Eavesdropping on Black-box Mobile Devices via Audio Amplifier's EMR

Huiling Chen[1], Wenqiang Jin[1*], Yupeng Hu[1*], Zhenyu Ning[1], Kenli Li[1,2], Zheng Qin[1], Mingxing Duan[1,2], Yong Xie[3], Daibo Liu[1], and Ming Li[4]
[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]National Supercomputing Center in Changsha, Hunan University, Changsha, China
[3]Nanjing University of Posts and Telecommunications, Nanjing, China
[4]The University of Texas at Arlington, Arlington, Texas, USA
{chenhl, wqjin, yphu, zning, lkl, zqin, duanmingxing}@hnu.edu.cn,
yongxie@njupt.edu.cn, dbliu.sky@gmail.com, ming.li@uta.edu

***Abstract*—Audio eavesdropping poses serious threats to user privacy in daily mobile usage scenarios such as phone calls, voice messaging, and confidential meetings. Headphones are thus favored by mobile users as it provide physical sound isolation to protect audio privacy. However, our paper presents the first proof-of-concept system, *Periscope*, that demonstrates the vulnerabilities of headphone-plugged mobile devices. The system shows that unintentionally leaked electromagnetic radiations (EMR) from mobile devices' audio amplifiers can be exploited as an effective side-channel in recovering victim's audio sounds. Additionally, plugged headphones act as antennas that enhance the EMR strengths, making them easily measurable at long distances. Our feasibility studies and hardware analysis further reveal that EMRs are highly correlated with the device's audio inputs but suffer from signal distortions and ambient noises, making recovering audio sounds extremely challenging. To address this challenge, we develop signal processing techniques with a spectrogram clustering scheme that clears noises and distortions, enabling EMRs to be converted back to audio sounds. Our attack prototype, comparable in size to hidden voice recorders, successfully recovers victims' private audio sounds with a word error rate (WER) as low as 7.44% across 11 mobile devices and 6 headphones. The recovery results are recognizable to natural human hearing and online speech-to-text tools, and the system is robust against a wide range of attack scenario changes. We also reported the *Periscope* to 6 leading mobile manufacturers.**

I. Introduction

Audio-driven applications (e.g., Skype, Zoom, Teams, and Phone Calls) provide comprehensive services enriching our daily experiences. By forecasting, the audio-driven application market is about to reach $2.84 billion by 2029 [1]. While enjoying high-quality audio services, audio eavesdropping has emerged as a preeminent threat to both personal privacy and commercial secrets. The scope of this phenomenon is far-reaching and has been projected to have catastrophic consequences, as the leakage of trade secrets alone is anticipated to result in annual losses exceeding billions of dollars [2]. In light of the potential risks associated with audio leakage, many people opt to connect their mobile devices with headphones[1], which provide a physical barrier between the user and the external environment, thereby rendering it challenging for surreptitious voice recorders to eavesdrop. Though using headphones seems to be a desired way of securing audio privacy, it is still essential to review the vulnerability of headphone-plugged mobiles and take a deep investigation of whether such audio systems are secure enough in defending against novel audio eavesdropping attacks.

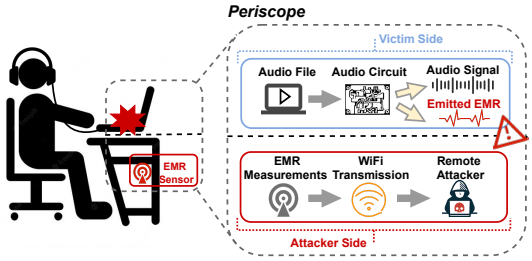


Fig. 1. *Periscope* exploits the EMR of mobile devices to recover the victim's private audio in daily headphone usage scenarios. We name our attack as *Periscope* as it can passively collect sensitive information at hidden places just like the periscopes deployed on submarines.

Eavesdropping on mobile devices secured by headphones poses harsh requirements to the attackers, especially since users may carry their headphones in a wide spectrum of usage scenarios. Thus, the attacker should carry a small-sized eavesdropping device that can be easily deployed in arbitrary hidden places, such as a nearby bush, underneath a table, or in a pocket, like the classic hidden voice recorders. Meanwhile, audio eavesdropping is essentially a *passive* process in which the attacker can only analyze the side-channel leakages for recovering audio. However, headphones are low-powered and the generated sounds are mostly restricted to the human ear, barely having sound leakages and hardly measurable especially under low volumes or in noisy backgrounds [3], [4]. These challenges render the state-of-the-art audio eavesdropping attacks, primarily targeting against human vocal voices [5]–[8] and loudspeakers [5], [9]–[20], inapplicable to headphone-plugged mobiles. Specifically, both human vocals

*Wenqiang Jin and Yupeng Hu are the corresponding authors.

[1]In this work, we refer the earphones and headsets as a unique term, i.e., headphone.

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.23240
www.ndss-symposium.org

and loudspeakers have loud acoustic sound leakages. Thus, to eavesdrop on users' private audio, these attacks analyze the vibrations of the audio source or its surrounding objects caused by sounds' airborne pressures. However, headphone-plugged devices barely have measurable acoustic leakages.

Alternatively, electromagnetic radiation (EMR) emitted by electronic devices was considered as an effective side-channel for covertly extracting secret information. For example, recent studies have demonstrated that EMR can be exploited to recover secret AES-128 keys [21], [22], screen contents [23], and users' keystrokes [24], [25]. These exploited EMR signals share favorable proprieties [26], [27] such as omnidirectional, obstacle-penetrable, and acoustic noise-resist. Motivated by these evidences, we present *Periscope*, the first proof-of-concept system that investigates the feasibility of exploiting EMR signals radiated from headphone-plugged mobiles to eavesdrop on users' private audio. Specially, we select wired headphones as our target since they are commonly considered more robust against eavesdropping attacks than wireless ones due to the use of physical wires to transmit audio signals. In contrast, wireless headphones transmit audio data via Bluetooth packets, which can be easily intercepted and decrypted by malicious attackers to obtain the underlying audio contents [28], [29]. To achieve the audio eavesdropping goal, we need to answer the following research questions (RQ).

- RQ1: How well the captured EMR can reflect audio sounds processed on a headphone-plugged mobile device?
- RQ2: What is the EMR source inside the device's audio circuits? And, how do the plugged headphones impact the radiations?
- RQ3: How to effectively recover intelligible audio sounds from an arbitrary device's EMR, especially when the attacker has no prior knowledge regarding the device in advance of the attack and its EMRs typically have low signal-to-noise ratios (SNRs)?

We conduct a feasibility study to investigate RQ1&2. To investigate the feasibility of recovering audio sounds from EMR, we first analyze the similarities between those two signals. The results show high signal cross-correlations across a wide range of test cases, i.e., different devices, audio types, and frequency ranges. To further locate the source of EMR, we conduct tests using sinuous test sounds with specific frequencies played through headphone-plugged mobile devices. The results show that the EMR signals emitted by the devices contained not only the played sounds' frequencies but also their cross-products and harmonics in the signal spectrum. Further analysis reveal that the primary source of EMR is the device's audio amplifier, and the unique spectrum characteristics are caused by signal distortions in the amplifier, which are recognized as a featured effect described in its datasheets. To investigate the impact of headphones on EMRs, we conducted controlled tests to compare the radiation spectrograms of the device when using headphones and when not using them, respectively. Our spectrum analysis suggests that the headphone's body parts, such as the wire and microspeaker, act as an antenna, enhancing the radiation strength. Overall, our feasibility study provides comprehensive evidences that recovering audio sounds from EMR is a viable approach, and we have identified the key source and characteristics of EMR that serve as the foundations for our recovery process.

Nevertheless, leveraging the device's EMR for audio eavesdropping is non-trivial. Especially, the radiated EMRs are susceptible to signal distortions from the audio amplifier and can be easily polluted by electromagnetic (EM) interference generated from nearby electronics. As a result, it is challenging to characterize the mapping relationship between captured EMRs and their corresponding original audio. Prior eavesdropping works [5], [9]–[13], [30]–[33] resolve the similar problem by employing machine/deep learning models. However, this approach requires collecting a substantial amount of training data from the target in advance of the attack, which limits its practicability in real-world deployments and renders the target as a "white-box".

Our goal is to develop an audio eavesdropping attack that can overcome these challenges without requiring prior knowledge of the target. Specifically, we assume the target to be a "black-box" meaning that the attacker has no information about the target's hardware specifics or access to any training data for preparing a learning-driven model (RQ3). As established in our feasibility study for answering RQ1&2, the EMRs radiated by mobile devices' audio amplifiers are primarily correlated with the original audio sounds and share a unique EMR spectrum characteristic for all amplifiers. Therefore, unlike previous EMR-based side-channel attacks [21], [22], [34], [35], we do not need knowledge of the device's hardware specifics or EMR characteristics to carry out our attack. Our focus, therefore, is to develop a training-free audio recovery scheme based solely on signal processing techniques. In particular, we first distill "useful" signals using bandpass filters and normalize the results to achieve higher signal SNR. The Discrete Wavelet Transform (DWT) [36] is then leveraged to recursively remove fast-varying ambient EM noises at each decomposed EMR signal level using a dynamically calculated threshold. We then develop a clustering-based scheme to clear signal distortions in the measured EMR signal's spectrogram, allowing us to extract signal components corresponding to the audio sound. Finally, the processed EMR signals are directly converted to the audio track to accomplish audio recovery. To assess the effectiveness of the proposed attack, we build a prototype using a 5.1cm-length ESP32 board [37] and Raspberry Pi 4, which cost approximately $85 and measures EMR signals at a distance as far as 1.05m. Benefiting from its properties of being small in size and long measurement distance, the attacker can easily deploy it at hidden places without the victim's notice, as shown in Fig. 1. By exploiting Wi-Fi transmission capability provided by Raspberry Pi, the attacker is able to remotely receive the device's EMRs and perform the audio recovery at 15m away. We conducted experiments on 11 different models of mobile devices and 6 models of headphones. Our results indicate that the proposed *Periscope* attack is robust against environmental dynamics, transparent to the attacker's eavesdropping angles, and remains effective even when launched behind obstacles. Furthermore, we evaluate the audio recovery performance comprehensively under diverse attack settings. The results demonstrate that our recovered speeches are intelligible to both human hearing and online recognition tools. Specifically, in an office room setting, the proposed attack achieves an average word error rate (WER) as low as 7.44% using the Microsoft speech-to-text API [38].

We report the threat to 6 leading mobile manufacturers including Apple, Lenovo, Huawei, Vivo, OPPO and Dell. Moreover, we also provide the countermeasures to mitigate the Periscope attack. As of write, Huawei has reproduced the proposed attack and plans to fix the issue by improving its hardware designs. We believe the Periscope attack may impact a wide range of mobile devices beyond those of reported manufacturers and filled a vulnerability report to CVE [39] and CNVD [40]. The contributions of this paper are summarized as follows.

- We demonstrate an alarming threat of leaking users' private audio contents via EMRs radiated from the mobile device's audio amplifier.
- We characterize the audio amplifier under audio processing tasks and perform feasibility studies to show that the amplifier's EMRs reflect audio inputs. Based on the analytic results, we design a "black-box" oriented audio recovery scheme to recover the audio sounds, which requires no prior knowledge of the target.
- We develop a prototype and demonstrate the severity of the threat on 11 mobile devices and 6 headphones. The proposed threats have been confirmed by a leading mobile manufacturer.

II. Related Work

Audio Eavesdropping Attacks. Prior works reveal the emerging threats of eavesdropping on human speeches [5]–[8] or loudspeakers [5], [9]–[20]. WiHear [5] shows that Wi-Fi signal can be affected by users' mouth and throat movements. In order to reconstruct the victim's speeches, they collect a considerable size of training data prior to the attack and map the spoken words with Wi-Fi signal measurements using machine learning models. In [6], KWong et al. propose to turn a magnetic hard drive into an acoustic microphone. They show that human speech can be restored by measuring the offset of the hard drive's read/write heads. SPEAKE(a)R [7] develops a malware that turns the speaker into an acoustic microphone. VibraPhone [8] measures the electromotive forces of a vibromotor in the smartphone to reconstruct human speeches. However, unlike human vocals, mobile devices plugged with headphones barely have significant sound leakages. Thus, these prior attacks are inapplicable to eavesdropping on private audios played via headphones.

On the other hand, another line of research demonstrates that audio sounds played via loudspeakers can be eavesdropped through various non-acoustic side channels. Smartphone's motion sensors, e.g., gyroscope [9], [33], and accelerometers [10]–[13], have good sensitivity in measuring subtle vibrations caused by the speaker's acoustic sounds. These attacks assume that the sensor readings from the victim device are accessible to the attacker by pre-installing malware. Davis et al. [14] exploits a 2200 FPS high-speed camera to measure the micro-vibrations of an object (e.g., tissue or teapot) near the loudspeaker. Lamphone [15] utilizes a telescope and photodiode sensor to observe a hanging light blub in the victim's room, which has micro-vibrations on its surface while the loudspeaker is playing acoustic sounds. These attacks recover the victim's sounds based on the fluctuations in their measured micro-vibrations. High-frequency radio signals, such as laser beam [16], software-defined radio [17], LiDAR [18], Wi-Fi channel state information (CSI) [5], mmEve [41], and RFID signals [19], are also exploited to reconstruct the loudspeaker's sounds. These attacks analyze the perturbation of the radio signals induced by airborne acoustic pressures. Note that high-frequency radio signals are susceptible to environmental dynamics. Hence, these attacks do not work with dynamic backgrounds, e.g., people walking by in the background. Data-driven training models are usually adopted to extract meaningful audio content. In this paper, we exploit the device's EMRs that have ultra-low frequency ranges, i.e., $\leq 5kHz$. Thus, it is more robust against background dynamics. MagEar [20] eavesdrops on audio sounds by leveraging near-field magnetic fluxes of speakers' voice coils. It requires a high volume of 80dB to achieve meaningful results. However, such sound levels can cause hearing damage, as indicated by CDC guidelines [42]. Moreover, it only achieves acceptable performance within a narrower range, i.e., $0° - 20°$, since speakers' magnetic fluxes are directional. In contrast, we leveraged the audio amplifier's EMRs which are omnidirectional radiated.

EMR-based Side-Channel Attacks. Electronic devices unavoidably generate EMRs while they are functioning. Van Eck [23] first shows that screen contents of a cathode ray tube (CRT) display can be reconstructed by analyzing its emitted EMRs received from a $15 TV receiver. It attracts many researchers' attentions of investigating the system vulnerabilities caused by leaking EMRs. Following this line of researches, EMRs have been exploited to infer victim's keystrokes on keyboards [24], [25], [43], profile device memory usages [44], identify the model of LCD monitors [34], recover the displayed information on mobile screens [35], and exfiltrate secret data by establishing electromagnetic covert channel [45]–[47]. In addition, the most recent studies show that fine-grained data processed on the device also have the leakage threats via EMRs. Screaming channels [21] shows that devices' secret AES-128 keys can be recovered from the coupled EMR signals radiated by mixed-SOC chips. Wang et al. [22] further extend the attack distance to 15m with the assistance of deep learning models. Cihan et. al. [48] shows the information printed from a laser printer can also be reconstructed via its EMRs. Tempest Comeback [49] indicates that audio processed by wireless devices can be extracted from the coupled EMR of the mixed-SOC chip. These prior works are parallel to our study.

III. Attack Model

Attack Scenario. The attacker seeks to recover audio sounds played on the victim's headphone-plugged devices by leveraging a miniaturized hidden EMR sensory device. Specifically, in the attack scenario, the victim uses headphones to setup a private conversation with her friends via Skype calls in a public space (e.g., subway cabin) or hold a confidential meeting at workplaces (e.g., office room). The attacker places the hidden EMR sensory device, e.g., underneath a table, in a bush nearby, or in an adjacent people's pocket, to stealthily eavesdrop on the victim's private audio. Such scenarios can be commonly found in daily life. For example, people often share the same table or room space in a public place like an office, library, or cafe. During rush hours, public transportation and elevators are usually crowded with passengers, and people inevitably come close to the others. The attacker can easily

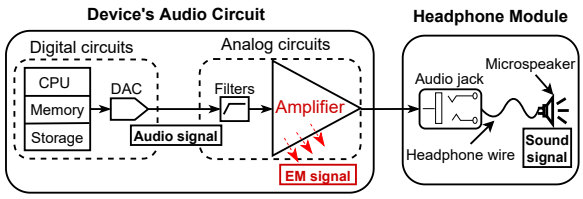


Fig. 2. The workflow of mobile device's audio circuit when connected with a wired headphone.

find a hidden place at the victim's nearby to deploy the eavesdropping device.

Attacker's Capability Limits. We assume that the attacker has no prior knowledge about the target device, including the device model, type, and hardware specifics. This is because, in a real-world attack scenario, the victim user may show up at any random time window and stay for only a short time period. It is barely possible for the attacker to learn well-detailed prior knowledge about its target. Therefore, the attacker cannot acquire the target device's hardware specifics or training datasets for facilitating the audio recovery process. Such assumptions treat the victim device as a "black-box". To launch a stealthy attack, the attacker is also assumed to avoid the usage of active measurement signals, such as LiDAR [18], mmWave [50], and WiFi [17].

EMR Sensory Device. We assume that the hidden EMR sensory device has enough storage to record the EMR readings locally. Meanwhile, it also has the wireless connectivity, e.g., Bluetooth or Wi-Fi, to forward the EMR readings to a remote PC for real-time audio recovery. Such that, the attack distance can be further extended.

IV. Preliminaries

In this section, we will discuss the working principle of mobile devices' audio processing circuits and investigate the relationship between their input audio and EMRs.

A. Working Principle of Device's Audio Circuits

There are two major audio electronic modules residing on the mobile device's printed circuit board (PCB) substrate including the digital circuits and analog circuits, as shown in Fig. 2.

The digital circuits mainly consist of CPU, memory, storage, and digital-to-analog converters (DAC). When audio-driven applications, like Skype, Teams, Zoom, and phone calls, run on the audio player's CPU, its underlying digital circuits are energized to decode the streamed audio files as digital signals by letting the transistors switch between low and high-voltage values to represent the 1s and 0s. This digitized audio signal is then passed through the DAC, which converts the switching digital signal into an alternating analog current flow.

The analog circuits mainly consist of filters and amplifiers. The audio processing chip utilizes a combination of low-pass filters (LPF) and high-pass filters (HPF) to remove circuitry white noises. Then, the denoised audio sound is passed through an amplifier to strengthen the signal amplitudes via automatic gain controls (AGC). Finally, the energized audio current flows through the headphone and powers its voice coil to vibrate

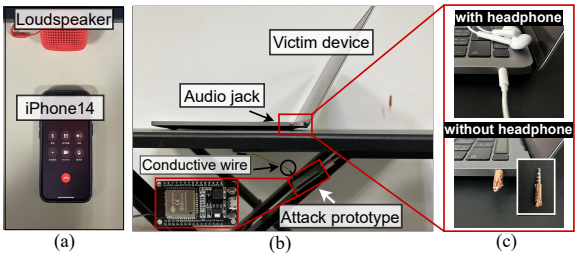


Fig. 3. We use iPhone 14 to call the target devices via Skype. The EMR sensory prototype collects the victim device' EMRs when it is processing the audio.

and generate audible sounds. Following Maxwell's equation and Lorentz force law [51], the intense fluctuations of audio currents will induce time-variant electromagnetic fields and continuously emit EMRs to the open space.

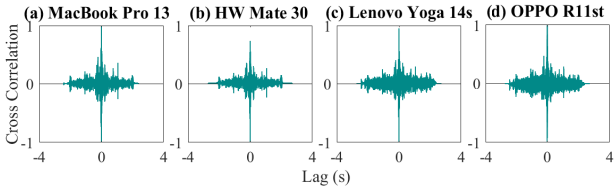


Fig. 4. Sharp peaks appear when $lag = 0$ indicating the strong correlations between the original audios and the device's EMR measures.

B. Can EMRs Reflect the Device's Audio Sounds?

We conduct a series of feasibility studies to characterize the EMRs and validate the feasibility of exploiting this side-channel for audio eavesdropping.

Experiment Setups. EMR can result in electric potential changes on a conductor within close proximity. As shown in Fig. 3, we build the sensory prototype using an ESP32 [37]. The analog input pin of the ESP32 is connected to a conductive wire to sense the electric potential changes caused by EMRs. The prototype is set with the highest measurement resolution, 12 bits, to ensure sensitivity to capture weak EMRs. The system samples the received signal with an A/D converter at a rate of 10k samples/sec, which is sufficient to recover comprehensible audio contents as suggested by previous works [9], [10], [12]. The ESP32 costs less than $10 and is only 5.1cm long, making it easy to conceal in unnoticeable places.

Correlation Analysis. To evaluate the feasibility of using EMRs for audio eavesdropping, we conducted a study to investigate the correlation between a device's EMRs and its audio inputs. For this study, we select four different mobile devices with varying models and shapes: Apple MacBook Pro 13, HW Mate 30, Lenovo Yoga 14s, and OPPO R11st. We set up voice calls using Skype between a remote caller and the target devices, with the victim plugging in HW AM115 headphones to prevent others from overhearing the conversation. As shown in Fig. 3(a), the remote caller plays a speech audio recording from the Harvard speech corpus [52] near her smartphone, while our attack prototype as shown in Fig. 3(b), is placed under the office table to collect real-time EMR emissions from the target device. We then calculate the cross-correlations $C(\tau)$ [53] between the EMR readings $E(t)$ and the device's audio $S(t)$ to quantify their similarity.

$$C(\tau) = \int_{-\infty}^{+\infty} S(t)E(t-\tau)d\tau, \quad (1)$$

in which τ denotes the time displacements between two signals. Large values of $C(\tau)$ represent the two given signals that are highly correlated. Since we have aligned the $E(t)$ and $S(t)$, the peak correlation value should exist when $\tau = 0$, if the two signals are correlated. Otherwise, $C(\tau)$ will be a flat curve with no significant peaks. Fig. 4 shows the normalized calculation results of $C(\tau)$. It can be seen that, for all tested devices, **the device's audio sounds and its corresponding EMR measures are significantly correlated. The results confirm the existence of audio leakages threat via devices' EMRs.**

Frequency Range of EMR Measurements. Human speech can cover a broad frequency spectrum. Therefore, we further investigate how well the EMR measurements can cover the audio frequency range to perform an effective side-channel attack. We keep the above settings unchanged and let the Skype caller play a chirp signal that linearly sweeps from the minimum frequency f_l to the maximum frequency f_h over the time duration T. Specifically, the chirp audio is:

$$S(t) = cos(\pi \frac{B}{T} t^2 + 2\pi f_l t), \tag{2}$$

where $B = f_h - f_l, t \in [0, T]$. We set $f_l = 0Hz$, $f_h = 6kHz$, and $T = 6sec$, respectively. Fig. 5 shows the EMR measurements of the four target devices. We find that the attack prototype can restore audio signals in the frequency range of 0Hz to 5kHz. It is noteworthy that English speeches consist of vowel and consonant sounds [15], which have a frequency range of 85Hz to 255Hz, and 2kHz to 4kHz, respectively. In addition, the intelligibility of human speeches is mainly determined by their low-frequency components [8]. Therefore, **EMR measurements with a frequency band of $[0Hz, 5kHz]$ is sufficient to capture abundant information of the victim's private speech audios.**

Furthermore, we observed additional frequency components in the EMR spectrogram (indicated by white arrows in Fig. 5) along with the chirp signal. These components were harmonic and cross-product[2] noises of the chirp signal and were crucial features indicating that the EMRs were generated from the device's audio amplifier. We will provide a detailed analysis of these noises and explain their root causes in Section IV-D. Moreover, as shown in Fig. 5, these signal distortions could damage the spectrum structure of the audio sounds, leading to reduced speech intelligibility. In Section V-B, we will employ advanced signal processing techniques to eliminate them.

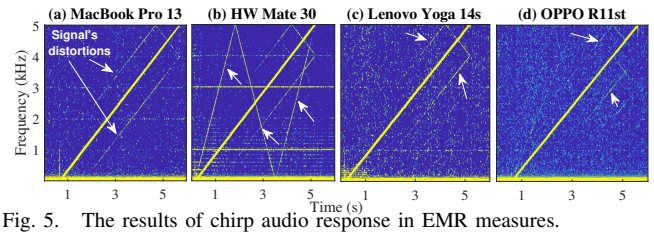


Fig. 5. The results of chirp audio response in EMR measures.

[2]Assuming the signal's fundamental frequency as f_1 and f_2, harmonics are multiple times of the fundamental frequency components, i.e., $N \times f_1$ and $N \times f_2$. Cross-products are multiplicative combinations of fundamental frequency components, i.e., $Nf_1 \pm Mf_2$. N and M are integer numbers.

C. How Do the Device's Headphones Impact EMRs?

Our previous measurements have demonstrated the presence of audio leakages through EMRs from devices. We further investigate the role of headphones in the radiation process. To remove the influence of headphones while maintaining the functionality of the laptop device's audio circuit, we cut off the headphone wire and covered the audio jack plug with copper to eliminate any potential EMRs associated with headphones[3], as shown in Fig. 3(c). We conduct the test using a MacBook Pro 13 laptop and have a Skype caller play an audio speech, "See you tonight at ten."

Fig. 6(a) illustrates the spectrogram of the original input audio, while Fig. 6(b)-(d) compare the measured EMR spectrograms of the MacBook Pro 13 laptop connected with/without HW AM115 headphones. Fig. 6(b) displays the measured EMR spectrogram when the remote user is silent and Skype is not transmitting audio content. Since the audio circuits are not working, no EMRs are produced. Fig. 6(c) displays the EMR spectrogram when the laptop is connected to a pair of headphones, and the victim listens to the speech on the Skype call. Comparing Fig. 6(c) with Fig. 6(b), there are groups of EMR signal components with intense magnitudes representing each word of the speech sentence, which matches against the signal components of the original audio spectrogram in Fig. 6(a). Surprisingly, a similar EMR spectrogram is also observed in Fig. 6(d), where the laptop does not have functional headphones, as shown in Fig. 3(c). Therefore, we identify that the source of EMRs is the mobile device. Additionally, we find that the EMR magnitudes are weaker compared to those measured (Fig. 6(c)) when the mobile device is plugged with a pair of functional headphones. The primary reason is that **the headphone body (including a conductive wire and the microspeaker part) can serve as an antenna and enhance the device's EMR strengths**. As a result, the attacker can eavesdrop on EMR leakages with higher signal-to-noise ratios (SNRs), making the proposed attack even more severe as the victim's audio contents can be restored at a longer distance.

D. What is the EMR Source inside the Device's Audio Circuits?

After identifying the EMRs originating from mobile devices, the focus of this part is to locate the EMR source within the device. To facilitate the analysis, we use a HW MateBook D14 laptop to play an audio sound $S(t)$ composed of two simple frequency tones, specifically $S(t) = cos(2\pi f_1 t) + cos(2\pi f_2 t)$, where $f_1 = 130Hz$ and $f_2 = 110Hz$. The frequency spectrum of its EMRs received at the attack prototype is shown in Fig. 7. In general, EMRs are expected to maintain linearity with respect to the input audio signal $S(t)$, however, additional cross-product frequency components, such as 150Hz, 410Hz, and 610Hz, are observed in addition to the original frequency components (110Hz and 130 Hz). To further investigate this phenomenon, the experiment is repeated with a Lenovo Yoga 14s laptop playing a single tone at 1kHz. In Fig. 8, harmonic frequency components, i.e., 3kHz, are observed in its EMR spectrogram.

[3]It is worth noting that unplugging the headphones would switch the laptop to speaker mode, generating acoustic sounds, whereas turning off the laptop speaker would deactivate the audio circuits. To avoid this, we remove the headphone body but leave its plug to keep the audio circuit functional.

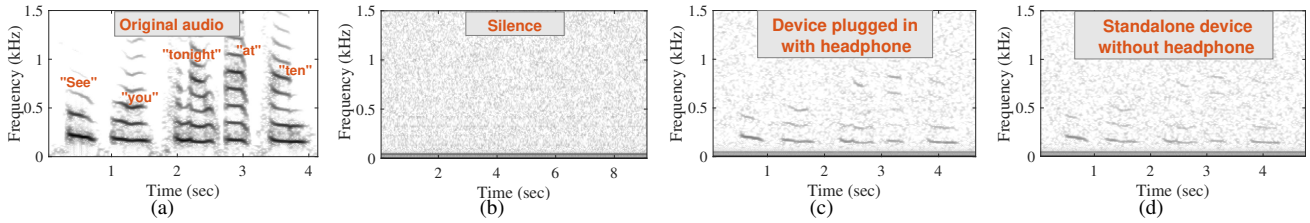
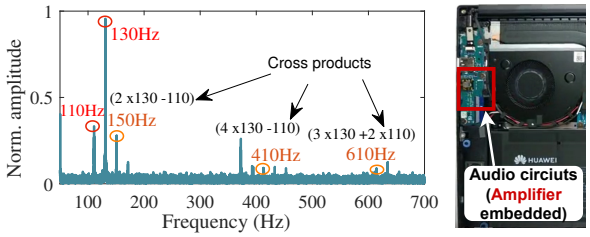
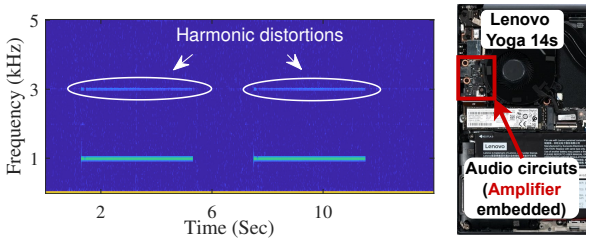


Fig. 6. Comparisons of EMR measurements when the device is connected w/o headphones. (a) Spectrogram of the original audio. (b) EMR measures when the device is in silence. (c) EMR measures of the device plugged in with headphones. (d) EMR measures of the standalone device.



(a) EMR measures show cross-products. (b) Device PCB substrate.

Fig. 7. Amplifier induces cross-products which are traced in EMR measures. (tested on MateBook D14).



(a) EMR measures show harmonic distortions. (b) Device PCB substrate.

Fig. 8. Amplifier induces harmonics. The 1kHz fundamental frequency results in the 3kHz harmonic. (tested on Lenovo Yoga14s).

It is well known that the nonlinearity between circuit inputs and outputs is an important property of amplifiers [54]–[60]. Amplifiers are capable of creating new frequencies as harmonics and cross products. Furthermore, these signal distortions are fundamental electrical characteristics that are described in the majority of audio amplifiers' datasheets, such as MAX98307 [61], LM4844 [62], TPA6166A2 [63], LME49710 [64], and OPA627 [65]. Additionally, we note that mobile devices' CPUs operate at clocks ranging from several GHz. Prior research [66] has also shown that EMRs emitted from ADCs fall within the frequency range of 3.2MHz to 24.3MHz. However, our measured EMR frequencies are in the range of a few KHz, directly correlated with audio processing. This suggests that neither the CPUs nor the ADCs are the EMR sources.

Based on the observations above, **it is inferred that the EMR signals are mainly derived from the mobile device's audio amplifier, and they will exist whenever the device is processing acoustic sounds.**

It is noteworthy that the integration of an amplifier in the audio circuits of mobile devices is essential for powering headphones and generating loud acoustic sounds [67]–[69]. To better understand how the audio amplifier's nonlinearity shapes the device's EMRs, we conducted a detailed investigation and developed a model based on existing literature [70], [71]. Specifically, we modeled the audio inputs of an amplifier and its corresponding outputs as follows. Let the input signal be denoted as $S(t)$, and the output signal be denoted as $S_o(t)$:

$$S_o(t) = AS(t) + B_1(S(t))^2 + ... + B_{n-1}(S(t))^n, \quad (3)$$

The gain for the input audio sounds is denoted as A, while the gain for high-power terms is represented by B. Linear electrical components show a linear correlation between their input and output signals, with only an attenuation in amplitude. However, as seen in Eq. (3), amplifiers not only increase the signal amplitudes but also introduce high-power terms (undesirable signal distortions). For example, if the input audio signal is a simple sum of two sinuous waves with frequencies f_1 and f_2, given by $S(t) = cos(2\pi f_1 t) + cos(2\pi f_2 t)$, the output $S_o(t)$ can be calculated as:

$$\begin{aligned} S_o(t) &= AS(t) + B_1(S(t))^2 + ... + B_{n-1}(S(t))^n \\ &= A[cos(2\pi f_1 t) + cos(2\pi f_2 t)] + B_1[cos^2(2\pi f_1 t) + \\ &\quad cos^2(2\pi f_2 t) + 2cos(2\pi f_1 t)cos(2\pi f_2 t)] + ..., \end{aligned}$$

in which the quadratic term generates new frequency components of $2f_1$, $2f_2$, $f_1 - f_2$, and $f_1 + f_2$. Further expansions of the high-power terms $B_{n-1}(S(t))^n$ reveal that non-linearity electronics result in a serial combination of the input frequencies as $N_1 f_1 \pm N_2 f_2$, $N_1, N_2 \in \mathbb{Z}$. Thus, the corresponding EMRs will also contain distortions, i.e., harmonics (e.g., $2f_1$ and $2f_2$) and cross products (e.g., $f_1 + f_2$ and $f_1 - f_2$), in addition to the audio's fundamental frequency components (e.g., f_1 and f_2). Therefore, assuming the input audio has a frequency spectrum in the range of $[f_b, f_t]$, the corresponding signal distortions can be expressed as $N_b f_b \pm ... N_i f_i \pm ... \pm N_t f_t$, where $f_i \in [f_b, f_t]$ and $N_i \in \mathbb{Z}$. This explains why the observed EMRs contain signal distortions, as shown in the Fig. 5-8, and why these distortions change the original spectrum of the audio signals, making it challenging to recover the original sound.

V. Audio Recovery

This section presents our technique details of recovering audio contents from mobile devices' EMRs. Leveraging the machine/deep learning models to map devices' side-channel signals with their original sounds could be an effective approach. In fact, such techniques are widely adopted by prior eavesdropping attacks [5], [9]–[12], [18], [19], [33], [50]. However, in real-world scenarios, the attacker can hardly have the chances to collect a large amount of training data from the target device in advance of the attack. Especially, the victim shows up at any random time-window and stays for uncertainty short time. There does not have sufficient time for the attacker to perform the data collection and prepare the attack. Therefore, we aim to develop the *Periscope* as a "black-box" attack. In such an attack, the attacker has no prior knowledge

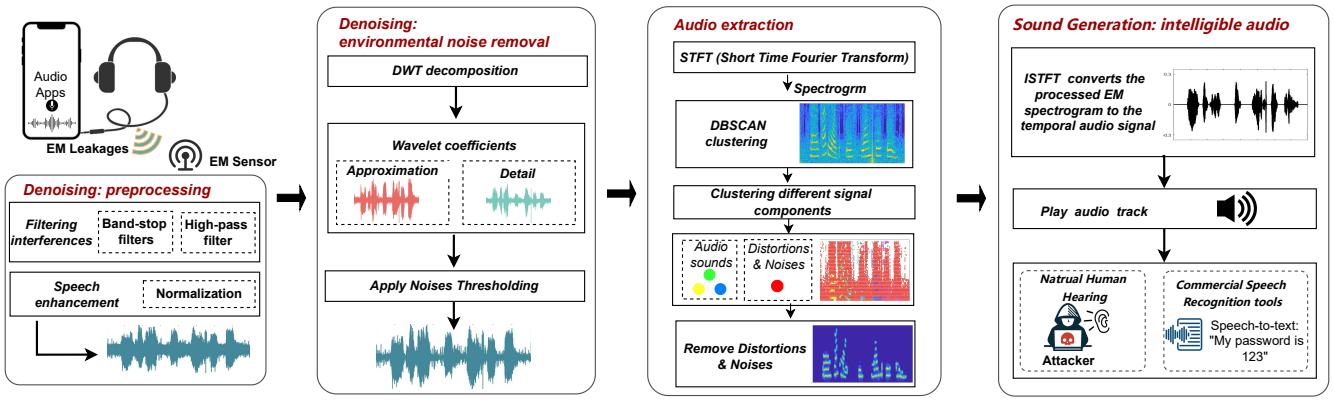


Fig. 9. Framework of Audio Recovery.

about the target device, including its hardware specifics and pre-measurement data. This is a crucial consideration since the attacker's lack of knowledge is a hallmark of real-world attacks.

As evidenced in Fig. 4-5, EMRs radiated from the amplifiers of different mobile devices exhibit similar signal patterns and are highly correlated with the corresponding audio inputs. This observation leads us to conclude that device variations do not significantly affect the EMR measurements. These side-channel signals are mainly associated with the audio inputs. Consequently, the attacker can launch the attack without prior knowledge of the target's audio hardware specifics. To complete a "black-box" attack, we also design our audio recovery scheme solely based on signal processing techniques to avoid the training hurdles that typically exist in prior works. As illustrated in Fig. 9, the audio recovery process consists of three main procedures, namely, *denoising*, *audio extraction*, and *sound generation*. The *denoising* procedure is responsible for eliminating environmental noise from the EMR measurements to enhance the signal's SNR. The *audio extraction* procedure isolates the fundamental audio components from the signal distortions induced by the amplifiers. Finally, the *sound generation* step takes the processed EMR signal and converts it into audio sounds. Fig. 10 presents the intermediate results of the EMR measures processed after each procedure. These procedures are critical in ensuring the successful recovery of audio signals from EMR, even in the absence of training data about the target device.

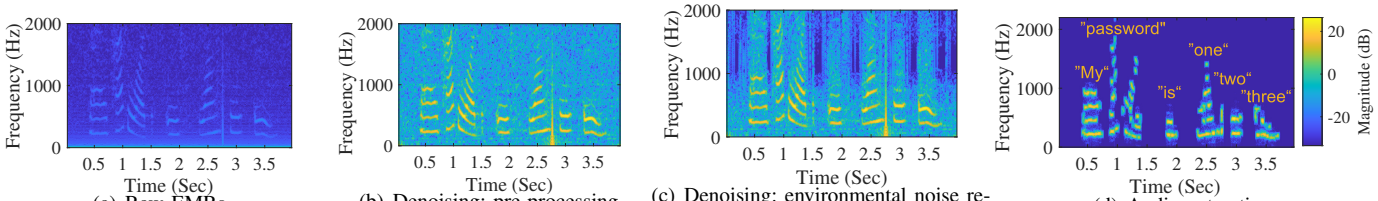
A. Denoising

Preprocessing. In indoor environments, ambient EMRs can also be observed at the frequency of 50Hz and its harmonic frequencies [72]. These frequencies are predominantly contributed by electronic power cables in the room, which generate AC currents at 50Hz. Consequently, they can cause significant interference in the measurements of mobile devices' EMRs. To mitigate this issue, we employ a sequence of band-stop filters to remove the cable EMRs in the received EMRs. Furthermore, since the human hearing range falls into the range of $[20Hz, 20kHz]$ [73], we use a high-pass filter with a cutting frequency of 20Hz to remove low-frequency noises. Additionally, to further improve the signal quality, the EMR measures are normalized into the range of $[-1, 1]$, as suggested by [15]. These pre-processing steps are important for enhancing the SNR of the EMR measures and enabling accurate audio recovery.

Environmental Noise Removal. Moreover, the recovery of intelligible audio from EMRs can be hindered by the presence of radiation noises originating from nearby electronic devices. These noises are typically random and cannot be easily eliminated using traditional filtering techniques due to their broad frequency spectrum, which can overlap with the audio's spectrum. To address this challenge, we employ Discrete Wavelet Transform (DWT) techniques [36], [74] to process the EMRs and remove environmental noises. DWT has been widely used in audio signal processing for effective noise removal.

It decomposes the EMRs into approximation coefficients and detail coefficients. The approximation coefficients represent the low-frequency components of the signal and retain the large-scale characteristics of the EMRs. As suggested by [8], these low-frequency components determine the intelligibility of spoken phonemes. Therefore, we keep the approximation coefficients unchanged during the audio recovery process. The detail coefficients, on the other hand, represent the high-frequency components of the signal and contain both fast-variant noises and fine details of the EMRs. In other words, some of the useful signals and environmental noises are intertwined in the detail coefficients. To remove its noise components, we apply multi-level wavelet decompositions. As shown in Fig. 11, a dynamic threshold method is applied to each level of detail coefficients recursively to remove the noise. The thresholds are calculated dynamically based on the estimated noise levels of the detail coefficients to ensure that the noise is appropriately removed without losing too many signal details. After the noise removal process is complete, we reconstruct the EMR signal from the approximation and denoised detail coefficients. By employing this DWT-based noise removal technique, we can significantly improve the SNR of the measured EMRs, leading to higher-quality of recovered audio signals.

In practice, we use a three-level wavelet decomposition with a Daubechies Wavelet basis function [75], which generates an approximation coefficient α^L (L = 3) and a sequence of detail coefficients $\beta^1, ..., \beta^l, ..., \beta^L$. The coefficients are



(a) Raw EMRs. (b) Denoising: pre-processing. (c) Denoising: environmental noise removal. (d) Audio extraction.

Fig. 10. Intermediate results of audio recovery procedures (The speech is "my password is one two three").

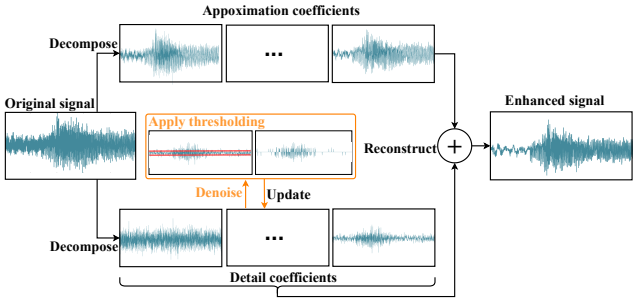


Fig. 11. Signal denoise with Discrete Wavelet Transform.

computed as follows:

$$\begin{cases} \alpha_k^L = \sum_{n\in \mathrm{N}} E(n)\phi_{n-2^L k}^L, & k \in \{1,2,...K_l\} \\ \beta_k^l = \sum_{n\in \mathrm{N}} E(n)\psi_{n-2^l k}^l, & l \in \{1,2,...,L\} \end{cases} \tag{4}$$

where $E(n)$ represents the discrete EMRs, N is the length of $E(n)$, ϕs and ψs are the wavelet basis functions, which are orthogonal to each other. K_l represents the length of coefficients at l-level decomposition. For each level of detail coefficients β^l, we compute its dynamic threshold thr^l by using the Birgé-Massart strategy [76] and use it to update the detail coefficients as β_k^{l*} below to remove the environmental noises,

$$\begin{cases} \beta_k^{l*} = \beta_k^l, & \textit{iff} \quad \beta_k^l \geq thr^l \\ \beta_k^{l*} = 0, & \textit{Others.} \end{cases} \tag{5}$$

Finally, by combining all the resulting coefficients (i.e., the approximation coefficients α^L and the updated detail coefficients $\{\beta^{1*},...,\beta^{l*},...,\beta^{L*}\}$), we reconstruct the final denoised EMRs with inverse DWT:

$$E(n) = \sum_{k\in \mathrm{K_L}} \alpha_k^L \phi_{n-2^L k}^L + \sum_{l=1}^{L} \sum_{k\in \mathrm{K_l}} \beta_k^{l*} \psi_{n-2^l k}^l. \tag{6}$$

B. Audio Extraction

The remaining task of audio recovery is to eliminate the signal distortions caused by the device's audio amplifier, as discussed in Section IV-D. We observe that the fundamental audio components form groups in the spectrogram, as illustrated in Fig. 5-8. These groups have significantly higher amplitudes than the distortions, making them distinguishable from the latter. Based on this observation, we propose a clustering algorithm, DBSCAN [77], [78], to recover the audio signal from the EMRs' STFT spectrogram. As shown in Fig. 12, DBSCAN groups the signal components whose STFT bins are located closely in the spectrogram. It outputs signal clusters in three categories: audio signal, distortions, and white noise. The audio signal, having a higher amplitude than the others, can be extracted by comparing the average amplitudes of the grouped signal clusters.

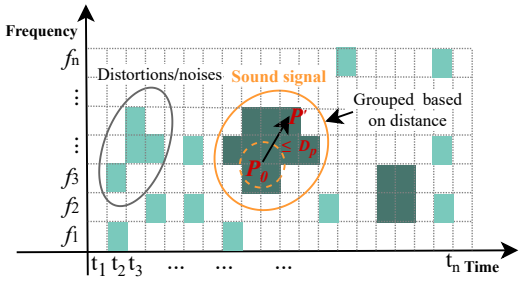
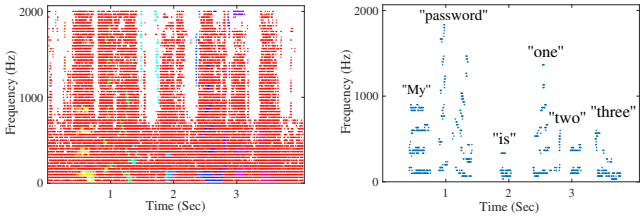


Fig. 12. DBSCAN extracts audio components from the EMR's STFT spectrogram: STFT bins (points) within the orange circle representing the audio signal.

Specifically, assuming the SD represents the EMR's STFT bins in the spectrogram, each bin (t, f, A) is regarded as a point denoted as $p \in SD$, in which A denotes the signal amplitude at time t and frequency f. As shown in Fig. 12, DBSCAN starts with an unlabeled point p_0 and screens all its neighbors within a pre-defined distance of D_p. If the number of neighbors is no less than the pre-defined parameter N_p, a new cluster C_0 is created by grouping the p_0 and its neighbors. DBSCAN then takes the point $p' \in C_0$ as a new start and expands the C_0 by incorporating its neighbors within the distance of D_p. If the C_0 cannot be expanded anymore, DBSCAN goes through the previous procedures over the remaining points in SD and forms new clusters until all points have been processed. We use Euclidean distance to measure the distance between two arbitrary points (p_i, p_j), i.e., $d(p_i, p_j) = \sqrt{(t_i - t_j)^2 + (f_i - f_j)^2 + (A_i - A_j)^2}$.



(a) Each color represents one signal cluster. (b) Isolated audio sounds from signal distortions and background noises.

Fig. 13. Audio extraction with DBSCAN.

DBSCAN generates signal clusters, i.e., $C = \{C_0, ...C_n..., C_N\}$. Fig. 13(a) shows signal clusters of the EMRs (audio sound: "My password is 123"). We then calculate the average amplitude of points (STFT bins) p_i of each cluster C_n, i.e.,

$$Avg_n = \frac{1}{|C_n|} \sum_{i=1}^{|c_n|} A_i, \quad p_i = (t_i, f_i, A_i) \in C_n, \tag{7}$$

and set A_{max} as the maximum value among the clusters. The signal components of audios are recognized as those clusters

whose average amplitudes Avg_n are no smaller than γA_{max}. γ is a coefficient falling into $[0, 1]$. Specifically, the audio's signal cluster is:

$$C_{audio} = \{C_n\},\ \textit{iff}\ Avg_n \geq \gamma A_{max}. \tag{8}$$

Afterward, we empirically set $\gamma = 0.1$ to achieve the best performance. To remove interference caused by distortions and white noises, we neutralize points other than those of audio signal clusters in the given STFT bin dataset SD by setting $A_i = 0$, *iff* $p_i \notin C_{audio}$. As shown in Fig. 13(b), the audio signal is successfully isolated after applying the neutralization. Finally, we conduct the inverse STFT over SD to obtain the clean EMRs representing the audio signal.

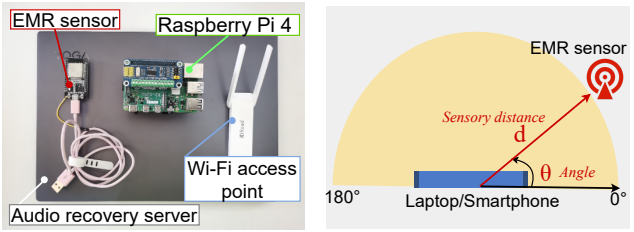
C. Sound Generation

We use the Matlab signal processing toolbox [79] to convert the processed EMRs as sound files. Specifically, the toolbox takes EMRs as a time-variant signal and ports its data points to a file with a "wav" header [80] added in advance. The attacker can directly play recovered "wav" files to hear the audio contents. In addition, he can also input them into an online speech-to-text recognition tool (e.g., Microsoft speech-to-text tool [38]) to transcribe the contents. Unlike prior works [5], [9]–[12], [18], [19], [33], [50], we pose no assumptions over the development of the speech-to-text recognition tool. The attacker does not need to build a specialized word/speech recognition model before the attack, which typically requires training data collection, device hardware pre-profiling, or highly-controlled attack scenarios. Our proposed attack considers the victim a "black-box", thus making the threat more practical.

VI. Evaluations

A. Experimental Setup

We build the *Periscope*'s eavesdropping system as presented in Fig. 14(a), which includes a portable Wi-Fi access point, an EMR sensor, a Raspberry Pi 4, and a laptop. We use a JDread 9600 portable access point to provide Wi-Fi converge for the EMR sensory device (ESP32 board), whose hardware specifics are described in Section IV-B. In the experiments, the EMR sensor is connected with the Raspberry Pi 4, which stores the measured EMR measurements and then forwards them to an audio recovery server (attacker's laptop) via the Wi-Fi connection. The audio recovery procedures run in real-time on the laptop to process the EMR measurements and output the recovered sounds. In the experiments, we evaluate the performances of the *Periscope* in terms of its ability to recover the victim's private audio contents under different settings. A wide spectrum of impact factors is examined, such as device heterogeneity, combinations of headphones and devices, devices' volumes, EMR sensory distances and angles, environmental dynamics, and audio applications. A total of 68 volunteers, 26 males and 42 females between 19 and 45 years old, were recruited for the experiments. Before each experiment, detailed instructions regarding experimental procedures are provided. The collected data are anonymized and properly stored locally from potential leakage. The IRB office of our institute has approved the entire research.



(a) (b)

Fig. 14. (a) The eavesdropping system of *Periscope*. (b) The definition of EMR sensor's relative position to the target device.

Speech Dataset. By default, we use speeches from the HARVARD speech corpus [52] as the audio eavesdropping contents. It contains 720 English sentences read by a female speaker. They are recorded as high-quality digital audio ".wav" files with a 48kHz sampling.

Default Setup. We use the following setups by default unless specific changes are mentioned. Without loss of generality, we use the HW MateBook D14, and iPhone SE2 plugged with Sennheiser PXC550 headphones as the target devices to represent the example of a laptop and a smartphone, respectively. The device volume is set as 80% by default. During the experiments, the attacker places the miniaturized EMR sensor underneath an office table to sense mobile devices' radiations, as shown in Fig. 3(b).

Evaluation Metrics. The performance of our proposed eavesdropping attack is evaluated via the following metrics: MOSNet, STOI, WER, and PSNR. In particular, **MOSNet** [81] is a commonly used subjective term in accessing audio qualities. It is built on a convolutional neural network (CNN) model, which predicts a human's subjective MOS (mean opinion score) rating. The predictions yield a score ranging from 1 (very poor) to 5 (very excellent). In addition, we also use an objective evaluation term, i.e., **STOI** [82] (Short-Time Objective Intelligibility), to validate how well humans can understand the recovered speech sentences. STOI ranges from 0 to 1. A larger STOI indicates better audio intelligibility. **Taal et al. [83] indicate that most people can recognize over 90% of the words within the speech sentence if the audio has an STOI index larger than 0.7.** We also tested the application of a voice recognition tool (Microsoft speech-to-text API [38]) on the restored audio. To evaluate the similarity between the transcribed and ground-truth tests, we adopted the word error rate (**WER**) [84], a standard metric for speech recognition. It is calculated as $WER = \frac{S+D+I}{N}$, in which S, D, and I represent the number of substitutions, deletions, and insertions, respectively, to match the transcribed speech against the ground-truth one. N is the number of words in the referenced ground-truth text. Moreover, we also select the Peak-Signal-to-Noise-Ratio (**PSNR**) [17], [85], as it is a basic metric to quantify the audio quality. A large value represents that the audio is clear with limited noise interference and vice versa.

In the following evaluations, we use all four metrics if the attack performance has significant variation tendencies. Otherwise, the MOSNet and STOI are selected as the primary metrics to present the results.

(a) Study Room (b) Subway Cabin

Fig. 15. *Periscope* attack is conducted in two daily headphone usage scenarios.

B. Real-world Case Study

Before going through the comprehensive evaluations (in the following sections), we first conduct two case studies to briefly demonstrate the effectiveness of the proposed attack. In the case studies, we made phone calls, sent voice messages, and streamed musics to the volunteer (victim) respectively and asked him to use a headphone plugged into the smartphone for listening to the audios. In particular, the phone caller and voice sender are located at remote distances, who said the following sentences via an iPhone 14 smartphone, i.e., "This verification code is ABC" (phone call 1), "My home address is near the central park" (phone call 2), "Your ID card number is one two three." (voice message 1), and "See you tonight at ten!" (voice message 2)". Additionally, three music segments, "My Heart Will Go On", "Summer Train", and "Grass Walk" are streamed on the victim's smartphone locally.

We studied two attack cases as shown in Fig. 15, in which we placed a hidden voice recorder and our EMR sensory device (ESP32 prototype) near the victim to compare their audio eavesdropping results. The EMR sensory distance and angle are defined as shown in Fig. 14(b).

Case 1: Study Room. The victim placed his smartphone (iPhone SE2) on the table. To launch a stealthy attack, the attacker placed the eavesdropping devices underneath a wooden table with 3cm by thick. The victim's smartphone was about 25cm and an angle within 30°-45° to the eavesdropping devices. The audio volume was set as 40%. The EMR sensory device continuously transmits the measurements to the attacker's laptop located at a distance of 5.5m to the victim for real-time audio recovery using the proposed *Periscope* designs.

Case 2: Subway Cabin. The victim sits on the chair and holds the smartphone with a natural body posture. The attacker placed the EMR sensory device in a bag of his colleague located at 15cm away by distance and 90°-140° by angle to the victim's smartphone. Since the subway cabin had loud acoustic noises, the victim used an audio volume of 60%. The audio recovery algorithm was running on the attacker's laptop at a distance of 6m to the victim.

Attack Results of Hidden Voice Recorder. Regrettably, our attempts to retrieve any discernible audio from the concealed voice recorder proved fruitless. We surmise that the ineffectiveness of the voice recorder to capture useful audio can be attributed to two principal factors. Firstly, headphones are known to possess good physical isolation properties that effectively prevent sound from leaking into the surrounding environment. Secondly, while there might be minimal sound leakage from the headphones, ambient noise levels generally exceed these negligible sounds, making it nearly impossible for the concealed voice recorder to record any meaningful audio that can be used for eavesdropping purposes. Our ambient noise measurements were conducted in two diverse settings, namely a study room and a subway cabin, revealing an acoustic noise level of 48.6dB and 94.8dB, respectively. As a reference, the loudness of normal human conversation is about 60dB [86].

Attack Results of *Periscope*. The attack performance of *Periscope* is summarized in Table I. *Periscope* recovers speeches with STOI scores over 0.7 for all test cases, indicating the recovery results have satisfactory intelligibility. We further validate this claim by inputting these recovered speech sounds to the Microsoft speech-to-text API [38]. Fig. 16(a) presents the correctly transcribed speech sentences. For the recovery results of music segments, we observed an average MOSNet of 2.57 for different test cases, which suggests that the attacker can well appreciate the music contents. We input the recovered audios into a music application, i.e., Shazam. Fig. 16(b) shows that it successfully recognized all recovered music segments. We believe all these successful audio eavesdropping are because *Periscope* exploits devices' EMRs as the side-channel which are naturally free from the impact of acoustic noises. Additionally, compared with the acoustic sounds, EMR signals can also easily penetrate obstacles (e.g., woods or plastics) without significant energy losses.

In sum, the case studies demonstrate that *Periscope* is an alarming threat of eavesdropping on users' private audio. In particular, the attack device is miniaturized and can be hind stealthily behind obstacles, while the attacker recovers users' audios at remote distances by receiving EMRs measurements via Wi-Fi connectivity.

TABLE I. PERFORMANCE OF THE *Periscope* ATTACK IN TWO DAILY HEADPHONE USAGE SCENARIOS.

Attack scenarios	Audio Type	STOI/MOSNet
Study Room	Phone call 1	0.73/2.62
	Phone call 2	0.71/2.45
	Voice Message 1	0.72/2.49
	Voice Message 2	0.73/2.63
	Music 1	0.70/2.31
	Music 2	0.71/2.35
	Music 3	0.72/2.56
Subway Cabin	Phone call 1	0.76/2.83
	Phone call 2	0.74/2.76
	Voice Message 1	0.75/2.80
	Voice Message 2	0.76/2.85
	Music 1	0.73/2.72
	Music 2	0.75/2.77
	Music 3	0.73/2.68

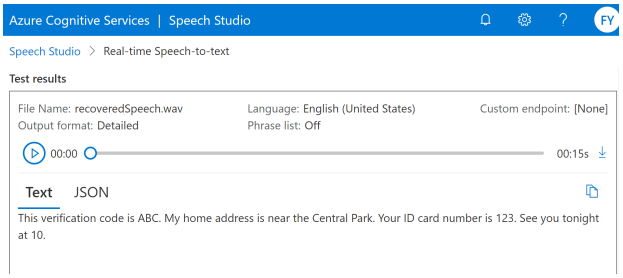
C. Comprehensive Evaluation of Periscope Attack

In this section, we evaluate the attack performances comprehensively by investigating the impact of different attack settings.

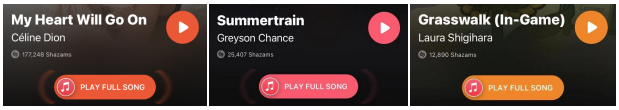
1) Impact of Target Devices' Diversity: In this experiment, we evaluated the proposed attack on both smartphones and laptops. Table II summarizes the recovered audio quality for different mobile devices. Among them, the maximum EMR

TABLE II. PERFORMANCE OF EAVESDROPPING ON DIFFERENT MOBILE DEVICES.

#	Mobile devices	Type	Manufacturer	Year	EMR sensory range(cm)	Attack range with Wi-Fi(m)	Evaluation metrics			
							MOSNet	STOI	WER (%)	PSNR(dB)
1	MacBook Pro 13	Laptop	Apple, USA	2020	82	15.82	2.80	0.74	19.50	20.48
2	MacBook Air 13	Laptop	Apple, USA	2019	81	15.81	2.89	0.75	18.43	21.28
3	RMN TPN-Q173	Laptop	HP, USA	2016	94	15.94	2.95	0.78	14.66	22.23
4	Yoga 14s	Laptop	Lenovo, China	2020	103	16.03	3.05	0.84	12.60	25.26
5	Inspiron 14 5410	Laptop	Dell, USA	2021	97	15.97	3.04	0.79	12.33	23.15
6	MateBook D14	Laptop	HW, China	2020	105	16.05	3.06	0.82	7.44	26.00
7	iPhone SE2	Smartphone	Apple, USA	2020	55	15.55	2.70	0.76	20.43	19.24
8	iPhone 14 Pro	Smartphone	Apple, USA	2022	45	15.45	2.62	0.71	23.77	17.44
9	iPhone 14	Smartphone	Apple, USA	2022	49	15.49	2.68	0.73	20.85	18.45
10	Mate 30	Smartphone	HW, China	2019	44	15.44	2.60	0.71	26.85	16.88
11	R11st	Smartphone	OPPO, China	2017	48	15.48	2.66	0.72	21.75	18.25



(a) Speech to text on Microsoft API [38]



(b) Music Recognition on Shazam

Fig. 16. Results of speech and music recognition tools to recognize the recovered audio of *Periscope*.

sensory range is derived by increasing the distance between the EMR sensor and its target device from 20cm to 120cm. For each experiment, the distance is gradually increased by 1cm. We asked the volunteers to listen to the audio restored under different distances. An attack is considered effective only if more than 50% volunteers can recognize the speech sentence. The maximum EMR sensory range is set as the one that triggered the first ineffective attack during the distance increments. It is observed that the longest EMR sensory distance is up to 105cm. In addition, smartphones have comparably lower functioning power than laptops, as do their EMRs. Consequently, the attacker achieves shorter EMR sensory ranges on smartphones. However, the EMR sensor is connected the Raspberry Pi, which provides the Wi-Fi connectivity. An attacker can exploit this by forwarding the EMR measures captured at close proximity to perform audio recovery at remote distances. This means that *Periscope* can attack devices from a maximum distance of 15.44m to 16.05m.

Table II also demonstrates the benchmark results of eavesdropping on the victim's speeches by hiding the EMR sensory device at a distance of 20cm (an office table typically has a thickness of less than 20cm). It is observed that all tested devices are vulnerable to the proposed attack. The average PSNR, WER, and MOSNet, are equal to 20.85, 17.83%, and 2.83, respectively. And, STOI values are all larger than 0.7, which indicates the resorted audios have good intelligibility [83].

TABLE III. SPECIFICS OF DIFFERENT HEADPHONES.

Headphone	Type	Wire length	Speaker diameter
Vivo XE160	In ear	1.25m	10mm
HW AM115	In ear	1.10m	11mm
Sennheiser PXC550	Over ear	1.40m	32mm
Samsung EO-IC100	In ear	1.20m	11mm
Bose NC700	Over ear	1.07m	40mm
Apple EarPods	In ear	1.20m	12mm

2) Impact of Combinations of Headphones and Devices: We further evaluate the attack performances when devices are plugged with different headphones. Six headphones are selected whose hardware specifics are summarized in Table III. During the experiments, we let the headphones be connected to six different mobile devices. Fig. 17 shows the WER matrix of each combination of headphones and devices. It can be seen that the attacker can successfully eavesdrop on the targets' speech sounds with WERs varying from 12.85% to 29.55%. In particular, for the Lenovo Yoga 14s, the lowest WER=12.85% is achieved when it connects with a Sennheiser PXC550. We infer the prominent performance is due to the headphone having a larger body, i.e., a 1.40m length of wire and a 32mm diameter of microspeaker unit. Therefore, it makes the headphone an excellent antenna to enhance the EMRs. Meanwhile, the large headphone body also requires higher power outputs from the mobile device's amplifier, thus stronger EMR strengths radiated.

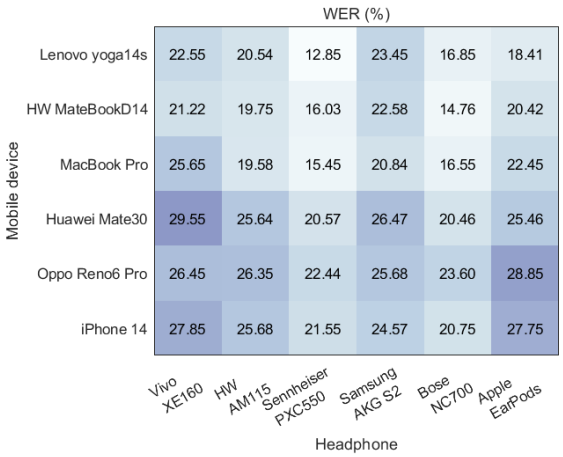
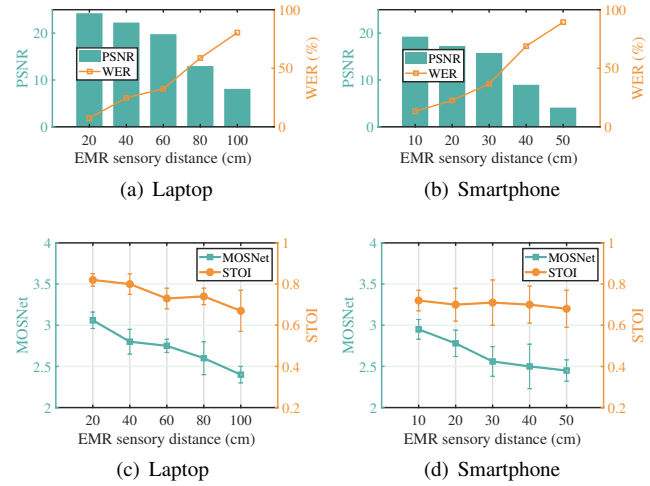


Fig. 17. Devices plugged with different headphones.

3) Impact of EMR Sensory Distances: In the experiments, we set the EMR sensory distance from 20cm to 100cm for the tested laptop (HW MateBook D14) and from 10cm to 50cm for the smartphone (iPhone SE2). As shown in Fig. 18, the recovered signal's PSNR is negatively correlated with increases of the distance. A longer distance leads to weaker EMRs received by the attacker. As a result, it becomes more challenging to recover audio sounds. The attacker achieves good speech recovery quality within 60cm and 30cm for the laptop and smartphone, respectively. It then drops. Specifically, Microsoft's Speech-to-Text tool has a WER of 32.45% at 60cm for the laptop, while it turns to 58.76% at 80cm. Meanwhile, MOSNet and STOI reflect human's subjective and objective evaluation of the recovered audio. Even at far distances, people still have a good understanding of recovered audio contents. For example, when the EMR sensory distance is set as 80cm, the laptop's MOSNet equals 2.6, and STOI equals 0.74> 0.7. Similar observations are found for eavesdropping on the smartphone at a distance of 40cm, in which MOSNet=2.5 and STOI=0.7. It is also worth mentioning that the target devices and the attacker are placed on two sides of a wood table, a non-LoS scenario shown in Fig. 3. Thus, we anticipate better signal quality under a LoS scenario. Besides, we only used one prototype in the experiment. In practice, many of them can be deployed such that the performances can be further enhanced by fusing EMR measures from multiple prototypes.



(a) Laptop (b) Smartphone

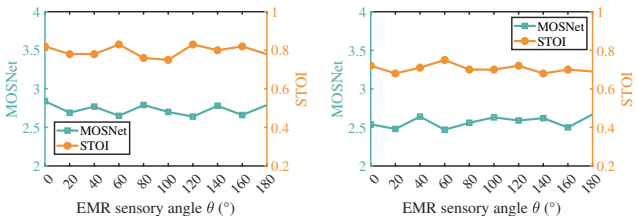
(c) Laptop (d) Smartphone

Fig. 18. Impact of EMR sensory distance.

4) Impact of Target Device's Volumes: In practice, victims may have different preferences in selecting their audio volumes. Thus, it is necessary to examine the impact of this factor on the attack performance. To provide a comprehensive evaluation, we connect the HW Matebook D14 laptop with a Sennheiser PXC550 headphone and vary its volume from 20% to 100%. For each volume setting, we measure its loudness using a SNDWAY SW-523 [87] decibel meter. Table IV summarizes the restored audio quality under different settings. We find out that PSNRs grow with the volume increases. For example, when the device has 20% volume, the attacker can recover the speech sound with PSNR=3.75dB, while it is increased to 16.42dB when the device sets volume as 60%. In addition, Microsoft's Speech-to-Text tool experiences a significant decrease in its WER values. Specifically, WER=94.10% when the volume is set as 20%, and it drops to 28.45% when volume=60%. This is because that the device's amplifier processed audios signals with larger amplitudes and generates stronger EMRs, which helps the attacker to recover audio sounds with higher qualities. As expected, the MOSNet and STOI are also promoted with the increase in volumes. We observe the *Periscope* recovers intelligible audios, i.e., STOI ≥ 0.7, when users select volumes higher than 40%. Following a recent study of 280 college students [88], 79% of them use their headphones for more than 1 hour daily with a volume of higher than 60%. Therefore, we believe *Periscope* is a piratical attack in real-world scenarios.

TABLE IV. IMPACT OF DEVICE VOLUME.

Volume (%)	20	40	60	80	100
Level (dB)	30.4	47.4	53.1	81.7	90.6
PSNR(dB)	3.75	9.50	16.42	25.75	36.00
WER(%)	94.10	53.44	28.45	8.45	5.01
MOSNet	2.10	2.56	2.85	2.90	3.10
STOI	0.56	0.72	0.78	0.80	0.86



(a) Laptop (b) Smartphone

Fig. 19. Impact of EMR sensory angle.

5) Impact of EMR Sensory Angles: We also examine if the relative angles between the EMR sensor and its target impact the audio recovery quality. We fix the EMR sensory distance as 40cm and vary their relative angles from 0° to 180°. Fig. 19 shows MOSNet and STOI measurements of eavesdropped audios at different relative positions. We find that MOSNets and STOIs are relatively stable across all test cases fluctuating around the average values: MOSNet equals 2.73 and 2.57 for the laptop and smartphone, respectively. STOIs equal 0.80 and 0.70 for the laptop and smartphone, respectively. This is because that mobile devices' EMRs are radiated omnidirectionally to all angles. Hence, the attacker's orientations are not the critical factor that impacts the eavesdropping performance. However, this is not the case for the attack, e.g., MagEar [20], that exploits speakers' magnetic filed variations, which can only be formed at a narrow-angle range. MagEar experiences about 21.7% performance degradation when two parties have a 20° displacement angle.

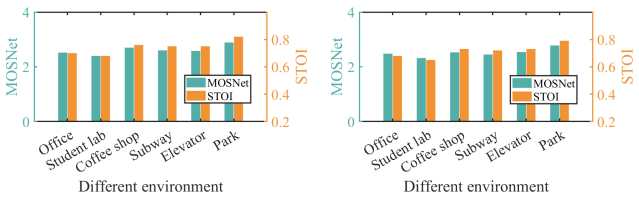
6) Impact of Physical Obstacles: We use five different mobile devices as the targets to evaluate the impact of physical obstacles. Usually, victims believe that physical isolation (such as soundproof wooden doors, plastic office desktop bezel, or concrete walls) can protect them from eavesdropping attacks since acoustic sounds will be blocked and cannot penetrate these obstacles. In this experiment, we demonstrate the effectiveness of our proposed attack by placing different obstacles between the attack prototype and the target. We used three kinds of materials with different thicknesses, i.e., a 12cm

wooden desk, a 10cm plastic bezel, and a 20cm concrete wall. The EMR sensory distance is fixed as 40cm. Table V compares the results for tested smartphones and laptops. It is observed that both evaluation metrics experience no significant decreases in the test cases of with/without the obstacles. For example, MacBook Pro 13 has MOSNet and STOI equal to 2.69 and 0.76, respectively, with no obstacles. When there is a wooden obstacle, the MOSNet=2.55 and STOI=0.72. Thus, these obstacles have a negligible effect on EMR side-channel eavesdropping. The reason is that they have low conductivity, such that EMR signals can penetrate them without significant energy loss.

TABLE V. IMPACT OF DIFFERENT PHYSICAL OBSTACLES.

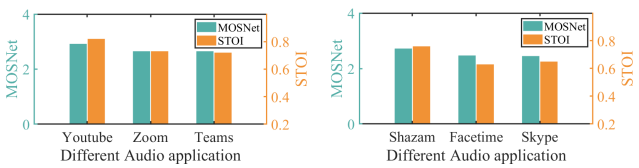
Mobile devices	MOSNet/STOI			
	No obstacle	Wood	Plastic	Wall
MateBook D14	2.80/0.80	2.75/0.78	2.72/0.75	2.62/0.74
Yoga 14s	2.76/0.78	2.70/0.74	2.72/0.73	2.55/0.70
MacBook Pro 13	2.69/0.76	2.55/0.72	2.50/0.73	2.43/0.69
iPhone 14	2.45/0.69	2.32/0.67	2.35/0.68	2.27/0.65
Mate 30	2.51/0.72	2.50/0.71	2.49/0.70	2.34/0.63

7) Impact of Environmental Dynamics: The victim user can show up at any random sites. In this experiment, we evaluate the audio eavesdropping performances in indoor and outdoor scenes, including an office, a student lab, a café shop, a subway cabin, an elevator cabin, and a park site. They have different ambient electromagnetic noises. For example, in the office or student lab, many people may use laptops, tablets, or smartphones for work or entertainment. The indoor appliances will also generate electromagnetic radiations that may cause additional interference to the audio recovery. However, the attack performances are less likely to be affected if the victim is in a place with fewer electronics around, e.g., a park. As shown in Fig. 20, the results meet our expectations. The performances vary along with the test sites' electromagnetic intensities. The MOSNet and STOI readings suffer significant degradation in the office and student lab scenarios. Still, we find out that the restored audios maintain good intelligibility with averaged STOI=0.74 and 0.72 for the laptop and the smartphone, respectively.



(a) Laptop (b) Smartphone

Fig. 20. Impact of different environment.



(a) Laptop (b) Smartphone

Fig. 21. Impact of different audio application.

8) Impact of Audio Applications: We perform experiments to evaluate the impact of the victim's audio application usages. There are six different applications included in the experiment, i.e., YouTube, Zoom, Teams, Shazam, Facetime, and Skype. The eavesdropping results are shown in Fig. 21. We find that better performances are achieved when the victim uses applications for entertainment, e.g., YouTube and Shazam (music player). For example, the attacker can recover audio sounds on YouTube with MOSNet=2.92 and STOI=0.82. It drops to 2.65 and 0.73, respectively, when the victim user is using the online conference tool Zoom. We infer the reason can be that entertainment applications typically provide high-quality audio sounds with high sampling rates, e.g., 48kHz, 96kHz, and 128kHz. In contrast, communication tools, such as Zoom and Teams, use lower audio sampling rates, e.g., 8kHz [89], which are sufficient for user hearings. Such designs also save the network bandwidth in transmitting these audio sounds. Since the applications have audio quality disparities, so as their corresponding EMRs and the restored audios.

VII. DISCUSSIONS

Disclosures. We are taking active steps to notify related vendors of the risk. Currently, we reported our eavesdropping threat to six leading enterprises in the mobile market, including Apple, Lenovo, Huawei, Vivo, OPPO, and Dell. The disclosure emails were sent to each of the manufacturers containing detailed explanations about the attack, the affected mobile products, and the corresponding eavesdropping evaluations. As of write, Apple, Huawei, Lenovo, and Dell have responded to our disclosure. Huawei has reproduced our attack results, and are processing our report under their high priorities. Lenovo informed us that they would like to follow our research and update us regarding their next steps. We believe that *Periscope* attack may impact a larger number of devices beyond those belonging to the reported manufacturers, and filled a vulnerability report to CVE and CNVD (ID: CNVD-C-2023-85063).

Countermeasures. In addition, we also provide defensive countermeasures that will remedy the threats. We hope our paper will be informative further to enhance the security of audio-driven devices and applications. One effective defense is to use EM shielding material inside the device's audio processing circuits, which can dampen the EM leakages. As discussed in Section IV-B, the frequency range of audio sounds is mainly below 4kHz. Thus, we can use copper metals to shield these radiating circuits. To investigate the effectiveness of shielding, we cover a Lenovo Yoga 14s laptop with a 1mm copper plate. With the shielding, the recovered speech's PSNR drops from 24.5dB to 4.25dB, and WER increases from 15.23% to 83.08%. Results show that it can effectively impede the propagation of EMRs and prevent the attacker from eavesdropping on meaningful audio content.

Potential Improvements. *Periscope* leverages the ESP32 board as a sensory device for collecting devices' EMRs and forwards them wirelessly to a remote attacker located tens of meters away. One potential improvement preferred by the attackers is to increase the EMR sensory range of the ESP32 board. Following Friis transmission equation [90], we believe such enhancement can be done by enlarging the ESP32 board's antenna size or connecting it with a signal amplifier to capture weaker EMR signals. However, it is worth noting that such

enhancements might increase the overall size of the EMR sensory device, making it more difficult to conceal in hidden places without being noticed by victims.

VIII. Conclusions

This paper presents the first proof-of-concept eavesdropping attack, *Periscope*, that leverages the EMRs emitted from mobile devices to recover victims' audio contents at remote distances. In particular, we find the signal distortion patterns exist in EMRs' spectrograms, thus identifying the primary radiation source as audio amplifies. Further studies show headphones enhance the radiation strengths, which further facilitates eavesdropping. We treat the target device as a "black-box" and design audio recovery schemes solely based on signal processing techniques. *Periscope* leverages a miniaturized prototype with the similar size of typically hidden voice recorders to collect the EMRs stealthily, and forward them to a remote attacker located at 15m away for audio recovery. Our evaluations show *Periscope* is able to recover private audio contents played on a wide range of mobile devices. The results are intelligible to both human hearing and recognition tools. We report the threat to leading manufacturers and hope to inspire them to rethink the security vulnerability of audio amplifiers embedded in mobile devices.

Acknowledgment

We would like to express our sincere gratitude to the anonymous reviewers for their valuable comments. Additionally, we would like to recognize the partial support for this work provided by the National Key R&D Project (2022YFB3103500), National Natural Science Foundation of China (Grant No. 62202150, 62302162, 62172234, U20A20174), Technology Projects of Hunan Province (2015TP1004), Science and Technology Key Projects of Changsha City (No. kh2103003), Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province, and Fundamental Research Funds for the Central Universities.

References

[1] maximize Market Research, "Audio plug-ins software application market: Global industry analysis and forecast (2021 - 2029)," https://www.maximizemarketresearch.com/market-report/global-audio-plug-ins-software-application-market/100413/.

[2] K. M. Robertson, D. R. Hannah, and B. A. Lautsch, "The secret to protecting trade secrets: How to create positive secrecy climates in organizations," *Business Horizons*, vol. 58, no. 6, pp. 669–677, 2015.

[3] Taotronics, "What is headphone sound leakage?" https://blog.taotronics.com/headphones/headphone-sound-leakage/.

[4] E. Festival, "How to stop headphones from leaking sound?" https://electricfieldsfestival.com/stop-headphones-from-leaking-sound/.

[5] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" *IEEE Transactions on Mobile Computing (IEEE TMC)*, vol. 15, no. 11, pp. 2907–2920, 2016.

[6] A. Kwong, W. Xu, and K. Fu, "Hard drive of hearing: Disks that eavesdrop with a synthesized microphone," in *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2019.

[7] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, "Speake(a)r: Turn speakers to microphones for fun and profit," in *USENIX Workshop on Offensive Technologies*, 2017.

[8] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *ACM International Conference on Mobile Systems, Applications, and Services (ACM Mobisys)*, 2016.

[9] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *USENIX Security Symposium (USENIX Security)*, 2014.

[10] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, "Spearphone: a lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers," in *ACM Conference on Security and Privacy in Wireless and Mobile Networks (ACM WiSec)*, 2021.

[11] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer." in *The Network and Distributed System Security Symposium (IEEE NDSS)*, 2020.

[12] R. Matovu, I. Griswold-Steiner, and A. Serwadda, "Kinetic song comprehension: Deciphering personal listening habits via phone vibrations," *arXiv preprint arXiv:1909.09123*, 2019.

[13] P. Hu, H. Zhuang, P. S. Santhalingam, R. Spolaor, P. Pathak, G. Zhang, and X. Cheng, "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2022.

[14] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: passive recovery of sound from video," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–10, 2014.

[15] B. Nassi, Y. Pirutin, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Passive sound recovery from a desk lamp's light bulb vibrations," in *USENIX Security Symposium (USENIX Security)*, 2022.

[16] R. P. Muscatell, "Laser microphone," *The Journal of the Acoustical Society of America*, vol. 76, no. 4, pp. 1284–1284, 1984.

[17] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *International Conference on Mobile Computing and Networking (ACM MobiCom)*, 2015.

[18] S. Sami, S. R. X. Tan, Y. Dai, N. Roy, and J. Han, "Lidarphone: acoustic eavesdropping using a lidar sensor," in *ACM Conference on Embedded Networked Sensor Systems (ACM Sensys)*, 2020.

[19] Y. Li, C. Duan, X. Ding, and C. Liu, "Tagmic: Listening through rfid signals," in *International Conference on Distributed Computing Systems (IEEE ICDCS)*, 2020.

[20] Q. Liao, Y. Huang, Y. Huang, Y. Zhong, H. Jin, and K. Wu, "Magear: eavesdropping via audio recovery using magnetic side channel," in *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, 2022.

[21] G. Camurati, S. Poeplau, M. Muench, T. Hayes, and A. Francillon, "Screaming channels: When electromagnetic side channels meet radio transceivers," in *ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, 2018.

[22] R. Wang, H. Wang, and E. Dubrova, "Far field em side-channel attack on aes using deep learning," in *Proceedings of the ACM Workshop on Attacks and Solutions in Hardware Security*, 2020.

[23] W. Van Eck, "Electromagnetic radiation from video display units: An eavesdropping risk?" *Computers & Security*, vol. 4, no. 4, pp. 269–286, 1985.

[24] L. Wang and B. Yu, "Analysis and measurement on the electromagnetic compromising emanations of computer keyboards," in *International Conference on Computational Intelligence and Security*, 2011.

[25] M. Vuagnoux and S. Pasini, "Compromising electromagnetic emanations of wired and wireless keyboards." in *USENIX security symposium*, 2009.

[26] E. C. Jordan, "Electromagnetic waves and radiating systems," *American Journal of Physics*, vol. 19, no. 8, pp. 477–478, 1951.

[27] T. J. Dishongh and M. McGrath, *Wireless sensor networks for healthcare applications*. Artech House, 2010.

[28] A. M. Lonzetta, P. Cope, J. Campbell, B. J. Mohd, and T. Hayajneh, "Security vulnerabilities in bluetooth technology as used in iot," *Journal of Sensor and Actuator Networks*, vol. 7, no. 3, p. 28, 2018.

[29] J. Padgette, K. Scarfone, and L. Chen, "Guide to bluetooth security," *NIST special publication*, vol. 800, no. 121, 2017.

[30] M. Gao, Y. Liu, Y. Chen, Y. Li, Z. Ba, X. Xu, and J. Han, "Inertiear: Automatic and device-independent imu-based eavesdropping on smartphones," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022.

[31] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022.

[32] J. Han, A. J. Chung, and P. Tague, "Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in *ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2017.

[33] S. Zhang, Y. Liu, and M. Gowda, "I spy you: Eavesdropping continuous speech on smartphones via motion sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–31, 2023.

[34] F. Mo, Y.-H. Lu, J.-L. Zhang, Q. Cui, and S. Qiu, "A support vector machine for identification of monitors based on their unintended electromagnetic emanation," *Progress In Electromagnetics Research M*, vol. 30, pp. 211–224, 2013.

[35] Z. Liu, N. Samwel, L. Weissbart, Z. Zhao, D. Lauret, L. Batina, and M. Larson, "Screen gleaning: A screen reading tempest attack on mobile devices exploiting an electromagnetic side channel," *arXiv preprint arXiv:2011.09877*, 2020.

[36] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in *Proc. conf. in acoustics and music theory applications*, vol. 66. Citeseer, 2001.

[37] ESP32, "Esp32," http://esp32.net/, 2022.

[38] Microsoft, "Microsoft speech to text tool," https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/#features, 2022.

[39] CVE, "Common vulnerabilities and exposures," https://cve.mitre.org/.

[40] CNVD, "Chinese national vulnerability database," https://www.cnvd.org.cn/.

[41] C. Wang, F. Lin, T. Liu, K. Zheng, Z. Wang, Z. Li, M.-C. Huang, W. Xu, and K. Ren, "mmeve: eavesdropping on smartphone's earpiece via cots mmwave device," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 338–351.

[42] C. for Disease Control and Prevention, "Loud noise can cause hearing loss," https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html.

[43] W. Jin, S. Murali, H. Zhu, and M. Li, "Periscope: A keystroke inference attack using human coupled electromagnetic emanations," in *ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, 2021.

[44] M. Dey, A. Nazari, A. Zajic, and M. Prvulovic, "Emprof: Memory profiling via em-emanation in iot and hand-held devices," in *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2018.

[45] C. Shen, T. Liu, J. Huang, and R. Tan, "When lora meets emr: Electromagnetic covert channels can be super resilient," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 1304–1317.

[46] M. Guri, "Air-gap electromagnetic covert channel," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–18, 2023.

[47] X. Ji, J. Zhang, S. Zou, Y. Chen, G. Qu, and W. Xu, "Magview++: Data exfiltration via cpu magnetic signals under video decoding," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2023.

[48] C. Ulaş, U. Aşık, and C. Karadeniz, "Analysis and reconstruction of laser printer information leakages in the media of electromagnetic radiation, power, and signal lines," *Computers & Security*, vol. 58, pp. 250–267, 2016.

[49] J. Choi, H.-Y. Yang, and D.-H. Cho, "Tempest comeback: A realistic audio eavesdropping threat on mixed-signal socs," in *ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, 2020.

[50] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *ACM International Conference on Mobile Systems, Applications, and Services (ACM Mobisys)*, 2019.

[51] J. A. Kong, "Theory of electromagnetic waves," *Wiley-Interscience*, 1975.

[52] P. Demonte, "Harvard speech corpus—audio recording 2019," 2019, university of Salford Collection.

[53] P. Bourke, "Cross correlation," *Cross Correlation", Auto Correlation—2D Pattern Identification*, 1996.

[54] C.-P. Liang, J.-H. Jong, W. Stark, and J. East, "Nonlinear amplifier effects in communications systems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 47, no. 8, pp. 1461–1466, 1999.

[55] M. Otala, "Non-linear distortion in audio amplifiers," *Wireless World*, 1977.

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

[57] P. A. Martinez and M. Lozano, "Nonlinear distortion in current-feedback amplifiers," *Microelectronics Journal*, vol. 16, no. 5, pp. 22–30, 1985.

[58] J. Choma, "Harmonic and intermodulation distortion in current-feedback bipolar transistor amplifiers," 1981.

[59] M. Otala, "Circuit design modifications and minimizing transient intermodulation distortion in audio amplifiers," in *Audio Engineering Society Convention 2ce*. Audio Engineering Society, 1972.

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

[61] M. Integrated, "Max98307/max98308," https://www.analog.com/media/en/technical-documentation/data-sheets/max98307-max98308.pdf, 2020.

[62] T. Instruments, "Lm4844," https://pdf1.alldatasheet.com/datasheet-pdf/view/113635/NSC/LM4844.html, 2005.

[63] ——, "Tpa6166a2 3.5-mm jack detect and headset interface ic," https://www.ti.com/lit/ds/symlink/tpa6166a2.pdf, 2015.

[64] ——, "Lme49710 highperformance,highfidelityaudiooperationalamplifier," https://pdf1.alldatasheet.com/datasheet-pdf/view/611613/TI1/LME49710.html, 2007.

[65] ——, "Opa627 and opa637 precision high-speed difet® operational amplifiers," https://www.ti.com/lit/ds/symlink/opa627.pdf, 2015.

[66] R. Zhou, X. Ji, C. Yan, Y.-C. Chen, W. Xu, and C. Li, "Dehirec: Detecting hidden voice recorders via adc electromagnetic radiation," in *IEEE Symposium on Security and Privacy (SP)*, 2023.

[67] W. M. Leach, *Introduction to electroacoustics and audio amplifier design*. Kendall/Hunt Publishing Company Dubuque, IA, USA, 2003.

[68] H. H. Scott, "The amplifier and its place in the high-fidelity system," *Journal of the Audio Engineering Society*, vol. 1, no. 3, pp. 246–254, 1953.

[69] K. Salahddine, L. Jalal *et al.*, "Design a new architecure of audio amplifiers class-d for a hardware mobile systems," *International Journal of Computer Science and Application*, vol. 1, no. 1, pp. 6–11, 2012.

[70] K. A. Simons, "The decibel relationships between amplifier distortion products," *Proceedings of the IEEE*, vol. 58, no. 7, pp. 1071–1086, 1970.

[71] A. Katz, "Linearization: Reducing distortion in power amplifiers," *IEEE microwave magazine*, vol. 2, no. 4, pp. 37–49, 2001.

[72] Z. Yan, Q. Song, R. Tan, Y. Li, and A. W. K. Kong, "Towards touch-to-access device authentication using induced body electric potentials," in *International Conference on Mobile Computing and Networking (ACM MobiCom)*, 2019.

[73] U. National Center for Biotechnology Information, "The audible spectrum," https://www.ncbi.nlm.nih.gov/books/NBK10924/, 2022.

[74] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[75] C. Vonesch, T. Blu, and M. Unser, "Generalized daubechies wavelet families," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4415–4429, 2007.

[76] L. Birgé and P. Massart, "From model selection to adaptive estimation," in *Festschrift for lucien le cam*, 1997, pp. 55–87.

[77] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[78] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *International Conference on the Applications of Digital Information and Web Technologies*, 2014.

[79] Matlab, "Signal processing toolbox," https://ww2.mathworks.cn/products/signal.html, 2022.

[80] WavFile, "Wavfile," https://docs.fileformat.com/audio/wav/, 2022.

[81] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[82] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010.

[83] ——, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[84] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[85] N. Verma and A. Verma, "Performance analysis of wavelet thresholding methods in denoising of audio signals of some indian musical instruments," *Int. J. Eng. Sci. Technol*, vol. 4, no. 5, pp. 2040–2045, 2012.

[86] C. for Disease Control and prevention, "What noises cause hearing loss?" https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html.

[87] SNDWAY, "Sw-523 decibel mete," https://www.aliexpress.com/item/4000703339570.html, 2022.

[88] B. K. Fasanya and J. D. Strong, "Younger generation safety: hearing loss and academic performance degradation among college student headphone users," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2018.

[89] I. T. U. R. G. (11/88), "Pulse code modulation (pcm) of voice frequencies," http://www.ktl.elf.stuba.sk/study/h320/Pdf/G711E.pdf, Nov. 1988.

[90] H. T. Friis, "A note on a simple transmission formula," *Proceedings of the IRE*, vol. 34, no. 5, pp. 254–256, 1946.