SurfingAttack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves

Qiben Yan*†, Kehai Liu*‡, Qin Zhou§, Hanqing Guo†, Ning Zhang¶
†Computer Science & Engineering, Michigan State University, {qyan, guohanqi}@msu.edu
‡Songshan Lake Materials Laboratory, Institute of Physics, Chinese Academy of Sciences, liukehai@sslab.org.cn
§Mechanical & Materials Engineering, University of Nebraska-Lincoln, zhou@unl.edu
¶Computer Science & Engineering, Washington University in St. Louis, zhang.ning@wustl.edu

***Abstract*—With recent advances in artificial intelligence and natural language processing, voice has become a primary method for human-computer interaction. It has enabled game-changing new technologies in both commercial sectors and military sectors, such as Siri, Alexa, Google Assistant, and voice-controlled naval warships. Recently, researchers have demonstrated that these voice assistant systems are susceptible to signal injection at the inaudible frequencies. To date, most of the existing works focus primarily on delivering a single command via line-of-sight ultrasound speaker or extending the range of this attack via speaker array. However, besides air, sound waves also propagate through other materials where vibration is possible. In this work, we aim to understand the characteristics of this new genre of attack in the context of different transmission media. Furthermore, by leveraging the unique properties of acoustic transmission in solid materials, we design a new attack called *SurfingAttack* that would enable multiple rounds of interactions between the voice-controlled device and the attacker over a longer distance and without the need to be in line-of-sight. By completing the interaction loop of inaudible sound attack, *SurfingAttack* enables new attack scenarios, such as hijacking a mobile Short Message Service (SMS) passcode, making ghost fraud calls without owners' knowledge, etc. To accomplish *SurfingAttack*, we have solved several major challenges. First, the signal has been specially designed to allow omni-directional transmission for performing effective attacks over a solid medium. Second, the new attack enables multi-round interaction without alerting the legitimate user at the scene, which is challenging since the device is designed to interact with users in physical proximity rather than sensors. To mitigate this newly discovered threat, we also provide discussions and experimental results on potential countermeasures to defend against this new threat.**

I. Introduction

Recent advances in artificial intelligence (AI) and machine learning have enabled new game-changing technologies for humans to interact with machines. Conversation with AI is no longer a scene in science-fiction movies, but day-to-day routines. It is now possible for everyday users to converse with voice assistants, such as Bixby, Siri, Google Assistant to arrange appointments on the calendar or to start the morning coffee brewing. While these new technologies significantly improve the living quality, they also change the landscape of cyber threats. Recent studies show that it is possible to exploit the non-linearity in microphone to deliver inaudible commands to the system via ultrasound signals [41], [42], [44], [52].

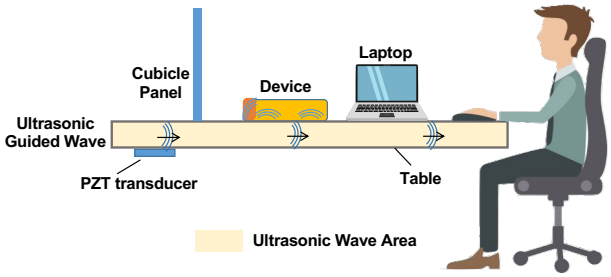


Fig. 1: *SurfingAttack* leverages ultrasonic guided wave in the table generated by an ultrasonic transducer concealed beneath the table.

DolphinAttack [52] by Zhang et al. was among the first to demonstrate inaudible attacks towards voice-enabled devices by injecting ultrasound signals over the air, which can launch from a distance of 5*ft* to the device. Recognizing the limitation in the range of the attack, LipRead [42] extends the attack range to 25*ft* by aggregating ultrasound signals from an array of speakers, which requires line-of-sight. While these two attacks demonstrate the feasibility of voice command injection via inaudible ultrasound, they focus solely on over the air transmission, which leads to several inherent limitations due to the physical property of ultrasound wave propagation in air, such as significant performance degradation when there is line-of-sight obstruction. However, sound wave is fundamentally the transfer of acoustic energy through a medium. It can propagate wherever vibration is possible, such as water and solid materials, in which the propagation characteristics are different from air. Furthermore, the current literature [42], [44], [52] focuses mostly on one-way interaction, i.e., they inject commands to voice assistants without expecting any feedback. However, voice-activated devices are designed to enable multiple rounds of interactions. While the previous literature has identified the new attack vector, its potential in multi-round communication has received little attention.

In this work, we aim to understand the new threats enabled by inaudible signal injection using ultrasound propagation in

*The first two authors contributed equally to this work. Dr. Liu conducted this work when he was a postdoctoral research fellow under the supervision of Dr. Yan.

Network and Distributed Systems Security (NDSS) Symposium 2020
23-26 February 2020, San Diego, CA, USA
ISBN 1-891562-61-4
https://dx.doi.org/10.14722/ndss.2020.24068
www.ndss-symposium.org

solid media, and the possibility of realizing multi-rounds of hidden communication with AI-based voice assistants. Using our proposed attack, *SurfingAttack*[1], we found that it is possible to deliver various inaudible voice commands in ultrasound to a wide range of target devices from different manufacturers via different solid media. Due to the unique properties of guided wave propagation, *SurfingAttack* not only enables attacks from a longer distance with a lower power requirement, but also eliminates the need to be in the line-of-sight for inaudible command injection attacks. By capitalizing on the capability to control feedback mechanisms via the initial injected command, *SurfingAttack* also enables inaudible multi-rounds of interactions between the attacker and the target device without alerting users in physical proximity. Fig. 1 illustrates one of the application scenarios of *SurfingAttack*, where a malicious device is hidden beneath the table to converse with the target device on the top. By injecting voice commands stealthily, attackers can instruct the voice assistants to leak various secrets, such as an authentication code for money transfer sent via an SMS message. The leaked secret can then be picked up by a malicious device hidden away and relayed back to the remote attacker. By leveraging unique guided wave propagation properties in solid media, *SurfingAttack* presents a new genre of inaudible attack on voice-activated systems.

While conceptually simple, there are several major challenges in realizing this attack: (a) how to design a hidden signal generator that can penetrate materials effectively and inject the inaudible commands without facing the victim's device? (b) How to engage in multiple rounds of conversations with the victim's device such that the voice response is unnoticeable to humans while still being recorded by a tapping device?

For the first question, while the characteristics of sound wave propagation in solid material is well studied for specific application domains such as structure damage detection [39], adapting the technique to deliver inaudible commands presents unique challenges, such as wave mode selection, vertical energy maximization, and velocity dispersion minimization. Traditional ultrasonic speakers, as used in previous attacks, are not suitable for exciting guided waves in table materials due to their transducer structures. In order to adapt to the solid medium, we utilize a special type of ultrasonic transducer, i.e., piezoelectric (PZT) transducer, to generate ultrasonic guided waves by inducing minor vibrations of the solid materials. However, due to the unique characteristics of ultrasound transmission in different solid materials, the selection of different modes of guided wave can lead to significant differences in the attack outcome, compared to the over-the-air delivery of manipulated signals. To enable *SurfingAttack*, we redesign a new modulation scheme that considers wave dispersal patterns to achieve optimal inaudible command delivery. *SurfingAttack* presents two unique features. First, the attack is omni-directional, which works regardless of the target's orientation or physical environment where the target resides. Second, the success of the attack is not impacted by objects on a busy tabletop. To the best of our knowledge, we are the first to deliver inaudible commands to a variety of mobile devices through ultrasonic guided waves in a busy environment. For the second question, to enable inaudible multi-rounds of interactions, a tapping device is added along with the ultrasound transducer to capture voice feedbacks from the device. In order to minimize the impact of the feedback on the environment, an injected command is used to tune the output of the device to the lowest volume setting, such that the feedback becomes difficult to notice by users, but can still be captured by a sensitive tapping device. We have conducted a series of experiments to understand the feasibility and limitation of such low-profile feedback.

Leveraging the low attenuation of guided waves in solid material and a place to hide the attack device, *SurfingAttack* can enable a variety of new attacks including not only the non-interactive attacks such as visiting a malicious website, spying, injecting fake information, and denial of service by turning on the airplane mode, but also interactive attacks that would require multiple rounds of conversations with the target device, such as unauthorized transfer of assets from the bank. To demonstrate the practicality of *SurfingAttack*, we build a prototype of the attack device using a commercial-off-the-shelf PZT transducer, which costs around $5 per piece. Using our prototype device, we conduct the following two attacks as a demonstration:

(1) *Hacking an SMS passcode.* SMS-based two-factor authentication has been widely adopted by almost all major services [17], which often delivers one-time passwords over SMS. A *SurfingAttack* adversary can activate the victim's device to read SMS messages in secret thereby extracting SMS passcodes.

(2) *Making fraudulent calls*. A *SurfingAttack* adversary can also take control of the owner's phone to call arbitrary numbers and conduct an interactive dialogue for phone fraud using the synthetic voice of the victim.

We have tested *SurfingAttack* on 17 popular smartphones and 4 representative types of tables. We successfully launch *SurfingAttack* on 15 smartphones and 3 types of tables. A website is set up (**https://surfingattack.github.io/**) to demonstrate the attacks towards different phones under different scenarios, and various new attacks such as selfie taking, SMS passcode hacking, and fraudulent phone call attacks. With the growing popularity of mobile voice commerce and voice payments [25], we believe the demonstrated interactive hidden attack opens up new attacker capabilities that the community should be aware of. In summary, our contributions are as follows,

- We present, *SurfingAttack*, the first exploration of attack leveraging unique characteristics of ultrasound propagation in solid medium and non-linearity of the microphone circuits to inject inaudible command on voice assistants. We validate the effectiveness of *SurfingAttack* on Google Assistant of 11 popular smartphones, and Siri of 4 iPhones. We also show the attack is resilient against verbal conversations.
- We evaluate *SurfingAttack* on 4 representative types of table materials. We find that *SurfingAttack* is most effective through 3 types of tables: aluminum/steel, glass, and medium-density fiberboard (MDF). Notably, *SurfingAttack* can achieve long-range attack of 30*ft* distance through a metal table (the longest table we can acquire is 30*ft*). We also validate the effectiveness of *SurfingAttack* on aluminum and glass tables

[1]The attack excites the guided waves that surf in the "ocean" of materials and reach the surface to launch attacks, and thus is dubbed *SurfingAttack*.

with different thicknesses (up to 1.5 inch aluminum and 3/8 inch glass).

- We further explore the possibility to pair command injection with a hidden microphone to enable hidden conversations between the attacker and the victim voice assistant. We demonstrate several practical attacks using the prototype we build, including hacking an SMS passcode and making a ghost fraud phone call without owners' knowledge.
- We provide discussions on several potential defense mechanisms, including using the high-frequency components of guided waves as an indication of intrusion.

II. Background and Threat Model

In this section, we introduce the background knowledge of inaudible voice attack and physics of ultrasonic guided waves.

A. Inaudible Voice Attack

Audio capturing hardware in voice-controllable systems generally includes a micro-electromechanical system (MEMS) sensor to convert mechanical vibration to a digital signal, one or more amplifiers, a low-pass filter (LPF), and an analog-to-digital converter (ADC) to retrieve the sound in the physical world. Inaudible voice attacks leverage the non-linearity of the microphone circuits to inject inaudible commands to these systems. The nonlinear response is due to the imperfection of microphone and amplifier circuits [16], [27]. Let the input sound signal be $s(t)$, the output of microphone can be written as:

$$s_{out}(t) = A_1 s(t) + A_2 s^2(t), \tag{1}$$

where A_i $(i = 1, 2)$ is the gain of $s^i(t)$, while the higher order terms are ignored as they are typically extremely weak. The non-linearity term $s^2(t)$ produces harmonics and cross-products. With carefully-crafted input signals based on the baseband signals of voice commands, the microphone with non-linearity can recover the baseband signals using the cross-product term at the low frequency. Let the baseband voice signal be $v(t)$, the modulated input signal for launching attack is designed as:

$$s(t) = (1 + v(t))cos(2\pi f_c t), \tag{2}$$

where f_c is ultrasonic carrier frequency. After passing through the microphone, the recorded signal by the microphone becomes:

$$r(t) = A_2(1 + 2v(t) + v^2(t))/2, \tag{3}$$

since the high frequency components will be filtered out by LPF. If the voice command component $v(t)$ dominates in the recorded signal, the voice controllable systems will recognize the command. Previous work [52] demonstrated that the nonlinear effect of MEMS microphone can be best incited by ultrasonic frequencies between 20 kHz and 40 kHz.

B. Ultrasonic Guided Waves

The ultrasonic guided waves propagating in free solid-material plates are known as *Lamb waves*, which have distinct characteristics compared with ultrasonic waves in air. Assuming the wave motion takes place in the x_1x_3 plane, propagating along the x_1 direction, as illustrated in Fig. 2. Assume the table plate is a stress-free plate with thickness of $2h$ (along the x_3 direction), i.e., the mechanical component of stress at the surfaces of the plate is zero. While Lamb waves, made up of a superposition of longitudinal and transverse modes, transmit in a thin plate, their propagation characteristics vary with excitation and structural geometry of the plate. Every Lamb waveform belongs to one of two modes: symmetric and anti-symmetric [23], as presented in Fig. 2.

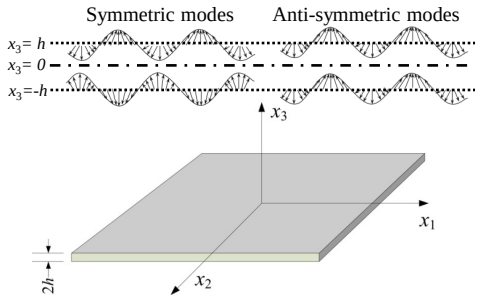


Fig. 2: Symmetric and anti-symmetric Lamb wave modes schematic in a free table plate. Symmetric Lamb wave is a family of waves whose motion (the arrows in the figure) is symmetrical with respect to the midplane of the plate (the plane x_3 = 0), and anti-symmetric wave is a family of waves whose motion is anti-symmetric with respect to the midplane.

For a symmetric mode waveform at a given frequency, the *displacement fields* u_1 (along x_1) and u_3 (along x_3) of the Lamb wave in the complex-value representation are given by [23]:

$$\begin{aligned} u_1(x_1, x_3, t) &= (ikA\cos\alpha x_3 + \beta B\cos\beta x_3)e^{i(kx_1-\omega t)} \\ u_3(x_1, x_3, t) &= (-\alpha A\sin\alpha x_3 - ikB\sin\beta x_3)e^{i(kx_1-\omega t)}, \end{aligned} \tag{4}$$

where A and B are given as an eigenvector of:

$$\begin{bmatrix} -2ik\alpha\sin\alpha h & \left(k^2-\beta^2\right)\sin\beta h \\ \left(k^2-\beta^2\right)\cos\beta h & -2ik\beta\cos\beta h \end{bmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{5}$$

For an anti-symmetric mode waveform, the displacement fields are:

$$\begin{aligned} u_1(x_1, x_3, t) &= (ikC\sin\alpha x_3 - \beta D\sin\beta x_3)e^{i(kx_1-\omega t)} \\ u_3(x_1, x_3, t) &= (\alpha C\cos\alpha x_3 - ikD\cos\beta x_3)e^{i(kx_1-\omega t)}, \end{aligned} \tag{6}$$

where C and D are given by:

$$\begin{bmatrix} 2ik\alpha\cos\alpha h & \left(k^2-\beta^2\right)\cos\beta h \\ \left(k^2-\beta^2\right)\sin\beta h & 2ik\beta\sin\beta h \end{bmatrix} \begin{pmatrix} C \\ D \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{7}$$

Here, ω and k are the angular frequency and the wavenumber of the Lamb wave, respectively. In addition,

$$\alpha = \sqrt{\left(\frac{\omega}{C_L}\right)^2 - k^2}, \beta = \sqrt{\left(\frac{\omega}{C_T}\right)^2 - k^2}, \tag{8}$$

where C_L and C_T are the longitudinal and transverse wave speeds that can be derived from:

$$C_L = \sqrt{\frac{E\left(1-v\right)}{\rho\left(1+v\right)\left(1-2v\right)}}, C_T = \sqrt{\frac{E}{2\rho\left(1+v\right)}}, \tag{9}$$

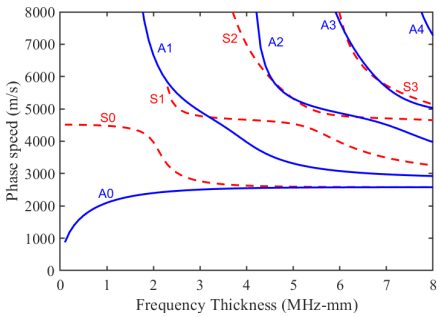


Fig. 3: The phase-velocity dispersion curve of a glass table.

which depends on three *material parameters*: E is the Young's module, v is the Poisson ratio, and ρ is the density of plate material.

The solutions of A, B, C, D in Eq. (5) and Eq. (7) require the determinants of the two matrices to vanish, which leads to the Rayleigh-Lamb equations (omitted for brevity) for the plate. The Rayleigh-Lamb equations quantify the relation between the angular frequency ω and the phase velocity C_p of the guided wave in the plate, where C_p is the propagation speed of the wave phase at a particular frequency within the guided wave signals. C_p relates to the wavenumber k: $C_p = \omega/k$. According to Rayleigh-Lamb equations, Lamb waves exhibit velocity dispersion; i.e., their propagation velocity depends on the frequency (or wavelength) and material parameters.

The phase-velocity dispersion curve, which depicts the correlation between the phase velocity and the frequency thickness product (i.e., the product of wave's angular frequency and material thickness, ωh), is plotted in Fig. 3 for a glass plate. We notice that with the increasing frequency thickness, more propagation modes appear, i.e., more types of waves with different transmission speeds will be propagating in the material and interfering with each other. As a result, the center frequency and guided wave mode of attack signals should be carefully selected based on attack target information extracted from guided wave sensitivity studies. Different guided wave modes introduce different wave propagation formats that could significantly impact the effectiveness of attack signal delivery, and the details of mode selection are presented in Section III-A.

C. Threat Model

The attacker's goal is to remotely converse with victim's voice controllable device to inject unauthorized voice commands or to access sensitive information without victim's knowledge. We assume the victim is familiar with his/her surrounding. This can be his/her own office or home, and he/she can notice any physical alteration.

Physical Access. We assume that adversaries can place a small attack device in the physical space of the device as long as it is not visible to the user. We assume that an adversary cannot physically touch the victim's devices, alter the device settings, or install malware apps. The activation commands ("Hey Siri", "OK Google") of voice assistants are generally voice fingerprinted, i.e., user verification is performed to authenticate these commands. We assume the attacker can synthesize the legitimate user's voice signals using known techniques [1], [33] to launch the attack when the target device is voiceprinted.

No Owner Interaction. We assume that the target device is placed on a medium that allows acoustic transmission, such as a tabletop, and it is not being actively used by the user. For smart home devices, the owner often interacts with it less than a dozen times a day. For mobile phones and tablets, it is also fairly common the owner is focusing on other activities not related to the device, such as reading books, having conversations with friends, working on a computer, etc.

Hidden Attack. One goal of the adversary is to attack voice assistants without being detected. The adversary will send the voice commands in ultrasonic frequencies that are inaudible to humans, and at the same time, turn down the volume of the device to the extent that it would be difficult for the users to notice the voice responses from the assistant, yet a hidden tapping device placed underneath the table can record them.

Attack Equipment. We assume that adversaries possess both the Piezoelectric (PZT) transducer designed for exciting ultrasonic guided wave and commodity devices for generating command signals. An ultrasonic signal source made of PZT transducer is relatively small and can generally be concealed and attached to a physical medium, such as the bottom of a table.

III. Key Elements of *SurfingAttack*

There are three necessary conditions for the success of *SurfingAttack*: (1) The ultrasonic wave in the table must be able to reach the device microphone embodied in the device enclosure. (2) Even when the microphone may not be in direct contact with the transmission medium, the wave should still be able to leverage the non-linearity of the device microphone on the tabletop to launch the inaudible command injection attacks. (3) The response from the victim device can be received by the attacker via the planted device without raising suspicion of the victim user. More specifically, the volume of victim's device can be tuned down such that user cannot notice it, yet the response can be recorded by a tapping device beneath the table.

A. Attack Wave Mode Selection and Generation

The first condition for the attack is the capability to deliver inaudible ultrasound waves to the target device effectively. Different from waves in air, the acoustic waves propagating in solid materials have acoustic dispersion phenomenon, during which a sound wave separates into its component frequencies as it passes through the material. Lower dispersion indicates a better concentration of acoustic energy. This implies that a proper Lamb wave mode for *SurfingAttack* should feature (1) low dispersion, (2) low attenuation, (3) easy excitability [46], and (4) high attack signal reachability. To achieve the aforementioned features, there are three key design decisions: the signal waveform, Lamb wave mode, and the ultrasound signal source.

First, guided wave signals can be generated via either windowed modulation or pulse signals. It has been shown that narrowband input signals are most effective in restricting

wave dispersal in large and thick plates [50]. As a result, *narrowband windowed modulation signals* is used to carry the attack command in *SurfingAttack* to minimize dispersion.

Second, different Lamb wave modes have different field distributions throughout the whole plate [11], depending on the different frequency-thickness and materials parameters as shown in Fig. 3. Since the attack frequency range from 20 kHz to 40 kHz has the best performance in stimulating the microphone's non-linearity effect, we are limited to the lower-order Lamb wave modes, i.e., A_0 or S_0 mode. In order to succeed in the attacks, the Lamb wave should be able to spread from a point of the table to the victim's device on the tabletop effectively. As a result, the generated Lamb waves need to produce a high out-of-plane displacement[2] on the table surface. As most of the displacement of the A_0 mode is out-of-plane, while most of the displacement of the S_0 mode is in-plane[3] with lower frequency-thickness products. A_0 wave mode below the cut-off frequencies of the higher order Lamb wave modes is selected to create the ultrasound commands.

Lastly, we choose to use a *circular piezoelectric disc* to generate the signal for its energy efficiency and omni-directivity. It applies a vertical force towards the table surface, resulting in a flexural wave propagating radially outwards and thus enabling an omni-directional attack through the table. The energy efficiency is important since the piezoelectric disc that is hidden under the solid materials needs to produce strong waves to reach extended distances with a minimal amount of energy. The omni-directivity is crucial because the attack should work regardless where the target's location and orientation are on the medium, i.e., wherever your phone is placed on the table. The omni-directivity of the attack is evaluated in Section VI-C. Furthermore, since objects on the table surface could change frequently, we need to make sure that the signal propagation still works regardless of whether there are objects on the table. The corresponding evaluation is presented in Section VI-G.

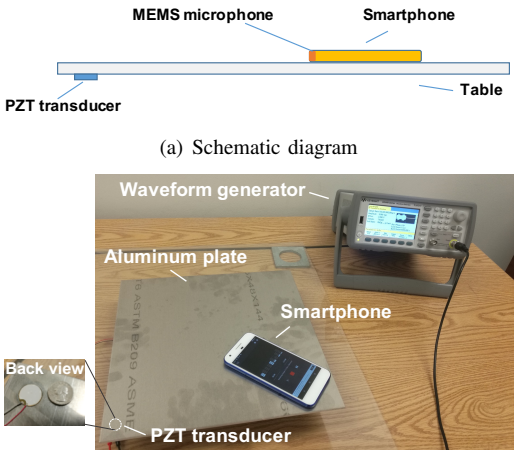
B. Triggering Non-linearity Effect via Solid Medium

While the non-linearity has been demonstrated for ultrasound wave that is directly delivered to the speaker via air, it is unclear if it is feasible to trigger the same effect when acoustic waves pass through the table materials to reach the external enclosure of the phone. We conduct extensive experiments to verify if the non-linearity effect of the voice capture hardware of a smartphone placed on the tabletop can be triggered by the ultrasonic guided waves that propagate in the table. The setup for one of the initial experiments is shown in Fig. 4.

We use a low-cost radial mode vibration PZT disc [45] (which only costs $5 per piece) with 22 mm diameter and 0.25 mm thickness to generate the ultrasonic guided wave. The disc is adhered to the underside of an aluminum plate with 3 mm thickness. The size of the PZT transducer is much smaller than the ultrasound speaker used in existing attacks [42], [44], [52], making the attacks more stealthy and economically accessible, as shown in Fig. 4(b). We use a chirp signal from 50 Hz to

[2]Out-of-plane displacement is defined as the displacement along the x_3 direction.

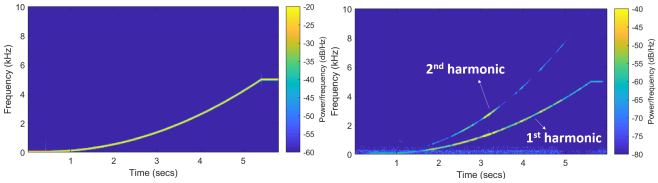
[3]In-plane displacement is defined as the displacement along the x_1 direction.



(a) Schematic diagram

(b) Experimental setup

Fig. 4: An illustration of the experimental setup for investigating the feasibility of *SurfingAttack*.



(a) (b)

Fig. 5: Results of non-linearity test: (a) spectrogram of the chirp baseband signal; (b) spectrogram of the recorded voice signal by the smartphone when the signal frequency is 25.3 kHz.

5 kHz as the baseband signal. The baseband signal is then imported to a Keysight 33500B series waveform generator and modulated onto a carrier. The 9V output is then supplied to the PZT transducer to excite Lamb waves. By analyzing the recorded signal of a smartphone (i.e., Google Pixel), the non-linearity of microphones could be evaluated.

Fig. 5 shows the spectrogram of the baseband signal and the recorded signal when carrier frequency $f_c = 25.3$ kHz. The ultrasonic guided wave propagates to the device microphone and any resulted sound is recorded. The results confirm the existence of the nonlinear response of the voice capture hardware incited by ultrasonic guided waves. Fig. 5(b) shows the recorded sound signal in the time-frequency domain, in which the first harmonic component is almost identical to the original signal displayed in Fig. 5(a). This result demonstrates the feasibility of attacking voice controllable systems placed on the tabletop through ultrasonic guided waves.

C. Unnoticeable Response

Unnoticeable response of a target phone is critical for keeping the attack under the radar. *Sound pressure level (SPL)* is used to quantify the sound pressure of a sound relative to a reference pressure at the eardrums of our hearing or on the diaphragms of the microphones. SPL is determined by the corresponding audio voltage, while standard reference sound pressure p_0 = 20 μ Pa $\equiv$ 0 dB is the quietest sound a human can perceive [49].

SPL depends on the distance between the area of measurement and point-shaped sound sources in the free field. We assume r_1 as the distance between the tapping device and the sound source, r_2 as the distance between the user and the sound source. L_1 and L_2 are SPLs at the tapping device and the user end, the relationship of which follows the inverse distance law as written below:

$$L_2 = L_1 - \left| 20 \cdot \log_{10} \left(\frac{r_1}{r_2} \right) \right|. \tag{10}$$

Approximately, an SPL drop of 6 dB is expected by doubling the difference of r_1 and r_2. When SPL at the user end drops below 0 dB, the voice response becomes essentially inaudible to the user. Thus, it becomes feasible to conceal *SurfingAttack* by adjusting the volume of the device via ultrasonic guide wave and placing a hidden tapping device closer to the victim's device underneath the table. Note that the inverse distance law is always an idealization because it assumes exactly equal sound pressure as sound field propagates in all directions. If there are reflective surfaces in the sound field, the reflected sounds will be added to the direct sound, resulting in a higher SPL at a field location than the inverse distance law predicts. If there are barriers between the source and the point of measurement, we may get a lower SPL.

To validate the feasibility, we evaluate the SPL of a Google Pixel phone at different volumes, the results of which are shown in Fig. 6. Here, we let the phone produce 1 kHz sinusoidal tones with low volume levels, and an A-weighting SPL meter is used to measure SPL at various distances. The experiment is conducted in a quiet office (about 400 square feet) with an average background noise of 40.5 dB. Although the SPL stays above 0 dB, it decreases with distance, and the signal is quickly overwhelmed by environmental noise after propagating 50∼100 cm at volume level 1∼3. We also deployed a microphone as a tapping device underneath the table, which is proven capable of recording the weak voice responses. The results show that it is feasible for the attacker to adjust the volume low enough to make the voice responses unnoticeable by the user from a moderate distance, while a hidden tapping device can still capture the sound. To enhance the sound capturing capability, we can deploy multiple tapping devices at different positions under the table to precisely capture the weak voice responses from the device speaker as well. In an environment with larger background noise, we can adjust volume even higher without alerting the owner. Lastly, the attacker can turn off the screen to further enhance the stealthiness of the threat. In Section V-D, we run extensive experiments to corroborate the stealthiness of *SurfingAttack* by measuring the responses of victim phones in different environments.

IV. Attack Design

To enable interactive hidden attacks, *SurfingAttack* generates well-crafted ultrasonic guided wave commands such that they can propagate along the table to control the voice assistants. The attack system is designed to initiate commands, record voice responses, and interact with victim devices. Without loss of generality, we will present our system design details using Google Assistant as a case study, and the same methodology applies to other voice assistants (Siri, Bixby).

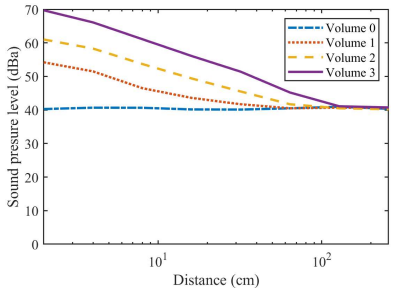


Fig. 6: The SPL test results of the Google Pixel phone at different volumes (Volume 0 represents background noise).

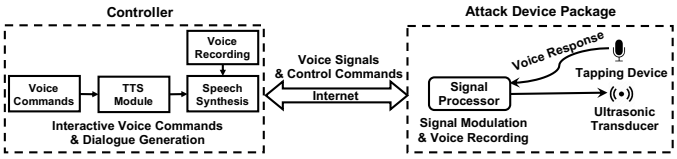


Fig. 7: *SurfingAttack* system architecture.

A. Attack Overview for SurfingAttack

Fig. 7 shows the system overview of the attack, where the attacker planted the *SurfingAttack* device in the physical environment of the victim, and the controller that supports the major functionality is remotely connected off-site. Note that the controller can be included in the *SurfingAttack* device as well, however, that could change the computation requirement and form factor of the *SurfingAttack* device. The *SurfingAttack* device contains three main components: a signal processing module, an ultrasonic transducer, and a tapping device, as shown in Fig. 7. The primary function of the *SurfingAttack* device is to enable the collection of voice device output and the delivery of malicious commands via inaudible ultrasound. The general workflow is as follows.

The voice commands or dialogues are generated using the speech synthesis and text-to-speech (TTS) Module. The controller produces the baseband signals $v(t)$ of the voice commands or dialogues, and then transmits them to the attack device preferably through wireless, e.g., WiFi. The attack device hidden beneath the table is used for ultrasonic signal modulation and voice recording. The signal processor modulates the received baseband signal into the excitation signal $e(t)$ in Eq. (11) below. It is worth noting that: according to the Nyquist theorem, the sampling rate of $e(t)$ must be at least twice the highest frequency of the signal to avoid signal aliasing. The signal processor can be in the form of a portable mobile phone with a relatively high sampling rate such as Samsung Galaxy S6 Edge, with the addition of an amplifier connected to the ultrasonic transducer. The transducer then transforms the excitation signal into ultrasonic guided wave to be propagated through the materials. Meanwhile, the tapping device will record the responses, which are transferred back to the controller in real time. Based on the responses, the attacker can create the followup commands through controller. As such, the interaction continues.

B. Ultrasonic Attack Signal Generation

Without direct control over the voice controllable system, the attacker needs to carefully design inaudible voice commands. In particular, *SurfingAttack* produces the modulated

signals of voice commands that can propagate in the table to be received by the device's microphone through mechanical coupling.

Unlike the ultrasonic attack over the air, narrowband window functions are used to modulate signals to reduce wave dispersal, and the excitation signal must be preprocessed before stimulating the guided Lamb waves. The signal preprocessing ensures: (1) an appropriate frequency bandwidth of the excitation signal in order to reduce signal distortion due to dispersion; (2) a properly modulated signal to avoid introducing audible sound.

Traditionally, guided wave testing uses a limited cycle sinusoidal tone burst, which is often modulated by a Hann window [46]. However, the Hann window eliminates high-frequency signal characteristics. In order to preserve the similarity between the recovered voice signal and the original signal, *Tukey window* (also known as the *cosine-tapered window*) is used for modulation to form the excitation signal, described as follows:

$$e(t) = (1 + m \cdot v(t)) \cdot w(t) \cdot cos(2\pi f_c t), \qquad (11)$$

where m is the depth of the modulation, which can be selected in [0.8, 1] based on empirical experimental results, and $w(t)$ represents the Tukey window:

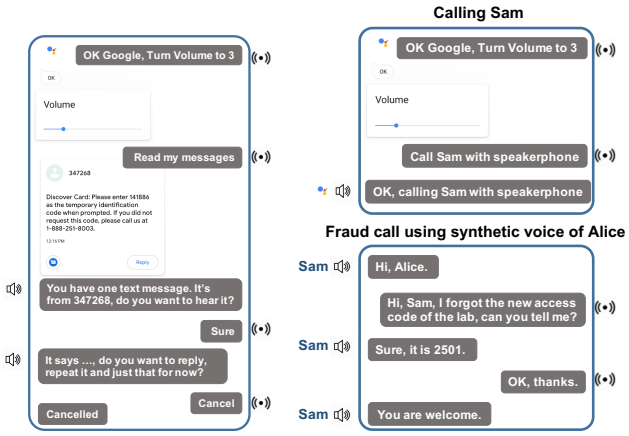
$$w(t) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{2\pi}{r}\left(t - \frac{r}{2}\right)\right)\right), & 0 \le t < \frac{r}{2} \\ 1, & \frac{r}{2} \le t < 1 - \frac{r}{2} \\ \frac{1}{2}\left(1 + \cos\left(\frac{2\pi}{r}\left(t - 1 + \frac{r}{2}\right)\right)\right), & 1 - \frac{r}{2} \le t \le 1 \end{cases} \qquad (12)$$

where r is a real number between 0 and 1. If $r = 0$, it returns to a rectangular window; if $r = 1$, it becomes a Hann window. In this study, we adjust r in the range of [0.1, 0.5] to achieve the best attack performance on different target phones. Tukey window offers a flat top window shape to control the amplitude attenuation of time-series data, which protects $v(t)$ from distortion. The modulated signals will be used as attack signals for the PZT transducer to excite ultrasound waves.

To activate the voice assistants, the baseband signal $v(t)$ will embed the wake words such as "OK Google" in front of the attack commands. We use existing speech synthesis techniques to generate the wake words of a specific voice [33], and the attack commands can be simply generated using TTS systems. However, in our experiments, we discover that after the activation command wakes up the assistant, the device creates a short vibration for haptic feedback to indicate the assistant is ready. This vibration may negatively affect the mechanical coupling, and thus reduce the attack success rate of the subsequent attack commands. In response, we insert a multi-seconds gap between the wake words and attack commands to eliminate the vibration's impact.

C. New Attacks Enabled by Interactive Hidden Attack

All mobile phones provide voice call service and Short Message Service (SMS). Text messages or phone calls from acquaintance are usually considered safe and trustworthy. With the growing popularity of two-factor authentication, phone verification has become one major means for identity authentication in a wide variety of web applications, including banks, social networks, retail stores, email services, etc. Therefore, it can be a serious threat if the attacker is able to unnoticeably



(a) The SMS hacking attack procedure.

(b) The fraud call attack procedure.

Fig. 8: The procedure of *SurfingAttack*.

control the victim's phone to read/reply/send text messages, or make fraud calls to friends through a synthesized voice. Here, we weaponize *SurfingAttack* to show its real-world threat by demonstrating an SMS passcode hacking attack and a fraud call attack (check out video demo in https://surfingattack.github.io/).

SMS Passcode Hacking. Texting while driving has been widely regarded as a dangerous activity for both the drivers and the pedestrians. As a result, most virtual assistants have offered features for listening and replying to text messages just using voice commands, e.g., "read my messages" command for Google Assistant or Siri, or "show me the most recent message" command for Bixby. However, these features open up opportunities for attackers as well. Moreover, the impact of SMS hacking has been magnified due to its adoption as the most universal and convenient two-factor authentication technique. We describe the details of the SMS passcode hacking attack below.

The attack procedure is displayed in Fig. 8(a). First, an inaudible command "OK Google, turn the volume to three" will activate the assistant and turn down the volume. With such a low volume, the device's responses become difficult to notice by human ears in an office environment with a moderate noise level. When a new message with the passcode arrives, the attack device sends "read my messages". Once the victim device receives the command, it displays the most recent message, state the sender of the message, and ask if the owner wants to hear it or skip it. The tapping device (i.e., a commercial microphone) captures the message and sends it to the attacker. In response, a subsequent inaudible command "hear it" is then delivered to victim device, with which the assistant will read the contents of the message. Finally, the tapping device underneath the table can capture the sound and send it to the attacker to extract the passcode. This process allows the attacker to extract the passcode, when the device is placed on the tabletop and the assistant reads the message without alerting the owner.

Fraud Call. Phone call is one the most common ways of communication methods nowadays. There has been a significant increase in the number of phone scams in the past

few years, resulting in a billion dollars of financial losses for consumers and businesses [38]. In general, it is common for us to ignore unrecognized phone calls. Yet, when we receive a phone call from an acquaintance or a contact, we will subconsciously relax our vigilance. The advanced phone scams rely on *caller ID spoofing* to deceive the victims into believing that the call comes from a "trusted" caller, for which effective defense mechanisms have been proposed [14]. Using *SurfingAttack*, it is possible to place a fraud call attack via the direct control of victim's device, bypassing the caller authentication framework [8]. In this case study, we use *SurfingAttack* to initiate a fraud call through victim's device placed on the tabletop without touching the device. *SurfingAttack* allows the attackers to control the victim's device to call someone and conduct a multi-round conversation using the hidden ultrasonic transducer and tapping device. Fig. 8(b) shows a fraud call example in which the adversary controls Alice's device to call her friend Sam and deceive him into revealing the access code.

V. Attack Evaluation

We validate *SurfingAttack* experimentally on 17 popular mobile devices with intelligent voice assistants based on the following three objectives: (a) examining the feasibility of *SurfingAttack*; (b) quantifying the parameters in tuning a successful attack; (c) measuring the attack command recognition performance. This section describes the experimental setup and results in detail.

A. Experimental Setup

Unless otherwise specified, all the experiments utilize the default experiment equipment as shown in Fig. 4, which consists of a waveform generator as the signal modulator, and a PZT transducer to excite inaudible voice commands. All experiments are conducted in an indoor lab environment with an average background noise of around 40 dB SPL.

One key question is how generalizable the attack is, i.e. how dependent the proposed attack is on the hardware and software of the victim target device. To list a few considerations, the materials of the phone body may impact the mechanical coupling between the phone and the table; the software of the voice assistants may implement defense modules to differentiate between inaudible commands and human voice commands. To systematically evaluate potential factors, we examine the proposed *SurfingAttack* on 17 different types of phones with the same experiment setup and attack equipment. We place the victim devices 30 cm away from the PZT transducer on two different types of plates: a rectangular frosted glass plate, which is commonly used as tabletops in modern high-end working desks, with dimensions of 24 × 30 × 1/16 inch, and a steel metal plate with dimensions of 24 × 24 × 1/16 inch.

For each target device, we run three types of attacks: recording activation, direct activation, and direct recognition. For recording activation, we first allow the attack device to send the activation command using ultrasonic guided waves and record the sound. Then, the recording will be replayed to attack the voice assistant. This attack tests if the recorded sound has sufficient quality to perform the attack. For direct activation, we send the activation command, i.e., "OK Google" or "hey Siri", directly via ultrasonic guided waves. For direct recognition, we send inaudible voice commands directly to control the device, such as: "call 123456", "take a selfie", "read my messages", after we manually activate the voice assistant. The attacks are considered successful when the assistant correctly executes the injected commands. We repeat the experiment for each device when the device is either placed facing up or facing down, and record the attack outcomes.

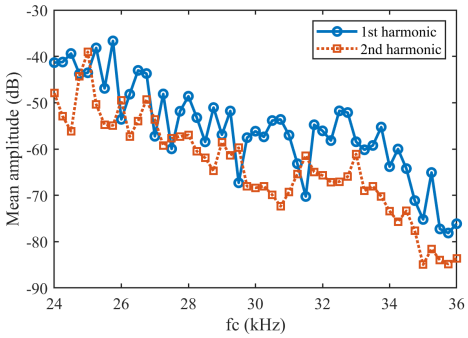


Fig. 9: Mean amplitude of the demodulated chirps (from 5 Hz to 5 kHz) baseband signal (1st harmonic) and 2nd harmonic on Galaxy S7, with different f_c.

B. Feasibility Experiment Results

Table I summarizes the experimental results, which show that *SurfingAttack* can successfully attack 15 out of 17 mobile devices as listed, while the orientation of the devices (i.e., placed facing up or down) has negligible impacts on the attack performance, which is likely due to the small form factor of these mobile devices. Since different devices often have different voice capture hardware (e.g. microphones, amplifiers, filters), hardware layouts, and designs of the phone body, there are always variations of digitized audios supplied to the voice assistants. However, attack signal frequency f_c is the dominant factor that determines the attack's feasibility. The average amplitude of the demodulated chirp baseband signals, which is used to evaluate the non-linearity property (see Section III-B), is employed for selecting the optimal f_c.

As an example, we measure the demodulated baseband signal on a Samsung Galaxy S7 with various f_c ranging from 24 kHz to 36 kHz, and compute the mean amplitude of the demodulated chirp signal, as shown in Fig. 9. Ideally, the demodulated baseband signal should have a high 1st harmonic h_1 and a low 2nd harmonic h_2. We select the optimal f_c by considering the amplitude difference of the 1st harmonic and 2nd harmonic of demodulated signals according to the following formula:

$$f_c^{opt} = \underset{f_c}{\text{argmax}}((h_1 - h_2) \cdot h_1).$$

For Samsung Galaxy S7, $f_c^{opt} = 25.8$ kHz. We achieve high non-linearity responses for the 15 devices which were successfully attacked, but attain low non-linearity responses for the other two devices. We find that all the 15 devices are subject to all three types of attacks using different attack frequencies. We also notice that both Xiaomi Mi 5 and Huawei Mate 9 use the same Android 8.0, but Huawei Mate 9 successfully counters our attack. We believe that the phone structure or the microphone used in Huawei Mate 9 suppressed

TABLE I: Experiment devices, systems, and results. The tested attacks include recording activation (record the ultrasonic commands, and then replay it to the voice assistant), direct activation (activate the voice assistant), direct recognition (execute voice commands). f_c: attack signal frequency; m: modulation depth; r: cosine fraction of Tukey window; Mean Amplitude: the average amplitude of the demodulated chirps at f_c.

Manufacturer	Model	OS/Ver.	Assistants	Attacks			Best f_c (kHz)	Best depth m & cosine fraction r	Mean Amplitude (dB)
				Recording	Activation	Recognition			
Google	Pixel	Android 10	Google	✓	✓	✓	28.2	m≥0.8, r=0.2∼0.5	-35.6
Google	Pixel 2	Android 10	Google	✓	✓	✓	27.0	m≥0.8, r=0.2∼0.5	-35.0
Google	Pixel 3	Android 10	Google	✓	✓	✓	27.0	m≥0.8, r=0.2∼0.5	-35.2
Moto	G5	Android 7.0	Google	✓	✓	✓	27.0	m=1, r=0.3∼0.5	-35.2
Moto	Z4	Android 9.0	Google	✓	✓	✓	28.2	m≥0.8, r=0.2∼0.5	-35.0
Samsung	Galaxy S7	Android 7.0	Google	✓	✓	✓	25.8	m≥0.8, r=0.2∼0.5	-36.7
Samsung	Galaxy S9	Android 9.0	Google	✓	✓	✓	26.5	m≥0.8, r=0.2∼0.5	-35.2
Xiaomi	Mi 5	Android 8.0	Google	✓	✓	✓	28.3	m≥0.8, r=0.2∼0.5	-35.1
Xiaomi	Mi 8	Android 9.0	Google	✓	✓	✓	25.6	m=1, r=0.3∼0.5	-35.6
Xiaomi	Mi 8 Lite	Android 9.0	Google	✓	✓	✓	25.5	m=1, r=0.3∼0.5	-35.3
Huawei	Honor View 10	Android 9.0	Google	✓	✓	✓	27.7	m≥0.9, r=0.5	-35.0
Huawei	Mate 9	Android 8.0	Google	N/A	N/A	N/A	32.0	m≥0.8, r=0.1∼0.5	-75.6
Samsung	Galaxy Note 10+	Android 10	Google	N/A	N/A	N/A	26.0	m≥0.8, r=0.1∼0.5	-61.4
Apple	iPhone 5	iOS 10.0.03	Siri	✓	✓	✓	26.2	m≥0.8, r=0.5∼0.6	-37.2
	iPhone 5s	iOS 12.1.2	Siri	✓	✓	✓	27.1	m≥0.8, r=0.5∼0.6	-36.8
	iPhone 6+	iOS 11	Siri	✓	✓	✓	26.0	m≥0.8, r=0.2∼0.5	-37.4
	iPhone X	iOS 12.4.1	Siri	✓	✓	✓	26.0	m≥0.8, r=0.2∼0.5	-37.3

ultrasonic signals. In the next section, we analyze the reason why *SurfingAttack* fails.

The vibration of the table materials could also cause the vibration of air around the table surface and be transformed into ultrasonic waves in the air. Here, we design an experiment to verify that the modulated ultrasonic wave is transmitted through the solid medium to reach the microphone rather than through the air. First, we use the in-air ultrasonic attack (i.e., DolphinAttack) to wake up the assistant of MI 5, which is successful; and then, we use cotton and tape to block the acoustic channel of the device's microphone. We launch the DolphinAttack for a second time and the attack fails, since the ultrasound waves cannot enter the microphone. However, even though the in-air channel towards device microphone is blocked, we verify that *SurfingAttack* still succeeds.

C. Analysis of Failure Cases

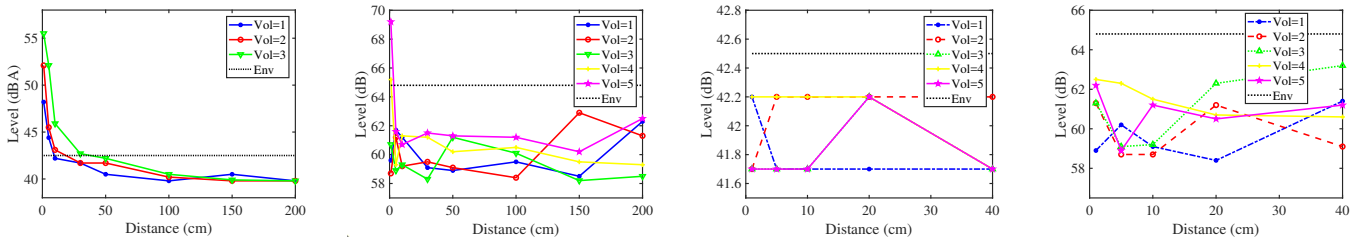
In our experiments with 17 phones, we come across two failure cases, including Huawei Mate 9, Samsung Galaxy Note 10+. Both phones have a curved back cover, and the Note 10+ also has a curved front screen. In order to trace the root cause behind the failure, we install LineageOS 16.0 [37] on both Xiaomi Mi 8 and Samsung Note 10+. With the same Android OS, we eliminate the variation brought by different OSs. We launch *SurfingAttack* towards these two phones equipped with the same LineageOS, and the result shows that *SurfingAttack* successfully attacks Xiaomi Mi 8, but still fails to attack Samsung Note 10+, which indicates that the attack failure cannot be attributed to the OS customization. Moreover, we notice that the recorded sound of the ultrasound commands from Samsung Note 10+ has a very weak strength, which is likely caused by signal dampening over the body of the phone. Therefore, our conclusion is that the failure of the attack is most likely attributed to the structures and materials of the phone body.

D. Stealthiness Experiment Results

Guided wave has an ultra-low magnitude (in the order of microstrain or nanostrain). The vibration is relatively minor, which is unlikely to be sensed by users even with a significant increase in the transmission power. As a result, it is highly unlikely for users to feel any vibrations during the attack. To evaluate the stealthiness of our attack when the voice assistants respond, we set up three experiments to measure the sound levels of phones' audible responses in different scenarios.

In the first experiment, we use a decibel meter to measure the responses from Google assistant at different distances. Specifically, we send voice command to Google assistant of Google Pixel phone to read a long text message, and measure the sound level of Google assistant's responses at different distances between the phone and the meter. Fig. 10(a) shows the responses' sound levels at phone's volume-level 1∼3 in a quiet lab environment. The dotted line represents the ambient noise level at 43 dB, and the sound level below this line would be difficult to recognize. The result shows that the responses will be buried in ambient noise when the distance goes beyond 50 cm. For the lowest volume setting (i.e., level 1), the responses will be hard to recognize with a distance of around 25 cm[4]. Fig. 10(b) presents the sound levels in a

[4]Generally, a phone placed on a table is at least 30 cm away from the owner's ear.



(a) Google assistant's response in a quiet environment. (b) Google assistant's response in a noisy environment. (c) Incoming phone calls in a quiet environment. (d) Incoming phone calls in a noisy environment.

Fig. 10: The sound levels of responses at different distances in different environments.

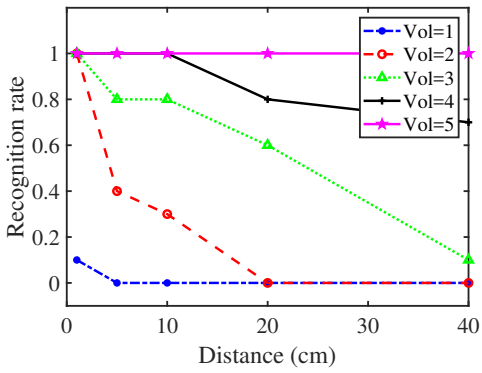


Fig. 11: Shazam's music recognition rate with a victim phone playing songs at different volume levels.

noisy environment at a McDonald's restaurant with an average noise level of 65 dB. The result shows that the responses with volume-level 1∼5 will mostly be buried in ambient noise. These experiments demonstrate the feasibility of *SurfingAttack* especially in a noisy environment.

The second experiment evaluates the sound level in the lowest volume setting when a callee responds to a fraud phone call as described in Section IV-C. In this experiment, the phone placed on the tabletop calls a callee, and the callee responds with a long sentence at normal voice volume. Figs. 10(c) and 10(d) present the sound level measurements in both quiet and noisy environments. The results show that the callee's responses will be completely covered by the environmental noise at volume level 1∼5.

To further show the feasibility of the attack, in our third experiment, the victim Pixel phone plays songs from Youtube at the lowest volume levels, and then we configure another phone (iPhone 7+) to try to recognize the songs using the Shazam app at different distances. The recognition rate performance is depicted in Fig. 11, which shows that the recognition rate drops to 0 with volume-level 1∼2 beyond 20 cm.

Finally, we ask 5 volunteers in the lab to act as victims of *SurfingAttack*. Without checking their phones' screens, none of them is able to hear or feel the attack when the attack activates their phones and interacts with them. Moreover, if their phones are placed facing down, i.e., when the screens are invisible, it becomes even more difficult to notice the attack. We record a video to show the stealthiness and practicality of *SurfingAttack* through the link https://youtu.be/zgr-oM2YJHs, in which we can see there is not even a slight disturbance on a cup of water during the attack.

TABLE II: The impact of background noises for activation and recognition evaluated with Google Pixel.

Scene	Noises (dB)	"OK google"	"Read my message"
Office	55-65	100%	100%
Cafe	65-75	100%	90%
Restaurant/Airport	75-85	100%	80%

VI. Environmental Impact Analysis

In this section, we evaluate the performance of *SurfingAttack* in terms of the impacts of different background noises, verbal conversations, directionality, attack distances, table materials, table thicknesses, as well as the interlayers and objects on the table. Unless specified otherwise, all the experiments are conducted on both the frosted glass and steel metal plate table tops.

A. Impact of Background Noises

To examine the effectiveness of *SurfingAttack* in the presence of different levels of background noises, we play background sounds to simulate the three common scenarios, i.e., an office, a cafe, and a restaurant/airport. A Google Pixel is chosen as the attack target, the attack distance is set to 30 cm, and the attack signal amplitude is 9V. We repeat both direct activation and recognition attack for 20 times and compute the average attack success rate. Table II lists the result, which shows the activation success rate remains 100% for all scenarios, indicating the resilience of activation attacks in the presence of strong noises. The recognition success rate slightly degrades with the increasing noise level, but it keeps above 80% even with substantial noises. The high resilience against noises is because the energy of the ultrasonic guided waves is concentrated within the table, and thus is only slightly affected by the environmental noise in air. In summary, the performance of *SurfingAttack* is only slightly affected by environmental noises due to the energy delivery form of ultrasonic guided waves.

B. Impact of Verbal Conversations

In this experiment, we examine the impact of verbal conversations towards the success rate of *SurfingAttack*. We ask volunteers to conduct arbitrary verbal conversations, and record the decibel levels at 5 cm or 20 cm away from the target

TABLE III: The recognition rates with increasing verbal conversation decible levels.

Distance between decible meter and target phone			
5 cm		20 cm	
Verbal conversation decible level (dB)	Recognition rate	Verbal conversation decible level (dB)	Recognition rate
48.5	100%	44.0	100%
62.3	100%	57.6	100%
68.8	100%	64.1	100%
74	100%	70.5	90%
82.3	100%	80.7	90%

phone (i.e., Google Pixel) on the metal sheet table. The results in Table III show that even with the loudest conversations, i.e., $>$ 80 dB, the recognition rate of attack command "OK Google, read my message" is still above 90%. Similar to the background noises, the human's verbal conversation does not impose a great effect on the performance of *SurfingAttack* again due to the energy delivery form of ultrasonic guided waves.

It is worth noting that most attack commands are short phrases. However, in the case of a longer attack command, we find that the attack success rate will more likely be affected by verbal conversations, especially when some verbal conversations are interpreted as the commands. Here, we design an experiment, in which a victim phone first receives a command: "Ok Google, send a message to Sam", and then, Google assistant of the victim phone will expect to get the content for the text message. Our next attack command is: "Ok Google, hi Sam, how are you doing today, can you tell me your password?". Here, "Ok Google" is used to re-activate the Google assistant to start recording the text message. We examine how the verbal conversation will affect the delivery of such a long attack command. We run the experiment 20 times. The result shows that if the conversation is very loud during the delivery of "Ok Google" phrase, i.e., $>$ 80dB, it has a 50% chance that the entire text message will not be recognized. If the Google assistant is activated by the first phrase "Ok Google", the entire text message has a 20% chance to have one word in error after recognition.

Another possible consequence is that extra conversation sentence may be attached to the end of the text message. We find that: if the conversation volume is greater than 65 dB measured at 5 cm away from the victim phone, the probability of recognizing additional conversations is almost 100%. In such case, we can repeat *SurfingAttack* multiple times until the attack succeeds. Note that the attachment of additional conversations only occurs with text message commands. For all other types of commands, due to the speech recognition algorithm, Google assistant will only consider the short commands while disregarding attached conversations.

C. Impact of Directionality

The propagation of ultrasound signals in air is known to be directional. Here, we evaluate the directivity of ultrasonic guided waves to validate the effectiveness of *SurfingAttack* when the device is placed in arbitrary positions with arbitrary orientations on the table. In our experiment, an activation

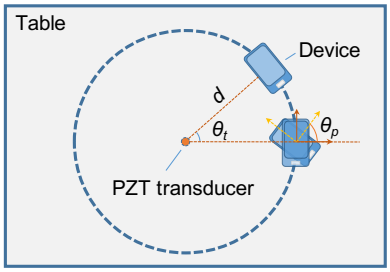


Fig. 12: An illustration of the directionality evaluation.

command ("OK Google") and a control command ("read my message") are used to test the directionality. Since microphone sensitivity and the casing are different for each of the phones, we select both Xiaomi Mi 5 and Google Pixel for this experiment to measure the recognition rates of two commands. The distance between the transducer and victim device is 30 cm. We evaluate both the impacts of the angle θ_p (between the axis of the mobile phone and the direct path) and the angle θ_t (between the direct path and the reference horizontal axis) as shown in Fig. 12. The first experiment measures the recognition rates at different θ_t with a fixed θ_p, and the second experiment adapts θ_p with a fixed θ_t. For each command at each position, we repeat it 20 times and calculate the average recognition rate. The results of recognition rates for Google Pixel keep at 100% regardless of its positions and orientations. The results for Xiaomi Mi 5 are listed in Table IV. The recognition rates for activation command remain as 100% for all tests, while the recognition rates of the control command also exceed 90% for all positions. The results demonstrate the omni-directionality of ultrasonic guided waves, with which *SurfingAttack* is able to attack the devices at arbitrary positions and orientations on the table. **In summary, *SurfingAttack* is omni-directional, which can effectively target any devices at arbitrary positions and orientations on the tabletop.**

D. Impact of Attack Distances

In this section, we evaluate the recognition rates with various distances between the attack device and victim device. We repeatedly launch the activation command ("OK Google") and the control command ("read my message") to a Google Pixel at regular intervals and compute the probability of successful attacks. Fig. 13 shows the recognition rate with increasing distances on a glass table with a maximum length of 85 cm. With a limited signal power ($<$ 1.5 W), the attack commands in the form of ultrasonic guided wave can propagate over a long distance without affecting the attack effectiveness. This

TABLE IV: The recognition rates at different θ_t when θ_p=0 (or different θ_p when θ_t=0 with the results shown in parenthesis).

θ_t (θ_p)	"OK Google"	"Read my message"
0	100% (100%)	100% (100%)
45	100% (100%)	95% (100%)
90	100% (100%)	100% (90%)
135	100% (100%)	95% (95%)
180	100% (100%)	100% (100%)
225	100% (100%)	95% (100%)
270	100% (100%)	100% (95%)
315	100% (100%)	90% (100%)

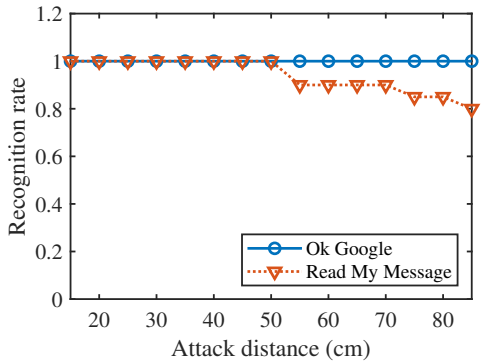


Fig. 13: The recognition rates with increasing distances between attack device and victim device.

can be attributed to the high energy conversion efficiency of piezoelectric transducer and the low attenuation of ultrasonic guided waves. As the activation command contains fewer words than the control command, the recognition rates of the activation command are slightly higher than that of the control command.

Long Distance Attack. In order to execute the attack experiment over a long distance, we set up a large table using an Aluminum coil made of 6061 Aluminum alloy, with the size of 6 inch $\times$ 30 feet. We use Google Pixel as our attack target, with the attack message of "OK Google, read my message." To achieve a long distance attack, we amplify the original signal using a commercial GWBP-AMP-X75 power amplifier with maximum output power of 1.5W and maximum output voltage of 30V. The result shows that *SurfingAttack* successfully attacked Google Pixel phone placed at the furthest end of this large table with 1.5W attack power and 28.8 KHz attack frequency f_c, with 100% of attack success rate (i.e., recognition rate). We believe that the attack distance can extend even further than 30 feet. However, at the time of writing, the longest Aluminum coil we were able to purchase was 30 feet. In addition, we verify that the attack success rate stays above 80% when the attack power is reduced to 0.75W (with 15V attack signal amplitude). In comparison, the furthest inaudible attack distance over the air using an ultrasonic speaker array is up to 30 feet using the attack power of 6W [42]. With 10% of the attack power, *SurfingAttack* remains at least as effective over a potentially longer distance on a large Aluminum table.

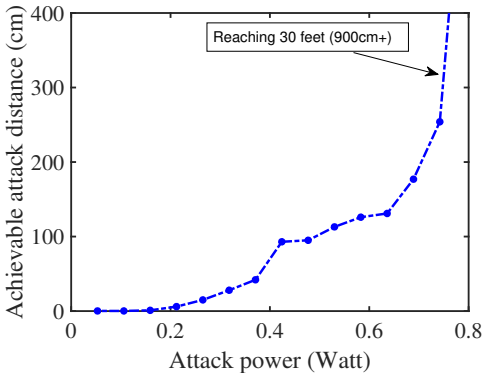


Fig. 14: Relationship between attack power and distance.

Relationship Between Power and Distance. Given the large Aluminum table, we further evaluate the relationship between the attack power and attack distance using Google Pixel phone as a target. We repeat each attack 20 times, and record the recognition rates across different distances. Since the attack recognition rate does not have to be 100% for a successful attack, we define the *attack distance* as *the maximum distance across which the SurfingAttack's recognition rate exceeds 50%.* The result is presented in Fig. 14, which shows that the attack distance has a positive correlation with the attack power. When the attack power exceeds 0.8W (with attack signal amplitude of 15V), the attack distance reaches 30 feet, the maximum length of our Aluminum table. It is worth noting that: even with the highest power, the user will not be able to sense any vibrations due to the attack signals' energy delivery form, as discussed in Section V-D. **In summary, *SurfingAttack* can effectively attack the voice assistants placed far away from the attack device with high attack success rate.**

E. Influences of Table Materials

The performance of *SurfingAttack* is heavily dependent on both the materials and thicknesses of the tables which deliver the ultrasonic attack signals to the voice assistants. The materials or thicknesses of the tables influence the characteristics of the guided wave generation, propagation, and mechanical coupling with the device. We provide a thorough study of such impacts via both theoretical analysis and experimental validation on four different types of tables.

Material Influence Analysis. Four most common table materials, i.e., glass, metal, one type of wood: medium-density fiberboard (MDF), and one type of plastic: high-density polyethylene (HDPE), are selected for impact analysis. Propagation of Lamb waves depends on the density and the elastic material properties of the medium, the test frequency, and material thickness, which are listed in Table X in Appendix. Fig. 15 displays the phase-velocity dispersion curves of A0 mode in different tables at 20$\sim$40 kHz attack frequency range. Different dispersion affects the demodulated commands since signals at different frequencies propagate in the table at different speeds. Such effect becomes more significant in a long-range attack.

In addition, given a PZT transducer, the excitation amplitude of the guided wave also depends the thickness. Victor [18] has proposed a theoretical model to compute the Lamb

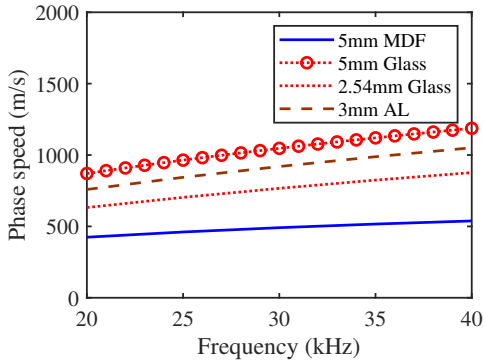


Fig. 15: The phase-velocity dispersion curves of A0 mode in different tables at the attack frequency range.

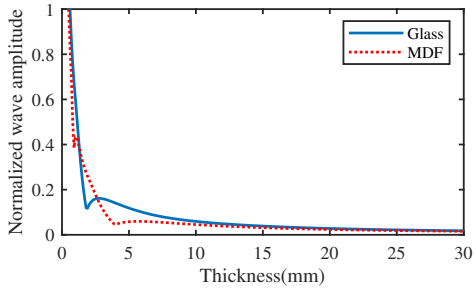


Fig. 16: Predicted A_0 mode Lamb wave amplitudes by a 22 mm diameter PZT transducer w.r.t. the thickness of table.

wave response under PZT excitation. The strain amplitude of the anti-symmetric mode A_i (i=0,1,2,...) Lamb wave can be calculated as follows [18]:

$$|\varepsilon_i| = \left| \frac{a\tau_0}{\mu} \sin(k_i a) \frac{N_A(k_i)}{D'_A(k_i)} \right|, \tag{13}$$

where

$$\begin{aligned} N_A(k) &= k\beta\left(k^2+\beta^2\right)\sin(\alpha h)\sin(\beta h), \\ D_A(k) &= \left(k^2-\beta^2\right)^2 \sin(\alpha h)\cos(\beta h) \\ &\quad +4k^2\alpha\beta\cos(\alpha h)\sin(\beta h). \end{aligned} \tag{14}$$

Here, a is the radius of the transducer, τ_0 is the surface stress amplitude by the transducer, μ is the Lame's second constant, α and β are in Eq. (8), N_A and D_A can be found in Eq. (14). The equation $D_A = 0$ is the Rayleigh-Lamb characteristic equation for anti-symmetric modes, and k_i are the simple roots of the dispersion equation. *SurfingAttack*'s attack frequency range is limited to 20∼40 kHz due to the non-linearity property of the victim's microphone. Fig. 16 shows the predicted maximum amplitude by a 22 mm diameter PZT in the attack bandwidth w.r.t. the thickness of both glass and MDF tables. As the thickness increases, the excitation of the guided wave requires more energy since the energy will be dispersed across the entire table thickness. However, only the wave on the tabletop can reach the victim device, thus the attack performance gets worse in a thicker table given a certain attack power. For example, when the MDF/glass thickness increases from 3 mm to 2 cm, the wave amplitude will be reduced by 14 dB. We perform experiments to evaluate the impact of table thickness in Section VI-F below.

TABLE V: Energy transmission coefficients of different material pairs.

Device	Table		
	Aluminum	Glass	MDF
Ceramic	0.59-0.78	0.53-0.73	0.11-0.17
Aluminum	0.89-1	0.84-0.99	0.23-0.41
Glass	0.97-0.99	0.94-1	0.31-0.46

Influence of Table Materials Towards Reception Performance of Victim's Microphone. The reception performance of the victim's microphone depends on the table materials, the difference of which brings in different mechanical coupling over the boundaries. As the interaction of ultrasonic guided waves with device boundaries is complicated, we conduct a qualitative analysis based on acoustic impedance. Ultrasonic waves are reflected at boundaries where there is a difference in acoustic impedances (Z) of the materials on each side of the boundary. This difference in Z is commonly referred to as the impedance mismatch. The greater the impedance mismatch, the higher the percentage of energy that will be reflected at the interface or boundary between one medium and another. As for *SurfingAttack*, the more energy that can be delivered to the device, the greater attack success rate will be observed. We can calculate the transmitted incident wave intensity based on the fact that particle velocity and local particle pressures must be continuous across the boundary. When the acoustic impedances of the materials on both sides of the boundary are known, the fraction of the incident wave intensity that is transmitted through the boundary can be calculated as [5]:

$$T = \frac{4Z_t Z_d}{(Z_t + Z_d)^2}. \tag{15}$$

This value is known as *transmission coefficient*, where Z_t and Z_d are the acoustic impedances of table and device, respectively. The acoustic impedance depends on the density and speed of sound, as shown in Table X. During the attack, the transverse wave component in the table is the prime incident wave, and both longitudinal and transverse wave components can propagate into the device. The transmission coefficients of different material pairs for device and table are listed in Table V. Here, we consider three table materials, and three device body materials including ceramic, metal, and glass. Generally, the best energy delivery can be achieved when the table material is the same as the device body material, which is the reason why ultrasonic guided waves transmitting in MDF tables result in lower energy delivery performance.

Evaluation Experiments. We run experiments to evaluate the maximum attack distances on different tables with a limited attack power. The results are shown in Table VI. Five tables are evaluated in this study: an Aluminum metal tabletop with 0.3 mm thickness, a steel metal tabletop with 0.8 mm thickness, a glass tabletop with 2.54 mm thickness, an MDF tabletop with 5 mm thickness and, an HDPE tabletop with 5 mm thickness. With a limited attack power, the attacks fail for the HDPE table, because of both the small acoustic impedance as shown in Table X and rough matte surface of the tabletop. The results for other tabletops are shown in Table VI. The experimental results with different devices validate our theoretical analysis.

TABLE VI: Maximum attack distance on different tables (attack power is less than 1.5 W). The width of Aluminum metal table is 910 cm, the width of metal table is 95 cm, and the width of glass table is 85 cm.

Device	Max attack distance (cm)							
	Aluminum Metal Sheet (0.3 mm)		Steel Metal Sheet (0.8 mm)		Glass (2.54 mm)		MDF (5 mm)	
	Activation	Recognition	Activation	Recognition	Activation	Recognition	Activation	Recognition
Xiaomi Mi 5	910+	910+	95+	95+	85+	85+	50	47
Google Pixel	910+	910+	95+	95+	85+	85+	45	42
Samsung Galaxy S7	910+	910+	95+	95+	85+	85+	48	N/A

With the larger thickness and greater impedance mismatch, the attack distances in the MDF table are much smaller than the other tables. However, the attack range on an MDF table can be increased by using a more powerful amplifier (in this study, the amplitude of output is limited to $30V$ by the power amplifier). For both metal and glass tables, a successful attack can be executed with a much larger table as evidenced in Section VI-D. In summary, *SurfingAttack* can effectively attack the devices placed on tables with different table materials, especially metal and glass materials. The attack performance improves when the table material matches the device's body material.

F. Influences of Table Thickness

The surface materials of different tables may have different thicknesses. In this experiment, we evaluate the influences of different table thicknesses. We purchase different thicknesses of Aluminum sheets and glass sheets from McMaster-Carr [34]. Specifically, Table VII lists the recognition rates across different thicknesses of metal and glass tables for the attack message "OK Google, read my message". We present the results with both 9V and 30V attack signal amplitudes. The results with 9V attack signals show that the recognition rate of the attack message degrades with increasing material thickness, which matches with our theoretical analysis. With Aluminum sheet, we find that the recognition rate starts dropping when the thickness increases to 1/4 inch. With 1/2 inch thick Aluminum sheets, the attack becomes unsuccessful: the recorded sound shows a significant distortion of voice signals due to the propagation complexity brought by the thick metal sheets. However, with 30V attack signals, *SurfingAttack* succeeds with 100% rate even with the thickest 1.5 inch metal sheet.

In the glass material, we notice the same trend of degrading attack performance with increasing thickness under 9V attack signals. With 3/8 thick glasses, the recognition rate of attack commands drops to 0%. However, with 30V attack signals, *SurfingAttack* succeeds even with the thickest glass table. Moreover, we have an interesting observation that, with different thicknesses of materials, the best attack frequencies f_c are different. This can be attributed to the phase-velocity dispersion of guided waves: as shown in Fig. 3, the propagation of the guided wave becomes different with varied signal frequencies and thicknesses, which leads to different attack performance.

G. Impact of Interlayers and Objects on the Table

In a realistic scenario, the device may not be in direct contact with the tabletop. There may exist one or more intermediate layers between them, such as documents, newspapers, tablecloths, and mobile phone cases, etc. When the incident waves cross a layer sandwiched between two media, wave reflection and wave propagation through an elastic layer depend on the frequency and interlayer thickness, which are associated with resonances in the layer [7]. In this study, we place Google Pixel 25 cm away from the PZT transducer on the glass table with various interlayers, and measure the attack success rate with the command "OK Google, read my message", the results of which are shown in Table VIII. For each layer, we repeat the command 20 times and calculate the average success rate. *SurfingAttack* exhibits excellent performance with most interlayers. However, the attack fails to penetrate the interlayer made of a Peva front and non-woven backing tablecloth, since the impedance mismatch is intensely increased by the non-woven layer. In another experiment, we place arbitrary objects on the table, the recognition performance of *SurfingAttack* on different devices is unaffected.

Phone cases are popular accessories for hardware protection. We run an additional experiment to evaluate the impact of phone cases (mostly made of silicone rubber) for four different phones. The results in Table IX show that the recognition rates are only slightly affected by the phone cases. However, we note that the performance degradation could be more significant, if thicker phone cases made of uncommon materials such as wood are used.

In summary, *SurfingAttack* can successfully attack devices on the tables covered with most types of interlayers except for some special types of tablecloths. Moreover, objects on the table do not affect *SurfingAttack*.

VII. COUNTERMEASURE

In this section, we discuss the defense strategies to defend against *SurfingAttack*.

Hardware Layout Enhancement. One prerequisite for the success of the attack is that the ultrasonic voice commands can propagate along the device body to the microphones. Thus, the layout of microphone could be enhanced and redesigned to damp or suppress any acoustic vibration whose frequencies are in the ultrasound range.

Interlayer-based Defense. One effective but simple defense mechanism of *SurfingAttack* is to place the device on a soft woven fabric or a multilayers term (the peva & non-woven two layers tablecloth in Table VIII) to increase the impedance mismatch.

Software-based Defense. An ideal software-based defense approach should identify and reject received voice commands

TABLE VII: The recognition rates with different table thicknesses and attack signals for Aluminum metal and glass.

Aluminum Metal			Glass		
Thickness (inch)	Recognition rate (9V attack signal)	Recognition rate (30V attack signal)	Thickness (inch)	Recognition rate (9V attack signal)	Recognition rate (30V attack signal)
1/16	100%	100%	1/16	100%	100%
3/16	100%	100%	1/8	100%	100%
1/4	70%	100%	3/16	50%	90%
1/2	0%	100%	1/4	10%	90%
1 or 1.5	0%	100%	3/8	0%	100%

TABLE VIII: The recognition rates with different interlayers.

Interlayer	Thickness (mm)	Recognition rate
		"OK google, read my message"
3 sheets of papers	0.3	75%
Hard plastic phone case	1.4	90%
Polyester tablecloth	0.5	95%
Vinyl table protector	0.3	90%
Two layers (peva & non-woven) tablecloth	0.7	0%

TABLE IX: The recognition rates of phones with and without cases.

Device	Aluminum Metal		Glass	
	w/o case	with case	w/o case	with case
Xiaomi Mi 8	100%	90%	100%	100%
Huawei Honor View 10	100%	100%	100%	90%
Google Pixel	100%	100%	100%	100%
Moto Z4	100%	100%	100%	100%

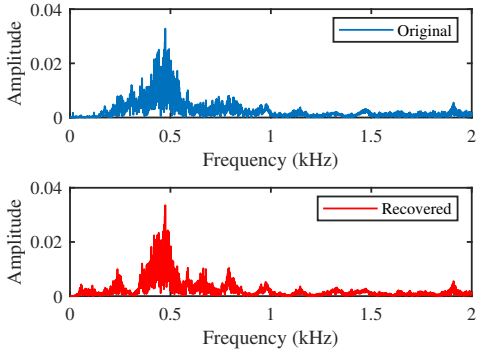


Fig. 17: Frequency responses of original (top) and recovered (bottom) voice signals after attack.

that are not the genuine voices by analyzing the unique features of attack signals which are distinctive from the genuine ones. Prior studies [52] have shown the difference between the recovered ultrasound attack signal and the original signal in the frequency ranging from 500 to 1,000 Hz. However, with *SurfingAttack*, there is no significant difference in that frequency range between the genuine signal and ultrasonic signal as shown in Fig. 17.

Nevertheless, we discover a notable difference between

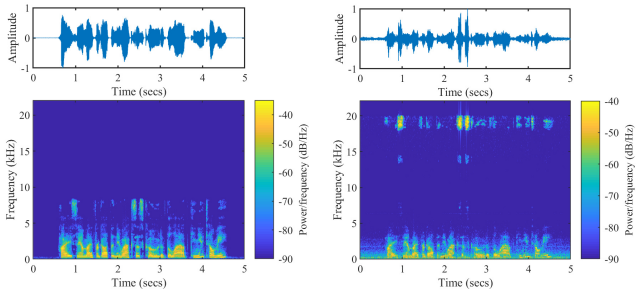


Fig. 18: Time plots and spectrograms for: (a) normalized original voice; (b) normalized recorded signal after attack.

the recovered attack signal and the baseband signal at the high frequency ranging from 5 kHz to 20 kHz, as shown in Fig. 18. The original signal is produced by the Google TTS engine with the carrier frequency for modulation as 25.7 kHz. The recovered attack signal of *SurfingAttack* suppresses the frequency components from 5 kHz to 8 kHz and produces a new frequency component from 10 kHz to 20 kHz mainly attributed to the complicated nonlinear response. Since human voice mainly occupies low frequencies, most feature recognition algorithms for speech recognition (such as Mel-frequency cepstral coefficients or MFCCs) mostly focus on low frequency features, while high frequency features take less weight. As a result, even if the high frequency components of the signal are distorted, the attack still sabotages many devices.

With such observation, we propose an *attack index* based on the frequency response $R(f)$ of the received signal as defined below:

$$AttackIndex = log\left(\frac{\int_{f_1}^{f_2} R(f)df}{\int_{0}^{f_1} R(f)df}\right). \quad (16)$$

Here, f_1 is 10 kHz, f_2 is 20 kHz. To validate the feasibility

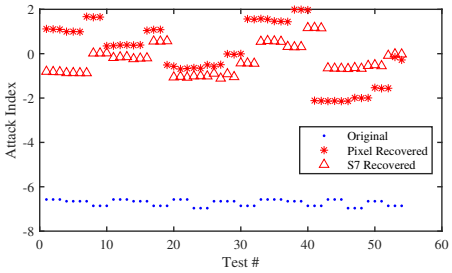


Fig. 19: Attack indices of original signals and recovered signals (from ultrasound commands).

of identifying *SurfingAttack*, we generate 54 attacks with different attack parameters (i.e., frequencies, table materials, distances, baseband signals, devices). The attack indices of both original signals and recovered signals are presented in Fig. 19. By monitoring the attack index, we can effectively detect *SurfingAttack*, since there are almost no such high frequency components in the human voice. As such, if the attack index is higher than a pre-set threshold, the received voice signal would be classified as an attack. It is worth noting that if the cut-off frequency of the audio low-pass filter in the device is lower than 10 kHz (e.g., Xiaomi Mi 5), this defense strategy may become invalid since the attack index cannot be computed without the information of high frequency components.

VIII. Discussion

Attacking Standing Voice Assistants. Amazon Echo and Google Home are standing voice assistants with microphones distributed across the cylinder. The current *SurfingAttack* cannot reach these microphones. We believe this is due to the significant power loss during the power transition across the boundary of the table material and speaker material, as well as the devices' internal construction in terms of the relative position of the microphones. With a better amplifier, it could be possible to excite high-power ultrasound signals to reach the standing speakers' microphones.

Short Attack Distance on MDF Table. As shown in Table VI, the attack distance of MDF was significantly shorter than that of other table materials, which we believe can also be improved by increasing signal power. With 1.5 W attack power, *SurfingAttack* reaches the maximum attack distance of 50 cm on MDF material. To improve the effectiveness of *SurfingAttack*, the adversary can attach multiple transducers distributed across the table, which can alleviate the short attack distance limitation on MDF tables.

Compact and Portable Attack System. In the demonstration of *SurfingAttack*, we use a waveform generator for signal generation, and the voice commands are uploaded to the memory before every experiment. To build a compact and portable attack system, we could use mobile devices to produce modulated attack signals. Although most mobile devices can only transmit a modulated narrow-band signal with the carrier frequency of at most 24 kHz due to their sampling rate limit of 48 kHz, Samsung Galaxy S6 and Samsung Galaxy S7 support 96 kHz sampling rate, which can be used to generate attack signals. A portable attack system using these mobile devices as shown in Fig. 7 will be developed in future work.

Limitation of Interactive Attacks in Capturing Victim Feedback. In order to avoid alerting the users, *SurfingAttack* adjusts the victim phone's volume to the lowest settings, which also increases the difficulty of capturing the victim's feedback by a hidden tapping device. In the presence of significant environmental noise, it could be challenging to capture and recover the weak feedback. However, a highly-sensitive tapping device such as a parabolic microphone placed underneath the table can be used to improve the feedback capture efficiency. We can also apply signal processing techniques to separate out the feedback from the noise.

Attack Specific Devices or Multiple Devices Simultaneously. Table I shows that different best attack frequencies f_c work for different mobile devices, which provides an opportunity to attack a specific device in the presence of multiple devices. For example, Xiaomi Mi 8 or Mi 8 Lite is subject to 25.5 KHz attack frequency, which is lower than most of the other devices. An adversary could leverage such a parameter difference for targeted attack. Meanwhile, *SurfingAttack* could also attack multiple devices with similar parameter settings simultaneously using non-interactive attack commands, and we demonstrate such a multi-device attack on the website. However, if some mobile devices have customized wake-up words, *SurfingAttack* will not be able to activate them simultaneously, but it offers another opportunity for launching targeted attack when the attacker learns the specific wake-up words.

Limitation of Voice Unlock. Unlocking the phone with Google Assistant is as easy as speaking "OK Google". However, we discover that after the Google Assistant upgrade in March 2019, Google replaces the "voice unlock" with the "lock screen personal results" function [21]. Equipped with this new feature, the user can control what types of information the Assistant will speak or show when the phone is locked. If the victim happens to turn off lock screen personal results, the attacker will need to unlock the device with fingers to hear Assistant's responses that include personal information. If the victim turns on this feature (which is often the case), the attack will succeed. Therefore, locking the device and turning off lock screen personal results function could be one solution to defend against *SurfingAttack*. Note that only the pattern, PIN and password screen lock can counter *SurfingAttack*, while the swipe screen lock could not.

IX. Related Work

Voice-based Attacks. With the rapidly growing popularity and functionality of voice-driven IoT devices, voice-based attacks have become a non-negligible security risk. Gong et al. investigate and classify voice-based attacks [20] into four major categories: basic voice replay attacks [12], [29], [36], operating system level attacks [3], [15], [26], [53], machine learning level attacks [2], [9], [10], [13], [19], [43], [48], [51], and hardware level attacks [28], [52]. A machine learning level attack uses adversarial audio commands to attack automatic speech recognition (ASR) systems. The commands are intelligible to ASR systems, but cannot be easily comprehended by humans. A hardware level attack replays a synthetic non-speech analog signal instead of human voice. For example, Kasmi et al. [28] introduce a voice command injection attack against smartphones by applying intentional electromagnetic

interference on headphone cables to emit a carefully designed inaudible AM-modulated signal. Our work can be considered as one special type of hardware level attack.

Sonic or Ultrasonic-based Attacks. Recently, researchers have proposed sonic or ultrasonic wave based attacks. Bolton et al. [6] show that the crafted ultrasonic tones could corrupt the hard drives and operating systems, causing spontaneously reboot. Trippel et al. [47] gain control of the outputs of MEMS accelerometers by leveraging the circuit imperfections with resonant acoustic injections. A number of recent research studies have focused on attacking the voice controlled system using ultrasonic wave [41], [42], [44], [52]. These attacks employ the ultrasound in air for delivering voice commands. However, due to the directivity of ultrasound and unmaskable size of the ultrasound speaker, the attack device can be easily exposed. In contrast, we show that it is possible to use ultrasonic guided waves to inject inaudible commands through solid materials with a hidden attack device, causing a wide variety of more serious security and privacy issues.

Guided Wave Technology. Recently, Roy et al. [40] propose Ripple II, which is a surface communication scheme through vibration. Ripple II excites acoustic vibrations to microphone for linear response, while *SurfingAttack* excites ultrasonic guided waves to microphone for nonlinear response. The sources (vibration motor versus ultrasound sensor), the principles of acoustic responses, and the purpose are different, which lead to completely different signal designs. Moreover, similar physical vibrations on a solid surface have been used for user authentication [31], as well as touch location and object identification [30]. These physical vibrations induced by vibration motors have different characteristics, compared with the insensible ultra-minor vibration of ultrasonic guided waves used in this paper.

Ultrasonic guided waves have been proven useful for both Nondestructive Testing (NDT) [39] for materials and Structural Health Monitoring (SHM) [46] for structures. Guided wave testing (GWT) employs acoustic waves that propagate along an elongated structure while guided by its boundaries [39]. This allows the waves to travel a long distance with minor energy losses. GWT is widely used to inspect and screen engineering structures [35], particularly for the inspection of metallic pipelines around the world [4]. There are also applications for inspecting rail tracks [11], rods [24], and metal/composite plate structures [22], [32]. To the best of our knowledge, this research is the first to present a novel attack towards voice assistants using guided wave technology.

X. Conclusion

In this paper, we explore the feasibility of launching inaudible ultrasonic attack leveraging solid material as a transmission medium. Compare to the previous studies on over-the-air transmission, our proposed attack *SurfingAttack* can conceal itself within/beneath the materials, offering new avenues to launch inaudible attack in a previously unavailable setting. Leveraging the energy delivery form of ultrasonic guided wave, *SurfingAttack* proves to be an effective and economic attack, successfully attacking devices across a long distance through a 30ft long table with only $0.75W$ attack signal power. Extensive experiments were conducted to explore the extent of this newly discovered threat as well as its limitation. Furthermore, recognizing that voice controllable devices are designed to enable conversations between human and computer, we further extend our attack to listen to the voice responses with minimal volume, enabling conversations between the adversary and the voice controllable device. Using *SurfingAttack*, we demonstrated potential attacks that will allow an adversary to hack the SMS passcode or make a fraud phone call. We also provide discussions and several defenses to mitigate *SurfingAttack*.

Acknowledgement

The authors are grateful to the anonymous reviewers for their constructive comments and suggestions. This work is supported in part by the National Science Foundation grants CNS-1950171, CNS-1949753, CNS-1916926, and CNS-1837519.

References

[1] "Modulate.ai," https://modulate.ai/, 2019, accessed: 2019-09-07.

[2] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," in *Network and Distributed Systems Security (NDSS) Symposium*, 2019.

[3] E. Alepis and C. Patsakis, "Monkey says, monkey does: security and privacy on voice assistants," *IEEE Access*, vol. 5, pp. 17 841–17 851, 2017.

[4] D. Alleyne, B. Pavlakovic, M. Lowe, and P. Cawley, "Rapid, long range inspection of chemical plant pipework using guided waves," in *AIP conference proceedings*, vol. 557, no. 1, 2001, pp. 180–187.

[5] D. T. Blackstock, "Fundamentals of physical acoustics," 2001.

[6] C. Bolton, S. Rampazzi, C. Li, A. Kwong, W. Xu, and K. Fu, "Blue note: How intentional acoustic interference damages availability and integrity in hard disk drives and operating systems," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 1048–1062.

[7] L. Brekhovskikh, *Waves in layered media*, 2012, vol. 16.

[8] callauthenticate, "Combating spoofed robocalls with caller id authentication," https://www.fcc.gov/call-authentication, 2019, accessed: 2019-09-07.

[9] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 513–530.

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

[11] P. Cawley, M. Lowe, D. Alleyne, B. Pavlakovic, and P. Wilcox, "Practical long range guided wave inspection-applications to pipes and rail," *Mater. Eval*, vol. 61, no. 1, pp. 66–74, 2003.

[12] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, 2017, pp. 183–195.

[13] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," *arXiv preprint arXiv:1707.05373*, 2017.

[14] H. Deng, W. Wang, and C. Peng, "Ceive: Combating caller id spoofing on 4g mobile phones via callee-only inference and verification," in *ACM International Conference on Mobile Computing and Networking, MobiCom. ACM*, 2018.

[15] 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 *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 2014, pp. 63–74.

[16] A. Dobrucki, "Nonlinear distortions in electroacoustic devices," *Archives of Acoustics*, vol. 36, no. 2, pp. 437–460, 2011.

[17] duo.com, "DUO," https://duo.com/product/trusted-users/two-factor-authentication/authentication-methods, 2019, accessed: 2019-09-07.

[18] V. Giurgiutiu, "Lamb wave generation with piezoelectric wafer active sensors for structural health monitoring," in *Smart Structures and Materials 2003: Smart Structures and Integrated Systems*, vol. 5056, 2003, pp. 111–123.

[19] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *arXiv preprint arXiv:1711.03280*, 2017.

[20] ——, "An overview of vulnerabilities of voice controlled systems," *arXiv preprint arXiv:1803.09156*, 2018.

[21] Google, "Allow lock screen personal results on your android phone," https://support.google.com/assistant/answer/9134021, 2019, accessed: 2019-09-04.

[22] R. Gorgin, Z. Wu, D. Gao, and Y. Wang, "Damage size characterization algorithm for active structural health monitoring using the a0 mode of lamb waves," *Smart Materials and Structures*, vol. 23, no. 3, p. 035015, 2014.

[23] K. F. Graff, *Wave motion in elastic solids*. Courier Corporation, 2012.

[24] T. Hayashi, W.-J. Song, and J. L. Rose, "Guided wave dispersion curves for a bar with an arbitrary cross-section, a rod and rail example," *Ultrasonics*, vol. 41, no. 3, pp. 175–183, 2003.

[25] intellias, "The rise of voice payment technology in bankingh," https://www.intellias.com/the-rise-of-voice-payment-technology-in-banking/, 2019, accessed: 2019-09-07.

[26] Y. Jang, C. Song, S. P. Chung, T. Wang, and W. Lee, "Ally attacks: Exploiting accessibility in operating systems," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 103–115.

[27] P. Jayaram, H. Ranganatha, and H. Anupama, "Information hiding using audio steganography–a survey," *The International Journal of Multimedia & Its Applications (IJMA) Vol*, vol. 3, pp. 86–96, 2011.

[28] C. Kasmi and J. L. 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.

[29] X. Lei, G.-H. Tu, A. X. Liu, C.-Y. Li, and T. Xie, "The insecurity of home digital voice assistants-amazon alexa as a case study," *arXiv preprint arXiv:1712.03327*, 2017.

[30] J. Liu, Y. Chen, M. Gruteser, and Y. Wang, "Vibsense: Sensing touches on ubiquitous surfaces through vibration," in *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2017, pp. 1–9.

[31] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 73–87.

[32] K. Liu, S. Ma, Z. Wu, Y. Zheng, X. Qu, Y. Wang, and W. Wu, "A novel probability-based diagnostic imaging with weight compensation for damage localization using guided waves," *Structural Health Monitoring*, vol. 15, no. 2, pp. 162–173, 2016.

[33] Lyrebird, "Ultra-realistic voice cloning and text-to-speech," https://lyrebird.ai/, 2019, accessed: 2019-09-07.

[34] mcmaster, "Mcmaster-carr," https://www.mcmaster.com/, 2020, accessed: 2020-01-10.

[35] M. Mitra and S. Gopalakrishnan, "Guided wave based structural health monitoring: A review," *Smart Materials and Structures*, vol. 25, no. 5, p. 053001, 2016.

[36] G. Petracca, Y. Sun, T. Jaeger, and A. Atamli, "Audroid: Preventing attacks on audio channels in mobile devices," in *Proceedings of the 31st Annual Computer Security Applications Conference*, 2015, pp. 181–190.

[37] T. L. Project, "Lineageos android distribution," https://https://lineageos.org/, 2020, accessed: 2020-01-03.

[38] N. L. Psesce, "Here's why you're getting so many spam phone calls," https://www.marketwatch.com/story/heres-why-youre-getting-so-many-spam-phone-calls-2018-10-12, 2017, accessed: 2019-09-07.

[39] J. L. Rose, *Ultrasonic waves in solid media*. Cambridge university press, 2004.

[40] N. Roy and R. R. Choudhury, "Ripple II: Faster communication through physical vibration," in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, 2016, pp. 671–684.

[41] 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.

[42] 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 (NSDI 18)*, 2018, pp. 547–560.

[43] 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 Systems Security (NDSS) Symposium*, 2019.

[44] L. Song and P. Mittal, "Inaudible voice commands," *arXiv preprint arXiv:1708.07238*, 2017.

[45] STEMINC, "Piezo disc 92 khz r 22x0.25mm wire lead," https://www.steminc.com/PZT/en/piezo-disc-7-mhz-r-22x025mm-wire-lead, 2019, accessed: 2019-09-07.

[46] Z. Su and L. Ye, *Identification of damage using Lamb waves: from fundamentals to applications*. Springer Science & Business Media, 2009, vol. 48.

[47] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2017, pp. 3–18.

[48] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," in *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, 2015.

[49] wiki, "Absolute threshold of hearing," https://en.wikipedia.org/wiki/Absolute_threshold_of_hearing, 2020, accessed: 2020-01-17.

[50] P. Wilcox, M. Lowe, and P. Cawley, "Mode and transducer selection for long range lamb wave inspection," *Journal of intelligent material systems and structures*, vol. 12, no. 8, pp. 553–565, 2001.

[51] 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.

[52] 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.

[53] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *IEEE Symposium on Security and Privacy (SP)*, 2019.

APPENDIX

A. Material Properties

The material properties of five different materials, including Ceramic, Aluminum, Glass, MDF, HDPE, are listed in Table X.

TABLE X: Material properties and acoustic impedances of different materials.

Materials	Young's module (Gpa)	Poisson ratio	Density (kg/m^3)	Acoustic impedance (MPa· s/m)
Ceramic	350	0.22	3,800	23.3
Aluminum	70	0.33	2,740	8.5
Glass	50	0.22	2,580	7.27
MDF	4	0.25	750	1.1
HDPE	0.6	0.46	930	0.437