EchoAttack: Practical Inaudible Attacks To Smart Earbuds

Gen Li
ligen4@msu.edu
Michigan State University

Zhichao Cao
caozc@msu.edu
Michigan State University

Tianxing Li
litianx2@msu.edu
Michigan State University

ABSTRACT

Recent years have shown substantial interest in revealing vulnerability issues of voice-controllable systems on smartphones and smart speakers. While significant prior works have leveraged inaudible signals to attack these smart devices, smart earbuds present unique challenges and vulnerabilities due to their extreme hardware constraints. In this paper, we present EchoAttack, a practical inaudible attack system for smart earbuds. The primary innovation of EchoAttack is the ability to leverage both indirect and direct paths to attack smart earbuds. To search for the optimal path, we design a path-searching algorithm based on the attenuation model of ultrasound. We also propose a novel approach to remove harmonics noise, which improves the attacking signal's SNR further. Finally, we propose using Zigbee radios to sniff the Bluetooth signal and enable a hidden feedback channel without the victim's awareness. We implement the EchoAttack prototype using off-the-shelf hardware components and evaluate the prototypes in four typical indoor and outdoor scenarios using six smart earbuds. Experimental results show that EchoAttack outperforms the pure direct-path attack by 75.8% on average in terms of attack success rate.

CCS CONCEPTS

• **Security and privacy → Security in hardware**; **Hardware attacks and countermeasures**; **Side-channel analysis and countermeasures**;

KEYWORDS

Voice Controllable Systems, Inaudible Attack, Earable Security and Privacy

ACM Reference Format:
Gen Li, Zhichao Cao, and Tianxing Li. 2023. EchoAttack: Practical Inaudible Attacks To Smart Earbuds . In *The 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23), June 18–22, 2023, Helsinki, Finland*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3581791.3596837

1 INTRODUCTION

Smart earbuds (e.g., Airpods, Pixel Buds) have become increasingly popular in recent years [1]. Once connected over Bluetooth, these earbuds often provide speech recognition technologies that allow users to control surrounding objects, such as smartphones, smart speakers, or smart home devices, by converting spoken words to machine-readable formats. A wide variety of voice-controllable systems, such as Apple's Siri, have been developed for smart earbuds.

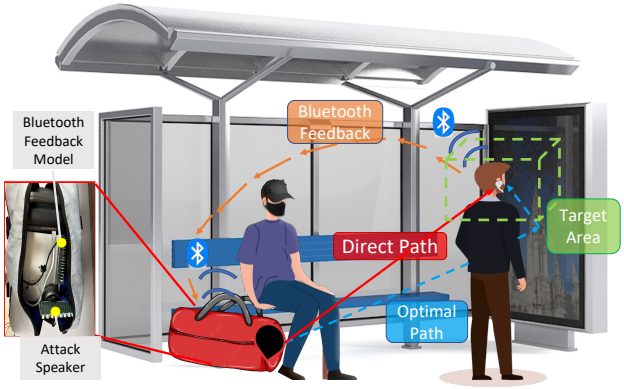


Figure 1: Illustration of EchoAttack at a bus stop.

Prior work has revealed several vulnerability issues of voice-controllable systems. For example, recent studies have shown the feasibility of using ultrasound to attack smartphone voice assistants based on its non-linearity effect [2, 3]. Although prior work has proposed various ultrasound-based attacks using smartphones, this paper mainly focuses on presenting unique challenges and vulnerability issues of smart earbuds due to the earbud's extreme hardware constraints.

Earbuds Vulnerability Compared to smartphones, smart earbuds are more vulnerable based on the two observations (§ 3). First, earbuds are worn in the ears, the location of which can be easily localized in some deterministic scenarios (e.g., gym, bus stop) by an attacker. In comparison, users may place their smartphones arbitrarily in their pockets or bags, making such ultrasound attacks hard to realize in practice. Second, the computation resources on earbuds are much weaker than those on smartphones [4]. Thus, software defense methods (e.g., Voice Match [5]) are hard to be implemented on earbuds.

Challenges While smart earbuds are more vulnerable than smartphones, carrying out successful attacks by exploiting these vulnerabilities is not trivial. First, the microphones on smart earbuds often have short sensing ranges and a narrow field-of-view (FoV) [6, 7], making it difficult to receive the attacking signal in real-world scenarios. Additionally, the existing ultrasound attack assumes that the microphone always faces the attacking speaker, which is not always true [2, 8]. Second, the non-linearity effect on smart earbuds is not as strong as on smartphones, meaning that small harmonic noise can compromise the quality of the attacking signals and reduce the attack success rate [9]. Finally, current ultrasound attack systems do not provide a feedback mechanism to inform the attacker about the

MobiSys '23, June 18–22, 2023, Helsinki, Finland

ACM ISBN 979-8-4007-0110-8/23/06…$15.00
https://doi.org/10.1145/3581791.3596837

success of their attack. Nevertheless, implementing such feedback channels is difficult as attackers typically have no control over the victim's devices [2, 8].

Overview of EchoAttack This paper addresses the above limitations, proposing EchoAttack, a practical ultrasound attack system for smart earbuds. The key idea is to leverage both indirect and direct paths to attack earbuds without access to the victim's devices. The attacking speaker can inject unauthenticated voice commands into ultrasound frequency, enabling earbuds to decode without the victim's awareness. Additionally, EchoAttack listens to surrounding Bluetooth signals to detect attack success. Figure 1 illustrates an attacking scenario where an attacker deploys a speaker near a victim at a bus stop, leveraging Bluetooth signals to enable a feedback channel and searching for the optimal attack path among in-direct (i.e., blue dashed line) and direct (i.e., red solid line) paths within the target area (i.e., green dashed box).

We design three key components to ensure the reliability of EchoAttack. First, we employ the ultrasound attenuation model to estimate the received signal strength (RSS) at the victim's device for optimal attack path searching. We assume the attacker knows the position and material of reflectors near the victim. Based on this information, we design a lightweight algorithm to compute all indirect and direct paths and construct a path table without knowing the precise location and orientation of the victim's device. We then exhaustively search for the optimal path in the path-searching table. Second, to further enhance the signal-to-noise ratio (SNR) of the attacking signal, we propose a harmonic noise removal algorithm to minimize the harmonic noise leakage from the modulated signal. Third, we sniff Bluetooth signals from the victim's device to determine whether the voice assistant is triggered. We design a Zigbee-based sniffer for recording Bluetooth signals and addressing unknown frequency hopping in Bluetooth. Then, we develop a decision-tree-based model to extract Bluetooth signal patterns as the feedback channel for the attacker.

We implement the EchoAttack prototype using cost-effective hardware components, including an ultrasound speaker array, an amplifier, a wave generator, and CC2530 Zigbee radios. We evaluate the EchoAttack prototype against six smart earbuds (i.e., Pixel Buds, Galaxy Buds, JBL Buds, AirPods 1, AirPods 2, Bose QuietComfort 45) in four scenarios: a public study area, bus stop, gym, and hallway. The experimental results show that EchoAttack achieves an 88.1% mean attack success rate across six earbuds, which is 75.8% higher than the pure direct-path attack in all scenarios.

Contributions We summarize our contributions as follows:

- We introduce EchoAttack, a practical inaudible attack system that exploits both indirect and direct paths to target smart earbuds without the victim's knowledge.
- We develop a path-searching technique that leverages the ultrasound attenuation model to identify the optimal attack path efficiently.
- We design a harmonics noise removal algorithm to minimize harmonic noise and enhances the SNR of the attacking signal.
- We present the idea of using Zigbee radio to intercept the Bluetooth signal and establish a covert feedback channel without the victim's awareness.
- We design and implement the EchoAttack prototype using commercial-off-the-shelf hardware components and conduct comprehensive experiments in real-world scenarios. Experimental results demonstrate that EchoAttack surpasses the pure direct-path attack in all tested scenarios.

2 BACKGROUND AND THREAT MODEL

2.1 Ultrasound Attack on Smart Device

Modern voice recording systems consist of four main components: 1) microphones converting sounds to electrical signals, 2) pre-amplifier amplifying electrical signals by a gain of 10–100×, 3) an analog-to-digital converter (ADC) converting analog signals to digital signals, and 4) low pass filters (LPFs) cutting off signals at 20 kHz [2]. These modules are linear systems within audible frequency (e.g., 20 Hz – 20 kHz), signifying that input signals are linearly proportional to outputs. However, beyond the audible frequency range, microphones, pre-amplifiers, and ADCs exhibit non-linearity, which can be modeled using Eq. 1:

$$S_{out}(t) = \Sigma_{i=1}^{\infty} A_i (S_{in}(t))^i \tag{1}$$

where S_{out} is the output signal, S_{in} is the input signal, and A_i represents the coefficient to the corresponding term. Although the output signal is theoretically an infinite sequence, third or higher terms can be ignored due to weak signal strength. The quadratic term $(S_{in}(t))^2$ produces harmonics and cross-products, generating new frequencies and recovering the audible signal. Audible signal f_a can be modulated on an inaudible frequency f_c using amplitude modulation in Eq. 2 [8]:

$$S_{in}(t) = (cos(2\pi f_a t) + 1)cos(2\pi f_c t) \tag{2}$$

After applying Eq.2 to Eq.1, both the audible frequency component f_a and inaudible frequency f_c, as well as their harmonics (e.g., $2f_a$, $3f_a$) and other cross products (e.g., $f_i - f_a$, $f_i + f_a$) can be extracted. After applying LPFs, high-frequency components (> 20 kHz) are eliminated, leaving the audible frequency component f_a. The nonlinearity effect of modern voice recording systems has been exploited to inject unauthenticated voice commands into voice control devices [2, 8, 10, 11]. Since the entire attack process is inaudible, it raises significant privacy concerns and can lead to serious security issues [3].

2.2 Threat Model

Attack Scenario: We consider a scenario when a victim wears smart earbuds that support hands-free voice assistant wake-up. The victim can freely move within the sensing area and turn their head in any direction. The attacker has a portable ultrasound speaker to attack the victim's earbuds.

Attack Goal: Considering the earbuds user can use the voice assistant to control their phone (e.g., opening voice recording, reading emails), the attacker's goal is to activate the voice assistant through the victim's earbuds and execute inaudible commands. For example, the attacker can force the phone to initiate a call to jam emergency services such as 911. In addition, the attacker can turn the victim's smartphone to airplane mode to make him/her lose connection or turn on the flashlight to drain the battery quickly [12, 13]. To

ensure stealthiness, the attacker should be away from the user to avoid the victim's awareness.

Assumptions: We assume that the attacker and victim are at the same scene and the attacker is out of the field of view of the victim. The victim is within a target area where the attacker knows the rough location and the information (including location and reflection coefficients) of surrounding reflection planes near the victim (like walls and screens). Note that the attacker does not know the exact location and orientation of the victim's device.

3 MOTIVATIONS

While a number of recent efforts have been exploring the problem of inaudible attacks on smartphones [2, 8, 10–13], we face unique challenges in the context of smart earbuds since both the hardware and the software design vary widely between smartphones and smart earbuds.

3.1 Microphone Sensing Capability

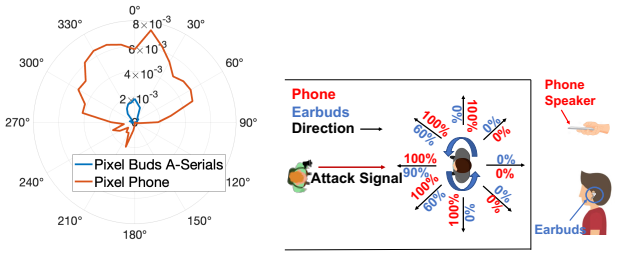
Smart earbuds differentiate themselves from smartphones in two key matrices: directionality and sensing range [6].

Directionality Smart earbuds have multiple built-in microphones facing various directions [7, 14]. For example, the second-generation AirPods Pro has two beam-forming microphones and an inward-facing microphone [15]. These microphones are either used as noise canceling or beamforming. Therefore, the earbuds will record only the signal from a small FoV.

Sensing Range Since smart earbuds are designed to sense audible sound from the mouth and inside the ear, the sensing area is significantly smaller than that of smartphones. Besides, due to the limited battery size, the gain of the pre-amplifier is also lower than that of smartphones [16]. Therefore, smart earbuds have a much shorter sensing range compared with smartphones.

We compare the directionality and sensing range between a pair of smart earbuds (Google Pixel Buds A-Serials) and a smartphone (Google Pixel 4). In this experiment, we put the earbuds and phone at the same distance (1 m) from the ultrasound speaker and changed their orientations (0° – 360°). The RSS is shown in Figure 2(a). The results show that smart earbuds have significantly smaller FoV and sensing ranges than smartphones.

Another underlying assumption in many existing inaudible attacks is that the victim device always faces the adversarial device [3]. However, in the context of smart earbuds, the RSS can be lower than the noise floor if the microphone does not face the speaker due to the victim's mobility. Thus, the victim's device will not recognize the unauthenticated voice command. To validate the hypothesis, we place a smartphone and a smart earbud 1 m away from a speaker (Figure 2(b)). We then leverage a direct-path attack [8] to activate the Google Voice Assistant by injecting voice commands into inaudible sound. We repeat the experiment 50 times under each setting. We observe that when the microphones are toward the speaker, both smartphone and smart earbuds achieve a high activation rate (100% for the smartphone and 90% for the smart earbuds). However, when the microphones are perpendicular to the speaker, the activation rate of the smart earbuds drops to 0% due to their narrow FoV and short sensing range. In contrast, the



(a) RSS of directionality and sensing range

(b) Comparison of the voice assistant activation rate using direct-path attack [8]

Figure 2: Comparison of smart earbuds' and smartphones' directionality and sensing capability.

smartphone can still be activated at a 100% success rate. Therefore, improving the SNR of inaudible attacks for smart earbuds in practical scenarios is challenging.

3.2 Software Vulnerability

Although smart earbuds have lower sensing capability, they have higher software vulnerabilities than smartphones [17–19]. To protect user privacy, many voice assistant systems (e.g., Siri and Google Voice Assistant) have voice match techniques that can tell the difference between legal users' voices and filter out unauthenticated users [20–22].

To measure the software vulnerability between smart earbuds and smartphones, we conduct a user study with eight participants (7 males and 1 female). We examined four voice assistants (Siri, Google Voice Assistant, Alexa, and Bixby) using six smart earbuds and four smartphones (Table 1). We first set up voice matches on eligible voice assistants (Siri and Google Vice Assistant) using $User1$'s voice. Then, we ask the rest of the participants to activate voice assistants. Each user repeats ten times to activate the voice assistant. Besides real human voice, we leverage Voicemaker [23] to generate 47 voice commands and then use the synthetic commands to activate voice assistants. Table 1 shows the activation rates using unauthenticated voice commands. We have two observations. First, Alexa and Bixby do not have voice match techniques, so they cannot prevent unauthenticated users from activating their voice assistants. Second, voice match technology can prevent unauthenticated users from activating voice assistants on smartphones. None of the unauthenticated voice commands can activate the voice assistant on voice-match-enabled smartphones. However, when the smart earbuds are connected to the phone, some unauthenticated voice commands can bypass the voice match and activate the voice assistant. For example, the activation rates using unauthenticated voice commands are 34%, 100%, and 26% on AirPods 3, Beats Studio, and Pixel Buds, respectively. We hypothesize that when smart earbuds are connected to the smartphone, the voice match algorithm will be offloaded to the earbuds. Due to hardware limitations (e.g., limited Digital Signal Processors (DSP)) on earbuds, some may not recognize unauthenticated voice commands. Therefore, it is important to investigate the potential inaudible attacks on earbuds.

Table 1: Software vulnerability of earbuds and phones when activating voice assistant using unauthenticated voice commands.

	Model	Phone OS	Voice Control	Voice Match	Activation Rate
Earbuds	AirPods 3	iOS 15.6	Siri	✓	34%
	Beats Studio	iOS 15.6	Siri	✓	100%
	Pixel Buds	Android 12	Google	✓	26%
	Sony WF1000XM4	Android 12	Alexa	✗	100%
	JBL Reflect Flow	Android 12	Alexa	✗	100%
	Galaxy Buds Live	Android 12	Bixby	✗	100%
Phones	iPhone X	iOS 15.6	Siri	✓	0%
	Galaxy A02s	Android 11	Google	✓	0%
	Galaxy A02s	Android 11	Alexa	✗	100%
	Galaxy s10e	Android 12	Bixby	✗	100%

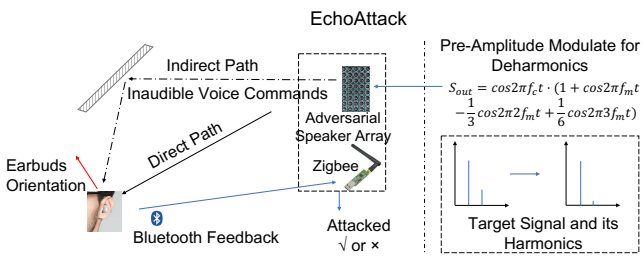


Figure 3: The system overview. EchoAttack can leverage both indirect and direct paths to attack smart earbuds. It has a harmonic removal module to enhance the signal's SNR and a passive feedback module to monitor the attacking results.

4 SYSTEM DESIGN

This section describes the three design components of EchoAttack. First, the key idea to enhance the signal strength of inaudible attacks on smart earbuds is leveraging the *indirect path* between the attacking speaker and the victim device so that the microphone can capture the inaudible signal within the small FoV [24]. Figure 3 shows a typical attacking scenario where the microphone is not toward the speaker. In this case, if we rely on the direct path to attack the victim's device, the SNR will be extremely low since the attacking signal is outside the microphone's FoV. To address the challenge, EchoAttack can leverage a reflective plane (e.g., wall, screen, table) to attack the victim's device via an indirect path. And we design a lightweight algorithm to search for the optimal path. Second, we design a harmonic removal algorithm to remove harmonic noises to improve the SNR of inaudible attacks further [25]. Finally, we propose a passive Bluetooth-based feedback module to collect hidden acknowledgment data from the victim device. The feedback module can tell the attacker whether the attack is successful or not without the victim's awareness.

4.1 Optimal Path Searching

This module aims to identify the optimal attack path between the attacking speaker and the victim's device. EchoAttack relies on several assumptions that need to be considered. Figure 4 shows the victim's location T is within a target area A and there are reflective surfaces $\mathcal{R} = \{\mathbf{R_i}\}$ surrounding the victim. The attacker

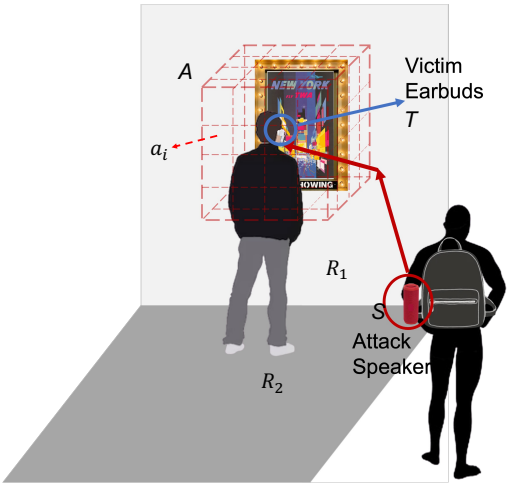


Figure 4: Attack victim when he/she is in the estimated area.

can determine A and the position of $\mathcal{R}$ relative to S, once they place the speaker at the location S. To launch an attack on a victim within area A, the attacker identifies specific paths that allow the speaker S to cover A. EchoAttack divides A into smaller areas $\{a_i, i \in \mathcal{N}, i > 0\}$ with size l^3, where l is the side length of each a_i. Also, EchoAttack does not know the exact victim's orientation, but it can estimate the range of orientation depending on attacking scenarios. For example, when a victim sits by a desk and looks at a screen, the victim's orientation is against the screen and moves within a small range (like 30° left or right). Suppose that the orientation is θ. Given $\{S, a_i, \theta, \mathcal{R}\}$, we can compute the direct path and all indirect paths (reflected on one or more reflect planes in $\mathcal{R}$). The path is denoted by its length, the orientation of the attack speaker, and the incident angle when it arrives a_i. We can derive all paths in Eq. 3:

$$\mathcal{P}(S, a_i, \theta, \mathcal{R}) = \{(l_j, \beta_j, \alpha_j), j <= \sum_{x=0}^{|\mathcal{R}|} C_{|\mathcal{R}|}^{x}\} \tag{3}$$

where l_j is the path length, β_j is the orientation of the attack speaker, and α_j is the incident angle of the ultrasound to victim earbuds.

We employ the ultrasound attenuation model to identify the optimal path in $\mathcal{P}$ for the given a_i and θ. Inaudible signal attenuation depends on distance, incidence/reflection angle, and reflective

surfaces. Attenuation due to distance follows the inverse-square law [26], making inaudible signal strength inversely proportional to the square of propagation distance. Microphone sensitivity is affected by the angle at which the inaudible signal hits it. Peak sensitivity occurs with the sound incident at right angles to the sensing surface and drops off following a cosine law. For instance, the received signal drops to 86% of peak sensitivity at a 30° incidence angle. When the inaudible signal encounters an object, some waves bounce off while others attempt to pass through or be absorbed. With flat surfaces (e.g., mirrors), incidence and reflection angles are equal. In practice, the attackers can intentionally select solid and flat reflection planes[1]. When obstacles are small (e.g., equal to or smaller than the inaudible signal wavelength), waves bypass the obstacle and spread into the shadow region behind it [27]. Diffraction amount is inversely proportional to acoustic frequency [3].

Assuming attacker speaker signal strength is I_s, the earbuds' RSS I_r can be formulated as Eq. 4:

$$I_r = \frac{C_L(\prod C_{r_i})C_\alpha cos(\alpha) \cdot I_s}{f \cdot L^2} + Noise \tag{4}$$

where C_L is the attenuation coefficient with path length, C_{r_i} is the reflection coefficient that the path passes, and C_α is a device-related incident angle coefficient which we can get with the pre-attack test.

In $\mathcal{P}(S, a_i, \theta, \mathcal{R})$, we search for the optimal path (either direct or indirect path) with the largest I_r, that is $p_i = \{p|max\{I_r(p)\}, p \in \mathcal{P}(S, a_i, \theta, \mathcal{R})\}$. In practice, the location of the S is often known and fixed, e.g., on a table, chair, or floor. Therefore, we only need to search for S's optimal orientation o ($o_i(S, a_i, \theta) = \beta, \beta \in p_i$). As the ultrasound speaker has FoV of at least 60° [28], we can simplify the calculation by dividing the 360° of victim orientation into eight directions, with a step of 45° to ensure coverage, and the path can cover the 360° orientation of the victim. After all the computation, there are $|a_i| \times N_\theta$ paths, where $|a_i|$ is the number of searching divisions and N_θ is the number of directions. We combine the adjacent paths to reduce the search space further. Finally, we apply the Sobol sequence [29] to draw a path from the path table. On average, for a path table with N paths, we need $\frac{N}{2}$ tries.

4.2 Harmonic Noise Removal

Existing inaudible attacks leverage Eq. 1 and Eq. 2 to inject audible voice commands into inaudible sound [3, 8]. However, not only the audible frequency component f_a but also its harmonics (e.g., $2f_a$, $3f_a$) will remain in the demodulated signal after applying LPFs [8]. Specifically, the audible component $S_{audible}$ is generated from the even power terms in Eq. 1, which can be derived in Eq. 5:

$$S_{audible} = \sum_{2i}^{\infty} A_{2i}(1 + cos(2\pi f_a t))^{2i} \tag{5}$$

While $S_{audible}$ is an infinite power series, the signal strength of the third and higher-order terms is negligible. Therefore, we only consider the second-order term $A_2(1 + cos(2\pi f_a t))^2$, as shown in Eq. 6

$$\begin{aligned} A_2(1 + cos(2\pi f_a t))^2 = A_2 + \frac{A_2}{2} + \\ 2A_2 cos(2\pi f_a t) + \frac{A_2}{2} cos(2\pi 2 f_a t) \end{aligned} \tag{6}$$

where the RSS of the second harmonic $cos(2\pi 2 f_a t)$ is $\frac{1}{4}$ of that of the target signal $cos(2\pi f_a t)$.

The harmonic can degrade the voice recognition accuracy of the victim device and lower the attack success rate. For example, when we modulated $f_a = 2$ kHz target signal to $f_c = 23$ kHz carrier, we observed that the RSS of $2f_a$ is only 3.755 dB lower than that of f_a, significantly degrading the signal-to-noise ratio of the f_a. To minimize the impact of the harmonics, we modulate the target signal onto the carrier by introducing new coefficient terms b_i in Eq. 2 and derive the enhanced target signal S' in Eq. 7,

$$S'_{in}(t) = cos(2\pi f_c t) \sum_{i=0}^{n} b_i cos(2\pi \cdot i \cdot f_a t) \tag{7}$$

where b_i is the coefficient of $cos(2\pi \cdot i \cdot f_a t)$, and by default $b_0 = 1, b_1 = 1$. By applying Eq. 7 to Eq. 1, we can derive the target signal and its harmonics in Eq. 8

$$\begin{aligned} RSS_{f_a} &= (2b_1 + \sum_{|i \pm j|=1, i,j \neq 0}^{n} b_i b_j) cos(2\pi f_a t) \\ RSS_{2f_a} &= (\frac{b_1^2}{2} + 2b_2 + \sum_{|i \pm j|=2, i,j \neq 0}^{n} b_i b_j) cos(2\pi \cdot 2 f_a t) \\ RSS_{3f_a} &= (2b_3 + \sum_{|i \pm j|=3, i,j \neq 0}^{n} b_i b_j) cos(2\pi \cdot 3 f_a t) \\ RSS_{4f_a} &= (\frac{b_2^2}{2} + 2b_4 + \sum_{|i \pm j|=4, i,j \neq 0}^{n} b_i b_j) cos(2\pi \cdot 4 f_a t) \\ \cdots &= \cdots \end{aligned} \tag{8}$$

Since the frequency range of human voice is from 100 Hz to 17 kHz [30], the fourth and the higher-order harmonics signal is likely above the audible frequency. Also, the fourth and higher-order terms are extremely weak. Therefore, we only remove the second and the third harmonic. Thus, we set n to 3 so that b_2 and b_3 can be derived in Eq. 8, e.g., $b_2 = -\frac{1}{3}$ and $b_3 = \frac{1}{6}$. Our final modulated signal can be derived in Eq. 9

$$\begin{aligned} S_{in} = cos(2\pi f_c t)(1 + cos(2\pi f_a t) + \\ b_2 cos(2\pi \cdot 2 \cdot f_a t) + b_3 cos(2\pi \cdot 3 \cdot f_a t)) \end{aligned} \tag{9}$$

To validate the effectiveness of the new modulation scheme, we conduct an experiment to compare the normalized RSS of the second harmonic and the third harmonic with and without the harmonic removal module. Figure 5 shows the average normalized RSS and the error bars covering the 90% confidence interval. We can see that the harmonic removal module decreases the RSS of the second harmonic and the third harmonic by 77.6% and 73.9%, respectively. However, there are still harmonic residues. The presence and strength of harmonics in a microphone can vary depending on the design and construction of the device. Moreover, the nonlinearity of a device is closely linked to its hardware design, and the choice of components used can significantly affect the resulting

[1]Considering that the reflection on these planes can keep the most signal energy, we omit the reflection coefficients of the reflectors during the path search.

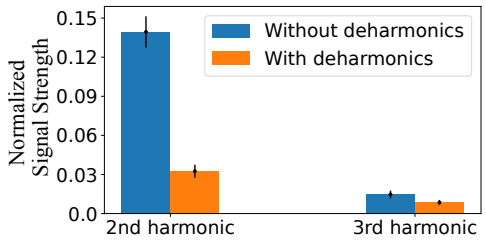


Figure 5: Normalized 2nd and 3rd harmonic signals with and without harmonic removal. The target signal frequency is 2 kHz, and the corresponding 2nd and 3rd harmonic signals are 4 kHz and 6 kHz.

non-linearity [9]. This means the degree of harmonics and non-linearity can differ between various earbud models. Therefore, we empirically state that the performance of our deharmonics varies among different earbuds, and the detailed performance difference will be shown in Section 6.

4.3 Hidden Feedback Module From Smart Earbuds

Earbuds are connected to smartphones with Bluetooth [31, 32], which operates at 2.4 GHz unlicensed band. We rely on the unique features of a voice assistant wake-up process to detect whether the voice assistant is triggered. First, after we finish ejecting the inaudible signals, we start to sniff the Bluetooth signals emitted during the voice handover between the earbud and the smartphone. Specifically, Zigbee [33] adopts IEEE 802.15.4 physical layer operating at the same 2.4 GHz band, which is divided into 16 non-overlapped channels. As shown in Figure 6, the 16 channels cover most Bluetooth channels. We leverage this overlapping to sniff Bluetooth signals without pairing. Specifically, for each attack, we use 16 Zigbee radios to sample the RSS of the 16 different channels continuously in a time window. The RSS sampling window lasts 1 ms. When we obtain 16 RSS vectors indicated as $v_{rss}(i)$ where i is the channel. Each vector contains n RSS samples $v_{rss}(i)[1]$ - $v_{rss}(i)[n]$.

For the RSS of 16 channels, we can get an RSS matrix $V_{16\times n}$. To extract the RSS feature of the Bluetooth channels, we need to remove the Wi-Fi and Zigbee noise from $V_{16\times n}$. Different from Wi-Fi and Zigbee, Bluetooth applies adaptive frequency hopping technology (AFH) [34]. AFH allows Bluetooth to switch frequency within Bluetooth channels (79 channels for classic Bluetooth, 40 channels for BLE) and can avoid interference to and from other devices that operate within its frequency band. Therefore, we first generate a static noise RSS matrix V_{noise} when the voice assistant is not activated and calculate the average noise level of each channel, Based on V_{noise}, we remove the RSS value that is lower than the average RSS of the static noise at each channel from the voice-assistant-activated sampling window matrix $V_{activate}$ as follows:

$$v_{ij} = \begin{cases} -110, & v_{ij} < average(V_{noise}[i]) \\ v_{ij}, & v_{ij} \geq average(V_{noise}[i]) \end{cases} \tag{10}$$

where $v_{ij} \in V_{activate}$, $i < 16$ and $0 \leq j < n$. This removes most noise from Wi-Fi and Zigbee without affecting the RSS trend of Bluetooth signals. After the noise removal, we need to extract

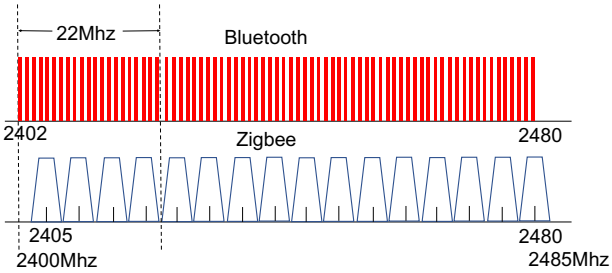


Figure 6: Channel overlap between Bluetooth and Zigbee.

the Bluetooth RSS changing pattern. First, we shrink the matrix $V_{activate}$ to a RSS vector $E_{activate}$:

$$E(j) = max(V_{activate}[i][j]), \quad 0 \leq i < 16, \quad 0 \leq j < n \tag{11}$$

After getting E, a sliding window w is applied to E to get smoothed RSS feature vector F:

$$F(i) = \frac{\sum_{i}^{i+w} E(i)}{w}, \quad 0 \leq i < i + w \tag{12}$$

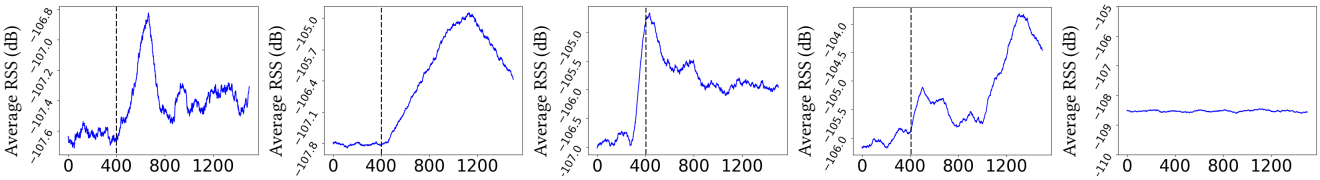
If Bluetooth signals appear, we will observe a higher RSS value in F. We demonstrate feature vectors obtained from a clean environment in Figure 7 for different voice assistants (Figure 7(a) – Figure 7(d)) and the clean environment noise (Figure 7(e), no voice assistant is activated). From the results, we can observe RSS changing after the wake-up command is sent (400 ms), and the change is obvious.

According to the noise-filtered feature vector in a sampling window, we use a decision tree to extract the unique features for attacking success classification. The decision tree is a common supervised machine learning model whose structure is tree-shaped, and each node of the decision tree makes a decision on a specific feature. Each input of the decision tree consists of a feature vector and a label. As shown in Figure 7, the RSS values before and after the voice command are sent have obviously different features. So we select the features as maximum, minimum, and average peak values of 300 ms before and after activation and the average RSS difference of 300 ms before and after activation. From our training results, the two most important features are the maximum peak value of 300 ms before activation and the minimum peak value of 300 ms after activation. It should be noted that the decision tree is a value-based and small machine learning model, so in environments that have different noise levels, we cannot use the same pre-trained decision tree. But we can easily train decision trees in different environments with a small amount of data or get decision trees at different noise levels in advance.

5 IMPLEMENTATION

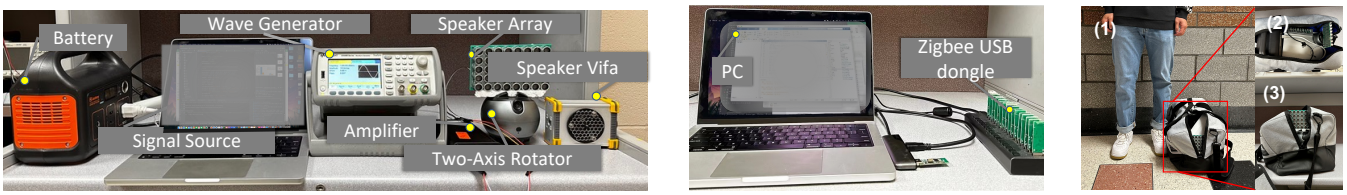
As shown in Figure 8, we implement our prototype, including the acoustic attacker system and the Bluetooth monitoring system, with commercial off-the-shelf hardware and devices.

For the attacker system shown in Figure 8(a), we make an 8 × 5 speaker array with 40 ultrasound speaks with 23 kHz central frequency [28] to transmit ultrasound signals in a long distance. In addition, we use high-quality full-band ultrasound speaker Vifa [35], supporting the signal frequency as high as 120 kHz, to attack those earbuds with non-23 kHz central frequency in general. We use



(a) Alexa (b) Bixby (c) Pixel (d) Siri (e) Clean

Figure 7: Pattern of different voice assistants being activated (a) – (d) and no activation (e). Wake-up words sent at 400 ms



(a) Acoustic attacker system (b) BT monitoring system (c) Portable system

Figure 8: The implementation details of EchoAttack. (a) acoustic attacker system, (b) Bluetooth monitoring system, and (c) portable EchoAttack system.

Table 2: The characteristics of the six earbuds in our experiment.

Earbuds	System	Hands-free Command	Central Frequency	Max. Distance
Pixel Buds A-Serials	Android 11	Google Assistant	23.5 kHz	3 m
Galaxy Buds Live	Android 12	Bixby	23.5 kHz	3 m
JBL Reflect Flow	Android 11	Alexa	23.5 kHz	1.5 m
AirPods 1st Gen	iOS 15.6	Not Supported	31.5 kHz	0.5 m
AirPods 2nd Gen	iOS 15.6	Siri	30.5–31.5 kHz	0.5 m
Bose QuietComfort® 45	Android 11	Not Supported	32 kHz	1.1 m

MATLAB running on a laptop computer to modulate the inaudible signals of attacking voice commands, and the modulated signals are generated and amplified by a Keysight waveform generator [36] and an Avisoft portable ultrasonic power amplifier [37], separately. We use a rotational holder [38] to enable the optimal attacking path search. The attacker system is powered by a portable battery.

For the Bluetooth monitoring system shown in Figure 8(b), we use 16 CC2531 USB dongles [39] (called *worker dongles*) to sample the RSS values across 16 Zigbee channels. Due to the limited storage resource on the worker dongles, we use another CC2531 USB dongle (called *master dongle*) to control the worker dongles to start channel sampling a little earlier than the beginning of an attack. The master dongle and worker dongles are connected to a laptop with a USB hub. Once the worker dongles are triggered, the sampling period will last for 1600 ms, and a total of 16×1600 RSS values are logged locally by each worker dongle. Then, the laptop pulls the data from all worker dongles through the USB port to determine whether an attack is successful.

As shown in Figure 8(c), we carefully package the attacker system and Bluetooth monitoring system, which share the same laptop, in a travel/sport bag. The USB hub and Zigbee dongles of the Bluetooth monitoring system are powered by the same portable battery.

6 EVALUATION

We evaluate the performance of EchoAttack with six different pairs of earbuds in four indoor and outdoor scenarios with diverse settings, including diverse attacking distances, background noises, and surrounding obstacles. We have received the IRB approval from our institute. In our experiments, we make sure that only the volunteer victim is in the attack range of EchoAttack. Others will not be affected.

6.1 Experimental Methodology and Setting

Our Earbuds: The characteristics of the six earbuds used for evaluation are listed in Table 2. We include all mainstream earbuds (e.g., Google, Apple, Samsung, Bose, JBL) supporting voice assistants on the market. During our experiments, we connect different earbuds to the corresponding smartphones. Specifically, we connect Google, JBL, and Bose earbuds to a Google Pixel smartphone with Android 11. And AirPods and Galaxy Buds are connected to an iPhone Xs with IOS 15.6 and a Samsung A02s with Android 12.

Attacking Speaker Selection: The 6 earbuds have different central frequencies ranging from 23 kHz to 32 kHz as shown in Table 2. For each pair of earbuds, we measure its central frequency via an enumeration search. Specifically, we modulate with a 2 kHz single

(a) S1 (b) S2 (c) S3 (d) S4

Figure 9: Our four experiment scenarios include (a) a public study area, (b) a bus stop, (c) the treadmill area in a gym, and (d) the hallway in our office building. In each scenario, our indirect attack path is indicated as the blue dashed line, and the direct attack path is shown as the red solid line.

tone on an ultrasound carrier, which is recorded by the earbuds with perfect direction alignment. Due to the non-linear effect, the 2 kHz single tone appears in the recorded signals. We search different frequencies of the ultrasound carriers from 20 kHz to 35 kHz with 0.1 kHz step length to find the strongest recorded 2 kHz single tone on the spectrum as the central frequency. Since the frequency range of the speaker array is 23 ± 2 kHz, we use it to attack Pixel, Galaxy, and JBL with 23.5 kHz central frequency. The full-band speaker Vifa is used to attack AirPods and Bose with 30–32 kHz central frequency.

Attacking Methodology and Performance Metrics: For four earbuds supporting hands-free command, we first conduct a voice-assistant-activation attack with "OK Google", "Hey Siri", "Hi Bixby", and "Alexa", respectively. After the voice assistant is woken up, we execute attacks with the voice command "open airplane mode" for all six earbuds. We define the *Attack Success Rate* as the ratio between successful attacks and total attacks to measure EchoAttack performance.

Experiment Scenarios: As shown in Figure 9, we evaluate our attack system in four scenarios. The first scenario (S1, Figure 9(a)) is a public study area, and a board is borrowed to reflect our attacking acoustic signals. The second scenario (S2, Figure 9(b)) is a bus stop, and the glass of the bus stop shelter provides our indirect attacking path. The third scenario (S3, Figure 9(c)) is the treadmill area in a gym, and the screen of a treadmill is used to build the indirect attacking path. The fourth scenario (S4, Figure 9(d)) is the hallway in our office building. Unlike the other three scenarios full of reflection surfaces, the hallway is a spacious environment where only the ground can be utilized as a reflection plane.

Moreover, we execute the attack with a portable bag shown in Figure 8(c) and invite a male volunteer to serve as the victim. The positions of the victim and the bag are shown in Figure 9(a)-(c), and our indirect attacking paths are illustrated with blue dashed lines. The volunteer behaves naturally in the four scenarios according to his habits. For example, his body and head may move, rotate, and shake in a certain area as usual.

Attacking Path Search: We measure each earbud's maximum distance under an attack over the direct path with the aligned speaker-earbud direction. As shown in Table 2, the maximum attacking distance of the speaker array reaches 3 m for Pixel and Galaxy. However, Vifa has less transmission power than the speaker array, so it only supports a maximum 1.1 m attacking distance for Bose.

Since the maximum attacking distances among different earbuds are different, we adjust the attacking position of the bag for each earbud. Specifically, we use the maximum distance in each experiment scenario to estimate the safest attacking position. After determining the attacking position and the victim's head position, we calibrate the indirect attacking path with our path-searching algorithm (Section 4.1). We use the attack success rate of the selected indirect path to identify the effectiveness of our path searching.

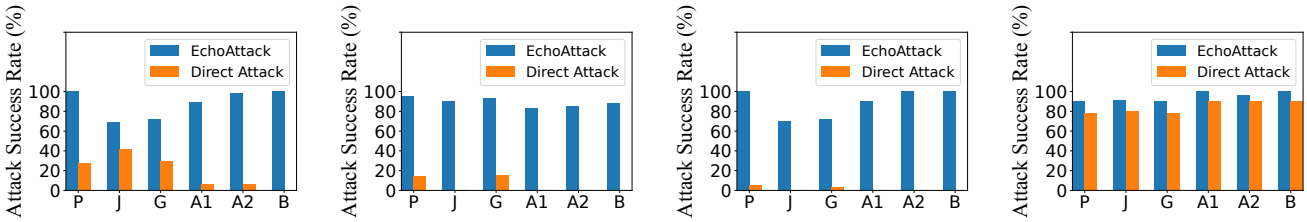
Bluetooth Feedback: For the Bluetooth feedback system (Section 4.3), we take an offline training process to generate the attack success classifier for each earbud in each scenario. Specifically, the attacker mimics the victim's behavior in similar scenarios in advance. The attacker will trigger worker dongles' recording 400ms before the wake-up words are played. We collect the RSS data of the voice assistant activated 400 times with positive labels. Then, we collect another 400 RSS sequences with negative labels when we do not launch any attack. The labeled RSS sequences are fed to train the classifiers with the "scipy" decision tree model [40].

Baseline Method: At the same attacking position of each scenario, we execute the attacks over the direct path, in which we put the speaker towards the victim's ear area without speaker-earbuds alignment. In our evaluations, we take the attacking via the pure direct path without applying the de-harmonic method as a baseline method, called "Direct Attack". The baseline direct attacking paths are shown as the red solid lines in Figure 9(a)–(d).

6.2 Overall Performance

In S1, S2, and S4, where the victim is relatively static, we conduct 150 attacks to calculate each earbud's corresponding attack success rate through the optimal indirect path in EchoAttack and the direct path in Direct Attack. In S3, with 150 attacks, we compare the attack success rate of EchoAttack when the victim is walking and running on the treadmill separately. Different earbuds are labeled: Pixel buds as P; JBL as J; Galaxy as G; AirPods1 as A1; AirPods2 as A2; and Bose as B.

Results: Figure 10 shows the performance comparison between EchoAttack and Direct Attack in four experimental scenarios shown in Figure 9. Overall, the mean attack success rate of EchoAttack is 75.8% higher than the Direct Attack. In S1 and S2, EchoAttack has 88.5% mean attack success rate, achieving almost 100% attack success rate using Pixel, AirPods2, and Bose earbuds. The mean attack success rate drops to 70.6% using Galaxy and JBL earbuds



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3 (d) Scenario 4

Figure 10: The performance comparison between EchoAttack and Direct Attack in the four scenarios from S1 to S4. (P: Pixel buds, J: JBL, G: Galaxy, A1: AirPods1, A2: AirPods2, B: Bose)

Table 3: The attack success rate of EchoAttack for different earbuds when the victim is walking and running in S3.

	P	J	G	A1	A	B
Walking	100%	70%	75%	90%	100%	100%
Running	100%	66.5%	64.8%	67.5%	80%	100%

due to the non-linearity effect difference among different hardware. In comparison, the maximum attack success rate of Direct Attack is only 42.0% and drops to 6.5% and even 0 using AirPods1 and Bose earbuds.

In S3, EchoAttack achieves 89.2% mean attack success rate, whereas the attack success rate is 0% for Direct Attack. Moreover, as shown in Table 3, when the victim runs on the treadmill, the mean attack success rate decreases to 79.8% due to the severe head vibration during running. However, for those scenarios where the victim's head is severely vibrating, EchoAttack can leverage the feedback module to detect the attack failure and then quickly issue a new attack. In S4, EchoAttack achieves 94.5% mean attack success rate, higher than the Direct Attack's 84.4% mean attack success rate. This verifies the effectiveness of utilizing ground, a commonly available reflection plane in a spacious environment, to attack the victim's earbuds with EchoAttack. Moreover, the attack success rate of Direct Attack in S4 is much higher than in S1 – S3 because the direction of the direct attack path is aligned with the orientation of the earbuds' microphone in S4, and no such alignment exists in S1 – S3. Thus, Direct Attack only works with a good direction alignment. Overall, the consistent attack success rate verifies that the voice assistant becomes more vulnerable under EchoAttack in practice.

6.3 Performance per System Module

Indirect Attack Path Effectiveness To evaluate the effectiveness of the indirect attack path, we conduct experiments in a controlled scenario where a victim stands 0.5 m away from a wall, and the earbud is 1.65 m above the floor. The earbuds' microphone port faces the ground at a 45° angle. We deploy the attacking speaker 2 m away from the wall and 1 m above the floor. The speaker and the victim are at a vertical line to the wall. In the experiments, the victim faces eight different directions, ranging from 0° to 325°, with a uniform 45° difference. The directions of 0° and 180° indicate that the victim faces the wall and the speaker. We apply the path-searching algorithm in each face direction to search for the optimal

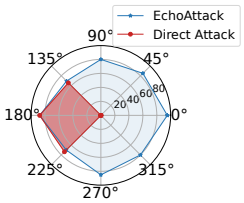


Figure 11: The effectiveness of indirect attack paths.

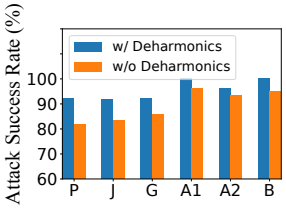


Figure 12: The efficiency of the de-harmonic method.

path by utilizing the wall and floor as reflectors. After determining the optimal indirect attack path, we use PixelBuds to conduct 20 attacks over the indirect and direct paths to calculate the attack success rate.

Results: Figure 11 shows the effectiveness of indirect attack paths. Overall, EchoAttack achieves 81.5% mean attack success rate when the victim faces the eight directions. Since the baseline method only leverages the direct path to attack the victim, the attack success rate is zero when the victim does not face the speaker.

Deharmonics Method We leverage the same experimental settings above to evaluate the efficiency of our harmonic removal method. In this experiment, the victim faces the wall. Then, we evaluate all six pairs of earbuds and conduct 50 attacks for each to calculate the attack success rate with or without the harmonic removal method.

Results: Figure 12 shows that applying the harmonic removal method improves the mean attack success rate by 6.1%. Due to the imperfection of different hardware's non-linearity effect, the improvement varies from 2.8% to 10.5% across the six earbuds.

Bluetooth Feedback System We leverage the pre-trained classifiers in Section 6.1 to evaluate the efficiency of the Bluetooth feedback system using four hands-free earbuds (Pixel, Galaxy, JBL, and Airpods) in S1–S3. For each pair of earbuds in each scenario, we collect 100 attack success RSS sequences and 100 RSS sequences without any attacks as the testing dataset. Notice the training dataset and testing dataset are collected on different dates, and co-existence interference (e.g., Wi-Fi) is relatively weak in all three scenarios.

Results: Table 4 shows the mean detection accuracy and F1-Score across the three scenarios. For all earbuds, EchoAttack achieves 95.19% mean detection accuracy, and the F1-Scores of non-activated

Table 4: The average accuracy and F1-Score of attack success feedback for four earbuds across 3 scenarios S1 – S3.

Earbuds	Accuracy (%)	F1-Score (%)	
		No Activated	Activated
Google Pixel Buds	91.20	92.93	90.07
Galaxy Buds Live	95.22	95.38	95.07
JBL Reflect Flow	99.06	98.11	99.17
AirPods 2nd Gen	95.26	95.87	94.32

and activated predictions are above 90%, indicating the effectiveness of our attack success feedback system.

6.4 System Practicability Study

Path searching Overhead The path searching overhead is determined by the side length of a_i within the target area A. In this experiment, we set the side length from 0.1 m to 0.3 m, and then defined the target areas in S1 – S3 as follows: In S1 (Figure 9(a)), the speaker is placed on a desk, 0.7 m from the floor and 0.15 m from the wall. The board is situated 2 m from the speaker. The center of the target area is 1.2 m from the wall and 1 m from the board. The left-to-right and front-to-back sizes are 0.8 m and 0.4 m. The victim sits randomly within the target area facing the board. The earbuds are 1.0 m above the floor. In S2 (Figure 9(b)), the speaker is placed on a 0.45 m high bench, 1.3 m from the glass wall. The center of the target area is 0.5 m from the glass wall. The left-to-right and front-to-top sizes are 0.6 m and 0.2 m. The victim stands within the target area, facing the glass wall within a 30° range left and right. The earbuds are 1.65 m above the floor. In S3 (Figure 9(c)), the speaker is placed on a 0.3 m high bench. The treadmill screen is 2 m from the speaker, and its height is from 1.4 m to 1.8 m. The victim walks/runs normally on the treadmill, 0.4 m from the screen. The target area extends 0.2 m left and right of the screen's center, with a height of 1.6–1.9 m. The earbuds are 1.65 m above the floor. In all three scenarios, the earbuds' microphone faces downwards at a 45° angle to the ground.

Results: Figure 13 shows the attack success rate and the number of paths in the path table. After combining the adjacent paths within a 5° range, the number of searching paths decreases by 85% without sacrificing attack performance. However, as the side length becomes larger, the speakers cannot search for the optimal path, degrading the attack performance. Therefore, we set the side length to 0.2 m to minimize the searching overhead while maintaining the attack efficiency. With a 0.2 m side length, path tables for S1, S2, and S3 have 8, 9, and 3 paths, respectively. The time required to generate the path table for the three scenarios is 42.8 ms, 10.7 ms, and 0.8 ms. The average number of attacks needed to perform a successful attack are 3.5, 4.2, and 1.4 in S1, S2, and S3.

Impact of Attacker-Victim Distance We use the Pixel Buds and the Galaxy Buds Live to evaluate the attack success rate of EchoAttack and Direct Attack with various attacker-victim distances. The earbuds are placed 1 m away from the reflection point, and the distance between the attacker speaker and the reflection point varies from 1.2 m to 2 m. The directions from attacker and target to the reflection point are perpendicular (the reflection angle is 45°), making the attacker-target distance vary from 1.56 m to 2.23 m.

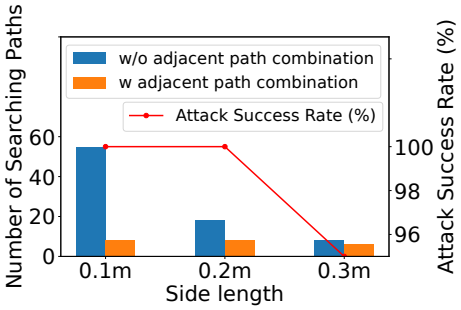
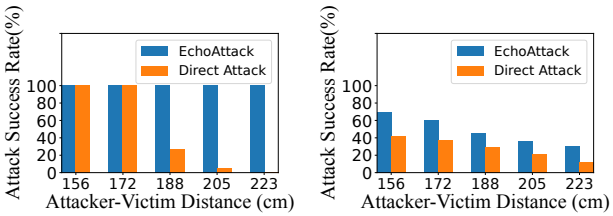


Figure 13: The impact of side lengths in path searching.



(a) Google Pixel Buds (b) Galaxy Buds Live

Figure 14: The impact of attacker-victim distance.

Results: Figure 14 shows the attack success rates when the attacker-victim distance varies. For Pixel Buds, when the attacker is close to the victim (< 1.7 m), both EchoAttack and Direct Attack achieve a 100% attack success rate. When the attacker-victim distance is greater than 2 m, Direct Attack's performance significantly drops to 0, whereas EchoAttack's performance remains the same. For Galaxy Buds, EchoAttack's attack success rate is consistently higher than Direct Attack by 30%. The results also indicate Galaxy Buds are more sensitive to signal energy change than Pixel Buds, exhibiting hardware diversity.

The Impact of Audible Noise We add two types of noises (white noises and human voices) in an office to study the impact of ambient noise on EchoAttack. We put a laptop next to the earbuds to play white noise and human voices, with the 55–70 dB signal strength.

Results: Figure 15 shows that the mean attack success rate drops from 100% to 43.33% when the white noise increases from 55 dB to 70 dB. Compared with white noise, the human voice has more impact on the performance of EchoAttack. The attack success rate under human voice noises from 58 dB to 70 dB is 18.79% lower than that of the white noise.

The Impact of Bluetooth Noise In practice, other Bluetooth devices can be around the victim and attacker, introducing noises to the Bluetooth feedback system. In this experiment, we put another Bluetooth device 1–5 m away from the victim. Specifically, we use the Pixel Buds as the victim and an iPhone as the noise source. The iPhone keeps playing music through AirPods2.

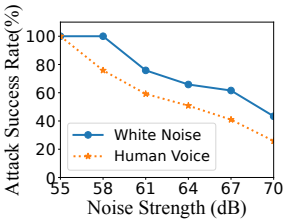


Figure 15: The impact of different noises.

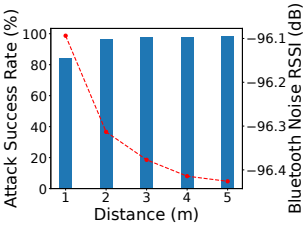


Figure 16: The impact of Bluetooth interference.

Results: The impact of Bluetooth noise levels on the classification accuracy is shown in Figure 16. We can see that the classification accuracy increases as the noise strength becomes weaker. At the shortest distance of 1 m, the average signal strength of the Bluetooth noise reaches -96.09dB, but the model accuracy is still over 83.95%. When is distance is greater than 1 m, the average signal strength of Bluetooth noises is low, which will not affect EchoAttack's performance.

The Impact of Attack Timing When a voice assistant is activated but does not hear a command in a short interval, it will emit an audible notification sound, such as a beep or tone, to indicate it is waiting for a command. Such notification sounds may draw the victim's attention and expose the attack. For example, Apple's Siri emits an "uh-huh" sound. If no further command is recorded during a subsequent time window, it will automatically deactivate itself. In this experiment, we illustrate the timing of EchoAttack and different voice assistant systems to show that EchoAttack can timely detect whether a victim's voice assistant is successfully activated and continuously issue an attack command to disable the notification sound.

We take Siri as an example to illustrate how to measure the timing of a voice assistant, including activation time, notification sound time, and time-out time. We activate the voice assistant by saying "Hey Siri" on an iPhone connected to AirPods 2. We use another iPhone to simultaneously record the screen of the activated iPhone and the audio from us and the AirPods 2. After "Hey Siri" is sent out, EchoAttack issues an attack command if the Bluetooth feedback system detects a successful activation. Then, we use Clipchamp [41], an online video/audio analysis tool, to extract the timing of EchoAttack, including the length of the RSS recording period, data pulling delay, and inference delay of Bluetooth feedback. We follow the same experimental procedure for other voice assistants (Google, Bixby, and Alexa) and earbuds (Pixel Buds, Galaxy Buds, and JBL Buds).

Results: Figure 17 shows the timing of "Hey Siri" and EchoAttack. For Siri, we observe that when the phrase is detected (0.6 s), an activation animation will be played on the phone screen. Then, Siri listens to user commands for another 1.5 s before emitting a notification sound. After 3.4 s timeout period, Siri deactivates itself. For EchoAttack, the RSS recording, the Bluetooth signal sampling [2], and detection take 0.6 s, 0.5 s, and 0.03 s, respectively. After detecting the activation phrase, EchoAttack has 1.27 s to issue an attack command to mute the notification sound without drawing the user's attention.

[2]We pull 9,600 data samples from the 16 slave USB dongles.

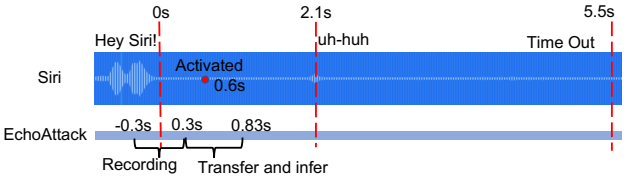


Figure 17: The timing of "Hey Siri" and EchoAttack.

Table 5: The timing of different voice assistants (VA) and earbuds

VA	Siri	Google	Bixby	Alexa
System	iOS 15.6	Android 12	Android 12	Android 12
Smart Phone	iPhone Xs	Galaxy A11	Galaxy S10e	Galaxy S10e
Earbuds	AirPods 2	Pixel Buds	Galaxy Buds	JBL Buds
Activated (s)	0.6	**1.5**	1.0	0.95
Beap (s)	2.1	**0.6**	3.0	—
Timeout (s)	5.5	10.0	5.0	6.0

Table 5 shows the results for other voice assistants (e.g., Google, Bixby, Alexa). For example, Bixby emits a beep to indicate its continued activation 3 s after detecting the activation phrase "Hi Bixby". This interval is long enough for EchoAttack (0.83 s) to issue an attack command without alerting the victim. Alexa does not emit any notification sound. After the activation phrase "Alexa" is spoken, it remains active and listens for commands for 6 s. Google's voice assistant emits a notification sound before detecting the activation phrase. In this case, EchoAttack cannot send the activation phrase and an attack command separately without incurring the victim's attention. Instead, EchoAttack send them simultaneously, so even if the victim notices the attack, it has been executed. However, it delays the Bluetooth feedback processing, increasing the time of the path searching in practice.

7 DISCUSSION

Smart Device Permission Control: Nowadays, the permission control of the voice assistant is used to balance the user experience and the potential security/privacy issues. For example, a smartphone's voice assistant cannot make a regular phone call without unlocking the smartphone. But EchoAttack can still compromise such permission control from three folds [42]. First, the users are allowed to configure the permission control settings. For example, with a configuration, the voice assistant can directly call people in the contacts without unlocking the smartphone. Thus, for those users who lower the permission control level of the voice assistant for easy use, EchoAttack can execute the attack successfully. Second, we can intentionally select victims over peeking in some cases (e.g., bus stop, study area). If we observe the victim's smartphone is unlocked and not in use (e.g., putting on a table), EchoAttack can execute many attacks. Third, even if the smartphone is locked and its permission control is enabled, EchoAttack can still force the device to make emergency calls (e.g., 911), configure the smartphone to airplane mode to disable service, or turn on the flashlight to drain the battery.

8 COUNTERMEASURES

We propose several possible ways to defend EchoAttack in practice.

Non-linearity signal cancellation: Similar to the defense methods [2, 8, 43] for smartphones, we can leverage extra ultra-sound speakers to cancel the attack acoustic signals. However, such approaches will bring extra overhead for users, making them less practical in daily use.

Security-aware Speech Recognition: Some methods [44–47] can train a personalized speech recognition model which can only recognize legal users' commands. This is a practical way without any hardware cost or modification. However, it will increase the computation complexity of the model, resulting in certain delays which might degrade the user experience. Moreover, the replay attack is hard to be completely mitigated with this approach.

Earable authentication: Because of the unique shape of human ears, the earable authentication that utilizes different modalities on earbuds to authenticate the voice command is another possible countermeasure. Liu et al. [48] proposed to use IMU sensors to do continuous authentication as the vibrations are different when a user is speaking. And Gao et al. [49] used the mapping between the in-air voice and the voice captured in the ear canal to authenticate. These two works somehow address the threat of ultrasound attack on earbuds, though they have limitations that need to be addressed. The vibrations while speaking may be affected by head or body motion, and the neural network in [49] is too heavy to run on a smartphone.

9 RELATED WORK

The design of EchoAttack combines the knowledge from several different research areas, including earable computing, ultrasound based privacy and security, and Bluetooth signal sniffing. We summarize some related works from the three categories as follows:

Earable Computing. Earable computing [50] is an emerging research area encompassing speech recognition, augmented reality, motion tracking, and security/authentication. For example, UNIQ [51] personalizes HRTF [52] using earbuds, enabling AR applications. Ear-AR [4] achieves acoustic AR in indoor environments, while EarSense [53] senses teeth-related gestures. EarGate [54] detects user gait patterns, and BioFace-3D [55] develops an earpiece biosensing system for 3D facial reconstruction. FaceSense [56] senses face touch behavior, and EarEcho [57] uses in-ear microphones for biometric authentication. And also some works [58, 59] use smartphone and earbuds to perform hearing test as wearing earbuds for longtime becomes a reason of hearing loss. Also, earbuds can work as a health information platform [60, 61] to monitor heart rate and so on. While these works focus on AR, motion tracking, authentication, health, EchoAttack addresses earbuds privacy/security issues and develops a practical attack system.

Ultrasound based Privacy and Security on Smart Devices. Ultrasound has been used to compromise smart devices by transmitting private or insecure information in various ways [2, 8, 62]. For example, DolphinAttack [8] and BackDoor [2] demonstrate the feasibility of using ultrasound and microphone non-linearity to attack voice assistants on smartphones. LipRead [63] achieves an extended attack range on smart voice-controlled devices. Metamorph [64] proposes a method utilizing ultrasound signals to deceive speech recognition systems. AIC [43] introduces a method to counteract the non-linearity effect at the microphone side to prevent potential attacks. Patronus [65] develops an ultrasound-based jamming system to protect voice recording privacy. However, these works do not consider attacking earbuds, which differ significantly from smartphones and other smart voice-controlled devices. Moreover, all these systems only use the direct path (i.e., the baseline method in the evaluation) to attack the smart devices. In contrast, EchoAttack establishes an efficient earbud-attacking system utilizing available in-direct paths in practice.

Bluetooth Sniffing. BlueEar [66] uses two Bluetooth radios to learn the frequency hopping of Bluetooth communication in the air. BlueSeer [67] and BlueDoor [68] propose method to sniff the BLE (Bluetooth Low Energy) communication. However, since the Bluetooth communication is very short (i.e., less than 1 second) during triggering a voice assistant and the earbuds do not use BLE, these methods cannot be directly applied for Bluetooth sniffing in our attack. Moreover, ZiSense [69] and SoNIC [70] propose to use Zigbee radio to classify Zigbee signals from other coexistent Wi-Fi and Bluetooth signals. ZiFi [71] uses Zigbee radio to detect the existence of Wi-Fi access point. However, these methods do not well address the frequency hopping issues in Bluetooth communication, thus cannot be directly adopted. In contrast, EchoAttack develops a system using 16 Zigbee radios to reliability detect short-period Bluetooth signals with frequency hopping.

10 CONCLUSION

To conclude, we found that the hand-free voice assistant function provided by off-the-shelf earbuds is vulnerable to ultrasound attacks since the position of the earbuds is explicitly known to an attacker. We propose EchoAttack, a practical inaudible earbuds attacking system to activate the voice assistant on smartphones, exposing victims' privacy in danger at offices, bus stops, or gyms. We mainly solve three challenges. First, we observe that the microphone sensing range on earbuds is directional. Thus, we develop a reflection model to establish the best direction of the ultrasound speaker. Second, we develop a harmonic canceling method to improve the quality of non-linearity sound recording with earbuds, thus improving the possibility of a successful attack. Third, we utilize Zigbee radios to design efficient feedback for whether an attack is successful by monitoring the Bluetooth signals of the communication between earbuds and smartphones. We implement EchoAttack with commercial ultrasound speakers and Zigbee dongles. We further conduct extensive experiments to evaluate the efficiency of EchoAttack in four real-world scenarios using six earbuds. The results show EchoAttack achieves 88.1% mean attack success rate across six earbuds in practical scenarios.

11 ACKNOWLEDGEMENTS

We sincerely thank our anonymous reviewers and shepherd for their invaluable feedback, which significantly improved our work. This research was partially funded by the National Science Foundation (NSF) via Award CNS-1909177.

REFERENCES

[1] Nick Hunn. The market for hearable devices 2016–2020. *WiFore Wireless Consulting*, 2016.
[2] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of ACM MobiSys*, 2017.
[3] Guoming Zhang, Xiaoyu Ji, Xinfeng Li, Gang Qu, and Wenyuan Xu. Eararray: Defending against dolphinattack via acoustic attenuation. In *NDSS*, 2021.
[4] Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury. Ear-ar: indoor acoustic augmented reality on earphones. In *Proceedings of ACM MobiCom*, 2020.
[5] Huan Feng, Kassem Fawaz, and Kang G Shin. Continuous authentication for voice assistants. In *Proceedings of ACM MobiCom*, 2017.
[6] Jerad Lewis. Understanding microphone sensitivity. *Analog Dialogue*, 46(2):14–16, 2012.
[7] Jelmer Tiete, F Domínguez, B da Silva, Abdellah Touhafi, and Kris Steenhaut. Mems microphones for wireless applications. In *Wireless MEMS Networks and Applications*, pages 177–195. Elsevier, 2017.
[8] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of ACM SIGSAC*, 2017.
[9] Xinyan Zhou, Xiaoyu Ji, Chen Yan, Jiangyi Deng, and Wenyuan Xu. Nauth: Secure face-to-face device authentication via nonlinearity. In *Proceedings of IEEE INFOCOM*, 2019.
[10] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of ACM ASIA-CCS*, 2020.
[11] Jian Mao, Shishi Zhu, Xuan Dai, Qixiao Lin, and Jianwei Liu. Watchdog: Detecting ultrasonic-based inaudible voice attacks to smart home systems. *IEEE Internet of Things Journal*, 7(9):8025–8035, 2020.
[12] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided waves. In *NDSS Symposium*, 2020.
[13] Yuanda Wang, Hanqing Guo, and Qiben Yan. Ghosttalk: Interactive attack on smartphone voice system through power line. *arXiv preprint arXiv:2202.02585*, 2022.
[14] John Eargle. Microphones: The basic pickup patterns. *Handbook of Recording Engineering*, pages 53–64, 2003.
[15] Apple. Airpods pro (2nd generation) - technical specifications. https://www.apple.com/airpods-pro/specs/.
[16] Douglas Self. *Audio power amplifier design*. Taylor & Francis, 2013.
[17] Vivian Genaro Motti and Kelly Caine. Users' privacy concerns about wearables: impact of form factor, sensors and type of data collected. In *Proceedings of FC*, 2015.
[18] Ke Sun, Chen Chen, and Xinyu Zhang. " alexa, stop spying on me!" speech privacy protection against voice assistants. In *Proceedings of ACM SenSys*, 2020.
[19] Aarthi Easwara Moorthy and Kim-Phuong L Vu. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction*, 31(4):307–335, 2015.
[20] Peng Cheng and Utz Roedig. Personal voice assistant security and privacy—a survey. *Proceedings of the IEEE*, 110(4):476–507, 2022.
[21] Chen Yan, Xiaoyu Ji, Kai Wang, Qinhong Jiang, Zizhi Jin, and Wenyuan Xu. A survey on voice assistant security: Attacks and countermeasures. *ACM Computing Surveys*, 55(4):1–36, 2022.
[22] Yuan Gong and Christian Poellabauer. An overview of vulnerabilities of voice controlled systems. *arXiv preprint arXiv:1803.09156*, 2018.
[23] Voice Maker In. Voice maker. https://voicemaker.in.
[24] Karl F Herzfeld. Reflection of sound. *Physical Review*, 53(11):899, 1938.
[25] Jody Kreiman and Bruce R Gerratt. Perceptual interaction of the harmonic source and noise in voice. *The Journal of the Acoustical Society of America*, 131(1):492–500, 2012.
[26] DW Choi and DA Hutchins. Ultrasonic propagation in various gases at elevated pressures. *Measurement Science and Technology*, 14(6):822, 2003.
[27] John S Rigden. *Physics and the Sound of Music*. 1985.
[28] CUI Devices. Cusa-t80-12-2200-th. https://www.mouser.com/datasheet/2/670/cusa_t80_12_2200_th-2306915.pdf.
[29] Stephen Joe and Frances Y Kuo. Constructing sobol sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30(5):2635–2654, 2008.
[30] WIKIPEDIA. Wideband audio. https://en.wikipedia.org/wiki/Wideband_audio#cite_note-2.
[31] C. Bisdikian. An overview of the bluetooth wireless technology. *IEEE Communications Magazine*, 39(12):86–94, 2001.
[32] Junjie Yin, Zheng Yang, Hao Cao, Tongtong Liu, Zimu Zhou, and Chenshu Wu. A survey on bluetooth 5.0 and mesh: New milestones of iot. *ACM Transactions on Sensor Networks*, 15(3):1–29, 2019.
[33] Sinem Coleri Ergen. Zigbee/ieee 802.15. 4 summary. *UC Berkeley, September*, 10(17):11, 2004.
[34] N. Golmie, O. Rebala, and N. Chevrollier. Bluetooth adaptive frequency hopping and scheduling. In *IEEE MILCOM*, 2003.
[35] Avisoft Biosacoustics. Ultrasonic dynamic speaker vifa. http://www.avisoft.com/playback/vifa/.
[36] KeySight. 33500b and 33600a series trueform waveform generators. https://www.keysight.com/us/en/assets/7018-05928/data-sheets/5992-2572.pdf.
[37] Avisoft. Portable ultrasonic power amplifier. http://www.avisoft.com/playback/power-amplifier/.
[38] AOOIIYSD. 360 tracking phone holder. https://www.amazon.com/Dual-axis-Automatic-Stabilizer-Recording-Streaming/dp/B09KZXQDLL?ref_=ast_sto_dp.
[39] Texas Instruments. Cc2531 usb hardware user's guide. https://www.ti.com/lit/ug/swru221a/swru221a.pdf?ts=1660903911334/.
[40] scipy. Scipy.decisiontree. https://scikit-learn.org/stable/modules/tree.html.
[41] ClipChamp. Clipchamp. https://clipchamp.com/en/.
[42] Hanqing Guo, Yuanda Wang, Nikolay Ivanov, Li Xiao, and Qiben Yan. Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition. In *Proceedings of ACM SIGSAC*, 2022.
[43] Yitao He, Junyu Bian, Xinyu Tong, Zihui Qian, Wei Zhu, Xiaohua Tian, and Xinbing Wang. Canceling inaudible voice commands against voice control systems. In *Proceedings of ACM MobiCom*, 2019.
[44] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proceedings of Interspeech*, 2019.
[45] Hanqing Guo, Chenning Li, Lingkun Li, Zhichao Cao, Qiben Yan, and Li Xiao. Nec: Speaker selective cancellation via neural enhanced ultrasound shadowing. In *Proceedings of IEEE DSN*, 2022.
[46] Hanqing Guo, Qiben Yan, Nikolay Ivanov, Ying Zhu, Li Xiao, and Eric J Hunter. Supervoice: Text-independent speaker verification using ultrasound energy in human speech. In *Proceedings of ACM ASIACCS*, 2022.
[47] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.
[48] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, and Kui Ren. Secure user verification and continuous authentication via earphone imu. *IEEE TMC*, 2022.
[49] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–25, 2021.
[50] Romit Roy Choudhury. Earable computing: A new area to think about. In *Proceedings of ACM HotMobile*, 2021.
[51] Zhijian Yang and Romit Roy Choudhury. Personalizing head related transfer functions for earables. In *Proceedings of ACM SIGCOMM*, 2021.
[52] Corentin Guezenoc and Renaud Seguier. Hrtf individualization: A survey. *arXiv preprint arXiv:2003.06183*, 2020.
[53] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. Earsense: earphones as a teeth activity sensor. In *Proceedings of ACM MobiCom*, 2020.
[54] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. Eargate: gait-based user identification with in-ear microphones. In *Proceedings of ACM MobiCom*, 2021.
[55] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. Bioface-3d: continuous 3d facial reconstruction through lightweight single-ear biosensors. In *Proceedings of ACM MobiCom*, 2021.
[56] Vimal Kakaraparthi, Qijia Shao, Charles J Carver, Tien Pham, Nam Bui, Phuc Nguyen, Xia Zhou, and Tam Vu. Facesense: sensing face touch with an ear-worn system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–27, 2021.
[57] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. Earecho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–24, 2019.
[58] Jessica Barczik and Yula C Serpanos. Accuracy of smartphone self-hearing test applications across frequencies and earphone styles in adults. *American journal of audiology*, 27(4):570–580, 2018.
[59] Justin Chan and Shyamnath Gollakota. Inner-ear cochlea testing with earphones. In *Proceedings of ACM MobiSys*, 2022.
[60] Steven F LeBoeuf, Michael E Aumer, William E Kraus, Johanna L Johnson, and Brian Duscha. Earbud-based sensor for the assessment of energy expenditure, heart rate, and vo2max. *Medicine and science in sports and exercise*, 46(5):1046, 2014.
[61] Assim Boukhayma, Anthony Barison, Serj Haddad, and Antonino Caizzone. Earbud-embedded micro-power mm-sized optical sensor for accurate heart beat monitoring. *IEEE Sensors Journal*, 21(18):19967–19977, 2021.
[62] Guangjing Wang, Yan Qiben, Wang Juexing Shane, Patrarungrong, and Huacheng Zeng. Facer: Contrastive attention based expression recognition via smartphone earpiece speaker. In *Proceedings of IEEE INFOCOM*, 2023.

[63] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *NSDI*, 2018.
[64] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *NDSS Symposium*, 2020.
[65] Lingkun Li, Manni Liu, Yuguang Yao, Fan Dang, Zhichao Cao, and Yunhao Liu. Patronus: Preventing unauthorized speech recordings with support for selective unscrambling. In *Proceedings of ACM SenSys*, 2020.
[66] Wahhab Albazrqaoe, Jun Huang, and Guoliang Xing. Practical bluetooth traffic sniffing: Systems and privacy implications. In *Proceedings of ACM MobiSys*, 2016.
[67] Tsvetelina Mladenova. Open-source erp systems: an overview. In *Proceedings of IEEE ICAI*, 2020.
[68] Jiliang Wang, Feng Hu, Ye Zhou, Yunhao Liu, Hanyi Zhang, and Zhe Liu. Bluedoor: breaking the secure information flow via ble vulnerability. In *Proceedings of ACM MobiSys*, 2020.
[69] Xiaolong Zheng, Zhichao Cao, Jiliang Wang, Yuan He, and Yunhao Liu. Zisense: towards interference resilient duty cycling in wireless sensor networks. In *Proceedings of ACM SenSys*, 2014.
[70] Frederik Hermans, Olof Rensfelt, Thiemo Voigt, Edith Ngai, Lars-Åke Nordén, and Per Gunningberg. Sonic: Classifying interference in 802.15. 4 sensor networks. In *Proceedings of ACM IPSN*, 2013.
[71] Ruogu Zhou, Yongping Xiong, Guoliang Xing, Limin Sun, and Jian Ma. Zifi: Wireless lan discovery via zigbee interference signatures. In *Proceedings of ACM MobiCom*, 2010.