Wireless Eavesdropping on Wired Audio with Radio-frequency Retroreflector Attack

Genglin Wang, Zheng Shi, Yanni Yang, Zhenlin An, Guoming Zhang, Pengfei Hu, Xiuzhen Cheng, *Fellow, IEEE*, Jiannong Cao, *Fellow, IEEE*

Abstract—Recent studies have demonstrated the feasibility of eavesdropping on audio via radio frequency signals or videos, which capture physical surface vibrations from surrounding objects. However, these methods are inadequate for intercepting internally transmitted audio through wired media. In this work, we introduce radio-frequency retroreflector attack (RFRA) and bridge this gap by proposing an RFRA-based eavesdropping system, RF-PARROT^2, capable of wirelessly capturing audio signals transmitted through earphone wires. Our system entails embedding a tiny field-effect transistor within the wire to establish a battery-free retroreflector, whose reflective efficiency is correlated with the amplitude of the audio signal. To preserve the details of audio signals, we designed a unique retroreflector using a depletion-mode MOSFET (D-MOSFET). This MOSFET can be triggered by any voltage level present in the audio signals, thus guaranteeing no information loss during activation. However, the D-MOSFET introduces a nonlinear convolution operation on the original audio, resulting in distorted audio eavesdropping. Thus, we devised an engineering solution which utilized a novel convolutional neural network in conjunction with an efficient Parallel WaveGAN vocoder to reconstruct the original audio. Our comprehensive experiments demonstrate a strong similarity between the reconstructed audio and the original, achieving an impressive 95% accuracy in speech command recognition.

Index Terms—Audio eavesdropping, radio-frequency retroreflector attack (RFRA), EM side-channel attack.

1 INTRODUCTION

Audio eavesdropping has been a prominent focus in the realm of security and privacy studies. Recent advancements have showcased that spoken audio signals can be wirelessly intercepted by capturing the physical surface vibrations of surrounding objects using RF signals [1]–[6], laser [7]–[9], and videos [10]. However, these methods have a notable limitation: they fail to intercept audio signals transmitted internally through wires, such as those in earphones and telephone lines. This gap poses significant security concerns, as wired audio transmissions are prevalent in various practical scenarios, ranging from personal earphones to professional audio systems. Furthermore, most audio signals transmitted through earphone wires are not encrypted. As such, eavesdropping can lead to severe privacy breaches.

A straightforward approach to eavesdropping on earphones involves conducting classic electromagnetic (EM) side-channel attacks [11]–[14], which passively measure the EM leakages from the target devices. Then, the attacker can reconstruct the original signal by analyzing the measured radio wave. Nonetheless, the EM leakages from the audio wire or speaker coils are too weak to be detected remotely. The advanced earphone eavesdropping system is limited to 50 cm [13]. Recent studies [15]–[17] have explored active radio-frequency retroreflector attacks (RFRA) for remote eavesdropping on wired digital signals. RFRA covertly embeds a retroreflector inside the target device. The retroreflector reflects the RF signal emitted by the attacker back to the colocated signal receiver for interception [16]. The reflective ability of the retroreflector varies as the digital signal switches between high and low voltage. For example, researchers [18] demonstrated the feasibility of implanting a tiny metal-oxide-semiconductor field-effect transistor (MOSFET) into the keyboard wire and employing commodity software-defined radio (SDR) to remotely eavesdrop on the keystrokes, as depicted in Fig. 1(a). Thus, RFRA holds significant potential for wireless eavesdropping on wired audio signals. In contrast to existing vibration-based audio eavesdropping and passive EM leakages measurement, RFRA directly acquires the unencrypted plain signals from signal lines with long eavesdropping distance. However, existing RFRA systems can only eavesdrop on digital signals but fail to deal with analog audio signals.

Thus, in this work, we come up with a key question: *Can we achieve wireless eavesdropping on analog audio signals using RFRA?* As depicted in Fig. 1(b), our goal is to employ RFRA for eavesdropping on analog audio signals transmitted through wires, remotely and through barriers like walls. Achieving this goal is, however, not an easy task due to the following issues. First, recovering every intricate detail of the analog audio signal, meanwhile disturbed by hardware noises, is far more complicated and difficult than simply discerning binary '0' or '1' in digital signals. Second, unlike digital signals that involve non-negative values, nearly half of the analog audio signals are negative, which cannot be perceived by the digital RFRA approach.

This work is supported by the National Natural Science Foundation of China (No. 62302274, 62422208, 62232010, 62202276, 62202274, 62232010), Shandong Provincial Natural Science Foundation (No. ZR2023QF113, 2023TSGC0105, ZR2022ZD02), Shandong Science Fund for Excellent Young Scholars (No. 2024HWYO-021, 2022HWYQ-038), and the Research Institute for Artificial Intelligence of Things at The Hong Kong Polytechnic University.

- *Genglin Wang, Zheng Shi, Yanni Yang, Guoming Zhang, Pengfei Hu and Xiuzhen Cheng are with Shandong University, Qingdao, China.*
- *Zhenlin An is with Princeton University, Princeton, United States.*
- *Jiannong Cao is with The Hong Kong Polytechnic University, Hong Kong SAR, China.*
- *Corresponding author: Yanni Yang*

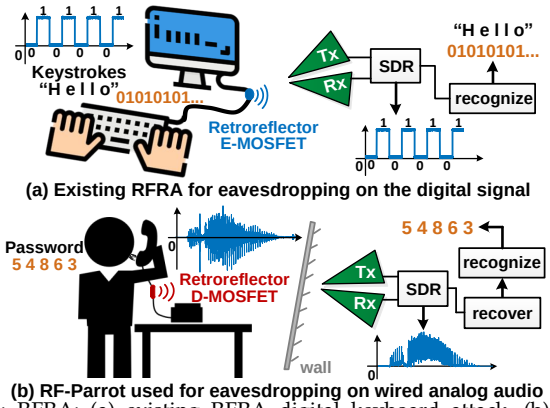


Fig. 1: RFRA: (a) existing RFRA digital keyboard attack, (b) wired analog audio eavesdropping using RF-PARROT^2

Regarding the above problems, we propose the first-of-its-kind RFRA-based audio eavesdropping system, named RF-PARROT^2, with a new retroreflector design. Specifically, we address the fundamental challenge of RFRA on the analog audio signal by employing the depletion-mode MOSFET (D-MOSFET) as the retroreflector. Previous RFRAs all use the enhancement-mode MOSFETs (E-MOSFETs), which can only react to positive voltage [15], [16], [18], thereby overlooking the negative component of audio signals and resulting in significant information loss. Instead, D-MOSFET offers a distinct advantage. The D-MOSFET, acting as a 'soft switch', can be turned on by both negative and positive audio signals.

Thus, we fabricate a tiny retroreflector (diameter < 3mm) using the D-MOSFET as a covert listening device, concealed within the audio wire. The D-MOSFET retroreflector functions as an analog modulator over the remotely transmitted RF signal. As such, the audio signal's voltage has a direct impact on the power of RF signals reflected back to the signal receiver, thus enabling the extraction of audio information from the received signal amplitude. Our analytical and experimental results confirm that the D-MOSFET-based retroreflector is superior to the E-MOSFET for maintaining the key information in analog audio signals. Furthermore, we carefully select the RF signal frequency for through-wall RFRA audio eavesdropping.

Despite the promising capability of D-MOSFET, it introduces a critical side effect on audio interception, which is caused by the nonlinear transfer characteristic curve of D-MOSFET. In specific, the received RF signal is nonlinearly distorted with unbalanced and inverted RF signal amplitudes between the negative and positive parts of the audio in the time domain. Accordingly, irrelevant additional frequency components are introduced in the frequency domain. Such distorted audio spectrograms hinder us from accurate reconstruction of the original audio.

To tackle the nonlinearity issue, we undertake a mathematical modeling approach to understand the impact of the nonlinear transfer curve on the received RF signal amplitude. Our analysis reveals that the nonlinear transformation is a convolution operation in essence. Thus, we may reconstruct the audio by deriving an analytical solution to reverse the convolution operation. However, it is computation-intensive yet impractical to obtain the analytical solution due to the numerous unknown parameters of the D-MOSFET. To this end, we investigate convolutional neural networks (CNNs) as an engineering solution. CNNs not only can simulate the convolution operation, but also adapt to various audio signals automatically. Thereby, we devise an encoder-decoder neural network with convolution layers to reconstruct the spectrogram. To enhance the reconstruction performance in face of RF noises, we combine the attention mechanism and devise a dedicated loss function. Finally, we leverage a novel neural vocoder, Parallel WaveGAN, to precisely recover the audio waveform.

In sum, our work makes the following contributions:

- We propose the first analog RFRA system, RF-PARROT^2, for eavesdropping on wired analog audio signals. RF-PARROT^2 presents a novel retroreflector design by employing the D-MOSFET.
- We demonstrate that RF-PARROT^2 can intercept analog audio signals at a distance of 1.5 m through the wall. We also develop a novel approach to address the audio distortion issue resulting from the nonlinear transfer curve of D-MOSFET.
- We evaluate RF-PARROT^2 over 24-hour long-context speech and 60000 speech commands from thousands of people in various environments. Experiments show we can achieve an impressive average Mel-cepstral distortion (MCD) of 3.66[1] on long-context speech. For speech commands, we achieve an average MCD of 4.2 and 95% command recognition accuracy.

2 Analog Audio Eavesdropping via RFRA: Attack Model & Attack Scenarios

This section introduces the attack model and attack scenarios of audio eavesdropping via RFRA.

2.1 Attack Model

As shown in Fig. 1(b), RF-PARROT^2 aims to eavesdrop on audio signals transmitted in the target audio device. We have the following assumptions for RF-PARROT^2:

Target Device. The potential victims mainly involve wired audio systems, for example, earphone wires, telephone wires, audio wires in both personal and professional acoustic devices.

Retroreflector Installation. We assume the attacker preinstalled the retroreflector in the target device. This can be done during fabrication or intermediate stages of the audio wire supply chain. The retroreflector can be made from high-compact um-sized MOSFET and effectively concealed inside the wire. In addition, the retroreflector only reflects RF signals without active emission; thus such installation would not be detected by victims.

Attack setup. The attacker is equipped with commodity SDR and a pair of transmitting and receiving antennas. During audio interception, the attacker does not need physical access to the target device. Moreover, the target device can be both acoustically and visibly isolated from the attacker.

2.2 Attack Scenarios

Many audio systems are vulnerable to the RF-PARROT^2 attack. We list several typical attack scenarios:

- *Personal Scenario.* Online meetings may touch upon private information, e.g. confidential phone numbers,

1. MCD below 8 indicates a high-fidelity of audio reconstruction.

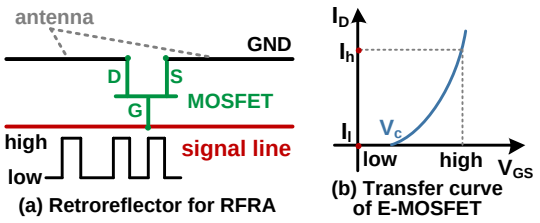


Fig. 2: (a) Retroreflector installation (b) E-MOSFET transfer curve

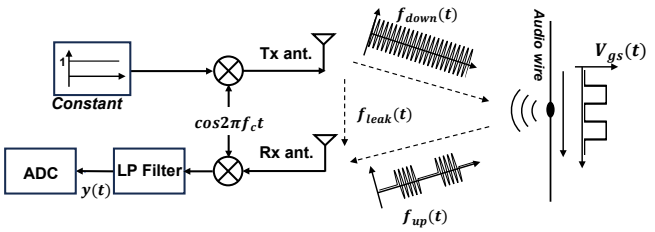


Fig. 3: Signal model in a digital RFRA system

passwords, credit card numbers, etc. In most cases, the remote attendees' voice is difficult to eavesdrop on if using earphones and headphones, because the voice is usually not played out loud through the speakers. However, attackers can secretly listen to attendees' voices from the earphone wire using RF-PARROT[2].

- *Industrial Scenario.* With the prevalence of voice control in industrial scenarios, the attacker can infer the speech commands with RF-PARROT[2] to steal important information of production from machines. The speech commands usually contain the confidential key words indicating the steps to produce products, some of which are business secrets.
- *Military Scenario.* Digit and action commands are commonly applied for message and task conveying in military systems. RF-PARROT[2] can eavesdrop on the sensitive action commands (e.g., 'go' and 'stop') and code number (i.e., 0 - 9) in audio transmission wires with a simple retroreflector without being noticed.

3 Analog Audio Eavesdropping via RFRA: Concept & Preliminaries

This section introduces the signal model of digital and analog RFRA, feasibility studies, and preliminary results.

3.1 Signal Model in a Digital RFRA System

MOSFETs contain three pins: Drain (D), Source (S) and Gate (G). In RF-PARROT[2], the G is connected to the signal line, while D and S are connected to the GND line, as shown in Fig. 2(a). Thus, we denote the digital signal, i.e., the relative voltage between G and S, as $V_{gs}(t)$,

Fig. 3 shows the signal model in a digital RFRA system consisting of a TX antenna, a RX antenna and an audio wire with an E-MOSFET retroreflector. The TX antenna transmits a single-tone continuous wave (CW) to the retroreflector, i.e., $f_{down}(t) = cos(2\pi f_c t)$, where f_c is the carrier frequency. When the CW arrives at the retroreflector, the GND line aside the D and S of the E-MOSFET will generate inductive current I_D and work like a dipole antenna which reflects the CW back. We denote the reflected signal as $f_{up}(t)$. Note that there is a leakage signal $f_{leak}(t) = A_l \cos(2\pi f_c t)$ propagated from TX antenna to the RX antenna directly, where A_l is the decayed amplitude. After down-conversion ($\times \cos(2\pi f_c t)$) and low-pass filter (LP filter), we acquire the raw eavesdropped audio signal $y(t)$. The E-MOSFET retroreflector switches between two states, i.e., non-reflective and reflective, as the binary signal varies between 0 and 1, making $y(t)$ as follows:

- **Non-reflective.** When a bit zero[2] passes through the signal line (i.e., $V_{gs}(t) = 0$), the induced current I_D is zero, making the retroreflector non-reflective. According to the transfer curve[3] in Fig. 2(b), the reflected signal $f_{up}(t)$ is 0. Thus, only the leakage signal $f_{leak}(t)$ is intercepted by the RX antenna.

$$\begin{aligned} y(t) &= LP[f_{leak}(t) \times \cos(2\pi f_c t)] \\ &= LP[A_l \cos^2(2\pi f_c t)] \\ &= LP[\frac{1}{2}A_l(1+\cos(4\pi f_c t))] = \frac{1}{2}A_l \end{aligned}$$

- **Reflective.** When a bit one passes through the signal line ($V_{gs}(t) > 0$), I_D is relatively high according to Fig. 2(b). Then, the retroreflector becomes reflective and $f_{up}(t)$ varies with $V_{gs}(t)$, i.e., $f_{up}(t) = V_{gs}(t)\cos(2\pi f_c t)$. Thus, y(t) is represented as:

$$\begin{aligned} y(t) &= LP[(f_{up}(t) + f_{leak}(t)) \times \cos(2\pi f_c t)] \\ &= LP[(V_{gs}(t) + A_l)\cos^2(2\pi f_c t)] \\ &= LP[\frac{1}{2}(V_{gs}(t) + A_l)(1+\cos(4\pi f_c t))] = \frac{1}{2}V_{gs}(t) + \frac{A_l}{2} \end{aligned} \quad (1)$$

Compared with the non-reflective state, $y(t)$ now contains an additional $\frac{1}{2}V_{gs}(t)$ desired by the attacker. Thus, they can eavesdrop on the wired digital signal by analyzing the amplitude changes of $y(t)$ between the above two states.

3.2 Signal Model in Our Analog RFRA System

3.2.1 Signal model of D-MOSFET retroreflector

RFRA becomes much more complicated when transmitting analog audio signals. First, different from the digital signal, analog audio contains negative voltage. However, the cutoff voltage V_c of E-MOSFET is positive, so the amplitude of the received signal's negative part is zero. As such, the audio signal below the cutoff voltage is completely lost.

To tackle this issue, we devise a new retroreflector using the D-MOSFET whose transfer curve is depicted in Fig. 4. D-MOSFET owns a preferable characteristic for RFRA on the analog audio signal, i.e., a negative cutoff voltage V_c. Then, I_D can be flexibly controlled by both negative and positive parts of the audio signal, making the received RF signal amplitude continuously change with the audio voltage. Thus, we have the potential to restore the original audio signal by employing the D-MOSFET.

To understand D-MOSFET, we carefully debunk its signal model. The signal model of an analog RFRA system shares the same RF front end (i.e., the TX and RX antennas, LP Filter and ADC) with the digital one. The difference is that, instead of switching between reflective and non-reflective states, the D-MOSFET works like a 'soft switch', modulating the reflected signal continuously. In other words, the reflected signal from the retroreflector is a transformed version of the audio signal, i.e., $f_{up}(t) =$

2. In digital systems, a bit zero/one is typically transmitted via low-level/high-level voltage.

3. The transfer characteristic curve refers to the drain current vs. gate to source voltage curve.

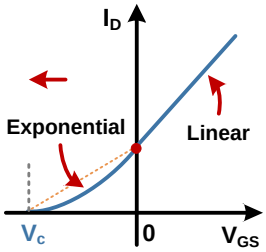


Fig. 4: D-MOSFET transfer curve

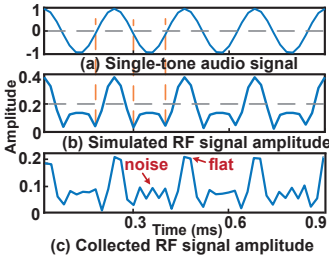


Fig. 5: Simulated & collected signals

$T(V_{gs}(t))\cos(2\pi f_c t)$, where $T(\cdot)$ represents the transfer curve of D-MOSFET. Similar to Eq. (1), $y(t)$ becomes:

$$\begin{aligned} y(t) &= LP[(f_{up}(t) + f_{leak}(t)) \times \cos(2\pi f_c t)] \\ &= \frac{1}{2}T(V_{gs}(t)) + \frac{A_l}{2} \end{aligned} \tag{2}$$

3.2.2 Nonlinearity of D-MOSFET retroreflector

Although the negative audio signal is preserved by the D-MOSFET retroreflector, its transfer curve incurs a non-negligible effect for effective audio eavesdropping. As shown in Fig. 4, part of the transfer curve is nonlinear, which means the captured RF signal would be distorted. Thus, we dive into the effect of the nonlinear transfer curve from different V_{gs} levels as follows.

- When $V_{gs} \geq 0$, I_D varies with the V_{gs} in an approximately linear way. Thus, the received RF signal amplitude proportionally changes with the audio signal voltage V_{gs}, i.e., $T(V_{gs}) = \alpha \cdot V_{gs}$.
- When $V_c \leq V_{gs} < 0$, I_D nonlinearly changes with the V_{gs}. The reflected RF signal amplitude will experience a nonlinear decay of the original audio signal. The nonlinearity in MOSFET is generally represented by a pseudo-exponential function [19], in which the exponent is controlled by the audio signal itself; then the received signal amplitude[4] is $T(V_{gs}) = -e^{\gamma \cdot V_{gs}} \cdot V_{gs}$. $T(V_{gs})$ will decay when V_{gs} is approaching V_c, and γ determines the decaying slope here.
- When $V_{gs} < V_c$, there is no current on the drain; thus the reflected RF signal amplitude $T(V_{gs}) = 0$. Note that the cutoff voltage of common D-MOSFETs are around -1 V, which is sufficient to react to the wire-transmitted audio signal voltage falling within a low voltage range of hundreds of millivolts [20].

Summarizing the above modeling, we can represent the relationship between the audio signal V_{gs} and the received RF signal amplitude $T(V_{gs})$ with the following equation.

$$T(V_{gs}) = \begin{cases} \alpha \cdot V_{gs}, & V_{gs} \geq 0 \\ -e^{\gamma \cdot V_{gs}} \cdot V_{gs}, & V_c \leq V_{gs} < 0 \\ 0, & V_{gs} < V_c \end{cases} \tag{3}$$

To validate this model, we first conducted an experiment where we transmitted a single-tone analog audio signal through a coaxial wire containing a D-MOSFET retroreflector (details provided in Section 4.2). Second, we applied Eq. (3) to simulate the received RF signal amplitude, setting α and γ to 1. We then compared the simulated signal with the collected one and obtain the following observations:

- The simulated signal amplitude (Fig. 5(b)) closely resembles the collected signal (Fig. 5(c)), demonstrating the effectiveness of our modeling approach.
- The periodicity of the single-tone audio signal is maintained, particularly in positive segments. However, due to nonlinear decay within $[V_c, 0)$, the negative part of the single-tone audio signal is flipped with reduced amplitudes. Besides, the weaker strength of this segment makes it more sensitive to noises.
- A flattening of the amplitude peak is observed in the collected RF signal amplitude. This discrepancy arises when the inductive current I_D exceeds the upper limit I_{max} induced by the TX antenna. When I_D exceeds I_{max}, the received signal amplitude will remain at the maximum. This issue can be potentially addressed by enhancing the antenna gain.

Note that a key issue when employing the retroreflector is whether it affects normal audio transmission. To quantify the audio distortion introduced by the retroreflector, we employ the total harmonic distortion (THD) [21]. The average THD of human speech transmitted through three retroreflector-embedded audio wires is consistently below -20 dB, indicating less than 1% distortion. Thus, the MOSFET retroreflector would not disturb normal audio transmission.

In summary, utilizing the D-MOSFET-based retroreflector enables comprehensive extraction of both negative and positive parts of the audio signal. This feature is advantageous for successful eavesdropping. However, the nonlinearity in the D-MOSFET transfer curve and unpredictable RF noises lead to the detrimental transformation of the original audio. In Section 4.5, we propose an effective solution to address the nonlinearity and noise issues.

3.3 Feasibility Study and Analysis

3.3.1 Feasibility Study

We conducted preliminary experiments to assess the feasibility of using the D-MOSFET for analog audio eavesdropping. First, we played English speech audio containing 10 digit commands (0 - 9) spoken by six individuals using a smartphone. The D-MOSFET assembled on a tiny PCB is embedded in the middle of a 1m-long wire connected to the phone's audio jack. Then, we transmitted CW from an SDR and collected the reflected signal using a pair of directional antennas. We quantitatively evaluated the similarity between the original audio and the received RF signal by calculating the Mel-cepstral distortion (MCD), which measures the difference between the eavesdropped and the original speech. Meanwhile, we obtained MCD values using an E-MOSFET for comparison. Note that a smaller MCD value indicates better audio eavesdropping performance. As listed in Table 1, the average MCD of D-MOSFET over 0 to 9 digits is 16-18, less than that of the E-MOSFET, demonstrating that D-MOSFET is more effective in keeping the audio information. On the other hand, efforts are required to tackle the nonlinearity problem to reduce the MCD within 8,

TABLE 1: MCD of eavesdropped signal by D-MOSFET and E-MOSFET

	0	1	2	3	4	5	6	7	8	9
D-MOS	16.9	17.0	16.8	17.7	17.0	17.4	17.7	18.4	16.2	18.5
E-MOS	20.6	23.4	23.9	22.9	22.3	22.6	22.1	24.2	25.9	25.4

4. The negative sign '-' is introduced to make $T(V_{gs})$ positive as I_D is positive.

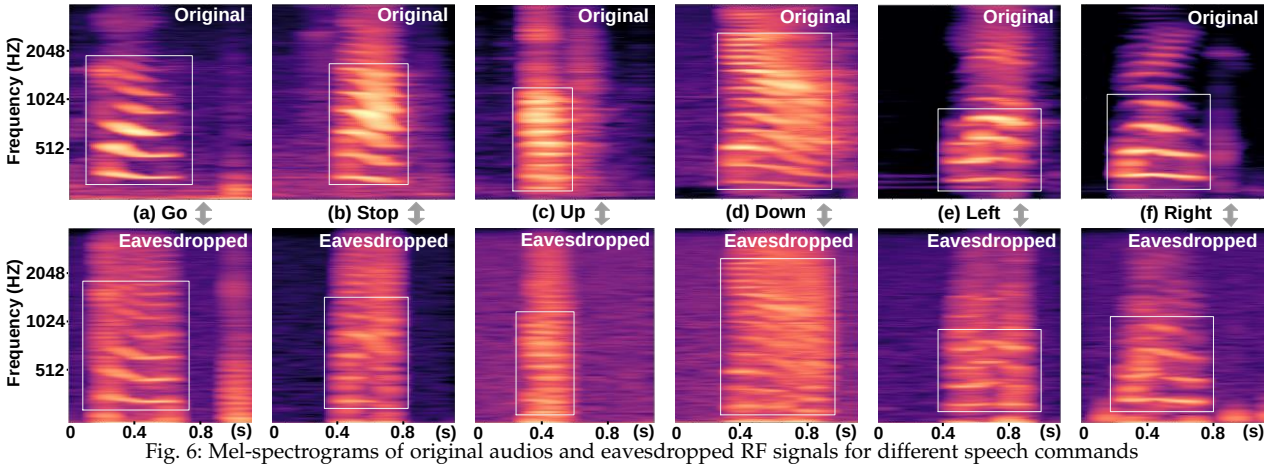


Fig. 6: Mel-spectrograms of original audios and eavesdropped RF signals for different speech commands

under which the eavesdropped audio can be recognized by common speech recognition systems [22]. Upon listening to the eavesdropped wav files, we observe that most stressed syllables are discerned, but the light plosive sounds exhibit distortion. Thus, further audio enhancement is demanded to improve interception quality.

Next, we extracted the Mel-spectrograms of the original and eavesdropped audio signals involving several typical speech commands (e.g. 'go', 'stop', 'up', 'down', 'left', 'right') in Fig. 6. First, we compare the similarity between Mel-spectrograms of the original audio and that of the received RF signal amplitude for each command. The primary frequency components present in the original audio are generally preserved, as shown in the boxed areas of each command's Mel-spectrograms. Second, we compare the eavesdropped signal's Mel-spectrogram of different commands. As depicted in the second row of Fig. 6, the eavesdropped Mel-spectrograms of the six commands exhibit distinct patterns, indicating that the eavesdropped signals of different speech audios are distinguishable.

Based on the above quantitative results in Table. 1 and visualization in Fig. 6, the eavesdropped signals show high similarity with the original signals. Therefore, these findings indicate the strong feasibility of using D-MOSFET for analog audio eavesdropping and speech recognition.

3.3.2 Analysis: degradation of eavesdropped signals

However, despite the similarity, the Mel-spectrograms of the eavesdropped signal still exhibit certain signal degradation in comparison to those of the original audio. This degradation primarily stems from the nonlinearity introduced by Eq. (3), which causes distortion between the eavesdropped and original signals. As we can see from Fig. 5 and Fig. 6, the negative outcomes of nonlinearity include (i) the 'flipped-and-attenuated' signal in the time domain, and (ii) the degraded frequency components in the frequency domain. The second issue (ii) is a direct consequence of the first (i), as the negative part of the audio signal is flipped over, which introduces additional frequencies to hamper the audio. To address this challenge, we will propose a mathematical analysis and an engineering solution in Section 4.5.

Another concern is noise interference. We can observe white noise in the eavesdropped spectrograms from Fig. 6. The intensity of white noise is near equal over the frequency and time axis. We will solve this problem in Section 4.4.

4 RF-PARROT^2 SYSTEM DESIGN

In this section, we introduce the details of RF-PARROT^2, including the retroreflector fabrication, RF signal setup, signal pre-processing, speech reconstruction and recognition.

4.1 Overview of RF-PARROT^2

The workflow of RF-PARROT^2 is depicted in Fig. 7 with five modules. The first module introduces the fabrication and integration of the D-MOSFET-based retroreflector into the audio wire. Second, we meticulously select an appropriate RF signal frequency to ensure effective reception of the backscattered signal emitted by the wire. Third, the signal preprocessing module eliminates noises and redundancy in the received RF signal and acquires the corresponding Mel-spectrogram. Then, we develop a two-phase audio reconstruction module, which first maps the distorted Mel-spectrogram to the original one and then transforms it to the audio waveform via a neural vocoder. Finally, the reconstructed Mel-spectrogram is fed into a speech recognition model to automatically decipher the speech content.

4.2 Design and Fabrication of Retroreflector

The retroreflector circuit diagram is illustrated in Fig. 8(a), comprising two primary components: an N-channel D-MOSFET and a 10 KΩ resistor. The resistor is to safeguard the MOSFET from potential damage due to overcurrent. The gate of the D-MOSFET is linked to the inner conductor of the audio wire, responsible for carrying the analog signal, while the drain and source are connected to the metallic shield (i.e., the GND) on each side of the wire after being split. In our implementation, as depicted in Fig. 8(b), the D-MOSFET and resistor are mounted on a small PCB for ease of assembly. The retroreflector's size can be further reduced by using high-compact nm-sized MOSFETs, which are much smaller than the diameters of common audio wires (3 mm to 5 mm) [23]. Thereby, the retroreflector can be discreetly installed inside the audio wire.

4.3 Transmitting and Receiving RF Signal

As shown in Fig. 2(a), the audio wire on each side of the retroreflector functions as a dipole antenna. To enhance the RF signal reflected from the dipole antenna, a meticulous RF signal tune-up is conducted, aligning it with the resonant frequency of the dipole antenna. For an audio wire of length of L, the dipole antenna would resonate at the odd

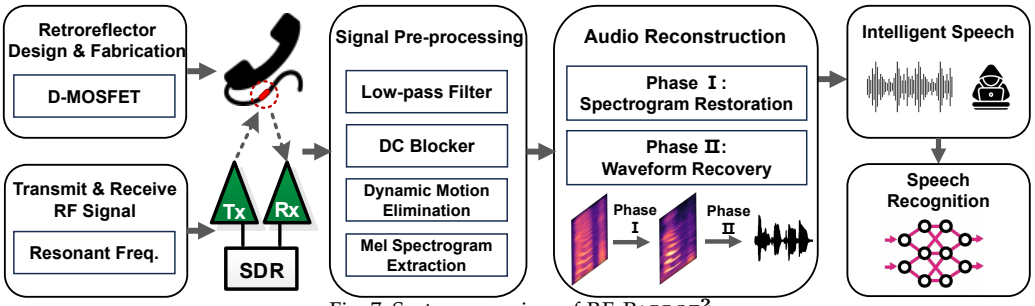


Fig. 7: System overview of RF-PARROT[2]

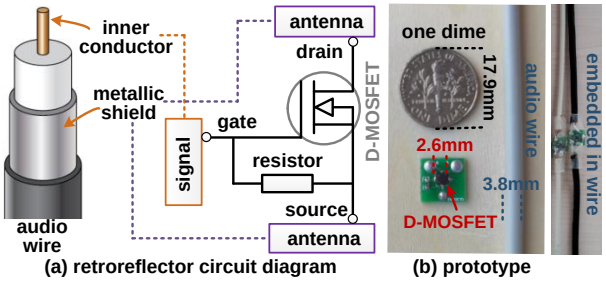


Fig. 8: Retroreflector fabrication: (a) circuit diagram and (b) prototype

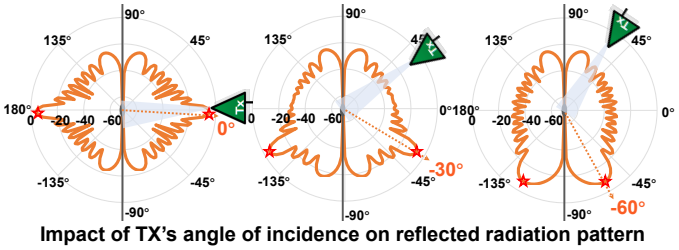


Fig. 9: Simulation of reflected radiation pattern on a straight wire with the TX antenna's angle of incidence being 0°, 30°and 60°. The red stars denote the direction of maximum reflected radiation.

multiples of half-wavelength for the RF signal [24], [25], i.e., $L = (2n-1)\cdot\lambda/2$, where $n \in \mathbb{N}_+$. Then, the candidate resonance frequency f_r becomes:

$$f_r = \frac{2n-1}{2L}\cdot c, n \in \mathbb{N}_+,$$

where c is the light speed. Typical audio wires are around 1 m-long, i.e., $L = 1$ m. Thus, f_r can be any odd multiple of 150 MHz. However, to minimize the risk of detection by defenses, it is preferable to emit the RF signal at commonly used frequencies, e.g., the popular 2.4 GHz band adopted by WiFi and Bluetooth. As such, an attacker's signal will be overwhelmed among the numerous signals around 2.4 GHz, minimizing the risk of being identified. The nearest resonant frequency to 2.4 GHz is 2.25 GHz = 150 MHz $\times$ 15, i.e., f_r = 2.25 GHz. Moreover, the RF signal at this frequency can penetrate through the wall. Thus, a CW at 2.25 GHz is transmitted from the TX antenna on the SDR, then reflected by the audio wire and received by the RX antenna with a sampling rate of 200 KHz.

To strengthen the signal reflected by the wire-form dipole antenna, we also investigate how to deploy the TX and RX antennas. Specifically, we employ CST Microwave Studio to simulate the radiation pattern of the 1m-long dipole antenna with a retroreflector inserted in the middle. The parameters of the dipole antenna align with the typical SVY75-2-1 wire. As shown in Fig. 9, the direction with maximum radiation is symmetrical with respect to the TX antenna direction (i.e., 0°, 30°and 60°). Based on the above

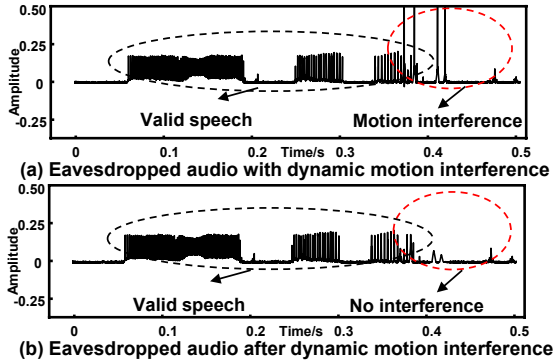


Fig. 10: The waveforms (a) before and (b) after elimination

observations, since attackers may not know the wire position in advance, it is more efficient for attackers to collocate the TX and RX antennas and gradually move the antenna pair to find the position with the strongest signal. Note that the audio wire is straight during simulation. However, assuming the wires to be ideally straight is not practical, because the shape may affect the characteristics of the dipole antenna. In Section 6.5.5, we investigate the performance of RF-PARROT[2] with several typical wire shapes.

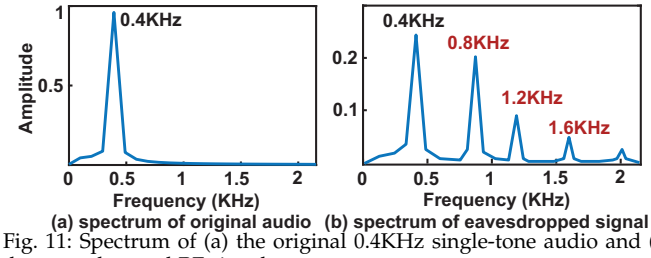
4.4 Signal Pre-processing

We pre-process the received RF signal as follows for further audio reconstruction.

Low-pass Filter. We apply a low-pass filter to eliminate high-frequency RF white noises. As the speech frequency band is mainly below 4 KHz [26]–[28], the cut-off frequency is set to 8 KHz to fully capture the audio signal.

DC Blocker. The received signal involves a DC component which comes from (i) the constant term $\frac{A_l}{2}$ from Eq. (2) (i.e., the leakage from TX antenna), and (ii) the multipath reflection from the static environment. As the DC component from multipath reflection changes significantly with different environments, it causes environment-dependent amplitude variance. So we apply a DC blocker with a sliding window to eliminate the DC component and normalize the eavesdropped signal with zero mean value.

Dynamic Motion Elimination. Due to the uncertainties in the environment, the received RF signal may contain the interference from users' dynamic motions around the target device. Fig. 10(a) shows the eavesdropped audio bothered by the body motion interference with sharp amplitude peaks. To eliminate the interference, we first apply a threshold to eliminate the motion interference signal. This threshold is set empirically at 1.5 times the maximum audio signal. Second, we apply a linear interpolation to fill the gap after elimination. The eavesdropped audio after dynamic motion elimination is depicted in Fig. 10(b).



(a) spectrum of original audio (b) spectrum of eavesdropped signal

Fig. 11: Spectrum of (a) the original 0.4KHz single-tone audio and (b) the eavesdropped RF signal

Mel-spectrogram Extraction. First, we re-sample the signal at a reduced sampling rate of 16 KHz to minimize the computation cost. As a common sampling rate for voice audio, 16 KHz ensures acceptable frequency perception with limited computation load. Since the upper limit frequency of audio is about 3400 Hz , the sampling rate that is higher than 6800 Hz (3400 Hz $\times$ 2) is enough to perfectly carry the full information of original audio, according to Nyquist sampling theorem. Therefore, such down-sampling would save the storage load while not affecting eavesdropping performance. Then, we divide the signal into fixed-length segments and calculate the Mel-spectrograms. The selection of Mel-spectrogram among other spectrograms lies in the alignment with human ear's perception, with a non-linear Mel frequency scale, expressed as $Mel(f) = 2595\log_{10}(1 + \frac{f}{700})$ [29]. The extracted Mel-spectrograms will be used for audio reconstruction.

4.5 Audio Reconstruction

The challenge of precise audio reconstruction, as discussed in Section 3.2.2 and Section 3.3.2, mainly stems from the nonlinearity $T(\cdot)$ brought by the transfer curve. Fig. 5 has shown the distortion in the time domain. The nonlinearity flips and attenuates the negative part of the original signal. It means for 'valleys' (i.e., the minima of the negative part) in the original signal, they flip over with reduced amplitude. The 'flip-and-attenuate' operation in the time domain causes harmonic distortion[5] in the frequency domain, leading to the additional frequency components. We compare the spectrum of the original pure 0.4 KHz single-tone audio signal with the eavesdropped signal in Fig. 11. Apart from the original 0.4 KHz frequency component, the spectrum of the eavesdropped RF signal also contains additional 0.8 KHz, 1.2 KHz, 1.6 KHz and higher components. Those high harmonics result in inaccurate perceiving of the audio content.

In response to the distortions above, we first derive the underlying mathematical mechanism of the nonlinear effect on the eavesdropped signal. Then, we propose a two-phase audio reconstruction module to precisely reconstruct original audio speech.

4.5.1 Mathematical modeling of the nonlinear effect

In the time domain, we need to find a reverse function $T^{-1}(\cdot)$ to fully recover the original audio signal $V_{gs}(t)$ from the distorted $T(V_{gs}(t))$, that is:

$$V_{gs}(t) = T^{-1}(T(V_{gs}(t)))$$

It seems we may derive $T^{-1}(\cdot)$ given the transfer curve. However, this is not easy for two reasons. First, the transfer curve parameters, α and γ, are unknown to us. Although one can pre-collect multiple traces of the RF signal to fit α and γ, it is quite labor-intensive. Besides, the fitting result could be error-prone due to signal noise. Second, for a positive $T(V_{gs}(t))$ value, we cannot decide whether $V_{gs}(t)$ is positive or negative since the function $T(\cdot)$ is not an injective function (i.e., for $x_1 \neq x_2$, $f(x_1) = f(x_2)$). Therefore, an alternative solution is required.

To this end, we propose an engineering solution in the frequency domain. First, we perform fast Fourier transformation (FFT) on the received RF signal amplitude $y(t)$ based on Eq. (2) and Eq. (3). Suppose the FFT of the original audio signal $V_{gs}(t)$ is $\mathscr{F}\{V_{gs}(t)\}$. When $V_{gs}(t) \geq 0$, we can represent the FFT of $y(t) = \frac{1}{2}\alpha \cdot V_{gs}(t)$ as $Y(f) = \frac{1}{2}\alpha \cdot \mathscr{F}\{V_{gs}(t)\}$ based on the linearity property of FFT; thus the audio fidelity is preserved for the positive part. However, when $V_c \leq V_{gs}(t) < 0$, the FFT of $y(t) = -\frac{1}{2}e^{\gamma \cdot V_{gs}(t)} \cdot V_{gs}(t)$ becomes complicated. We decompose $y(t)$ as follows:

$$a_1(t) = -\frac{1}{2}e^{\gamma \cdot V_{gs}(t)},\ a_2(t) = V_{gs}(t),\ y(t) = a_1(t) \cdot a_2(t)$$

Based on the convolution theorem, $Y(f)$ is expressed as:

$$\begin{aligned} Y(f) &= \mathscr{F}\{a_1(t) \cdot a_2(t)\} \\ &= A_1(f) * A_2(f) = A_1(f) * \mathscr{F}\{V_{gs}(t)\} \end{aligned}$$

To losslessly recover $\mathscr{F}\{V_{gs}(t)\}$ out of $Y(f)$, we need to find $A_1^{-1}(f)$, in which $A_1^{-1}(f) * A_1(f) = \delta(f)$, where $\delta(f)$ is the FFT of the constant signal 1. Then, we can conduct convolution between $A_1^{-1}(f)$ and $A(f)$ as below.

$$\begin{aligned} A_1^{-1}(f) * Y(f) &= A_1^{-1}(f) * A_1(f) * \mathscr{F}\{V_{gs}(t)\} \\ &= \delta(f) * \mathscr{F}\{V_{gs}(t)\} = \mathscr{F}\{V_{gs}(t)\} \end{aligned} \quad (4)$$

If $A_1^{-1}(f)$ is acquired, we can get the precise spectrum $\mathscr{F}\{V_{gs}(t)\}$ from $Y(f)$ by performing convolution operation between $A_1^{-1}(f)$ and $Y(f)$. Then, after inverse FFT on $\mathscr{F}\{V_{gs}(t)\}$, we can restore the original audio signal $V_{gs}(t)$. In a word, the audio reconstruction task lies in how to compute $A_1^{-1}(f)$. Similar to obtaining $T^{-1}(\cdot)$, the analytical expression of $A_1^{-1}(f)$ is also hard to obtain due to the unknown parameters.

Other than the analytical solution, we may use convolutional neural networks (CNNs) as an engineering solution. Specifically, we employ the CNN to substitute the convolution operation in Eq. (4). In addition, neural networks also show notable advantages in automatically adapting to different speech signals and combating the signal noises and harmonic distortion.

4.5.2 Phase I: spectrogram restoration

Our goal is to build a CNN model for restoring the nonlinearly deformed spectrogram to its original version. To achieve this, we utilize the encoder-decoder U-net [30], which is widely applied for translation, spectrogram recovery and other mapping purposes [31]–[36].

However, directly employing U-net may not effectively restore the original spectrogram due to the RF signal noises during silent periods. As shown in Fig. 6, the darker areas in the Mel-spectrogram are primarily from the inherent noises in the RF transceiver. These noises are not conducive to training the restoration network. Thus, we adopt two main improvements on U-net to enhance the model. First,

5. Harmonic distortion is defined as additional signal content at multiples of the original frequency.

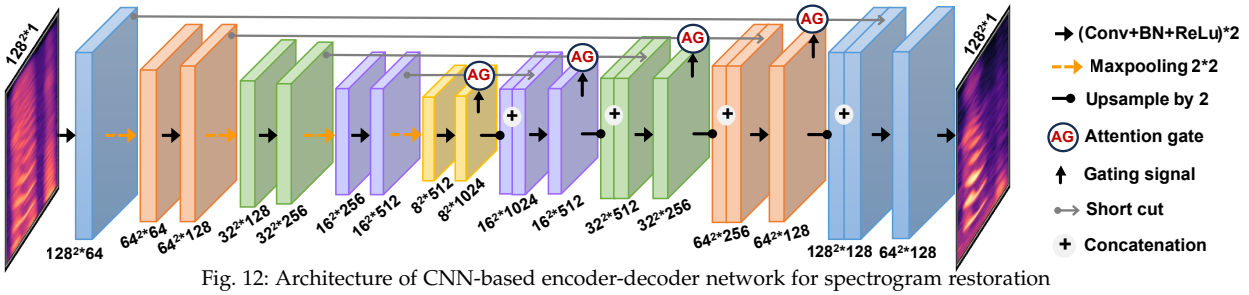


Fig. 12: Architecture of CNN-based encoder-decoder network for spectrogram restoration

inspired by the attention mechanism, which enables selective focus on the most significant and pertinent portions of the input, we adopt attention U-net architecture [37]. It can learn to focus on target structures of varying audio frequency components. In this way, the network will pay more attention to the audio component over background noise during training.

Second, we modify the loss function of U-net. Traditional U-net networks typically employ the mean squared error (MSE) loss, treating every pixel equally important. However, in our case, the audio component should have a greater significance than the background noise during training. Thus, we propose a new loss function that assigns a higher weight to the squared error of the audio signal part. Specifically, we calculate the weight w_i for each pixel based on its normalized strength in decibels. Since the audio part has a higher amplitude than the background noise, the weights are assigned proportionally. The formulated loss function is as follows.

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot (p_i - \hat{p_i})^2, \; w_i = \sqrt{\frac{p_i - p_{min}}{p_{max} - p_{min}}}, \qquad (5)$$

where p_i and $\hat{p_i}$ represent the pixel values in the Mel-spectrograms of the original audio and reconstructed audio, respectively. p_{min} and p_{max} denote the minimum and maximum pixel strength. N is the total number of pixels.

The restoration network architecture is depicted in Fig. 12. It can achieve a high-quality reconstruction of the Mel-spectrogram. As depicted in Fig. 13, compared with the raw Mel-spectrogram of the eavesdropped RF signal in Fig. 13(b), the restored Mel-spectrogram in Fig. 13(c) exhibits almost the same distinct frequency components with the Mel-spectrogram of the original audio signal in Fig. 13(a).

4.5.3 *Phase II: waveform recovery*

After obtaining the reconstructed Mel-spectrogram, we need to transform the spectrogram into the audio waveform. To do this, we utilize a novel neural vocoder, known as Parallel WaveGAN [38]. The 2-D Mel-spectrogram of the speech signal is the input into the trained vocoder, enabling the conversion into 1-D natural-sounding audio signal. Parallel WaveGAN contains a generator (G) and a discriminator (D). The generator, which is composed of 30 stacked residual blocks with dilated non-causal convolution layers, recovers the audio waveform as the output. The dilated convolutions allow networks to achieve extensive receptive fields with minimal layers [39]. The discriminator is composed of ten layers of non-causal dilated 1-D convolutions. It takes the original speech and generated speech as input, and tries to classify the generated speech as real or fake.

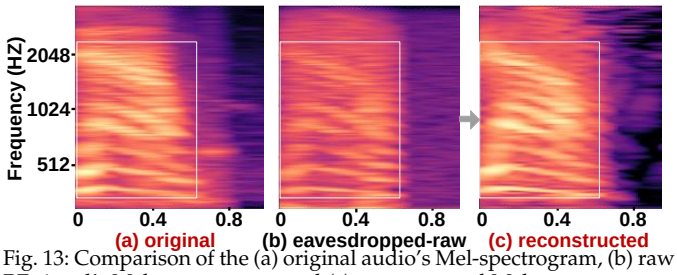


Fig. 13: Comparison of the (a) original audio's Mel-spectrogram, (b) raw RF signal's Mel-spectrogram, and (c) reconstructed Mel-spectrogram

The reason for choosing parallel WaveGAN lies in accurate phase estimation from restored Mel-spectrogram. The input Mel-spectrogram only involves the magnitude without phase, so preserving its quality is difficult. Traditional methods include (i) performing inverse Short Time Fourier (iSTFT) on input spectrograms, and (ii) applying Griffin-Lim algorithm [40], which iteratively estimates the phase spectrum with a random phase initialization. However, neither of these methods can accurately estimate the audio phase due to the lack of capability in perceiving amplitude-phase relationship. As such, we adopt Parallel WaveGAN. It can precisely learn the amplitude-phase relationship between the waveform and Mel-spectrogram of audio clips.

4.6 Speech Command Recognition

We also utilize the Mel-spectrograms of the reconstructed audios to automatically recognize various speech commands transmitted in the audio wire. Speech commands, e.g., digit and action commands, are widely applied for message and task conveying in military systems and large-scale industrial manufacturing, which involve critical information under spying risks. Thus, we demonstrate the ability of RF-PARROT^2 in speech command recognition. In specific, we employ the ResNet-50 convolutional network [41] to train the command recognition model on the digit and action commands. ResNet-50 is selected as it has been widely used in classifying speech spectrograms with high accuracy [42], [43].

5 IMPLEMENTATION

This section presents the implementation details, including experiment setup, datasets and training details.

5.1 Experiment Setup

The experiment setup is shown in Fig. 14. We conduct the experiment with a through-wall condition. The wall is made of double-panel tempered glass, with each panel' width being 5 mm. Also, there is a plastic shutter inside the two panels. As shown in Fig. 14, the shutter is closed to maximize the signal attenuation, as well as to create an invisible target environment from the attacker's view.

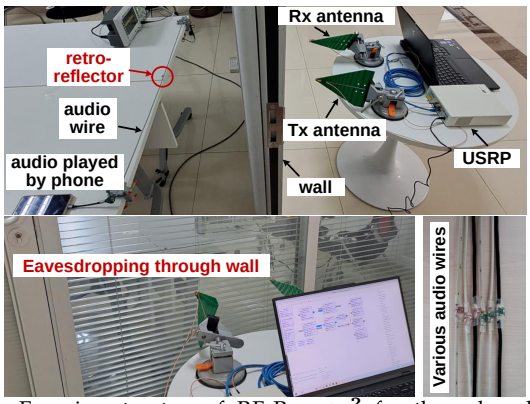


Fig. 14: Experiment setup of RF-PARROT^2 for through-wall audio eavesdropping

We implement RF-PARROT^2 with commercial devices, including a USRP N210 as SDR, a pair of SX200-150(P)C log-periodic antennas, and CE3512K2 D-MOSFETs. We select multiple types of audio wires and embed our retroreflector inside. The original audio files are played by a smartphone whose audio jack is connected to the audio wire. The distance between antennas and audio wire is 1 m as default. The RF signal is controlled by GNURadio, whose default TX and RX gains and RF signal sampling rate are set to 10 dB and 200 KHz, respectively. Moreover, the default volume of the smartphone is 80%. We will discuss the impact of distances, antenna gains, and volume in Section 6.5.

5.2 Speech Audio Datasets

We employ three public speech audio datasets, including Speech Commands Dataset (SCD) [44], Free Spoken Digit Dataset (FSDD) [45], and The LJ Speech Dataset (LJSpeech) [46]. FSDD contains 3,000 recordings of 10 English-spoken digits (0 - 9) from 6 subjects. SCD consists of 64,727 audio files collected from thousands of subjects, each containing an English-spoken command, i.e., 10 action commands ('Up', 'Down', 'Left', 'Right', 'On', 'Off', 'Stop', 'Go', 'Yes', 'No') and 10 digits commands (0 - 9). Different from FSDD and SCD, LJSpeech contains continuous long-context audio clips. It includes 13,100 audio clips of a single speaker with a total length of approximately 24 hours. The first two datasets are used for evaluating speech command recognition, and the third one is employed for evaluating long-context eavesdropping performance.

The speech audios are resampled at a rate of 16 KHz, with each sample quantized to 16 bits. After calibration, the shift between original audio and eavesdropped audio is less than 10 ms. Every audio file is aligned to 1.28 s. Mel-spectrograms are extracted with a hop length of 256 points. Prior to training, the Mel-spectrogram is normalized with zero mean and unit variance.

5.3 Training

The spectrogram restoration and audio waveform recovery were trained via Pytorch using the RTX 3090 GPU. They were separately trained but validated as a whole.

The spectrogram restoration model was trained with an SGD optimizer for 300 epochs. The learning rate and batch size were set at 1e-4 and 128, respectively. The waveform recovery model was trained with a Rectified Adam (RAdam) optimizer for 400k epochs. The initial learning rate for generator and discriminator were first set at 1e-4 and 5e-5, respectively, then reduced by half every 200k epochs.

6 EVALUATION

This section provides evaluation results of RF-PARROT^2 in audio eavesdropping.

6.1 Evaluation Metrics

We employ multiple metrics to comprehensively evaluate the performance of RF-PARROT^2 as follows:

• **Mel-cepstral distortion (MCD)** is an objective measure of the difference between the original audio's Mel-frequency cepstral coefficients and that of the reconstructed one. The MCD is calculated as:

$$MCD = \frac{10}{\ln 10} \sum_{t=0}^{T-1} \sqrt{2 \sum_{m=1}^{M} (V_m^{targ}(t) - V_m^{ref}(t))^2},$$

where V_m^{targ} and V_m^{ref} are the Mel frequency-scaled cepstral coefficients from the original and reconstructed audio, respectively. M is the length of cepstral coefficients. Smaller MCDs indicate better audio reconstruction performance. Note that the speech recognition system can perceive the reconstructed audio with an MCD less than 8.

• **Word error rate (WER)** is an objective measure of word recognition accuracy. The wrong words after reconstruction need to be substituted, deleted, or inserted with correct ones. Then, WER is calculated as follows:

$$WER = \frac{S + D + I}{N},$$

where S, D, I represent the number of substitutions, the number of deletions, and the number of insertions, respectively. N is the number of words in the original audio content. We recruited 20 volunteers to listen to 20 randomly selected original and corresponding reconstructed audio samples and identify the words spoken. A lower WER indicates higher comprehensibility of the reconstructed audio.

• **Signal-to-noise ratio (SNR) and Peak signal-to-noise ratio (PSNR)** are used to assess the quality of the audio signal. SNR is calculated by $SNR = 10\log(\frac{E_{ref}}{E_{noise}})$. PSNR is calculated as follows:

$$PSNR = 10\log_{10}(\frac{V_{max}^2}{\frac{1}{T}\sum_{t=1}^{T}[y^{targ}(t) - y^{ref}(t)]^2}),$$

$$V_{max} = \max(\max(y^{targ}(t)), \max(y^{ref}(t))),$$

where y^{targ} and y^{ref} are the waveforms of the original and reconstructed audios. A higher SNR value means less noise interfering with the signal. A higher PSNR value represents a higher quality of compressed signal compared with the original audio signal.

• **Short-time objective intelligibility (STOI)** is an objective measure of the speech intelligibility (from 0 to 1). Higher STOIs mean better speech enhancement performance from the distorted audio.

• **Mean opinion score (MOS)** is a subjective measure to evaluate the reconstructed audio quality. We recruited 20 volunteers, including 10 males and 10 females aged from 20 to 30, to listen to 20 randomly selected reconstructed audio and score the audio similarity compared with the original one. The score ranges from 1 to 5, where a higher score means a higher similarity.

• **Accuracy and F1-score** are common metrics evaluating the performance of classification. They are used to evaluate the 10 spoken digits and 10 action commands classification.

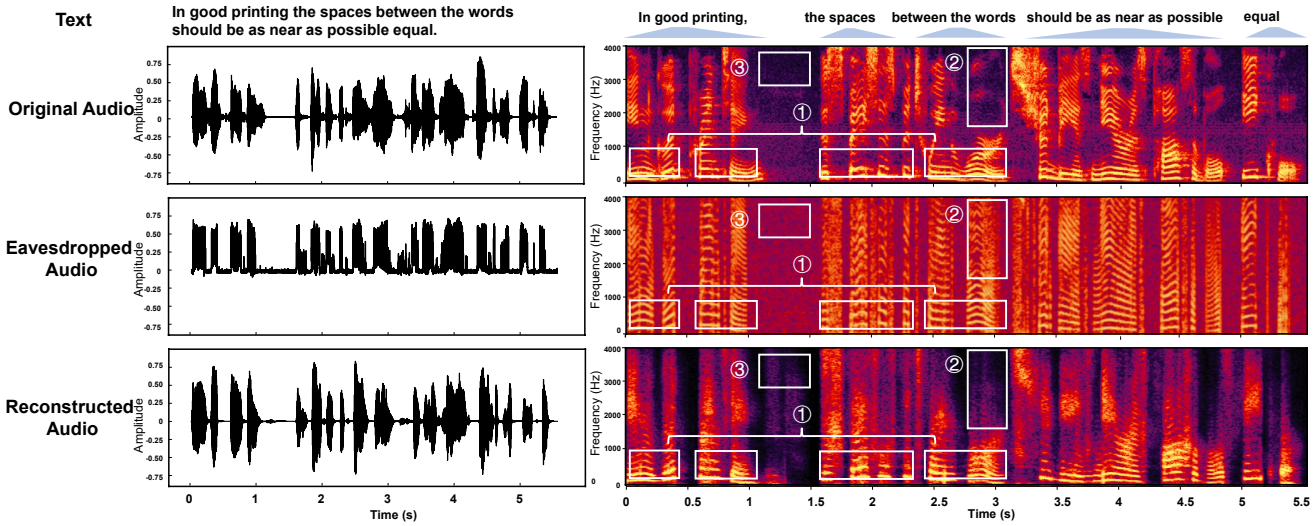


Fig. 15: The long-context audio of the LJ001-0134.wav from the LJ Speech dataset, resampled at 8 KHz. More results on speech commands and long-context audios are released on **https://genglinwang.github.io/RFRA4audio/**.

6.2 Performance: Long-context Audio Reconstruction

In this experiment, we investigate the audio reconstruction performance on LJSpeech that contains long-context audio clips. First, we quantitatively calculate the metrics of long-context audio reconstruction, as shown in Table 2. The average MCD value is 3.66, indicating that the speech recognition systems can detect and perceive the reconstructed audio. The average STOI, WER, and MOS are 0.68, 24.2%, and 3.43, respectively. This means the reconstructed speech audio is intelligible, semantically accurate, and subjectively understandable. The SNR of 7.39 means that the audio is clear and easy to detect or interpret, and the PSNR 25.31 usually correlates with a high quality reconstruction,

Next, we visualize and compare the original audio, the eavesdropped RF signal amplitude, and the reconstructed audio waveform and spectrogram in Fig. 15. We have four observations as follows:

The problem of 'flip-and-attenuate' operation in the time domain is largely addressed. As shown in the waveforms of Fig. 15, the negative amplitude of eavesdropped audio undergoes flipping and attenuation, affected by the nonlinearity of D-MOSFET retroreflector. After reconstruction, the negative audio re-flips and returns to its original amplitude. Notably, the reconstructed audio waveform shows high similarity with the original one. This indicates that our system can effectively learn the mapping from distorted eavesdropped waveforms to the original.

The low-frequency component of the reconstructed audio shows high similarity with the original audio. The low-frequency components of reconstructed audio (lower than 1500 Hz), especially those lower than 1000 Hz in the boxed area ① in Fig. 15, are highly similar to the original one. As the human voice is heavily dominant below 1500 Hz, the reconstructed audio contains most of the speaking voice. This demonstrates that our system can effectively capture the human voice information from the distorted audio.

The harmonic distortion is reduced after spectrogram restoration. The boxed area ② on the spectrogram of eavesdropped audio in Fig. 15 is full of harmonic frequency components, while the same area of box ② on the original one is much weaker. After spectrogram restoration, the high frequency components are significantly reduced, as shown in the boxed area ② of the reconstructed audio. This means RF-PARROT^2 is capable of eliminating the harmonic frequency components.

The system demonstrates strong resistance to background noises. Different from the original audio, the eavesdropped audio is plagued by background noises. As shown in the boxed area ③ of the eavesdropped audio in Fig. 15, the background is filled with white noise, whose frequency components are near-constant from 0 to 4 kHz. Yet, through audio reconstruction, the noise is eliminated. This suggests that our system can determine and mitigate noises.

TABLE 2: Quantitative evaluation results of long-context audio reconstruction. It is with 95% confidence intervals.

Dataset	MCD↓	STOI↑	WER↓
LJSpeech	3.66±0.06	0.68±0.002	24.2±4.8%
	SNR↑	PSNR↑	MOS↑
	7.39±0.13	25.31±0.25	3.43±0.15

* ↓/↑ means lower/higher value of the metric indicates better performance.

6.3 Performance: Speech Command Reconstruction

In this experiment, we investigate the audio reconstruction performance on speech commands. We first calculate the MCD value for the reconstructed audios with 10 digits and 10 action commands, respectively. As depicted in Fig. 16, the MCD values of the 20 commands' reconstructed audios are all below 8. The lowest MCD of 2.82 is achieved for the command 'down', while the highest MCD of 6.08 corresponds to the command 'two'. Meanwhile, we also compare the MCD of the audio eavesdropped by E-MOSFET, raw audio eavesdropped by D-MOSFET, and reconstructed audio eavesdropped by D-MOSFET in Table 3. The reconstructed audio with D-MOSFET achieves the lowest average MCD of 4.2, revealing the effectiveness of our proposed audio reconstruction method.

Second, we calculate the SNR and PSNR for the reconstructed audio of all 20 commands in Fig. 17. The average SNR and PSNR values of different commands fall within 0-5 dB and 15-25 dB, respectively, indicating that the essential audio information is more prominent compared to background noises. The command 'one' exhibits the lowest PSNR whereas the command 'down' attains the highest

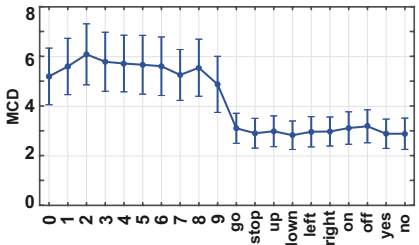


Fig. 16: MCD of reconstructed audio

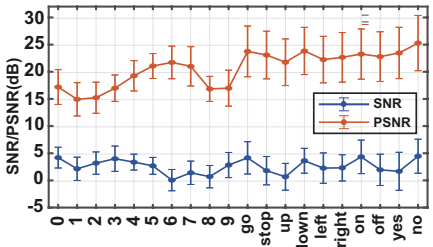


Fig. 17: SNR and PSNR of reconstructed audio

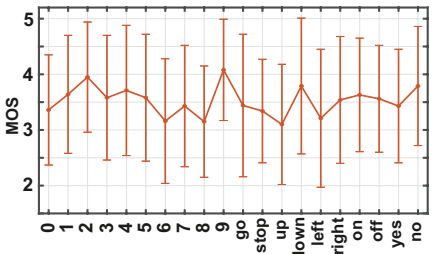


Fig. 18: MOS of reconstructed audio

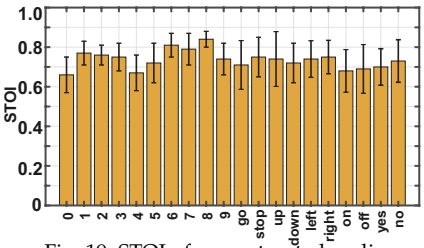


Fig. 19: STOI of reconstructed audio

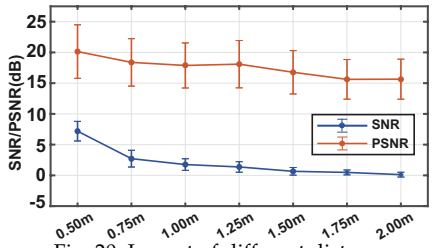


Fig. 20: Impact of different distance

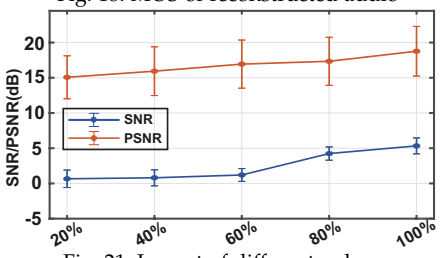


Fig. 21: Impact of different volume

TABLE 3: MCDs on speech commands using different MOSFETs

MOSFET type	E-MOSFET	D-MOSFET	
		Raw	**Reconstructed**
MCD	23.1	17.0	**4.2**

TABLE 4: F1-score of speech command recognition

Command	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
F1-score	0.97	0.95	0.87	0.86	1.00	0.95	0.87	0.91	0.97	0.94

Command	**go**	**stop**	**up**	**down**	**left**	**right**	**on**	**off**	**yes**	**no**
F1-score	0.95	1.00	0.98	0.95	0.97	1.00	0.90	0.95	0.97	1.00

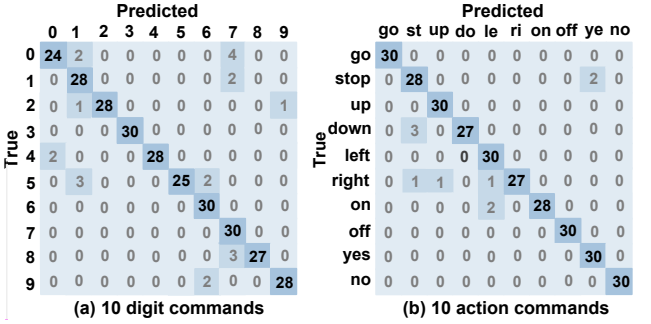


Fig. 22: Confusion matrix of (a) digit and (b) action commands

PSNR, which also aligns with the MCD results. Overall, the results show that the reconstructed audio signal can achieve a relatively high quality.

Third, we show the average MOS of each speech command among all volunteers in Fig. 18. The MOS values of all 20 commands range from 3.1 to 4.4 with an average of 3.6. The highest MOS can reach 4.1 for the command 'nine', and the lowest MOS of 3.4 for the command 'up' is still above the average level of 3, which means that volunteers think that over half of the original speech is recovered.

Finally, we calculate the average STOI of each speech command in Fig. 19. The average STOI values of every command vary from 0.66 to 0.84 with an average of 0.73. The command 'eight' achieves highest STOI, while the lowest is command 'zero'. Remarkably consistent is the performance within each specific command, with a maximum standard deviation of 0.14 on command 'up'. These findings underscore the stable intelligibility of the reconstructed audio across various commands, as well as within the same command. In a word, the above objective and subjective evaluation results validate that RF-PARROT^2 can realize high-quality speech commands eavesdropping.

6.4 Performance: Speech Command Recognition

In this experiment, we illustrate the speech command recognition performance. The F1-scores of all 20 commands are listed in Table 4. The highest F1-score can reach 1 for the 'four', 'stop', 'right', and 'no' commands, and the lowest F1-score is 0.86 for the 'three', yet still above 0.85. The detailed confusion matrices are shown in Fig. 22. Commands that have similar vowels, e.g., 'on' and 'down' (all contain the /ɔ/), are more likely to be mixed up with each other. Apart from using the D-MOSFET retroreflector, we also use the E-MOSFET retroreflector to intercept the audio signal. Then, we train and test the command recognition model using audio samples from the E-MOSFET, whose average accuracy is only around 65%. By contrast, the average recognition accuracy using D-MOSFET is 95% for the 20 speech commands, indicating that the reconstructed commands by RF-PARROT^2 can be accurately differentiated from each other.

6.5 Impact of Practical Factors

6.5.1 *Impact of the distance*

This experiment aims to examine the impact of the through-wall distance between the target wire and RF transceivers on eavesdropping performance. We set different distances from 0.5 m to 2.0 m. The SNR and PSNR of the eavesdropped audio for different distances are shown in Fig. 20. Generally, both SNR and PSNR decline as the distance increases, indicating greater signal attenuation in the air. Specifically, when the distance increases from 0.5 m to 1 m, the SNR drops from 7.19 to 1.76; the PSNR declines from 20.14 to 18.38, due to greater signal attenuation in the air. Despite the observed decline in SNR and PSNR values, they are still maintained at a relatively high level that can be perceived. Beyond 1.5 meters, the eavesdropped signal gradually becomes indistinguishable. Note that the default TX and RX antennas' gains are set at only 10 dB here, in Section 6.5.3, we witness the increase of SNR and PSNR values with higher antenna gains. Thus, by lifting the antenna gain, we can enhance the eavesdropping quality at longer distances.

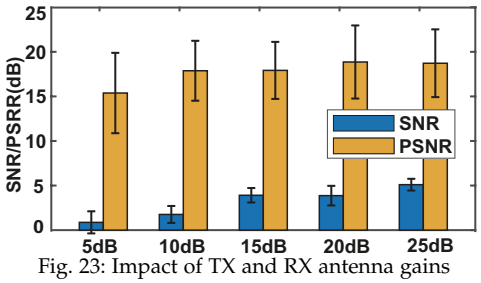


Fig. 23: Impact of TX and RX antenna gains

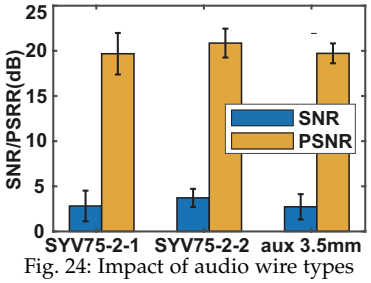


Fig. 24: Impact of audio wire types

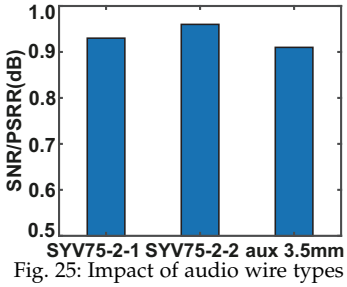


Fig. 25: Impact of audio wire types

6.5.2 Impact of the audio volume

In this experiment, we investigate the effect of audio volume of the smartphone on the eavesdropping performance. Various levels of audio volume ranging from 20% to 100% were set for evaluation. The SNR and PSNR values of the eavesdropped audio signal at different volume levels are shown in Fig. 21. As the audio volume increases, both SNR and PSNR values exhibit a steady growth because a higher volume can resist more of the RF signal attenuation effect and background noises. Although there is a slight degradation in the eavesdropping quality at lower volume levels, the SNR and PSNR still stand out from the underlying noises when the volume is not less than 60%. For volume between 0-40%, the SNR and PSNR will undergo a slow drop until the audio signal is totally none at 0%.

6.5.3 Impact of the antenna gain

In this experiment, we investigate the effect of the antenna gain on the eavesdropped audio. In GNURadio, we configure various antenna gains for both the TX and RX antennas, ranging from 5 dB to 25 dB. As depicted in Fig. 23, the SNR and PSNR values generally increase with the gain increasing. This trend occurs because higher TX gain results in stronger currents on the GND line, which in turn induces stronger reflected RF waves.

6.5.4 Impact of the audio wire type

This experiment explores the impact of different types of audio wires on eavesdropping performance. We choose three types of audio wires: SYV 75 wire series (SYV 75-2-1 and SYV 75-2-2) and 3.5 mm aux audio wire. The SYV 75 wire series is commonly utilized for analog signal transmission. Different models of the SYV 75 wire series may vary in terms of diameter, weight, attenuation coefficient, and other properties. The SNR and PSNR of the eavesdropped audio signal using different wires are presented in Fig. 24, which remain at similar and high levels, i.e., around 4 dB in SNR and 20 dB in PSNR. This shows the effectiveness of D-MOSFET in recovering audio signals transmitted through different types of audio wires. We then use eavesdropped audio from different wires to recognize speech commands. Remarkably, the recognition accuracy for all wire types exceeds 90%, providing further evidence of the effectiveness of RF-PARROT^2 across audio wires.

6.5.5 Impact of wire shapes

In this experiment, we investigate the impact of different shapes of audio wires on the eavesdropping audio. We choose five typical shapes of audio wires in practical scenarios, as shown in Table 5. Shape 1 is the straight audio wire, while the shape 2-3 are bent circles with different number of winding turns and the shape 4-5 are twisted helical coils with different lengths. The corresponding SNR and PSNR of each shape is depicted in Fig. 26. The highest SNR and PSNR belong to Shape 4, exceeding that of Shape 1 with 4.72 dB and 2.37 dB, respectively. Notably, we find the performance of helical coil shapes (Shape 4 - 5) performs better than circular shapes (Shape 2 - 3), and the straight wire behaves the worst among all the five shapes. This indicates RF-PARROT^2 likely perform better with meandered wire shapes (e.g., circular and helical coil shape) compared with the straight wire, because the meandered dipole antenna can improve the its radiation resistance, in turn enhancing the radiation efficiency [47].

TABLE 5: Description of Shape 1 - 5 in Fig. 26

	Description
Shape 1 (default)	A straight and one-meter long cable.
Shape 2	Shape 1 bent into a circular shape.
Shape 3	Shape 1 bent into a circular shape and wound twice.
Shape 4	Shape 1 bent into a helical coil shape, with the total length being 0.75 m.
Shape 5	Shape 1 bent into a helical coil shape, with the total length being 0.5 m.

TABLE 6: Description of settings in Fig. 29 and Fig. 30

	Description
Setting 1	No extra objects except for eavesdropping devices.
Setting 2	Existence of working WiFi router.
Setting 3	Existence of clutter containing clothes, plastic corner reflector, water bottle, wound cable coil, etc.
Setting 4	Add metal objects (<30cm each side) to Setting 3.
Setting 5	Existence of metal sheet (70×70×1cm in size)

6.5.6 Impact of different users

In this experiment, we investigate the impact of different users on the system performance. We first eavesdrop on the SCD dataset, which is treated as the train set in model training. Then, we test our model on another dataset FSDD to see if the model trained on the specific group of people also works effectively on another group. The speakers in FSDD dataset includes George, Jackson, Lucas, Nicolas, Theo, and Yweweler. Over the six speakers, we mainly study (i) the overall user generality performance, and (ii) whether the model achieves relatively balanced generalization over new speakers. We plot the results in Fig. 27. Firstly, the overall performance indicates that nearly all reconstructed speech is highly intelligible, with average MCD values from the FSDD dataset being below 8. The highest average MCD value is for 'Lucas', which is still under 5. Secondly, our model demonstrates relatively balanced reconstruction dealing with new speakers. This is evidenced by the MCD values of the six speakers, which are consistently around 4, indicating robust generalization across various new users.

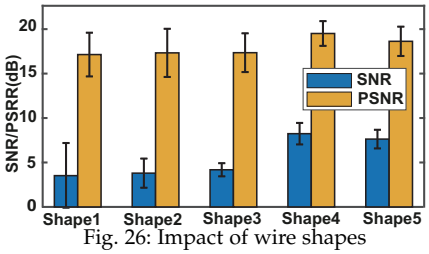


Fig. 26: Impact of wire shapes

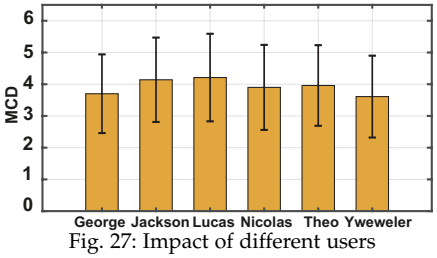


Fig. 27: Impact of different users

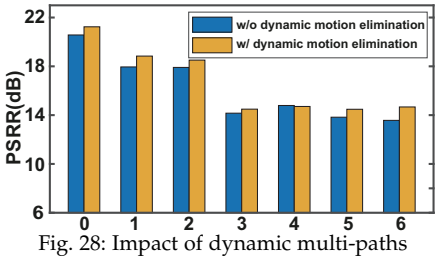


Fig. 28: Impact of dynamic multi-paths

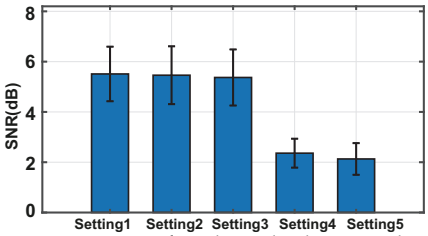


Fig. 29: Impact of multi-path objects and RF device

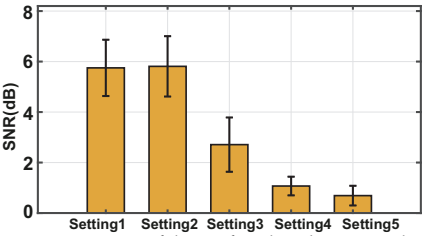


Fig. 30: Impact of line-of-sight objects and RF device

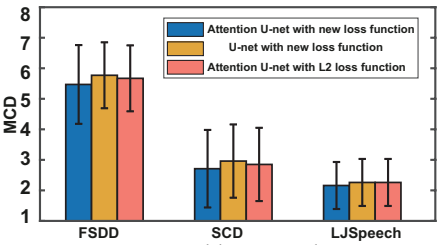


Fig. 31: Ablation study

6.5.7 *Impact of dynamic multi-paths*

In this experiment, we intentionally induce dynamic multi-paths by asking 1 - 6 participants to move around the TX and RX antenna at a normal walking speed (<1.5m/s). The distance between antennas and participants are about 3m or even less, though still maintaining no line-of-sight block. The results are shown in Fig. 28. We observe the decrease of PSNR values of eavesdropped audio when more participants walk around our device. We believe our system maintains high resistance to dynamic multi-paths with 3 participants or less. With more participants, the PSNR values converge to 15 dB. Though, the eavesdropped audio still can be detected and perceived. Additionally, we study the effect of dynamic motion elimination by conducting an ablation study, as shown in Fig. 28. The PSNR values increase by an average of 0.59 dB. Dynamic motion elimination increases the overall performance in face of dynamic multi-paths when pre-processing the RF signal.

6.5.8 *Impact of object multi-path and RF device*

In this experiment, we measure the SNR values of the eavesdropped audio in several settings. The environmental settings are depicted in Table. 6, with all extra objects evenly placed around the eavesdropping devices to create multi-path-rich environments. Notably, WiFi frequency is close to the working frequency of our system, so we select WiFi in the setting 2 to study whether radio-frequency interference influences the eavesdropping performance. The results are depicted in Fig. 29. Compared with Setting 1, The average SNR values of Setting 2 and 3 only decrease by 0.05 and 0.14, indicating that WiFi devices and clutter do not significantly influence the eavesdropping performance. After adding metal objects in the clutter, the SNR value is reduced to 2.36. The SNR value is even lower when a metal sheet is placed around in Setting 5. This implies that our system shows resilience to RF interference and multi-paths like clothes, plastic and water, but can be sensitive to metal objects. This can inspire a potential countermeasure towards our attack via electromagnetic shielding. We will discuss it in Section 8.

6.5.9 *Impact of nearby line-of-sight objects and RF device*

The environmental settings are depicted in Table. 6. Different from the settings in Section 6.5.8, all extra objects are placed between the audio wire and the antenna, i.e., the line-of-sight path. We also set a working WiFi router as line-of-sight RF interference source. The results are depicted in Fig. 30. The WiFi interference does not affect eavesdropping, since the average SNR values of Setting 1 and 2 are approximately equal. The reason is twofold. First, the WiFi router's size is too small to create enough line-of-sight obstacle. Second, the working frequency of WiFi has no overlap with RF-PARROT[2]. The SNR value significantly falls to 2.71 in Setting 3, indicating that line-of-sight clutter does influence the eavesdropping performance. The SNR value is even lower when a clutter with metal or a metal sheet is placed around in Setting 4 and 5. Based on this phenomenon, we find our attack method is sensitive to line-of-sight metal objects. Inspired by this, we will discuss a possible countermeasure to this attack in Section 8.

6.6 Ablation Study

In this experiment, we conduct an ablation study to figure out whether our proposed methods in model design improve the eavesdropping performance as expected. We mainly apply two methods in model design, including the introduction of the attention mechanism into U-net (i.e., the attention U-net) and the new loss function. To study whether the two methods work effectively, we intentionally remove one and compare the performance with the original model. We calculate the MCD values for the three datasets (FSDD, SCD and LJSpeech) in Fig. 31. Please note the lower MCD value indicates better audio reconstruction. As we can see from Fig. 31, compared with U-net network without attention, the introduction of attention mechanism help to decrease the MCD values by 0.22. Moreover, the attention U-net achieves lower MCD values when using the new loss function instead of L2 loss. This demonstrates the effectiveness of attention mechanism and the new loss function in model design.

Besides, we also use the dynamic motion elimination to eliminate interference from dynamic motion when doing signal pre-processing. As illustrated in Fig. 28, the PSNR values improve by an average of 0.59 dB using dynamic motion elimination. Thus, Eliminating dynamic motion enhances overall performance when pre-processing the RF signal, particularly in the presence of dynamic multi-paths. This

TABLE 7: The eavesdropping approaches in literature and the comparison with our RF-PARROT²

	Solution	Capability		
		wire -transmitted	audio eavesdropping	thru-wall (opaque)
Vibration -based	Accelerometer [48]–[50]	✗	✔	✗
	Laser [7]	✗	✔	✗
	RFID [6]	✗	✔	✔
	WiFi [5]	✗	✔	✔
	mmWave [2], [3]	✗	✔	✔
	video [10]	✗	✔	✗
Passive EM	MagEar [13]	✔	✔	✗
	VoiceListener [52]	✔	✔	N/A
Active EM	'The Thing' [53]	✗	✔	✔
	keyboard [16]	✔	✗	N/A
	Suplyspawn [54]	✔	✗	N/A
	Ragemaster [54]	✔	✗	N/A
	RF-PARROT²	✔	✔	✔

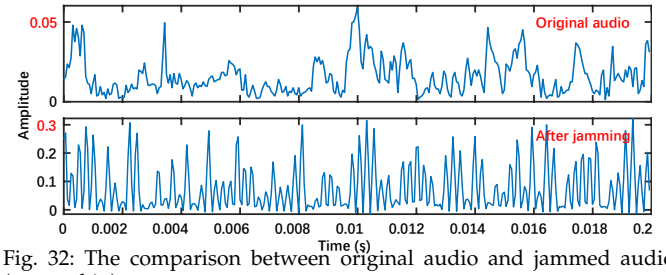


Fig. 32: The comparison between original audio and jammed audio (zoomed in)

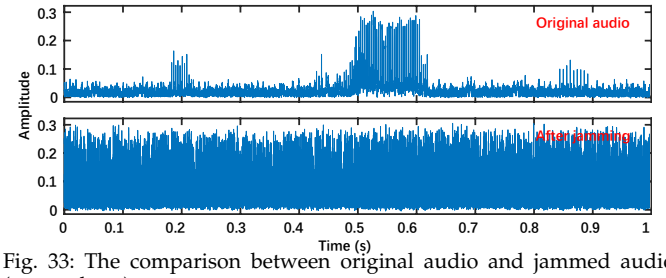


Fig. 33: The comparison between original audio and jammed audio (zoomed out)

demonstrates the effectiveness of dynamic motion elimination in signal pre-processing.

7 RELATED WORKS

In this section, we categorize related research on audio eavesdropping. Table 7 provides a summary of these studies and compares them with RF-PARROT².

7.1 Audio Eavesdropping via Mechanical Vibration

Acoustic signals are mechanical waves that induce continuous vibration in surrounding elastic objects. Researchers harness such phenomenon to investigate various vibration sensors, e.g., accelerometer [48]–[50], laser [7], RFID [6], WiFi [5], mmWave [2]–[4] and videos [10], for audio eavesdropping. AccEar [49] leverages the smartphone's built-in accelerometer, which is sensitive to mechanical waves generated by the loudspeaker, to reconstruct the audio being played. Tag-Bug [6] utilizes RFID tags placed on objects surrounding the loudspeaker and collects the backscattered RFID signal for eavesdropping. mmMIC [51] achieves speech recognition directly from the mouse and throat-reflected mmWave signal. mmEve [4] successfully recovers the audio signal at smartphone's earpiece mode with a COTS mmWave sensor. However, these vibration-based eavesdropping methods require the audio to be played through the loudspeaker in order to induce vibrations in nearby objects. By contrast, our RF-PARROT² overcomes this constraint via a new design of retroreflector. Upon embedding of the retroreflector into the targeted device, the attacker gains the capability to directly eavesdrop on the audio as it is being transmitted through the wire.

7.2 Audio Eavesdropping via EM Side-channel Attack

Passive EM. Since 1980s, researchers has been passively collecting EM side-channel leakage and eavesdropping on various IoT devices [**?**], [35], e.g., televisions [55], CRT monitors [56], LCD screens [57], keyboards [58], etc. Only a few of these passive methods focus on audio eavesdropping. To eavesdrop on audio signals, some researchers find speakers may emit the unintended magnetic field when playing audio, so attackers may infer the original audio from the magnetic side channel. MagEar [13] designs a coil to capture the magnetic field variation from the earpiece and reconstruct the audio. VoiceListener [52] employs the magnetometer as the eavesdropper and formulates a training-free mechanism to effectively reconstruct the audio from low-resolution sensors. Despite their effectiveness in restoring sound from devices with low volume, such magnetism-based techniques encounter a limitation in distance, typically within 50 cm. Another limitation is that the working principles of magnetic audio eavesdropping heavily rely on the magnetic leakages from specific components (e.g., coils) on speakers. This prevents the magnetic audio eavesdropping from generalizing to other acoustic systems. In contrast, RF-PARROT² is capable of achieving audio eavesdropping through walls at greater distances. And RF-PARROT² targets almost all the wires for audio transmission, not limited to specific acoustic devices.

Active EM and RFRA. Different from passively analyzing the RF leakages, active EM side-channel attack actively transmits RF signals to induce errors or information leakages. RFRA is one of the most important techniques among those active approaches. It pre-installs a retroreflector within the device, intentionally transmits RF signal to trigger the RF reflection, and analyzes the reflected RF signal. It was first revealed in the NSA Advanced Network Technology catalog (NSA ANT catalog) [54] in 2013. However, RFRA actually has historical roots dating back to 'The Thing' [53], a listening device covertly installed in a great seal as a gift to an ambassador in the 1940s. This listening device would reflect RF signal when illuminated at a particular radio frequency, and the reflected audio is amplitude-modulated by the talking speech of the ambassador. Its working principle is quite similar to that of NSA's RFRA and our RF-PARROT².

Till now, RFRA has been designed to conduct eavesdropping on the wire-transmitted signal, but only for digital signals, like keyboards [16]–[18] and VGA cables [54]. For instance, researchers have demonstrated that keystrokes sent over a USB cable can be reconstructed with less than a 5% error rate using digital RFRA techniques at distances of up to 10 meters [16]. However, these endeavors have largely disregarded the prevalence of analog audio signals in practi-

cal scenarios. Therefore, we enhance previous retroreflector designs in RF-PARROT^2 employing the D-MOSFET and advanced audio reconstruction methods.

8 COUNTERMEASURES

We recommend the following countermeasures to defend against potential eavesdropping via RF-PARROT^2 .

Electromagnetic shielding. Since RF-PARROT^2 requires transmitting a single-tone RF signal to activate the retroreflector, a possible countermeasure is electromagnetic shielding, which requires shielding the victim device with metal, faraday cage, and specialized metal paints. This prevents the victim device from emitting unintentional EM leakages, as well as blocking the intentional RF illumination from attackers. To validate whether electromagnetic shielding can protect the audio wire from being eavesdropped, we test the SNR value with and without a metal sheet board (70*70*1cm) on the line-of-sight path between the antennas and the audio wire. The results show the SNR value decreases from 5.75 to 0.69 after adding a shielding board, and the SNR could further decrease with a more comprehensive EM shielding coverage. This provides evidence of the anti-eavesdropping capability of electromagnetic shielding, while electromagnetic shielding will also affect normal communication, e.g., WiFi and Bluetooth.

Interference. We can generate irregular RF jamming signals on the frequency band of the attack signal to disturb the received RF signal on the attacker's side. A potential method is to generate a square wave sequence with random amplitude, and mount it on the frequency band of the attack signal, i.e., 2.25 GHz for RF-PARROT^2. To validate the feasibility of interference, we intentionally use an additional SDR USRP X310 to generate 1 KHz square wave with random amplitude, ranging from 0 to 1.0. The carrier wave is 2.25 GHz, aligned with the working frequency of RF-PARROT^2. As shown in Fig. 32 and 33, the received signal is overwhelmed by interference. Apart from this method, there are also two potential methods. The first is to conduct channel randomization [59], which can be effective against RF eavesdroppers. Another one is to jam unauthorized wireless signals via intelligent reflecting surface (IRS) [60].

Detection. If the wire has been embedded with a D-MOSFET retroreflector, detecting it becomes challenging as the retroreflector does not affect the normal audio transmission. However, the resistance between the D-MOSFET's source and drain is not consistently 0 Ω, occasionally leading to common-mode noise during long-range audio transmission, as the GND of the signal line may not perfectly balance with that of connected devices [61]. So the common-mode noise in the signal line could potentially serve as an indicator of being attacked. However, it may fail since the common-mode noise is quite rare when audio signal is transmitted with short wires, like earphones wires.

9 CONCLUSION

This paper proposes RF-PARROT^2 as the first wired analog audio eavesdropping attack. It presents a novel approach that involves modifying an earphone wire by embedding a carefully designed retroreflector. With this setup, RF-PARROT^2 is capable of remotely eavesdropping on audio signals transmitted through the wire. To reconstruct the intercepted audio, RF-PARROT^2 utilizes an encoder-decoder neural network with convolutional layers. This approach achieves impressive results, including low MCD in long-context audio reconstruction and high accuracy of 95% in identifying speech commands. We believe our work will raise awareness regarding the potential safety risks associated with earphone systems.

REFERENCES

[1] A. Xu, Y. Jiang, Y. Cao, G. Zhang, X. Ji, and W. Xu, "Addp: Anomaly detection for dtu based on power consumption side-channel," in *2019 IEEE 3rd Conference on Energy Internet and Energy System Integration*. IEEE, 2019, pp. 2659–2663.

[2] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu, and K. Ren, "mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect," in *IEEE International Conference on Computer Communications*. IEEE, 2022, pp. 820–829.

[3] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *IEEE International Conference on Computer Communications*. IEEE, 2022, pp. 11–20.

[4] 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.

[5] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, 2014, pp. 593–604.

[6] C. Wang, L. Xie, Y. Lin, W. Wang, Y. Chen, Y. Bu, K. Zhang, and S. Lu, "Thru-the-wall eavesdropping on loudspeakers via rfid by capturing sub-mm level vibration," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–25, 2021.

[7] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: eavesdropping via lidar sensors," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 354–367.

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

[9] "Argo-a security," https://argoasecurity.com/index.php?route=product/product&product_id=263, 2020.

[10] 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, jul 2014. [Online]. Available: https://doi.org/10.1145/2601097.2601119

[11] J. Choi, H.-Y. Yang, and D.-H. Cho, "Tempest comeback: A realistic audio eavesdropping threat on mixed-signal socs," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1085–1101.

[12] M. Schulz, P. Klapper, M. Hollick, E. Tews, and S. Katzenbeisser, "Trust the wire, they always told me! on practical non-destructive wire-tap attacks against ethernet," in *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, ser. WiSec '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 43–48. [Online]. Available: https://doi.org/10.1145/2939918.2940650

[13] Q. Liao, Y. Huang, Y. Huang, Y. Zhong, H. Jin, and K. Wu, "Magear: eavesdropping via audio recovery using magnetic side channel." in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2022, pp. 371–383.

[14] Z. Han, J. Ma, C. Xu, and G. Zhang, "Ultrajam: Ultrasonic adaptive jammer based on nonlinearity effect of microphone circuits," *High-Confidence Computing*, p. 100129, 2023.

[15] "Gbppr vision 26: Overview of the nsa's tawdryyard radar retroreflector," https://www.youtube.com/watch?v=KDQxDxiflyo.

[16] S. Wakabayashi, S. Maruyama, T. Mori, S. Goto, M. Kinugawa, Y.-i. Hayashi, and M. Smith, "A feasibility study of radio-frequency retroreflector attack," in *12th USENIX Workshop on Offensive Technologies*, 2018.

[17] S. Wakabayashi, "Investigation of radio frequency retroreflector attacks," Ph.D. dissertation, Waseda University, 2019.

[18] M. Ossmann, "The nsa playset: Rf retroreflectors," *DEF CON*, vol. 22, no. 8, 2014.

[19] C.-C. Chang and S.-I. Liu, "Pseudo-exponential function for mosfets in saturation," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 11, pp. 1318–1321, 2000.

[20] A. Sattar, "Depletion-mode power mosfets and applications," *IXYS Corporation (10 pages)*, 2014.

[21] "Total harmonic distortion," https://en.wikipedia.org/wiki/Total_harmonic_distortion.

[22] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang, and W. Xu, "The feasibility of injecting inaudible voice commands to voice assistants," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, pp. 1108–1124, 2019.

[23] "Common coaxial cable sizes," http://www.snakebitdrill.com/commoncoaxialcablesizes.aspx.

[24] K. S. Rao, P. V. Nikitin, and S. F. Lam, "Antenna design for uhf rfid tags: A review and a practical application," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 12, pp. 3870–3876, 2005.

[25] K.-C. Kim, S.-M. Kim, J.-Y. Kwon, T.-W. Kang, and J.-H. Kim, "The design of calculable standard dipole antennas in the frequency range of 1˜ 3 ghz," *Journal of the Korean Institute of Electromagnetic and Science*, vol. 12, no. 1, pp. 63–69, 2012.

[26] "Voice frequency," https://en.wikipedia.org/wiki/Voice-frequency.

[27] C. Cai, R. Zheng, and J. Luo, "Ubiquitous acoustic sensing on commodity iot devices: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 432–454, 2022.

[28] Z. Qin, X. Zhang, and S. Li, "A robust adversarial attack against speech recognition with uap," *High-Confidence Computing*, vol. 3, no. 1, p. 100098, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667295222000502

[29] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[31] J. Chen, M. Ma, R. Zheng, and L. Huang, "Specrec: An alternative solution for improving end-to-end speech-to-text translation via spectrogram reconstruction." in *International Speech Communication Association (INTERSPEECH)*, 2021, pp. 2232–2236.

[32] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, "Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 13 198–13 206.

[33] 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 *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 14–26. [Online]. Available: https://doi.org/10.1145/3307334.3326073

[34] R. Zhao, J. Yu, T. Li, H. Zhao, and E. C. Ngai, "Radio2speech: High quality speech recovery from radio frequency signals," *arXiv preprint arXiv:2206.11066*, 2022.

[35] L. P. Rachakonda, M. Siddula, and V. Sathya, "A comprehensive study on iot privacy and security challenges with focus on spectrum sharing in next-generation networks (5g/6g/beyond)," *High-Confidence Computing*, p. 100220, 2024.

[36] X. Feng, J. Han, R. Zhang, S. Xu, and H. Xia, "Security defense strategy algorithm for internet of things based on deep reinforcement learning," *High-Confidence Computing*, vol. 4, no. 1, p. 100167, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266729522300065X

[37] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[38] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[39] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[40] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[42] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Speaker gender recognition based on deep neural networks and resnet50," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–13, 2022.

[43] L. Le, A. N. M. Kabir, C. Ji, S. Basodi, and Y. Pan, "Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images," in *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*. IEEE, 2019, pp. 106–110.

[44] "Speech commands dataset," https://dagshub.com/kingabzpro/Speech-Commands-Dataset/src/master.

[45] "Free spoken digit dataset (fsdd)," https://github.com/Jakobovski/free-spoken-digit-dataset.

[46] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[47] O. O. Olaode, W. D. Palmer, and W. T. Joines, "Effects of meandering on dipole antenna resonant frequency," *IEEE Antennas and Wireless Propagation Letters*, vol. 11, pp. 122–125, 2012.

[48] 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 *Network and Distributed System Security Symposium*, 2020, pp. 1–18.

[49] 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, 2022, pp. 1757–1773.

[50] 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 *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, pp. 288–299.

[51] L. Fan, L. Xie, X. Lu, Y. Li, C. Wang, and S. Lu, "mmmic: Multimodal speech recognition based on mmwave radar," in *IEEE International Conference on Computer Communications*, 2023.

[52] L. Wang, M. Chen, L. Lu, Z. Ba, F. Lin, and K. Ren, "Voicelistener: A training-free and universal eavesdropping attack on built-in speakers of mobile devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–22, 2023.

[53] "The thing," https://cryptomuseum.com/covert/bugs/thing/index.htm#ref/.

[54] "Nsa ant catalog," https://www.eff.org/files/2014/01/06/20131230-appelbaum-nsa-ant-catalog.pdf.

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

[56] M. G. Kuhn, "Optical time-domain eavesdropping risks of crt displays," in *Proceedings 2002 IEEE Symposium on Security and Privacy*. IEEE, 2002, pp. 3–18.

[57] ——, "Compromising emanations: eavesdropping risks of computer displays," University of Cambridge, Computer Laboratory, Tech. Rep., 2003.

[58] Y.-L. Du, Y.-H. Lu, and J.-L. Zhang, "Novel method to detect and recover the keystrokes of ps/2 keyboard," *Progress In Electromagnetics Research C*, vol. 41, pp. 151–161, 2013.

[59] H. Hassanieh, J. Wang, D. Katabi, and T. Kohno, "Securing rfids by randomizing the modulation and channel," in *12th USENIX Symposium on Networked Systems Design and Implementation*, 2015, pp. 235–249.

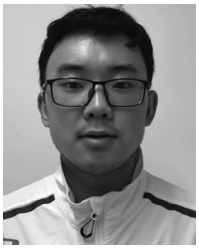
[60] P. Staat, S. Mulzer, S. Roth, V. Moonsamy, M. Heinrichs, R. Kronberger, A. Sezgin, and C. Paar, "Irshield: A countermeasure against adversarial physical-layer wireless sensing," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1705–1721.

[61] "Common-mode signal," https://https://en.wikipedia.org/wiki/Wikipedia:Citing_Wikipedia.

Genglin Wang received the B.E. degree from the School of Information Science and Engineering at Shandong University, Shandong, China. He is currently working toward the PhD degree at the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China. His current research interests include mobile computing and wireless sensing.

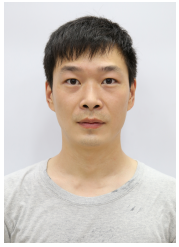
Zheng Shi is going to receive the B.E. degree in communication engineering from Shandong University, Shandong, China. His current research interests include wireless sensing, especially intelligent reflecting surface.

Yanni Yang received her Ph.D. degree in computer science from The Hong Kong Polytechnic University in 2021. Before that, she received the B.E. degree and M.Sc. degree from the Ocean University of China in Qingdao, in 2014 and 2017, respectively. She is currently an assistant professor in the School of Computing Science and Technology at Shandong University. She visited the Media Lab at MIT in 2019 as a visiting student. Her research interests include wireless human sensing, pervasive and mobile computing, and Internet of Things. She has published over 20 papers in top academic conferences and journals.



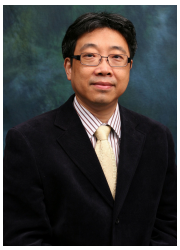
Zhenlin An received the B.E. degree from the School of Information Communicating Engineering at Dalian University of Technology, China, in 2017, and the Ph.D. degree from the Department of Computing at The Hong Kong Polytechnic University, in 2021. He is currently a postdoctoral research associate at Princeton University. Prior to that, he was a postdoc fellow at The Hong Kong Polytechnic University from 2021 to 2023. His research interests include wireless and backscatter communication, wireless security, indoor localization, and mobile computing.

Guoming Zhang is an assistant professor at the School of Computer Science and Technology of Shandong University. He received his M.S. degree in Mechanical and Electronic Engineering from Beijing Institute of Technology in 2016, and his Ph.D. degree in Electrical Engineering from Zhejiang University in 2021. His research interests are in the areas of IoT security, and AI security, especially for the security of speech recognition systems. He won the best paper awards of ACM CCS 2017, Qshine 2019, and WASA 2022.



Pengfei Hu is a professor in the School of Computer Science and Technology at Shandong University. He received the Ph.D. degree in Computer Science from UC Davis. His research interests include cyber security, data privacy, and mobile computing. He has published more than 40 papers in premier conferences and journals on these topics. He served as TPC for numerous prestigious conferences, and associate editors for IEEE TWC and IEEE IoTJ. He is the recipient of 2022 ACM SIGBED China Rising Star Award.

Xiuzhen Cheng received her MS and Ph.D. degrees in computer science from the University of Minnesota – Twin Cities, in 2000 and 2002, respectively. She was a faculty member at the Department of Computer Science, The George Washington University, from 2002-2020. Currently, she is a professor of computer science at Shandong University, Qingdao, China. Her research focuses on blockchain computing, IoT Security, and privacy-aware computing. She is a Fellow of IEEE.



Jiannong Cao received his M.Sc. and Ph.D. degrees in computer science from Washington State University. He is currently a Chair Professor with the Department of Computing at The Hong Kong Polytechnic University (PolyU). He is also the dean of Graduate School, the director of Research Institute of Artificial Intelligent of Things, and the director of the Internet and Mobile Computing Lab at PolyU. His research interests include distributed systems and blockchain, wireless sensing and networking, big data and machine learning, and mobile cloud and edge computing. He has co-authored 5 books, co-edited 9 books, and published over 500 papers in journals and conference proceedings. He is a member of Academia Europaea, a fellow of IEEE, and an ACM distinguished member.