RF-Parrot: Wireless Eavesdropping on Wired Audio

Yanni Yang*, Genglin Wang†, Zhenlin An‡, Guoming Zhang*, Xiuzhen Cheng*, Pengfei Hu*
*School of Computer Science and Technology, Shandong University, China.
†School of Information Science and Engineering, Shandong University, China.
‡Department of Computer Science, Princeton University, USA.

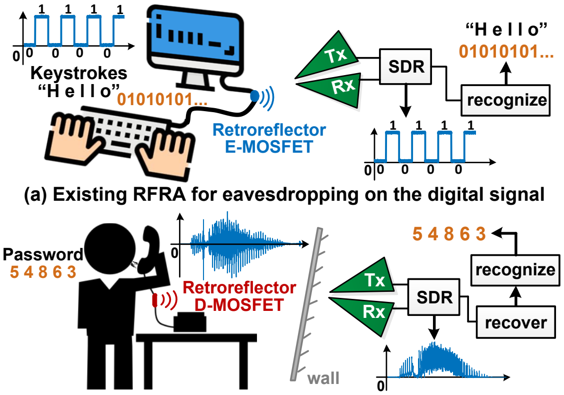
***Abstract*—Recent works demonstrated that we can eavesdrop on audio by using radio frequency signals or videos to capture the physical surface vibrations of surrounding objects. They fall short when it comes to intercepting internally transmitted audio through wires. In this work, we first address this gap by proposing a new eavesdropping system, RF-Parrot, that can wirelessly capture the audio signal transmitted in earphone wires. Our system involves embedding a tiny field-effect transistor in the wire to create a battery-free retroreflector, with its reflective efficiency tied to the audio signal's amplitude. To capture full details of the analog audio signals, we engineered a novel retroreflector using a depletion-mode MOSFET, which can be activated by any voltage of the audio signals, ensuring no information loss. We also developed a theoretical model to demystify the nonlinear transmission of the retroreflector, identifying it as a convolution operation on the audio spectrum. Subsequently, we have designed a novel convolutional neural network-based model to accurately reconstruct the original audio. Our extensive experimental results demonstrate that the reconstructed audio bears a strong resemblance to the original audio, achieving an impressive 95% accuracy in speech command recognition.**

***Index Terms*—Audio eavesdropping, active retroreflector attack, RF side-channel attack.**

I. Introduction

Audio eavesdropping has long been a subject of interest in security and privacy research. Recent advancements have demonstrated the interception of audio signals wirelessly by capturing the physical surface vibrations of surrounding objects using RF signals [1]–[6], laser [7], and videos [8]. However, these methods have a significant limitation: they are incapable of intercepting audio signals transmitted internally through wires, such as those in earphones and telephone lines. This gap in capability presents considerable potential security issues, as wired audio transmissions are prevalent in numerous practical scenarios, from personal earphone use to professional audio systems. Moreover, most audio signals in earphone wires are not encrypted, and eavesdropping can result in more serious privacy breaches.

A naive solution to eavesdrop on the earphone is to perform classic electromagnetic side-channel attacks [9]–[11], which passively measure the electromagnetic emission from the target devices. Then, the attacker can reconstruct the original signal by analyzing the measured radio wave. However, the EM emission from the audio wire or speaker coils is too weak to be detected remotely. The advanced earphone eavesdropping system can only work within 50 cm [10]. Recently, there have been some studies [12]–[14] on active radio-frequency retroreflector attacks (RFRA) to eavesdrop on wired digital signals remotely. RFRA involves a retroreflector secretly embedded inside the target device, which reflects the incoming RF signal emitted by the attacker back to an RF signal receiver for interception [13]. The reflective efficiency of the retroreflector varies as the digital signal switches between high and low voltages. For example, researchers [15] demonstrated that we could implant a tiny metal-oxide-semiconductor field-effect transistor (MOSFET) into the keyboard wire and employ the commodity software-defined radio (SDR) to remotely eavesdrop on the kicked keystroke via the backscattered RF signals from MOSFET, as depicted in Fig. 1(a). However, existing RFRA systems can only eavesdrop on digital signals but can not deal with analog signals (i.e., audio signals).



(a) Existing RFRA for eavesdropping on the digital signal

(b) RF-Parrot used for eavesdropping on wired analog audio

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

In this work, we ask - *Can we eavesdrop on the wired analog audio signal wirelessly?* Particularly, as shown in Fig. 1(b), we aim to perform RFRA eavesdropping on analog audio signals at a long range and through the wall via only embedding a tiny transistor (i.e., diameter <3 mm) inside the earphone wire. Such a secret bug overwhelms microphone-based and coil-based tapping devices with a covert design and longer eavesdropping distance. Yet, it is a daunting task to answer the above question. First, recovering every detail of the analog signal, meanwhile bothered by noises, is much more difficult than simply retrieving either '0' or '1' in digital signals. Second, unlike the digital signal, almost half of the analog audio signal is negative, containing essential information. Regarding the above problems, we propose the first-of-its-kind analog RFRA system, named RF-Parrot, by re-designing the retroreflector to fit the analog requirements in audio eavesdropping. We carefully select the RF signal frequency for through-wall RFRA audio eavesdropping

*Pengfei Hu is the corresponding author.

and develop advanced signal processing algorithms for high-quality audio recovery and speech command recognition.

As a pioneer attempt, RF-PARROT addresses the fundamental challenge of RFRA on the analog audio signal by utilizing the depletion-mode MOSFET (i.e., D-MOSFET) as the retroreflector. Previous RFRA works all employ the enhancement-mode MOSFET (E-MOSFET), which requires a positive voltage on the gate for RF signal radiation [12], [13], [15]; thus, they fail to capture the negative part in audio signals, leading to much information loss. On the contrary, the D-MOSFET can be turned on by both negatively and positively charged gate, offering the potential for audio eavesdropping. On top of such a principle, we fabricate a tiny retroreflector using the D-MOSFET as a covert listening device that can be hidden in the audio wire. The retroreflector functions as an analog modulator over the remotely transmitted RF signal. As such, the audio signal's voltage change directly affects the strength of radiated RF wave back to the signal receiver; thus, the audio information can be revealed from the received RF signal amplitude. Our analytical and experimental results show that the D-MOSFET-based retroreflector is more effective in reserving the key information in analog audio signals than the E-MOSFET. Attackers can secretly substitute the audio wire with the one embedded with our designed tiny bug for eavesdropping without being noticed.

Despite the promising feature of the D-MOSFET, it still leaves detrimental effects on precisely recovering the original audio signal. First, the transfer characteristic curve of D-MOSFETs is nonlinear, resulting in the captured RF signal being a distorted version of the original audio signal. Particularly, by investigating the effect of the D-MOSFET's transfer curve on the audio signal, we find that the nonlinear transfer curve causes unbalanced and flipped RF signal amplitudes between the negative and positive parts of the analog audio signal. Such imbalance and deformation lead to a serious attenuation and distortion of the eavesdropped audio signal, impeding the successful reconstruction of the original audio.

To tackle the nonlinearity issue caused by the D-MOSFET, we mathematically model the effect of the nonlinear transfer curve on the received RF signal amplitude. We find that the nonlinear transformation is essentially a convolution operation on the spectrum of the original audio. Based on this key finding, we intuitively aim to reconstruct the audio by deriving an analytical solution to reverse the convolution operation. However, the above convolution computation highly depends on the audio signal in various forms. Besides, the nonlinearity parameters are also unknown to us. Thus, it is extremely difficult to obtain the desired solution. To this end, we propose to harness the convolutional neural network (CNN), which not only provides the required convolution operation but also can adapt to various audio signals automatically. As such, we devise an encoder-decoder neural network with convolution layers for audio reconstruction. We further introduce the attention mechanism and dedicatedly design a loss function for the network to remove the effect of RF signal noises and achieve better reconstruction performance.

In summary, our work makes the following contributions:

- We propose the first analog RFRA system, RF-Parrot, for eavesdropping on the wire-transmitted analog audio signal remotely via a new design of the retroreflector made from the D-MOSFET.
- We demonstrate that RF-PARROT can intercept analog audio signals at a distance of 1 m through the wall. We also devise a novel method to tackle the audio deformation problem incurred by the nonlinear transfer curve of D-MOSFET.
- We evaluate RF-Parrot using over 65,000 speech commands from thousands of people in various environments. Extensive experiments show we can achieve an average mel-cepstral distortion (MCD) of 6.8[1], signal-to-noise ratio (SNR) of 4.7, and 95% command recognition accuracy.

II. ANALOG AUDIO EAVESDROPPING VIA RFRA: CONCEPT & PRELIMINARIES

This section introduces the attack model, the concept of RFRA, the principle of RFRA for the analog audio signal, feasibility studies, and preliminary results.

A. Attack Model

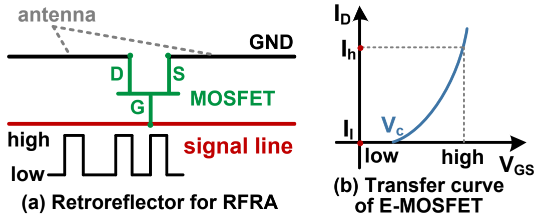
The attacker aims to eavesdrop on the wire-transmitted audio signal in the target device via the backscattered RF signal from the retroreflector embedded in the wire, as shown in Fig. 1(b). We make the following assumptions for RF-Parrot attacks. First, we assume the attacker can secretly replace the victim's audio wire with the one embedded with a tiny retroreflector sealed inside. Since the MOSFET-based retroreflector is commonly within 3 mm and does not actively emit signals, such a replacement would not be noticed by victims. Second, there is acoustic isolation, e.g., soundproof insulation or wall, between the attacker and the victim so that the attack can be conducted out of the victim's sight. Third, the attacker has no prior information about the content of the audio emitted from the victim. Finally, the SDR and antennas used for eavesdropping are commercially affordable and portable for the attacker. The cost to reproduce the audio wire with a retroreflector embedded is also within tens of US dollars.

B. Traditional RFRA on Digital Signal

RFRA conducts attacks by embedding a battery-free retroreflector into the data transmission wire of the target device. Existing works mainly employ the E-MOSFET as the retroreflector for digital RFRA. As shown in Fig. 2(a), the MOSFET gate (G) is connected to the wire's signal line, and the drain (D) and source (S) are tied to the GND line. The high and low levels of the digital signal vary the gate voltage V_{GS}. From the transfer characteristic curve[2] of E-MOSFET in Fig. 2(b), the high $V_{GS} > 0$ incurs a higher current I_h on the drain than that of the low $V_{GS} = 0$, i.e., I_l. When the remote attacker sends the RF signal to the retroreflector, the GND line aside the drain and source works like a dipole antenna which can reflect the RF signal back to the attacker's RF receiver. As

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

[2] The transfer characteristic curve refers to the drain current vs. gate to source voltage curve.



(a) Retroreflector for RFRA (b) Transfer curve of E-MOSFET

Fig. 2. (a) MOSFET as retroreflector and (b) transfer curve of E-MOSFET

such, the received RF signal has a higher amplitude due to I_h, whereas the amplitude decreases to zero for I_l. In sum, the radiated RF wave is AM-modulated by the digital signal transmitted via the wire. By demodulating the received RF signal, we can recover the original digital information from the low- and high-level RF signal amplitude.

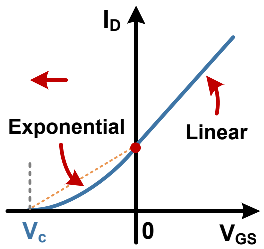
C. A New Design of RFRA on Analog Audio Signal

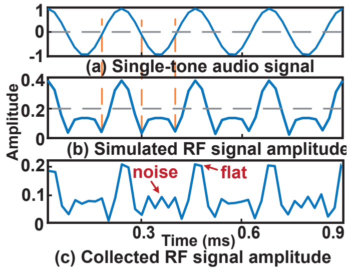
RFRA becomes much more complicated when transmitting analog audio signals. First, different from the digital signal, analog audio contains negative voltages. However, the cutoff voltage V_c of E-MOSFET is positive, resulting in zero amplitude of the received RF signal for the negative part of the audio. As such, the audio signal below the cutoff voltage is completely lost. Second, compared with the binary '1' or '0', the analog audio signal is continuous over time and can take any value within a certain range. It is intractable to precisely reconstruct every detail of the analog audio signal via RFRA in the face of signal noise and attenuation in the air.

The root cause of the information loss using the E-MOSFET stems from the inappropriate transfer curve for eavesdropping on the analog audio signal. To tackle this issue, we devise a new retroreflector using the D-MOSFET whose transfer curve is depicted in Fig. 3. D-MOSFET owns a preferable characteristic for RFRA on the analog audio signal, i.e., the negative cutoff voltage V_c. Then, I_D can also 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 the principle of using D-MOSFET for analog audio eavesdropping, we carefully debunk the transfer curve of the D-MOSFET and model its effect on the backscattered RF signal's amplitude from the audio wire. First, we divide the transfer curve in Fig. 3 into three parts based on the gate voltage V_{GS}: $V_{GS} \geq 0$, $V_c \leq V_{GS} < 0$, and $V_{GS} < V_c$.

- When $V_{GS} \geq 0$, I_D varies with the V_{GS} in an approximately linear way within the audio voltage. Thus, the received RF signal amplitude $a(t)$ proportionally changes with the audio signal voltage $s(t)$, i.e., $a(t) = \alpha \cdot s(t) + \beta$.
- When $V_c \leq V_{GS} < 0$, I_D nonlinearly changes with the V_{GS}. The reflected RF signal amplitude $a(t)$ will experience a nonlinear decay of the original audio signal $s(t)$. The nonlinearity in MOSFET is generally represented by a pseudo-exponential function [16], and the decay magnitude is controlled by the audio signal itself, making the received signal amplitude[3] as $a(t) = -e^{\gamma \cdot s(t)} \cdot s(t)$.





(a) Single-tone audio signal (b) Simulated RF signal amplitude (c) Collected RF signal amplitude

Fig. 3. D-MOSFET transfer curve Fig. 4. Simulated & collected RF signal

- When $V_{GS} < V_c$, there is no current on the drain; thus the reflected RF signal amplitude $a(t) = 0$. Note that the cutoff voltages of common D-MOSFETs are below -1 V, which is sufficient to react to the wire-transmitted audio signal voltage falling within the low voltage range of hundreds of millivolts [17].

Summarizing the above modeling, we can represent the effect of the audio signal $s(t)$ on the received RF signal amplitude $a(t)$ with the following equation.

$$a(t) = \begin{cases} \alpha \cdot s(t) + \beta, & s(t) \geq 0 \\ -e^{\gamma \cdot s(t)} \cdot s(t), & V_c \leq s(t) < 0 \\ 0, & s(t) < V_c \end{cases} \tag{1}$$

The parameters α, β, and γ in Eq. (1) are constant values, which are determined by the D-MOSFET.

To validate the modeling result, we transmit a single-tone analog audio signal via a coaxial wire in which a D-MOSFET retroreflector is embedded inside (see details in Section III-B) and apply Eq. (1) to simulate the received RF signal amplitude[4] and compare it with the real collected signal from an SDR, as depicted in Fig. 4. We have the following observations from Fig. 4: (i) The simulated signal amplitude in Fig. 4(b) and the collected one in Fig. 4(c) exhibit similar patterns, verifying the effectiveness of our modeling result. (ii) The main periodicity of the single-tone audio is maintained; particularly, the positive part of the audio is well preserved. However, due to the nonlinear decay between $[V_c, 0)$, the negative part of the single-tone audio flips over with smaller amplitudes. Besides, the weaker strength also causes the negative part to suffer more from noise fluctuations. (iii) The peak in the simulated amplitude becomes flat in the collected RF signal amplitude. Such a breach is caused by the upper limit of the induced current i_{max} provided by the RF transmitter. Once I_D exceeds i_{max}, the received RF signal amplitude will stay at the maximum. This issue can be fixed by increasing the RF transmitter gain. Another concern is whether the retroreflector affects the normal audio transmission. We employ the total harmonic distortion (THD) to quantify the audio distortion introduced by the retroreflector. The average THD of human speech transmitted in three common audio wires are all below -20dB, meaning less than 1% distortion.

In sum, using the D-MOSFET-based retroreflector can capture the information in both negative and positive parts of the audio signal without disturbing normal audio transmission, which is beneficial for successful eavesdropping. However,

[3] The negative sign '-' is introduced to make $a(t)$ positive as I_D is positive.

[4] In this simulation, α, β, and γ are all set to 1.

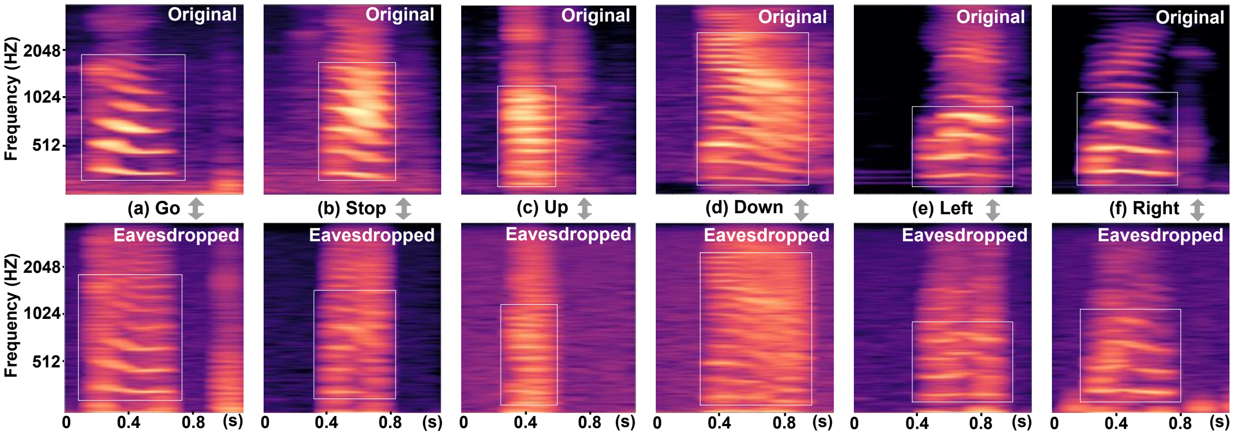


Fig. 5. Mel-spectrogram of original audios and eavesdropped RF signals for different speech commands

TABLE I
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

the nonlinearity in the D-MOSFET transfer curve and unpredictable RF noises result in the detrimental transformation of the original audio. In Section III-E, we propose solutions to address the nonlinearity and noise issues.

D. *Feasibility Study and Analysis*

We conduct preliminary experiments to show the feasibility of using the D-MOSFET for analog audio eavesdropping. First, we use a smartphone to play the English speech audio of 10 digits from 0 to 9 spoken by six persons. The D-MOSFET is assembled on a tiny PCB and embedded in the middle of 1 m-long wire which is connected to the phone audio jack. Then, we emit the RF signal from an SDR and collect the reflected signal using a pair of directional antennas. The received RF signal amplitude is saved as a wave file. Then, we quantitatively evaluate 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 also obtain the MCDs using an E-MOSFET for comparison. Note that a smaller MCD value indicates better audio eavesdropping performance. As given in Table I, the average MCD of D-MOSFET over 0 to 9 digits is around 6-7 less than that of the E-MOSFET, demonstrating that D-MOSFET is more effective in reconstructing the analog audio signal. On the other hand, we still need to tackle the nonlinearity problem to lower the MCD within 8, under which the eavesdropped audio can be recognized by the speech recognition system [18]. We also listen to the eavesdropped wave files: most stressed syllables can be perceived but mixed with background noises; besides, the light plosive sounds are distorted. Thus, the eavesdropped audio needs further enhancement and reconstruction.

Next, we calculate and observe 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. 5. First, we compare the similarity between the original audio’s mel-spectrogram and that of the received RF signal amplitude for each command. The main frequency components in the original audio can be generally preserved in the eavesdropped signal, as shown in the boxed areas of each command’s mel-spectrograms in Fig. 5. Second, we compare the eavesdropped signal’s spectrogram of different commands. As depicted in the second row’s figures of Fig. 5, the six commands illustrate a significant difference among their eavesdropped mel-spectrograms, indicating that the eavesdropped signals of different speech audios are distinguishable.

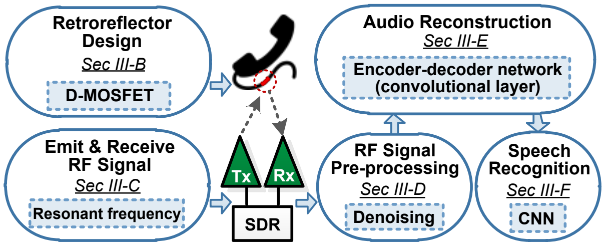


Fig. 6. System overview of RF-Parrot

According to the above preliminary results, the D-MOSFET retroreflector demonstrates great potential in eavesdropping on the wire-transmitted analog audio signal; meanwhile, the eavesdropped signal can be used to recognize different speech commands. However, we also observe that the eavesdropped signal’s mel-spectrograms are blurred and noisy compared with those of the original audio. The key frequency components in the audio become less dominant in strength due to the presence of other disturbing and noisy frequencies. This may arise from the nonlinear effect of the transfer curve, as the negative part of the audio signal is flipped over, which introduces high-frequency components in the mel-spectrogram. We will tackle this problem in Section III-E.

III. RF-PARROT SYSTEM DESIGN

In this section, we introduce the design details of RF-PARROT, including the fabrication of the retroreflector, RF signal setup, signal pro-processing, and audio reconstruction.

A. *Overview of* RF-PARROT

The system overview of RF-PARROT is shown in Fig. 6, which contains five key modules. In the first module, the

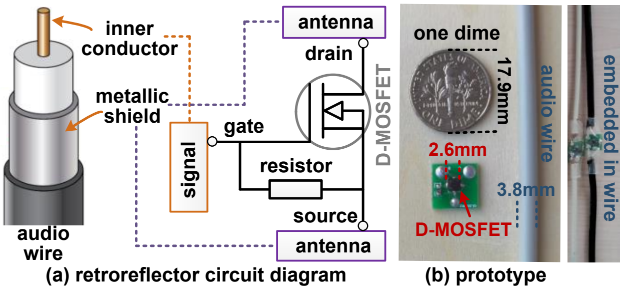


Fig. 7. Retroreflector fabrication: (a) circuit diagram and (b) prototype

D-MOSFET-based retroreflector is fabricated and embedded into the audio wire. Second, in the RF signal emission and receiving module, we carefully select a proper RF signal frequency for effectively receiving the backscattered signal from the audio wire. Third, the signal pre-processing module removes noises and redundancy in the received RF signal and obtains the corresponding mel-spectrogram. Fourth, in the audio reconstruction module, we develop an encoder-decoder network with convolutional layers to tackle the nonlinear effect on the eavesdropped signal's mel-spectrogram which is then converted back to the audio signal. Finally, we input the reconstructed signal's mel-spectrogram into a speech command recognition model to automatically obtain the speech content.

B. Fabrication of D-MOSFET-based Retroreflector

The circuit diagram of the retroreflector is depicted in Fig. 7(a). It consists of two main components, i.e., a N-channel D-MOSFET and a 10KΩ resistor. The resistor is used to protect the MOSFET from being destroyed by the overcurrent. The gate of the D-MOSFET is connected to the inner conductor of the audio wire which carries the analog signal. The drain and source are connected to the metallic shield (i.e., the GND) on each side of the wire after being split, respectively. In our implementation, the D-MOSFET and resistor are assembled on a tiny PCB in Fig. 7(b) for the convenience of welding. Note that the size of the retroreflector can be further reduced by directly embedding the D-MOSFET and resistor into the audio wire. Particularly, the size of D-MOSFET is usually within 3 mm-long, which is smaller than the diameters of common audio wires (3 mm-5 mm) [19]. Thus, the retroreflector can be installed inside the audio wire without being noticed.

C. RF Signal Emission and Receiving

As shown in Fig. 2(a), the audio wire on each side of the retroreflector naturally acts as a dipole antenna. To obtain a stronger RF signal backscattered from the dipole antenna, we carefully tune the RF signal to work on the resonant frequency of the dipole antenna. For the audio wire with a length of L, the dipole antenna would resonate at the odd multiples of half-wavelength for the RF signal [20], [21], i.e., $L = (2n-1)\cdot\lambda/2$, where $n \in \mathbb{N}_{+}$. Then, the candidate resonance frequency f_r can be expressed as follows:

$$f_r = \frac{1}{2}\cdot(2n-1)\cdot c/L,$$

where c is the light speed of 3×10^8 m/s. Common audio wires are around 1 m-long, i.e., $L = 1$ m. Thus, f_r can be any odd multiple of 150 MHz. However, considering that the attacker can be more easily detected by the defense side if emitting the signal at an uncommon frequency, we select the f_r around the 2.4 GHz, which is the most broadly utilized frequency band. As such, the signal emitted by the attacker can be drowned out in the congested 2.4 GHz spectrum, decreasing the attacker's risk of being identified. The nearest resonant frequency to 2.4 GHz is 2.25 GHz = 150 MHz × 15, i.e., f_r = 2.25 GHz. Moreover, the RF signal at this frequency is capable of penetrating through the wall. Thus, we send out a 2.25 GHz continuous wave from the transmitting antenna on the SDR, which is then reflected by the audio wire and received by the receiving antenna with a sampling rate of 200 KHz. Finally, we save the received signal amplitude into a wave file.

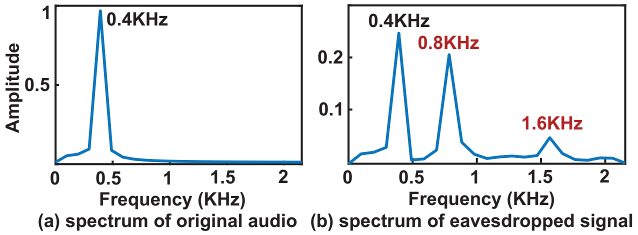


Fig. 8. Spectrum of (a) the original 0.4KHz single-tone audio and (b) the eavesdropped RF signal

D. RF Signal Pre-processing

The received RF signal amplitude undergoes the following pre-processing steps. First, we apply a low-pass filter to eliminate high-frequency RF noise from the signal. The cut-off frequency is set to 8 KHz since the usable voice frequency band in telephony is basically below 4 KHz [22]. According to the Nyquist sampling theorem, a sampling rate of 8 KHz is necessary to adequately represent the voice frequency band. Then, we remove the DC component from the signal to prevent the voice frequencies from being masked by the DC offset. Next, we re-sample the signal at a reduced sampling rate of 8 KHz to minimize the computation and storage demands in subsequent stages. Finally, we divide the signal into fixed-length segments, each containing a single word, and calculate the mel-spectrogram of each segment.

E. Audio Reconstruction

Recall that the nonlinear transfer curve of D-MOSFETs in the negative voltage range incurs the deformation of the original audio signal. Such deformation leads to extra frequency components in the spectrum of the received RF signal. For instance, we compare the spectrum of the original 0.4 KHz single-tone audio signal with that of the eavesdropped RF signal using a D-MOSFET-based retroreflector in Fig. 8. Apart from the original 0.4 KHz frequency component, the spectrum of the eavesdropped RF signal also contains additional 0.8 KHz and 1.6 KHz components, which could result in inaccurate perceiving of the audio content. To solve this problem, we first derive the underlying mathematical mechanism of the nonlinear effect on the eavesdropped signal. Then, we propose an audio reconstruction model to tackle the nonlinearity issue.

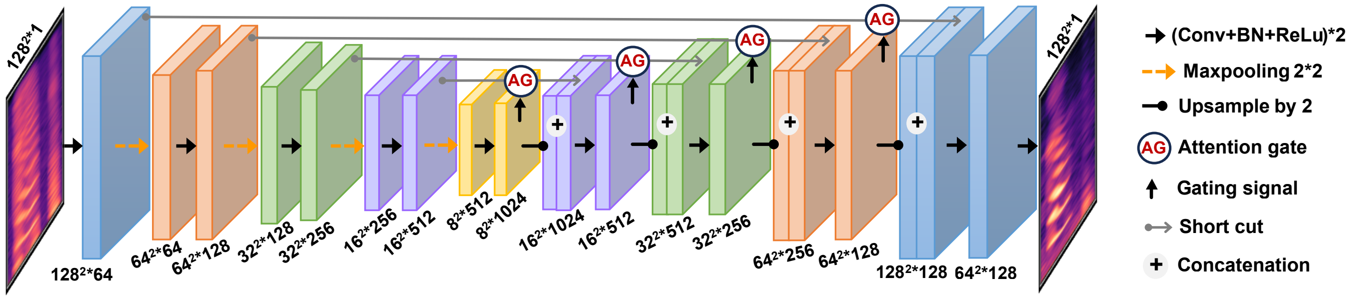


Fig. 9. Architecture of encoder-decoder network with convolutional layers for audio reconstruction

1) Mathematical modeling of the nonlinear effect: We perform fast Fourier transformation (FFT) on the received RF signal amplitude $a(t)$ based on Eq. (1). Suppose the FFT of the original audio signal $s(t)$ is $S(f)$. When $s(t) \geq 0$, we can express the FFT of $a(t) = \alpha \cdot s(t) + \beta$ as $A(f) = \alpha \cdot S(f) + \beta \cdot \delta(f)$ according to the linearity property of FFT; thus the audio fidelity is preserved for the positive part of the audio signal. However, when $V_c \leq s(t) < 0$, the FFT of $a(t) = -e^{\gamma \cdot s(t)} \cdot s(t)$ becomes complicated. To solve this problem, we first decompose $a(t)$ as the product of two parts:

$$a_1(t) = -e^{\gamma \cdot s(t)},\ a_2(t) = s(t),\ a(t) = a_1(t) \cdot a_2(t)$$

Based on the convolution theorem, $A(f)$ can be expressed as:

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

To losslessly recover $S(f)$ out of $A(f)$, 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.

$$A_1^{-1}(f)*A(f) = A_1^{-1}(f)*A_1(f)*S(f) = \delta(f)*S(f) = S(f) \quad (2)$$

By performing inverse FFT on $A_1^{-1}(f) * A(f)$, i.e., $S(f)$, we can restore the original audio signal $s(t)$. However, it is extremely difficult to deterministically obtain $A_1^{-1}(f)$ due to the following two reasons. First, $A_1^{-1}(f)$ is decided by the audio signal $s(t)$. In consideration of various speech signals in practice, we can hardly find the analytical expression of $s(t)$, let alone the further complex exponential function, FFT, and convolution related to $s(t)$. Second, the decay factor γ in $a_1(t)$ is unknown to us. Although one can collect multiple traces of the RF signal in advance to fit γ, it is quite a labor-intensive process. Meanwhile, the fitting result could be error-prone due to signal noise. Therefore, an alternative solution is required.

2) Audio reconstruction via convolutional neural network: The operation in Eq. (2) is essentially a convolutional process. Instead of explicitly deriving the $A_1^{-1}(f)$, we may employ the CNN with convolutional layers as the core to accomplish the above computation task. In addition to the required convolution operation, neural networks are also good at automatically adapting to various speech signals and resisting signal noise.

In specific, we take the mel-spectrogram of the eavesdropped RF signal amplitude as the input. Our goal is to build a model for converting the nonlinearly deformed mel-spectrogram to its original version. To reach this goal, we adopt the encoder-decoder model, which is widely applied for translation purposes [23]–[25], to “translate” the deformed mel-spectrogram. Thus, we select U-Net, a classical encoder-decoder convolutional network [26], for audio reconstruction.

Directly employing the U-Net network may not be able to recover the original audio mel-spectrogram well due to the disturbing RF signal noises in the silent period within the received signal. As shown in Fig. 5, the darker areas in the mel-spectrogram are mainly incurred by the inevitable noises in the RF transceiver, which is not helpful in training the reconstruction network. Thus, we make two main improvements on U-Net to achieve better reconstruction performance. First, inspired by the attention mechanism which has the ability to selectively focus on the most relevant parts of the input, we introduce attention gates into the U-Net. In this way, the network will pay more attention to the audio part rather than the background noise during training.

Second, we revise the loss function of U-Net. The existing U-Net networks use the mean squared error (MSE) loss between pixels of the input and output mel-spectrogram, under the assumption that every pixel is equally important. However, the audio part should play a more critical role than the background noise during network training. Thus, we propose a new loss function, which assigns a higher weight to the squared error of the audio signal part. Since the audio part has a larger amplitude than the remaining noises, each pixel's weight w_i is assigned proportional to its normalized strength in decibel. Then, the loss function is formulated as follows.

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot (p_i - \hat{p}_i)^2,\ w_i = \sqrt{\frac{p_i - p_{min}}{p_{max} - p_{min}}}, \quad (3)$$

where p_i and $\hat{p}_i$ refer to the pixel's value 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 architecture of the whole audio reconstruction model is depicted in Fig. 9. Our designed model can achieve a high-quality reconstruction of the audio signal. As shown in Fig. 10, compared with the raw mel-spectrogram of the eavesdropped RF signal in Fig. 10(b), the key frequency components of the audio become more outstanding in the reconstructed mel-spectrogram in Fig. 10(c). After obtaining the recovered mel-spectrogram from the enhanced U-net network, we need to transform the spectrogram into the audio signal. Thus, we employ the Griffin-Lim algorithm which is popularly used to

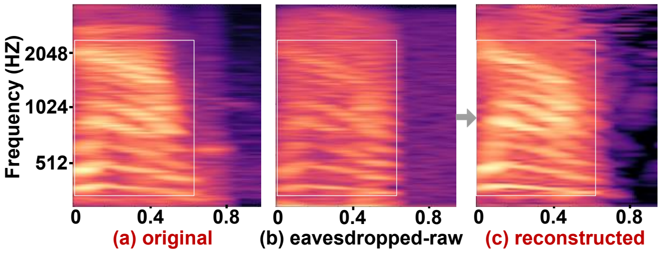


Fig. 10. Comparison of the (a) original audio's mel-spectrogram, (b) raw RF signal's mel-spectrogram, and (c) reconstructed mel-spectrogram

reconstruct the audio waveform [27]. The Griffin-Lim algorithm reconstructs the audio signal by randomly initializing a complex-valued phase mel-spectrogram and iteratively updating it to minimize the difference between the mel-spectrogram of the reconstructed signal and the original one.

F. Speech Command Recognition

We utilize the reconstructed mel-spectrogram of the eavesdropped RF signal 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 in speech command recognition. In specific, we employ the ResNet-50 convolutional network [28] 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 [29], [30].

IV. Evaluation

In this section, we present the implementation details, employed metrics, and evaluation results of RF-Parrot in audio eavesdropping and speech command recognition.

A. Implementation

1) Experiment setup: We implement RF-Parrot using commercial devices, including a USRP N210 as the SDR, a pair of SX200-150(P)C log-periodic antennas, and CE3512K2 D-MOSFETs. The experiment setup is shown in Fig. 11. We select multiple types of audio wires and embed our designed retroreflector inside. The original audio files are played by a smartphone whose audio jack is connected to the audio wire. The RF signal is controlled by GNURadio. The transmitting power and RF signal sampling rate are set to 10 dBm and 200 KHz, respectively. The received RF signal amplitude is saved as a wave file. The audio mel-spectrogram reconstruction and speech command recognition models are implemented via Pytorch using the RTX 4090 GPU. For mel-spectrogram calculation, we set the hop length and the number of mel bands to 80 and 128, respectively. All audio files are aligned to 1.28 s, and the mel-spectrogram size is 128*128.

2) Speech audio dataset: We employ two public speech audio datasets, i.e., Free Spoken Digit Dataset (FSDD) [31] and Speech Commands Dataset (SCD) [32]. FSDD includes 3,000 recordings of 10 English-spoken digits ('zero' to 'nine') 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 ('zero' to 'nine'). In audio reconstruction and command recognition tasks, data samples used for training and testing are allocated in a ratio of 4:1.

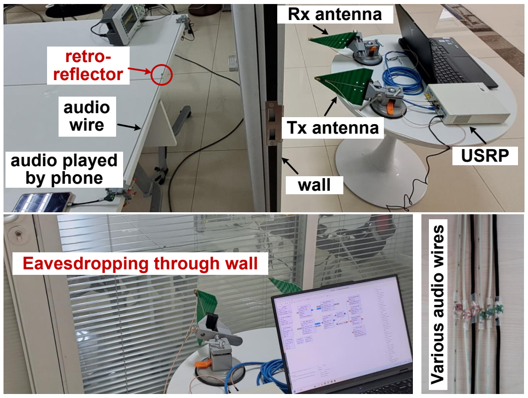


Fig. 11. Experiment setup of RF-Parrot for through-wall audio eavesdropping

3) Evaluation metrics: We employ four types of metrics to evaluate the performance of RF-Parrot on audio reconstruction and speech command recognition 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. Smaller MCDs indicate better audio reconstruction performance. Note that the speech recognition system can perceive the reconstructed audio that has an MCD of less than 8.
- **Signal-to-noise ratio (SNR) and Peak signal-to-noise ratio (PSNR)** are used to assess the quality of the audio signal. 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 one.
- **Mean opinion score (MOS)** is a subjective measure to evaluate the reconstructed audio quality. We recruit 20 volunteers, including 10 males and 10 females aged from 20 to 30, to listen to the 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 used to evaluate the performance of speech command recognition, including 10 spoken-digit classifications and 10 action command classifications.

B. Performance of audio reconstruction

In this experiment, we investigate the audio reconstruction performance from various aspects. We first calculate the MCD value for the reconstructed audios with 10 digits and 10 action commands, respectively. As depicted in Fig. 12, the MCD values of the 20 commands' eavesdropped audios are all below 8. The lowest MCD of 5.9 is achieved for the action command 'down', while the highest MCD of 7.9 corresponds to the digit command 'eight'. 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 II. The D-MOSFET reconstructed audio by the convolutional

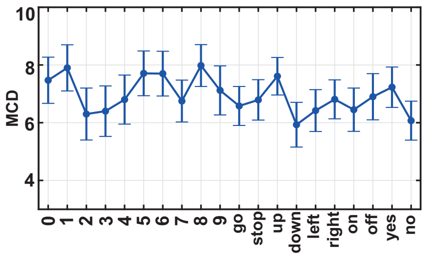


Fig. 12. MCD of reconstructed audio

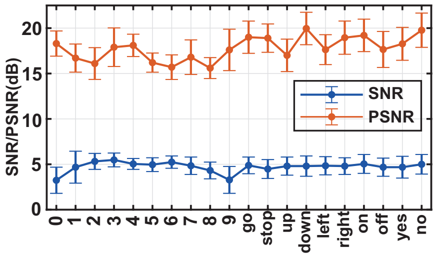


Fig. 13. SNR and PSNR of reconstructed audio

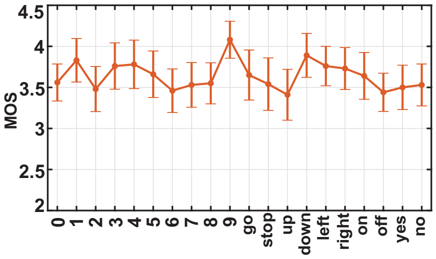


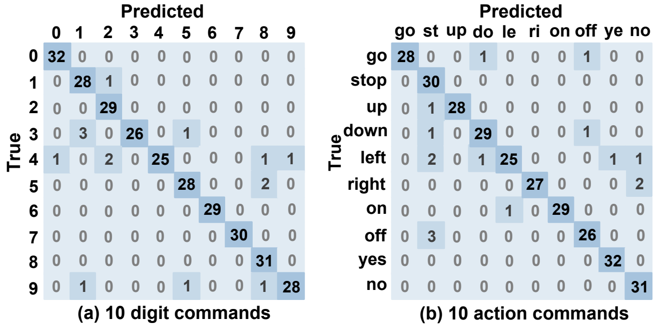
Fig. 14. MOS of reconstructed audio

TABLE II
COMPARISON OF MCD USING DIFFERENT MOSFETs

MOSFET type	E-MOSFET	D-MOSFET	
		Raw	**Reconstructed**
MCD	23.1	17.0	**6.8**

TABLE III
F1-SCORE OF SPEECH COMMAND RECOGNITION

Command	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
F1-score	0.98	0.92	0.95	0.93	0.91	0.93	1.0	1.0	0.94	0.93
Command	**go**	**stop**	**up**	**down**	**left**	**right**	**on**	**off**	**yes**	**no**
F1-score	0.97	0.91	0.98	0.94	0.91	0.96	0.95	0.91	0.98	0.95



Predicted

True	0	1	2	3	4	5	6	7	8	9
0	32	0	0	0	0	0	0	0	0	0
1	0	28	1	0	0	0	0	0	0	0
2	0	0	29	0	0	0	0	0	0	0
3	0	3	0	26	0	1	0	0	0	0
4	1	0	2	0	25	0	0	0	1	1
5	0	0	0	0	0	28	0	0	2	0
6	0	0	0	0	0	0	29	0	0	0
7	0	0	0	0	0	0	0	30	0	0
8	0	0	0	0	0	0	0	0	31	0
9	0	1	0	0	0	1	0	0	1	28

(a) 10 digit commands

Predicted

True	go	st	up	do	le	ri	on	off	ye	no
go	28	0	0	1	0	0	0	1	0	0
stop	0	30	0	0	0	0	0	0	0	0
up	0	1	28	0	0	0	0	0	0	0
down	0	1	0	29	0	0	0	1	0	0
left	0	2	0	1	25	0	0	0	1	1
right	0	0	0	0	0	27	0	0	0	2
on	0	0	0	0	1	0	29	0	0	0
off	0	3	0	0	0	0	0	26	0	0
yes	0	0	0	0	0	0	0	0	32	0
no	0	0	0	0	0	0	0	0	0	31

(b) 10 action commands

Fig. 15. Confusion matrix of (a) digit and (b) action commands

neural network achieves the lowest MCD of 6.8, revealing the effectiveness of our proposed audio reconstruction method. The above MCD results indicate that the eavesdropped audio signal closely resembles the original speech signal.

Next, we calculate the SNR and PSNR for the 20 commands' reconstructed audios in Fig. 13. The average SNRs and PSNR of different commands are 3-5 dB and 15-20 dB, respectively, indicating that the critical audio information is more dominant than the background noise. The digit command 'eight' achieves the lowest PSNR whereas the action command 'down' achieves the highest PSNR, which also matches with the MCD results. The results show that the reconstructed audio signal can achieve a relatively high quality.

Finally, we illustrate the average MOS of each speech command among all volunteers in Fig. 14. 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 digit command 'nine', and the lowest MOS of 3.4 for the action command 'up' is still above the borderline level of 3, which means that volunteers think that over half of the original speech is recovered. In a word, the above objective and subjective evaluation results validate that RF-PARROT can realize high-quality reconstruction of the eavesdropped audio signal.

C. *Performance of speech command recognition*

In this experiment, we show the speech command recognition performance using the reconstructed mel-spectrogram. The F1-scores of all 20 commands are illustrated in Table III. The highest F1-score can reach 1 for the 'six' and 'seven' commands, and the lowest F1-score is 0.91 for the 'four' and 'off' commands, which are all above 0.9. The detailed confusion matrices are shown in Fig. 15. Commands that have similar vowels, e.g., 'five' and 'nine' (all contain the /ɑɪ/), 'off' and 'stop' (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-based retroreflector to eavesdrop on the audio signal, then train and test the command recognition model, 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 audios by RF-PARROT can be accurately distinguished among different speech commands.

D. *Impact of practical factors*

1) Impact of the distance: In this experiment, we investigate the effect of the through-wall distance between the target wire and RF transceivers on eavesdropping performance. We set different distances, i.e., 0.5 m, 0.7 m, and 1 m. The SNR and PSNR of the eavesdropped audio for different distances are shown in Fig. 16. The SNR and PSNR generally decrease with a longer distance. In specific, when the distance increases from 0.5 m to 1 m, the SNR drops from 4.6 to 2.1; the PSNR declines from 17.7 to 16.3, as a result of more signal attenuation in the air. Although we witness the decline of SNR and PSNR, they are still maintained at a relatively high level. In addition, the current gain of RF antennas is set as 10 dB. We find the eavesdropping audio quality for longer distances can be further enhanced by increasing the gain of RF antennas.

2) Impact of the audio volume: In this experiment, we investigate the effect of audio volume from the smartphone on the eavesdropping performance. We set different levels of audio volume, i.e., 60%, 80%, and 100%. The SNR and PSNR of the eavesdropped audio signal under different volumes are shown in Fig. 17. By the incline of audio volume, both SNR and PSNR grow steadily because a higher volume can resist more of the RF signal attenuation effect and background noises. Albeit the minor degradation of reconstructed audio for the lower volume, the SNR and PSNR exceed 0 dB and 15 dB, respectively, stilling outstanding from the underlying noises.

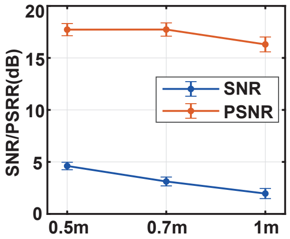


Fig. 16. Impact of distance

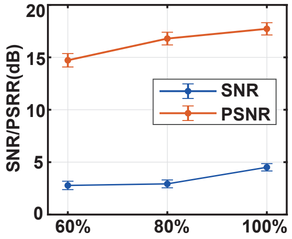


Fig. 17. Impact of audio volume

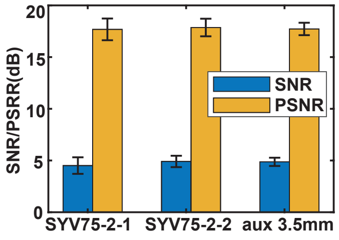


Fig. 18. Impact of different audio wires

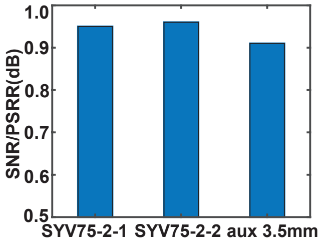


Fig. 19. Recognition accuracy of wires

3) Impact of the audio wire type: In this experiment, we investigate the effect of different types of audio wires on eavesdropping performance. We choose three types of audio wires, i.e., SYV75 wire series (SYV75-2-1 and SYV75-2-2) and daily 3.5 mm aux audio wire. SYV75 wire series are widely applied in the transmission of analog signals. Different models of the SYV75 wire vary in diameter, weight, attenuation coefficient, etc. The SNR and PSNR of the eavesdropped audio signal using different wires are shown in Fig. 18, which remain at similar and high levels, i.e., around 5 dB in SNR and 17 dB in PSNR. This shows that our designed retroreflector can effectively recover the audio signal transmitted via different audio wires. We then use the audio signal eavesdropped from different wires to recognize the speech command. The recognition accuracy is all above 90%, providing more evidence of the effectiveness of RF-PARROT across audio wires.

E. Countermeasures

We recommend the following countermeasures to defend against potential eavesdropping on wired audio. **(1) 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. Past works have shown that channel randomization can effectively against RF eavesdroppers [33]. **(2) Electromagnetic shielding.** We can also add RF shielding, e.g., metal foil or wire mesh, on the audio wire to impede the attacker from activating the retroreflector.

V. RELATED WORKS

A. Vibration-based Audio Eavesdropping

Acoustic signals are mechanical waves that force surrounding elastic objects to vibrate continuously. Researchers harness such phenomenon to investigate various vibration sensors, e.g., accelerometer [34]–[36], laser [7], RFID [6], WiFi [5], mmWave [2]–[4] and videos [8], for audio eavesdropping. AccEar employs the built-in accelerometer in the smartphone, which is sensitive to the mechanical wave caused by the phone's loudspeaker, to reconstruct the played audio [35]. Tag-Bug places RFID tags on the surrounding objects around the loudspeaker and collects the backscattered RFID signal for eavesdropping [6]. mmMIC realizes speech recognition directly from the mouse and throat reflected mmWave signal [37]. However, the above vibration-based eavesdropping methods require the audio to be played out by the loudspeaker so that the nearby object's surface can be driven to vibrate. By contrast, our RF-Parrot releases the above constraint with the help of the designed retroreflector, which can directly eavesdrop on the audio during transmission through the wire.

B. Magnetism-based Audio Eavesdropping

When playing audio, the speaker radiates a changing magnetic field, which can be used to infer the audio content. MagEar designs a coil to capture the magnetic field variation and reconstruct the audio from the earpiece [10]. VoiceListener employs the magnetometer as the eavesdropper and develops a training-free and universal mechanism to effectively reconstruct the audio from low-resolution sensors [38]. Magnetism-based approaches manage to restore the sound of speakers with smaller volumes, e.g., headphones and earpieces. However, they face a critical limit of the eavesdropping distance within 50 cm due to the sharp attention of the magnetic field. In comparison, RF-Parrot can achieve through-wall audio eavesdropping at a longer distance.

C. RFRA-based Audio Eavesdropping

RFRA practices can be traced back to "The Thing" [39], also known as the Great Seal bug, which is secretly installed in an ambassador's office for spying on talking speech and other outbound audios. Recently, RFRA schemes have been designed to conduct eavesdropping on the wire-transmitted signal, but only for digital signals [12]–[15]. For example, researchers have shown that digital keystrokes transmitted via a USB cable can be restored with less than a 5% error rate by the digital RFRA method at the 1 m distance [13]. However, existing works ignore the prevalent presence of analog audio signals in practice. Thus, we improve the previous design of retroreflectors in RFRA using the D-MOSFET and effectively eavesdrop on the audio signal with high quality.

VI. ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Program of China (No. 2021YFB3100400), National Natural Science Foundation of China (Grant No. 62302274, 62202276, 62202274, 62232010), Shandong Science Fund for Excellent Young Scholars (No. 2022HWYQ-038), and Natural Science Foundation of Shandong (No. ZR2023QF113),

VII. CONCLUSION

This paper proposes the first wired analog audio eavesdropping attack, RF-PARROT. With a modified earphone wire embedded with a tiny and judiciously designed retroreflector, RF-PARROT can remotely eavesdrop on audio signals through the wire. By leveraging the encoder-decoder neural network with convolutional layers for audio reconstruction, RF-PARROT can achieve 95% accuracy in identifying speech commands. We believe this work will raise awareness of the potential safety hazards of 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] 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

[9] 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

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

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

[18] 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.

[19] "Common coaxial cable sizes," http://www.snakebitdrill.com/common coaxialcablesizes.aspx.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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