victim audio, and dispersion features. Additionally, SUAD Defense developed a universal defense against IVAs based on microphone nonlinearity, but without affecting VAs. An innovative UAP training method was designed to generate perturbation signals capable of blocking IVAs randomly sent at arbitrary frequencies. Extensive experiments demonstrate the effectiveness SUAD Attack and Defense.

REFERENCES

[1] P. Garai, A. Nikam, S. Kerkar, S. Deshmukh, A. Mane, and S. Bhise, "Voice ai-intelligence based voice assistant," in *IITCEE*, 2025, pp. 1–6.

[2] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, "Compute trends across three eras of machine learning," in *IEEE IJCNN*, 2022, pp. 1–8.

[3] R. Jampala, D. S. Kola, A. N. Gummadi, M. Bhavanam, and I. Rani Pannerselvam, "The evolution of voice assistants: From text-to-speech to conversational ai," in *IDCIoT*, 2024, pp. 1332–1338.

[4] J. Li, C. Chen, M. R. Azghadi, H. Ghodosi, L. Pan, and J. Zhang, "Security and privacy problems in voice assistant applications: A survey," *Comput. Secur.*, vol. 134, p. 103448, 2023.

[5] Y. Tu, V. S. Tida, Z. Pan, and X. Hei, "Transduction shield: A low-complexity method to detect and correct the effects of EMI injection attacks on sensors," in *Proc. of the 2021 ACM Asia CCS*, 2021, pp. 901–915.

[6] Z. Xu, R. Hua, J. Juang, S. Xia, J. Fan, and C. Hwang, "Inaudible attack on smart speakers with intentional electromagnetic interference," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 5, pp. 2642–2650, 2021.

[7] T. Liu, F. Lin, Z. Wang, C. Wang, Z. Ba, L. Lu, W. Xu, and K. Ren, "Magbackdoor: Beware of your loudspeaker as a backdoor for magnetic injection attacks," in *2023 IEEE Symposium on Security and Privacy (SP)*, 2023, pp. 3416–3431.

[8] D. Dai, Z. An, and L. Yang, "Inducing wireless chargers to voice out for inaudible command attacks," in *IEEE symposium on security and privacy (SP)*, 2023, pp. 1789–1806.

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

[10] C. Kasmi and J. Lopes Esteves, "Iemi threats for information security: Remote command injection on modern smartphones," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 6, pp. 1752–1755, 2015.

[11] Y. Wang, H. Guo, and Q. Yan, "Ghosttalk: Interactive attack on smartphone voice system through power line," in *29th Annual Network and Distributed System Security Symposium*, 2022.

[12] 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*, 2020, pp. 2631–2648.

[13] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proc. of the 24th ACM CCS*, 2017, pp. 103–117.

[14] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 2–14.

[15] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *15th USENIX Symposium on Networked Systems Design and Implementation*, 2018, pp. 547–560.

[16] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *27th Annual Network and Distributed System Security Symposium*, 2020.

[17] X. Ji, J. Zhang, S. Jiang, J. Li, and W. Xu, "Capspeaker: Injecting voices to microphones via capacitors," in *In Proc. of the 28th ACM CCS*, 2021, pp. 1915–1929.

[18] V. I. Erofeev and A. Malkhanov, "Nonlinear acoustic waves in solids with dislocations," *Procedia IUTAM*, vol. 23, pp. 228–235, 2017.

[19] T. Liu, F. Lin, Z. Ba, L. Lu, Z. Qin, and K. Ren, "Micguard: A comprehensive detection system against out-of-band injection attacks for different level microphone-based devices," in *Proc. of the 33rd USENIX Conference on Security Symposium*, 2024, p. 3963–3978.

[20] S. Wang, J. Cao, X. He, K. Sun, and Q. Li, "When the differences in frequency domain are compensated: Understanding and defeating modulated replay attacks on automatic speech recognition," in *Proc. of the ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1103–1119.

[21] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proc. of the 26th ACM CCS*, 2019, pp. 1215–1229.

[22] M. Zhou, Z. Qin, X. Lin, S. Hu, Q. Wang, and K. Ren, "Hidden voice commands: Attacks and defenses on the vcs of autonomous driving cars," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 128–133, 2019.

[23] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proc. of ACM MobiCom*. ACM, 2017, pp. 343–355.

[24] J. Guan, L. Pan, C. Wang, S. Yu, L. Gao, and X. Zheng, "Trustworthy sensor fusion against inaudible command attacks in advanced driver-assistance systems," *IEEE Internet of Things Journal*, vol. 10, no. 19, pp. 17 254–17 264, 2023.

[25] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "Eararray: Defending against dolphinattack via acoustic attenuation," in *28th Annual Network and Distributed System Security Symposium*, 2021.

[26] Z. Li, C. Shi, T. Zhang, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array," in *Proc. of the 28th ACM CCS*, 2021, pp. 1884–1899.

[27] Y. Meng, J. Li, M. Pillari, A. Deopujari, L. Brennan, H. Shamsie, H. Zhu, and Y. Tian, "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *31st USENIX Security Symposium*, 2022, pp. 1077–1094.

[28] J. Mao, S. Zhu, X. Dai, Q. Lin, and J. Liu, "Watchdog: Detecting ultrasonic-based inaudible voice attacks to smart home systems," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8025–8035, 2020.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations*, 2015.

[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations*, 2018.

[31] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, "Reshape dimensions network for speaker recognition," in *25th Annual Conference of the International Speech Communication Association*, 2024.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[33] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1992.

[34] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *IEEE ICASSP*, 2018, pp. 4779–4783.

[35] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[36] P. Neekhara, S. Hussain, P. Pandey, S. Dubnov, J. J. McAuley, and F. Koushanfar, "Universal adversarial perturbations for speech recognition systems," in *20th Annual Conference of the International Speech Communication Association*, 2019, pp. 481–485.

[37] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[38] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proc. of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 2014, pp. 63–74.

[39] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. A. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium, USENIX Security*, 2016, pp. 513–530.

[40] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *Proc. of the 26th NDSS*, 2019, pp. 1–18.

[41] 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 *Proc. of the 27th USENIX Conference on Security Symposium*, 2018, pp. 49—-64.

[42] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Trans. Intell. Technol.*, vol. 6, no. 1, pp. 25–45, 2021.

[43] Y. Huang, B. Obada-Obieh, and K. Beznosov, "Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks," in *ACM CHI*, 2020, pp. 1–13.