RFSpy: Eavesdropping on Online Conversations with Out-of-Vocabulary Words by Sensing Metal Coil Vibration of Headsets Leveraging RFID

Yunzhong Chen
Shanghai Jiao Tong University
Shanghai, China
chenyzh@sjtu.edu.cn

Jiadi Yu*
Shanghai Jiao Tong University
Shanghai, China
jiadiyu@sjtu.edu.cn

Yingying Chen
Rutgers University
New Brunswick, NJ, USA
yingche@scarletmail.rutgers.edu

Linghe Kong
Shanghai Jiao Tong University
Shanghai, China
linghe.kong@sjtu.edu.cn

Yanmin Zhu
Shanghai Jiao Tong University
Shanghai, China
yzhu@cs.sjtu.edu.cn

Yi-Chao Chen
Shanghai Jiao Tong University
Shanghai, China
yichao@sjtu.edu.cn

ABSTRACT

Eavesdropping on human sound is one of the most common but harmful ways to threaten personal privacy. As one of the most essential accessories, headsets have been widely used in common online conversations, such as online calls, video meetings, etc. The metal coil vibration patterns of headset speakers/microphones have been proven to be highly correlated with the speaker-produced/microphone-received sound content. This paper presents an online conversation eavesdropping system, *RFSpy*, which uses only one RFID tag attached on a headset to alternately sense the metal coil vibrations of headset speaker and microphone for eavesdropping on speaker-produced and microphone-received sound. In some accessible scenarios, such as meeting rooms, offices, etc., assuming attackers secretly attach a small, battery-free RFID tag under one ear cushion of an eavesdropped user's headset without being noticed. Meanwhile, RFID readers are camouflaged as decorations placed in/out of rooms to transmit and receive RF signals. When the eavesdropped user talks with other users online by using the headset, *RFSpy* first activates the RFID tag attached on the headset to capture the metal coil vibration patterns of headset speaker and microphone upon RF signals. Then, *RFSpy* reconstructs sound spectrograms from the RF signal-based vibration patterns for not only trained words but also untrained (i.e., out-of-vocabulary) words by utilizing a designed Sound Spectrogram Reconstruction (SSR) network. Finally, *RFSpy* converts the sound spectrograms to conversation content through a sound recognition API. Extensive experiments in real environments demonstrate that *RFSpy* can eavesdrop on online conversations with out-of-vocabulary (OOV) words effectively.

[*] Jiadi Yu is the corresponding author

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06…$15.00
https://doi.org/10.1145/3643832.3661887

CCS CONCEPTS

• **Security and privacy → Hardware attacks and countermeasures**.

KEYWORDS

RFID, conversation eavesdropping, coil vibration, out-of-vocabulary word

ACM Reference Format:
Yunzhong Chen, Jiadi Yu, Yingying Chen, Linghe Kong, Yanmin Zhu, and Yi-Chao Chen. 2024. *RFSpy*: Eavesdropping on Online Conversations with Out-of-Vocabulary Words by Sensing Metal Coil Vibration of Headsets Leveraging RFID. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3643832.3661887

1 INTRODUCTION

In recent years, with the rapid development of Internet of Things (IoT) sensing technologies, sound eavesdropping has become a major security concern due to the non-encrypted characteristic of human-speaking sound, which makes it vulnerable for attackers to acquire sensitive information, such as personal privacy, business secrets, and even military secrets, etc. Given the potential risks, existing acoustic sensor-based eavesdropping methods, such as using microphones to record sound, have been extensively researched for establishing mature countermeasures, e.g., setting up soundproof rooms, placing anti-recording devices [5], etc. In addition, acoustic sensing methods can only record human-speaking sound but fail to eavesdrop on speaker-produced sound during online conversations through headsets. Hence, side-channel attack methods for sound eavesdropping are becoming increasingly attractive.

Some recent works explore using various non-acoustic sensors, such as motion sensors and RF signals, to eavesdrop on speaker-produced sound [1, 22, 44–46, 59] or human-speaking sound [7, 39, 51]. However, RF signal-based sound sensing methods [22, 44, 45, 51, 59] generally require employed sensors, e.g., mmWave-radars, facing towards eavesdropped subjects, such as users' throats, phones, and speakers, etc. Hence, they are not robust to users' positions, directions, etc., during sound eavesdropping. As one of the most common and essential accessories, headsets/earphones are widely used in various online conversations, such as online calls, video meetings, and online business services, etc. Hence, some researchers use customized headsets/earphones equipped with metal coil [31] or motion sensor [6] to eavesdrop on speaker-produced sound [31] or microphone-received sound, i.e., human-speaking sound [6]. However, due to the limitations of employed sensors, the aforementioned works enable sensing either speaker-produced sound or human-speaking sound, but fail to eavesdrop on both speaker-produced and human-speaking sound through only one sensor, which limits their application scenarios greatly.

Nowadays, Radio-frequency identification (RFID) technologies have been developed rapidly. RFID tag has the characteristics of metal sensitivity [14, 21], flexibility, small size, and low cost, so it is suitable to be attached on common headsets for sensing the metal coil vibration of speakers and microphones. Since the metal coil vibration patterns of headsets speakers/microphones are highly correlated with the speaker-produced/microphone-received sound content [26], we consider leveraging RF signals to eavesdrop on online conversations through a headset with an attached RFID tag.

In some easily accessible or public scenarios, such as offices, meeting rooms, labs, etc., attackers can secretly attach a small and battery-free RFID tag under the ear cushion of an eavesdropped user's headset when the eavesdropped user is absent temporarily, making it difficult to be noticed by the eavesdropped user during online conversations through the headset attached an RFID tag. Due to the close distance between the speaker and microphone of headsets, it is feasible for attackers to use only one RFID tag to alternately sense the metal coil vibrations of speaker and microphone. Furthermore, owing to the outstanding penetration ability of RFID signals, a compact RFID reader and antennas can be camouflaged as decorations, such as flower pots, framed paintings, and bulletin boards, placed in rooms, or outside walls to transmit and receive RF signals for achieving online conversation eavesdropping. However, to implement online conversation eavesdropping through headsets using RF signals, we face several challenges in practice. First, the eavesdropping scheme needs to utilize only one RFID tag attached on an headset for sensing the metal coil vibration of both the headset speaker and microphone in/out of rooms. Second, the eavesdropping scheme should be able to remove various interference from the received RF signals for recovering conversation content accurately. Third, the eavesdropping scheme should be able to implement sound eavesdropping for not only trained words but also untrained words, i.e., out-of-vocabulary (OOV) words.

In this paper, we first investigate the feasibility of online conversation eavesdropping leveraging an RFID tag, and find that RF signals can capture the metal coil vibration patterns of headset speakers and microphones. Motivated by the observations, we propose an online conversation eavesdropping system, *RFSpy*, which utilizes one RFID tag secretly attached on a headset to eavesdrop on speaker-produced and microphone-received sound. In *RFSpy*, when a user talks with other users online by using a headset with an attached RFID tag, the RFID tag is first activated by RF signals sent by a camouflaged signal-transmitting antenna to sense the metal coil vibration of the headset speaker and microphone alternately. Then, *RFSpy* receives RF signals back-scattered by the RFID tag, and pre-processes the received RF signals to remove human body motion interference and echo interference using designed Variational Mode Decomposition (VMD) algorithm and Adaptive Echo Removing (AER) algorithm, respectively. After the signal preprocessing, *RFSpy* obtains RF signal spectrograms corresponding to the metal coil vibration patterns of the speaker/microphone. Next, *RFSpy* trains a mapping relationship model between RF signal spectrograms and sound spectrograms through a designed Sound Spectrogram Reconstruction (SSR) network. Based on the trained mapping relationship model, *RFSpy* further constructs a phoneme-based pixel-column mapping relationship to reconstruct sound spectrograms for not only trained words but also untrained (i.e., out-of-vocabulary) words. Subsequently, *RFSpy* converts the reconstructed sound spectrograms to time-domain sound waveforms leveraging Griffin-Lim algorithm. Finally, the time-domain sound waveforms are converted to sound content through sound recognition API for achieving online conversation eavesdropping. Experiments in real environments show that *RFSpy* can effectively eavesdrop on online conversations with OOV words.

We highlight our main contributions as follows:

- We propose a *RFSpy* system, which utilizes only one RFID tag attached on a headset to alternately sense the metal coil vibration of headset speaker and microphone for eavesdropping on speaker-produced and microphone-received sound.
- We present a Variational Mode Decomposition (VMD) algorithm and an Adaptive Echo Removing (AER) algorithm to remove human body motion interference and echo interference on RF signals, respectively.
- We design a SSR network to train a mapping relationship model between RF signal spectrograms and sound spectrograms, and further construct a phoneme-based pixel-column mapping relationship to reconstruct sound spectrograms for not only trained words but also untrained (OOV) words.
- We conduct extensive experiments in real environments, and the results show that *RFSpy* achieves an average Mel-Cepstral Distortion (MCD) of 6.37 and Word Error Rate (WER) of 17.88% for eavesdropping on online conversations with OOV words.

2 ATTACK MODEL AND EAVESDROPPING SCENARIO

People often make online conversations through headsets, such as online calls, video meetings, and online business services, etc. The metal coil vibration patterns of headset speakers/microphones have been proven to be highly correlated with the speaker-produced/microphone-received sound content [26]. Since RFID tags are sensitive to nearby metal [14, 21], it is possible for attackers to attach an RFID tag on headsets for conversation eavesdropping.

Some online conversation scenarios, such as offices, meeting rooms, labs, etc., are often accessible to attackers, so attackers can secretly attach a small, battery-free RFID tag under the ear cushion of headsets during eavesdropped users' temporary absence. Typically, people tend to use a larger headset for online conversation to obtain a comfortable experience and superior sound quality. Current mainstream RFID tags are only about ten millimeters in size, which can be easily attached under ear cushion of the headset. So, it is hard for the eavesdropped users to notice the RFID tag. Even for smaller earphones, there exist small RFID tags, e.g., 2mm × 3mm

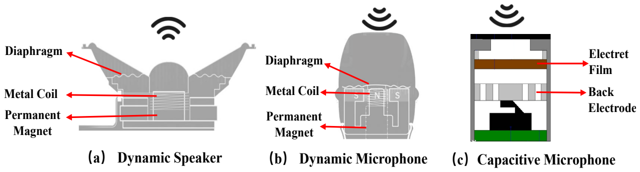


Figure 1: Illustration of speakers and microphones.

in size[36], which can be attached under the ear cushion of these earphones without raising attention.

Meanwhile, a compact RFID reader and antennas can be camouflaged as decorations, such as flower pots, framed paintings, and bulletin boards, etc., placed in rooms to transmit and receive RF signals for eavesdropping on speaker-produced and microphone-received sound during online conversations. Moreover, owning to the outstanding wall-penetrating ability of RFID signals [12, 57], the RFID reader and antennas can also be placed outside walls to enhance their camouflage capabilities for achieving more natural online conversation eavesdropping.

Since the passive sensing capability of RFID tags, as long as eavesdropped users wear headsets attach an RFID tag for online conversations within the sensing range of RFID antennas, attackers can always conduct conversation eavesdropping attacks.

3 BACKGROUND AND PRELIMINARY

RFID tag has high metal sensitivity, making it suitable for sensing the metal coil vibration of headset speakers and microphones for online conversation eavesdropping.

3.1 Background

To satisfy the needs of online conversations, a headset usually contains speakers and microphones. Speakers serve the purpose of producing sound, while microphones are utilized for capturing sound. Dynamic speakers are the most common type of speakers in headsets, which contains three main parts, i.e., permanent magnet, diaphragm, and metal coil, as shown in figure 1(a). It converts electrical energy into sound energy by interaction between the magnetic field and electric current. The sound energy further drives the metal coil vibration of speakers, thereby producing sound.

Figure 1(b) and 1(c) show dynamic and capacitive microphones. Dynamic microphones consists of three main parts, i.e., permanent magnet, diaphragm, and metal coil, as shown in figure 1(b). Surrounding sound drives the metal coil vibration of microphones, and sound energy is converted into electrical energy by interaction between magnetic field and electric current, thereby receiving sound. Capacitive microphones use principle of capacitors to receive sound. As shown in figure 1(c), it consists of an electret film (i.e., metal film) and a back electrode, which combines as a capacitor. When surrounding sound drives the electret film vibration of a capacitive microphone, the capacitance of capacitor changes, which converts sound into electrical signals, thereby receiving sound.

During sound receiving/producing process of microphones/speakers, metal coils or electret films undergo high-frequency vibration, which are correlated with corresponding sound contents [26]. Usually, metal coil in speakers/microphones contains metal, which exhibit significant reflection characteristics for electromagnetic waves [14, 21]. For simplicity, we refer to both metal coil and electret film collectively as "metal coil" in the following. When both parties of online conversations talk by using a headset with an attached RFID tag, the speaker-produced/microphone-received sound drives the metal coil vibration of headset speaker/microphone. If there are electromagnetic waves hitting the vibrating metal coil, it brings time-varying electromagnetic waves. Hence, it is possible for reflected electromagnetic waves capture metal coil vibration patterns of speakers/microphones.

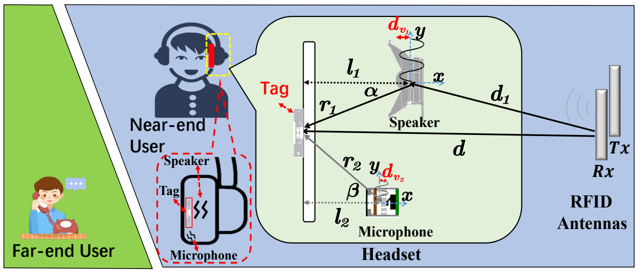


Figure 2: Illustration of modeling the metal coil vibration of headsets upon RF signals.

3.2 Modeling the Metal Coil Vibration of Headsets upon RF Signals

When a near-end user (i.e., an eavesdropped user) talks with a far-end user (i.e., a user that talks online with the eavesdropped user) online by using a headset with an attached RFID tag, the sound produced by the near-end user/far-end user drives the metal coil vibration of headset microphone/speaker. As shown in figure 2, we attach an RFID tag on one side of the near-end user's headset. Supposing the signal-transmitting antenna T_x and signal-receiving antennas R_x are located at the same position, T_x first sends electromagnetic signals to activate the RFID tag on the headset, and the activated tag feedbacks RF signals through back-scattering to R_x. For the eavesdropped user, the entire conversation process involves two alternating stages, i.e., the listening stage and speaking stage.

In the listening stage, the sound of far-end user drives the metal coil vibration of headset speaker, which is captured by the RFID tag attached on the headset. Specifically, the received RF signals from the RFID tag contain three parts, i.e., RF signals reflected by all stationary objects, RF signals propagating through line-of-sight (LOS) path, and RF signals reflected by the vibrating metal coil of speaker. Phase of RF signals reflected by all stationary objects are constant, which is $\Delta\varphi_1 = \Phi$. Phase of RF signals propagating through LOS path is $\Delta\varphi_2 = 2\pi\frac{d}{\lambda}$, where d is the distance between RFID antennas and tag, λ is wavelength of RF signals. Phase of RF signals reflected by the metal coil of speaker is

$$\Delta\varphi_3 \approx -\left(2\pi \cdot \frac{d_1}{\lambda}\right)\gamma + 2\pi\frac{\frac{l_1}{\cos\alpha} - \frac{d_{v1}}{\cos\alpha}\cos 2\pi f t}{\lambda}, \tag{1}$$

where d_1 represents the distance between the speaker and RFID antenna, r_1 represents the linear distance between the speaker and RFID tag, l_1 represents the vertical distance between the speaker

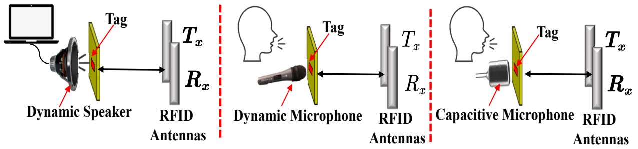


Figure 3: Illustration of sensing disassembled speaker/ microphone vibration using RF signals.

and RFID tag, d_{v1} represents the maximum vibration amplitude of speaker metal coil, α is the angle between r_1 and l_1, t is time, f is the metal coil vibration frequency of speaker, and γ represents the reflection coefficient of metal on RF signals. The reflection coefficient γ is $\frac{z_m - z_a}{z_m + z_a}$, where z_m is impedance of metal and z_a is impedance of air. Since z_m is much smaller than z_a, $\frac{z_m - z_a}{z_m + z_a}$ is approximately equal to 1. Based on the above analysis, the received RF signals $\Delta\varphi$ is represented as

$$\begin{aligned}\Delta\varphi &= \Delta\varphi_1 + \Delta\varphi_2 + \Delta\varphi_3 \\ &= \frac{cos\alpha\,(\lambda\Phi + 2\pi d - 2\pi d_1\gamma) + 2\pi l_1}{cos\alpha\lambda} - \frac{2\pi d_{v_1}\cos(2\pi f t)}{\lambda\cos\alpha} \\ &= k - \frac{2\pi d_{v_1}\cos(2\pi f t)}{\lambda\cos\alpha},\end{aligned} \tag{2}$$

where $k = \frac{\cos\alpha(\lambda\Phi + 2\pi d - 2\pi d_1\gamma) + 2\pi l_1}{\cos\alpha\lambda}$. k is a constant because α, Φ, λ, d, d_1, and γ are all constants. Based on Eq. 2, we can infer the metal coil vibration frequency f and amplitude d_{v_1} of the speaker from the received phase of RF signals.

During the speaking stage, the sound of near-end user drives the metal coil vibration of headset microphones, which is sensed by the RFID tag on the headset. Similar to the modeling process in the listening stage, we can infer the metal coil vibration frequency and amplitude of the headset microphone from the received phase of RF signals when the near-end user speaks.

Based on the inferred metal coil vibration frequency and amplitude of headset speaker and microphone, we construct the metal coil vibration model of headset speakers and microphones upon RF signals.

3.3 Feasibility Analysis of Online Conversation Eavesdropping using RF Signals

Based on the constructed metal coil vibration model of headset speakers/microphones, we consider using an RFID tag attached on a headset to alternately capture the metal coil vibration patterns of speakers and microphones for eavesdropping on speaker-produced and microphone-received sound during online conversations.

To explore the feasibility of conversation eavesdropping by sensing the metal coil vibration of headsets leveraging RF signals, we conduct experiments to analyze the spectrograms of received RF signals. Since the third-order harmonic signals achieve a trade-off between sensing sensitive and signal intensity [42], we choose the USRP N210 reader to monitor the third-order harmonic RF signals at the frequency of 2761.89MHz with the sampling frequency of 2MHz, and utilize the phase acquisition method in [30] for improving the accuracy of sound eavesdropping. Furthermore, we utilize the method in [55] to magnify the received RF signals for sensing the subtle metal coil vibration more effectively. To avoid the echo interference between the speaker and microphone, a disassembled dynamic speaker, dynamic microphone, and capacitive microphone are separately employed in the experiments. As shown in figure 3, we use a hard plastic plate attached an RFID tag to imitate the headset shell, which is placed in front of the RFID antennas, and the speaker/microphone is placed close to the RFID tag.

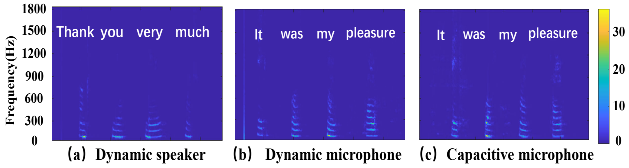


Figure 4: Examples of spectrograms of RF signal phase when speaker produces/microphone receives sound.

Figure 4 (a) shows the spectrogram of RF signal phase corresponding to the dynamic speaker when it produces words "Thank you very much". Figure 4(b) and (c) shows the spectrograms of RF signal phase corresponding to the dynamic microphone and capacitive microphone when they receive words "It was my pleasure". It can be observed from the figures that there are obvious differences between spectrograms of different words for the speaker and two microphones. The structural similarity is a common metric to measure the similarity of two images [19], so we calculate the structural similarity of spectrograms corresponding to these words. The results show that the structural similarity of different words' spectrograms are all less than 0.5, and that of the same word's spectrograms are all larger than 0.8 for the dynamic speaker. And the structural similarity for both dynamic and capacitive microphones reveal the same characteristics.

Therefore, different words can be effectively distinguished based on the spectrograms of RF signals for speakers and microphones. The above encouraging results demonstrate the potential of eavesdropping on speaker-produced and microphone-received sound by sensing the metal coil vibration of headset speakers and microphones using RF signals.

4 SYSTEM OVERVIEW

We propose a online conversation eavesdropping system, *RFSpy*, which utilizes an RFID tag attached on a headset to alternately capture the metal coil vibration patterns of headset speakers and microphones for eavesdropping on online conversations. Figure 5 shows the architecture of *RFSpy* system. In *RFSpy*, signal-transmitting antenna first sends RF signals to activate the RFID tag on headset, and then RF signals back-scattered by the tag are received by signal-receiving antenna for recovering online conversation content. *RFSpy* system is composed by the following modules:

Pre-processing signal. After receiving the RF signals back-scattered by the RFID tag, *RFSpy* first removes human body motion interference through Variational Mode Decomposition (VMD) algorithm. Then, *RFSpy* removes the echo interference between the speaker and microphone leveraging Adaptive Echo Removing (AER) algorithm on the received RF signals. After signals pre-processing, *RFSpy* obtains the RF signal spectrograms corresponding to the metal coil vibration patterns of speaker/microphone.

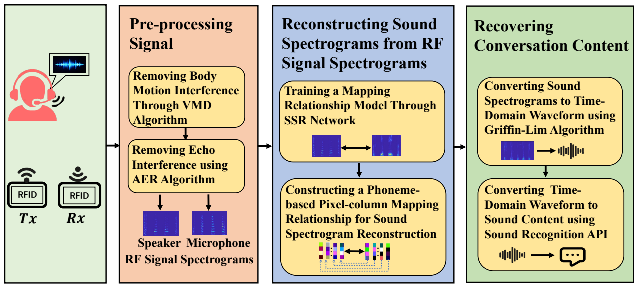

RFID
Tx
RFID
Rx
Pre-processing Signal
Removing Body Motion Interference Through VMD Algorithm
Removing Echo Interference using AER Algorithm
Speaker Microphone
RF Signal Spectrograms
Reconstructing Sound Spectrograms from RF Signal Spectrograms
Training a Mapping Relationship Model Through SSR Network
Constructing a Phoneme-based Pixel-column Mapping Relationship for Sound Spectrogram Reconstruction
Recovering Conversation Content
Converting Sound Spectrograms to Time-Domain Waveform using Griffin-Lim Algorithm
Converting Time-Domain Waveform to Sound Content using Sound Recognition API

Figure 5: Architecture of *RFSpy* System.

Reconstructing sound spectrograms from RF signal spectrograms. *RFSpy* first trains a mapping relationship model between RF signal spectrograms and sound spectrograms through Sound Spectrogram Reconstruction (SSR) network. Based on the trained mapping relationship model, *RFSpy* further constructs a phoneme-based pixel-column mapping relationship to reconstruct sound spectrograms for not only trained words but also untrained (i.e., out-of-vocabulary) words.

Recovering conversation content. *RFSpy* first converts the sound spectrogram to time-domain sound waveforms using Griffin-Lim algorithm. After that, the time-domain sound waveforms are converted into sound content through sound recognition API for achieving conversation eavesdropping.

5 SIGNAL PRE-PROCESSING

To eavesdrop on online conversations through a headset with an attached RFID tag, *RFSpy* first collects RF signals reflected from the tag to capture the metal coil vibration patterns of headset speaker and microphone. Since background environments usually affect the received RF signals, a spectral subtraction algorithm [4] is utilized to remove the environmental interference. Besides, the received RF signals usually suffer from human body motion interference and echo interference between the speaker and microphone during online conversations, so *RFSpy* needs to further remove these interference.

5.1 Removing Human Body Motion Interference

When a near-end user talks with a far-end user online by using a headset with an attached RFID tag, the received RF signals usually suffer from body motion interference, such as facial movements, walking, and bone vibration, etc.

Generally, there are three kinds of motions that can be captured by RF signals: (1) *Bone vibration*, caused by human speech, which has a relatively high frequency ($100 \sim 500Hz$ [13]); (2) *Body movements*, such as facial movements, walking, and waving hands, etc., during online conversations, which has a relatively low frequency ($1 \sim 10Hz$ [41]); (3)*Metal coil vibration of headset speakers/microphones*, caused by speaker-produced/ microphone-received sound, which has a frequency range from $50 \sim 800Hz$ [35].

RFSpy employs Variational Mode Decomposition (VMD) [15] algorithm to extract RF signals corresponding to metal coil vibration and remove RF signal interference corresponding to bone vibration and body movements. The VMD algorithm can adaptively decompose the phase of RF signals into multiple frequency components, and each frequency component corresponds to a kind of specific motion. Specifically, VMD assumes that any complicated signal $f(x)$ consists of k sub-signals, i.e., intrinsic mode functions (IMFs). Each IMF is described as an analytic signal $u_k(t)$, which has a limited frequency bandwidth with center frequency w_k.

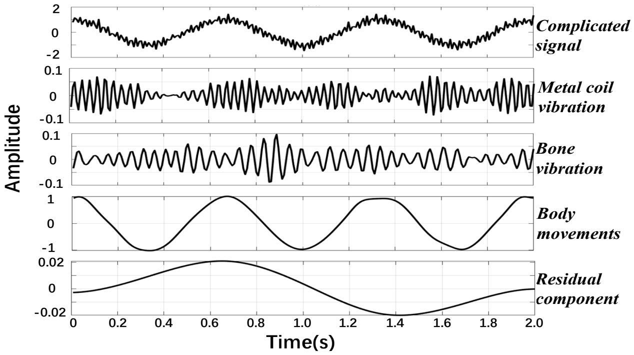


Figure 6: An example of VMD decomposition.

To separate the received RF signal $f(x)$ into k IMFs, VMD constructs an objective function. The optimization of the objective function satisfies two constraints: (1) the sum of estimated bandwidth for all analytic signal $u_0(t), u_1(t), u_2(t), ..., u_{k-1}(t)$ is minimized; (2) the sum of all estimated analytic signal $u_0(t), u_1(t), u_2(t), ..., u_{k-1}(t)$ is equal to the received RF signal $f(x)$. The initial center frequencies of $u_0(t), u_1(t), u_{k-1}(t)$ are set close to the average of their frequency ranges. By utilizing Alternate Direction Method of Multipliers (ADMM) [15], VMD updates the analytic signal $u_i(t)$ and corresponding center frequency w_i iteratively until the analytic signal $u_i(t)$ converges for all k IMFs. Finally, RF signal $f(x)$ is decomposed into k IMFs.

Figure 6 shows the VMD decomposition results of a period of RF signal phase when a user speaks through a headset microphone during walking. It can be observed from the figure that after VMD decomposition, *RFSpy* obtains four IMFs, which is the metal coil vibration of headset speakers/microphones, bone vibration, body movements, and residual component, respectively.

To verify the effectiveness of VMD algorithm, we compare the decomposed metal coil vibration patterns through VMD with that from a disassembled microphone, as shown in figure 3(b) (i.e., without human body motion interference), when a user speaks the same content, and the results are shown in figure 7. We calculate the Pearson Correlation Coefficient[11] between the decomposed metal coil vibration patterns through VMD and that from a disassembled microphone, and the value reaches 0.85. These positive results demonstrate the high similarity between them. Hence, by utilizing the VMD algorithm, *RFSpy* removes body motion interference from the received RF signals effectively.

5.2 Removing Echo Interference

Besides human body motion interference, RF signals also suffer from echo interference, i.e., the speaker-produced sound drives the metal coil vibration of not only headset speakers but also headset microphones, and the metal coil vibration of the headset microphones interfere with the received RF signals. Similarly, the metal

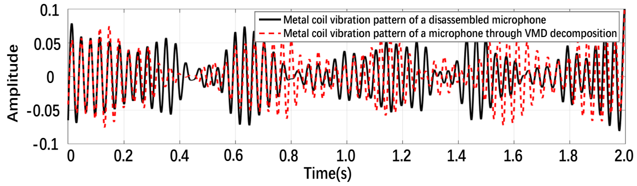


Figure 7: Example of metal coil vibration patterns decomposed via VMD and from a disassembled microphone.

coil vibration of headset speakers interferes with RF signals for sensing the metal coil vibration of headset microphones.

To remove echo interference of headset speakers/microphones, we design an Adaptive Echo Removing (AER) algorithm based on Least Mean Square (LMS) filter [32] to extract RF signals only corresponding to the metal coil vibration of headset speakers/microphones, which involves two stages, i.e., the speaker eavesdropping stage and microphone eavesdropping stage.

In the speaker eavesdropping stage, the sound of far-end user drives the metal coil vibration of not only the headset speaker but also the headset microphone, and RF signals reflected from the metal coil of speaker and microphone are recorded as $v(n)$ and $x(n)$, respectively, where n represents the number of RF signal sampling points. When RF signal $x(n)$ propagates in space through multiple paths $w(n)$, it causes echo interference $y(n)$ on RF signals $v(n)$, i.e., $y(n) = x(n) * w(n)$, where $*$ denotes the signal convolution operation, and $y(n)$ is the echo interference on speaker caused by the metal coil vibration of microphone. Hence, RF signals $d(n)$ reflected from the speaker contains echo interference, which are represented as

$$d(n) = x(n) * w(n) + v(n). \tag{3}$$

RFSpy utilizes the AER algorithm to extract RF signals only corresponding to the metal coil vibration of speaker. Specifically, the optimization of AER algorithm is an iterative process, each iteration consists of three steps. In the first step, AER uses the LMS adaptive filter to estimate propagation paths $\hat{w}(n)$ of RF signals reflected from the microphone, which is represented as $\hat{w}\,(n) = [w\,(0)\,, w\,(1)\,, ..., w\,(\mathrm{N})]^T$. In the second step, AER first records RF signals $x(n)$ reflected from the microphone, and then estimates the RF signal echo interference $\hat{y}\,(n)$ from the microphone to speaker, which is represented as $\hat{y}\,(n) = x\,(n) * \hat{w}\,(n) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n$. In the third step, AER subtracts $\hat{y}\,(n)$ from $d(n)$ to obtain $\hat{v}(n)$ that eliminates the echo interference after once iteration.

To complete the whole iterative optimization process, we define two objective functions. The first objective function is

$$S_{min}(\mathbf{w_{n+1}}) = \|\mathbf{w_{n+1}} - \mathbf{w_n}\|^2\,, \tag{4}$$

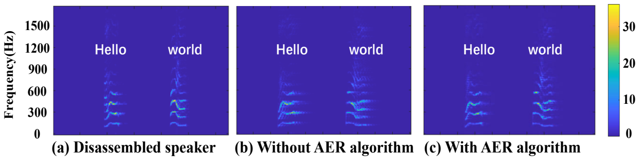
which represents that the iteration step size is minimized for iteration stability. The second objective function is

$$E_{min}(\mathbf{w_{n+1}}) = d(n) - \mathbf{w}_{\mathbf{n+1}}^{\mathbf{T}}\mathbf{x_n}, \tag{5}$$

which represents that the error between the estimated and actual RF signal paths is minimized.

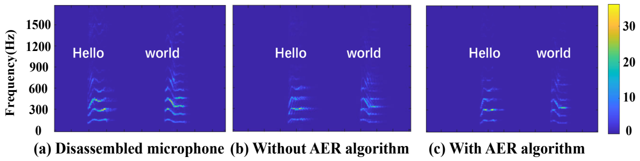
Based on the above two objective functions, a joint optimization objective function is constructed as

$$J_{\min}\,(\mathbf{w}_{\mathrm{n+1}}) = \|\mathbf{w}_{\mathrm{n+1}} - \mathbf{w}_{\mathbf{n}}\|^2 + \lambda\left(d(n) - \mathbf{w}_{n+1}^{T}\mathbf{x}_n\right), \tag{6}$$



(a) Disassembled speaker (b) Without AER algorithm (c) With AER algorithm

Figure 8: An example of removing the echo interference in speaker eavesdropping stage.



(a) Disassembled microphone (b) Without AER algorithm (c) With AER algorithm

Figure 9: An example of removing the echo interference in microphone eavesdropping stage.

where λ is a constant parameter. When the joint optimization objective function converges, *RFSpy* obtains the estimated echo paths $\hat{w_n}(n)$. After that, we bring $\hat{w_n}(n)$ into Eq.(3) to get RF signals $v(n)$ reflected by the metal coil vibration of speaker. Based on the above AER algorithm, *RFSpy* eliminates echo interference on the headset speaker from the headset microphone.

Moreover, for the microphone eavesdropping stage, by leveraging the echo interference removing algorithm similar to the speaker eavesdropping stage, *RFSpy* can also eliminate the echo interference when the near-end user speaking.

To verify the effect of AER algorithm, we conduct experiments to compare the spectrogram of RF signal phase with/without AER algorithm with that from a disassembled speaker and microphone (as shown in figure 3(a) and figure 3(b)) when the speaker produces or microphone receives sound "hello world", and the results are shown in figure 8 and figure 9. We calculate the structural similarity between the spectrogram of RF signal phase with/without AER algorithm and that from the disassembled speaker and microphone. By utilizing AER algorithm, the structural similarity increases from 0.58 to 0.85 in speaker eavesdropping stage, and increases from 0.65 to 0.82 in microphone eavesdropping stage. Hence, *RFSpy* removes the echo interference of speakers and microphones through designed AER algorithm effectively.

6 RECONSTRUCTING SOUND SPECTROGRAMS FROM RF SIGNAL SPECTROGRAMS

After the signal pre-processing, RF signals are able to capture the metal coil vibration patterns of speakers and microphones when headset speakers/microphones produce/receive sound. Based on the captured patterns, *RFSpy* further reconstructs sound spectrograms from RF signal spectrograms for online conversation eavesdropping.

6.1 Training a Mapping Relationship Model Through SSR Network

To reconstruct sound spectrograms from RF signal spectrograms, we design a Sound Spectrogram Reconstruction (SSR) network to

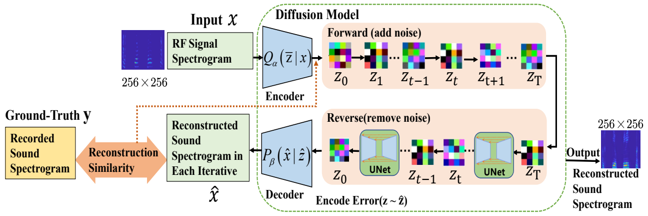

Input x
256 × 256
RF Signal Spectrogram
Diffusion Model
$Q_\alpha(\bar{z} \mid x)$
Encoder
Forward (add noise)
z_0 z_1 z_{t-1} z_t z_{t+1} z_T
Ground-Truth y
Recorded Sound Spectrogram
Reconstruction Similarity
Reconstructed Sound Spectrogram in Each Iterative
$\hat{x}$
$P_\beta(\hat{x} \mid \hat{z})$
Decoder
Reverse(remove noise)
z_0 UNet z_{t-1} z_t UNet z_T
Encode Error(z ~ ẑ)
256 × 256
Output
Reconstructed Sound Spectrogram

Figure 10: Architecture of SSR network.

train a mapping relationship model between RF signal spectrograms and sound spectrograms. Figure 10 shows the architecture of SSR network, which consists of three main parts, i.e., encoder, diffusion model, and decoder.

Specifically, RF signal spectrogram x of a headset speaker/microphone is first segmented into 2s pieces, and then normalized to a 256 × 256 pixel size for reducing the complexity of model construction. Then, the normalized piece is input to an encoder. The encoder performs nonlinear mapping to compress high-dimensional data as low-dimensional latent space representation $\bar{z}$, i.e., the encoder models a probability distribution $Q_\alpha(\bar{z} \mid x)$ from x to $\bar{z}$, mapping the input x to the latent representation $\bar{z}$.

After that, the latent space representation $\bar{z}$ is sent to a diffusion model [37] for estimating a feature distribution $\hat{z}$, which aims to learn the real feature distribution z. The diffusion model contains two processes, i.e., the forward and reverse diffusion process, and each process contains T steps. In the forward diffusion process, *RFSpy* adds Gaussian distribution noise $q(z_t \mid z_{t-1}, V_s)$ on feature distribution z_{t-1} to obtain a new feature distribution z_t in each step. V_s is a variable gating parameter, which is determined by the input category, i.e., $V_s = 1/0$ represents RF signal spectrograms corresponding to the metal coil vibration of microphone/speaker. After T steps in the forward diffusion process, a feature distribution z_T is obtained. In the reverse diffusion process, based on the feature distribution z_T in the forward process, *RFSpy* estimates a Gaussian noise distribution at each step using U-Net architecture, and the Gaussian noise distribution is represented as $p_\theta(z_{t-1} \mid z_t, V_s)$. *RFSpy* removes the estimated Gaussian noise distribution from the feature distribution z_t to obtain a new feature distribution z_{t-1}. After T steps, *RFSpy* obtains a low-dimensional latent space representation z_0, i.e., the estimated feature distribution $\hat{z}$ of the input RF signal spectrogram.

Based on the estimated feature distribution $\hat{z}$, *RFSpy* outputs the reconstructed sound spectrogram $\hat{x}$ corresponding to the RF signal spectrogram through a decoder, and the output process takes the recorded sound spectrogram (i.e., ground truth) as the goal. The decoder models a probability distribution $P_\beta(\hat{x} \mid \hat{z})$ from the feature distribution $\hat{z}$ to the reconstructed sound spectrogram $\hat{x}$ in each iterative. After above reconstruction step, SSR network completes once iterative. Next, *RFSpy* calculates reconstruction similarity between the output $\hat{x}$ of previous iterative and the recorded sound spectrogram, which is used to adjust the latent space representation z_0 of diffusion model in next iterative. *RFSpy* repeats above steps until mapping relationship model between the RF signal spectrogram and corresponding sound spectrogram is established. Based on the established mapping relationship model, *RFSpy* outputs the final reconstructed sound spectrogram from the input RF signal spectrogram.

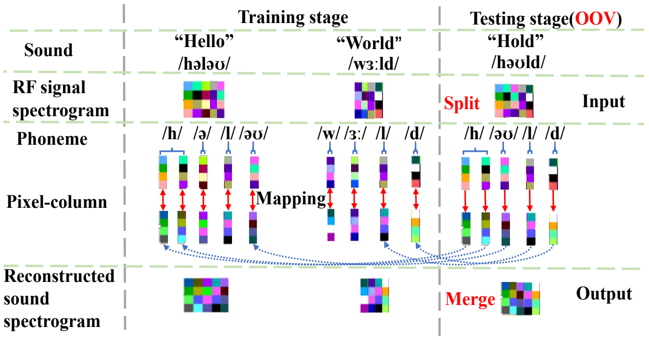


Figure 11: An example of reconstructing sound spectrograms from RF signal spectrograms for OOV words.

To train the SSR network, we define a loss function as

$$\mathcal{L}(\alpha, \beta; x, y) = \mathbb{E}\left[\log P_\beta(\hat{x} \mid \hat{z}, y)\right] - D_{KL}\left[Q_\alpha(\hat{z} \mid x) \| P(z)\right], \quad (7)$$

where $\mathbb{E}\left[\log P_\beta(\hat{x} \mid \hat{z}, y)\right]$ is the likelihood probability between the reconstructed sound spectrogram and recorded sound spectrogram, and $D_{KL}\left[Q_\alpha(\hat{z} \mid x) \| P(z)\right]$ represents the feature encode error between estimated feature distribution $\hat{z}$ and real feature distribution z. The loss function aims to minimize the encode error of probability distribution $Q_\alpha(\hat{z} \mid x)$ between the estimated feature distribution $\hat{z}$ and the real feature distribution z while maximize the likelihood function of probability distribution $P_\beta(\hat{x} \mid \hat{z}, y)$ between the recorded sound y and the reconstructed sound spectrogram $\hat{x}$. When the loss function converges, *RFSpy* trains the mapping relationship model between RF signal spectrograms and corresponding sound spectrograms.

6.2 Constructing a Phoneme-based Pixel-Column Mapping Relationship for Sound Spectrogram Reconstruction

To recover the complete conversation content, a straightforward method is to train a model, which is able to establish the mapping relationship between RF signal spectrograms and corresponding sound spectrograms for all possible words involved in online conversations. However, it is not realistic to train such a model, because the conversation content could contain arbitrary words. Hence, we consider training a model that can construct the mapping relationship between RF signal spectrograms and sound spectrograms for not only trained words but also untrained words, i.e., out-of-vocabulary (OOV) words.

To eavesdrop on conversations with OOV words, we construct a phoneme-based pixel-column mapping relationship. Since the normalized RF signal spectrogram has a size of 256 × 256 pixels containing M (i.e., 256) pixel-columns, the size of corresponding reconstructed sound spectrogram through the SSR network is also set as 256×256 pixels. Hence, *RFSpy* is able to construct the M pixel-column mapping relationship between RF signal spectrograms and reconstructed sound spectrograms one pixel-column to one pixel-column, which forms a M pairs pixel-column mapping relationship.

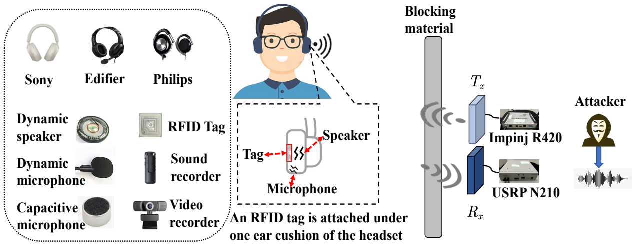
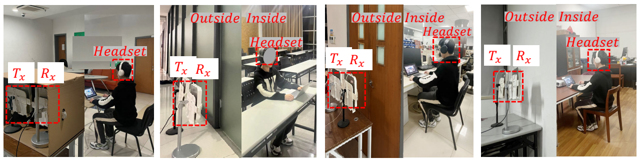


Figure 12: Illustration of experimental settings.



(a) Hall with cardboard (b) Meeting room with soundproof glass (c) Lab with wood door (d) Office with brick wall

Figure 13: Environments and blocking materials.

The M pairs pixel-column mapping relationship contains N non-repetitive pixel-column mapping relationship between RF signal spectrograms and sound spectrograms.

Usually, the pronunciation of each English word consists of several phonemes, and there are total of 48 phonemes in all English words [10]. Since each phoneme consists of limited sound spectrogram pixel-columns, the number of pixel-column categories is a finite constant for all 48 phonemes. Generally, a few hundred common English words contain all the word phonemes [10], which can be used to train a mapping relationship between RF signal spectrograms and corresponding phoneme spectrograms for all the 48 phonemes through the SSR network. Since the number of pixel-columns for different phonemes are different, *RFSpy* cannot accurately define the number of pixel-columns for each phoneme. Therefore, we consider utilizing the pixel-column as the mapping unit to reconstruct sound spectrograms from the corresponding RF signal spectrograms.

As shown in figure 11, assuming that *RFSpy* constructs the pixel-column mapping relationship between RF signal spectrograms and reconstructed sound spectrograms for words "hello" and "world" through the SSR network in the training stage. In the testing stage, *RFSpy* captures the RF signal spectrogram corresponding to an OOV word "hold", which is first split into pixel-columns and input to the phoneme-based pixel-column mapping relationship. Based on the constructed pixel-column mapping relationship for words "hello" and "world" in the training stage, *RFSpy* maps each pixel-column of RF signal spectrogram for OOV word "hold" to the corresponding sound spectrogram pixel-column. Finally, all the mapped sound spectrogram pixel-columns are merged as the reconstructed sound spectrogram for OOV word "hold".

7 RECOVERING CONVERSATION CONTENT

Based on the phoneme-based pixel-column mapping relationship, *RFSpy* is able to reconstruct sound spectrograms from corresponding RF signal spectrograms. Since the mapping unit of phoneme-based pixel-column mapping relationship is a pixel-column instead of a phoneme, *RFSpy* only gets sound spectrograms without knowing the actual sound phonemes. Hence, we consider converting sound spectrograms into time-domain sound waveforms for conversation content recovery.

To recover conversation content, *RFSpy* employs Griffin-Lim algorithm [20] to convert sound spectrograms to corresponding time-domain sound waveforms. Since there is no phase information in the sound spectrograms, Griffin-Lim algorithm is able to estimate the phase information of sound spectrograms by iterations for obtaining time-domain sound waveforms. Specifically, there are totally four steps in the Griffin-Lim algorithm. In the first step, *RFSpy* calculates amplitude matrix A_1 of a reconstructed sound spectrogram, and randomly initialize a phase matrix ϕ_1. In the second step, based on the amplitude matrix A_1 and phase matrix ϕ_1, *RFSpy* conducts Inverse Short-Time Fourier Transform (ISTFT) to obtain a time-domain sound waveform. In the third step, *RFSpy* conducts Short-Time Fourier Transform (STFT) for the time-domain sound waveform to obtain a new reconstructed sound spectrogram, during which the phase matrix ϕ_2 is saved and the amplitude matrix of new reconstructed sound spectrogram is replaced by A_1. In the forth step, *RFSpy* repeats the second and third step until the value of $\left|A_i - A_j\right|$ is smaller than a threshold τ. Based on the above method, *RFSpy* is able to convert sound spectrograms into corresponding time-domain sound waveforms. Finally, a Sound-to-Text API [9] is employed to convert the time-domain sound waveforms into conversation content.

8 EVALUATION

In this section, we conduct experiments to evaluate the performance of *RFSpy* in real environments.

8.1 Setup and Methodology

Environmental settings. Figure 12 shows the illustration of experimental settings. An Impinj Speedway R420 Commercial off-the-shelf (COTS) RFID reader is utilized to transmit RF signals and a USRP N210 with a SBX daughter-board is utilized to capture harmonic RF signals [25]. The COTS reader and USRP device utilize an antenna with the gain of $15dBi$ working at $920MHz$ and an antenna with the gain of $16dBi$ working at $2.4GHz$, respectively. The signal-transmitting antenna (T_x) continuously emits RF signals and the signal-receiving antenna (R_x) receives RF signals back-scattered by RFID tags. We conduct experiments in four environments with various blocking materials, i.e., a hall with cardboard, a meeting room with soundproof glass, a lab with wood door, and an office with brick wall, as shown in figure 13.

We choose several commonly used headsets in our experiments, i.e., Sony with dynamic speakers and capacitive microphones, Edifier with dynamic speakers and dynamic microphones, and Philips with dynamic speakers and capacitive microphones. Each headset is attached a $12mm \times 12mm$ in size, battery-free RFID tag under one ear cushion to avoid being noticed, which works at the frequency of about $900MHz$.

Data collection and implementation. We recruit 9 volunteers (4 males and 5 females, aged between 20 and 50 years old) to participate in the experiments. Each volunteer randomly selects a headset each time to talk online in an in-the-wild manner, i.e., volunteers

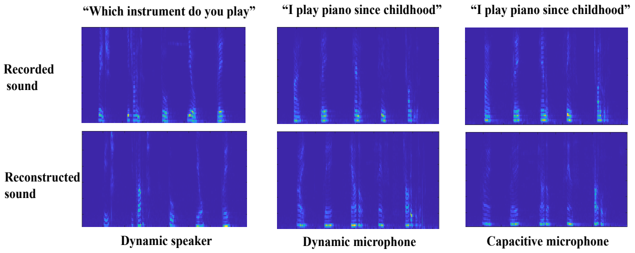


Figure 14: Examples of reconstructed sound spectrograms and recorded sound spectrograms.

perform daily activities and online conversations naturally, and they are not aware that their conversations are eavesdropped.

RFSpy chooses hundreds of common conversations from LibriTTS corpus[61] that cover all the phonemes in English as the training data, which involves about 2500 English words. To demonstrate the capability of our model to recover OOV words, we utilize other conversations involving about 1000 words to evaluate the sound eavesdropping performance of *RFSpy*, which contains trained and untrained words. The RF signals are processed by a computer with NVIDIA RTX 3090 Ti GPU, and the corresponding sound (i.e., ground truth) is recorded by sound recorders. RF signals and the corresponding sound are configured to start recording at the same time so that the two signals are aligned, and a video recorder is used to record the status of volunteers during online conversations.

We utilize 3-fold cross validation to evaluate the performance of *RFSpy*. Specifically, the 9 volunteers are divided into three groups, and each group contains three users. We choose data of two group volunteers to construct a conversation eavesdropping model, and another one group volunteers to evaluate the performance of *RFSpy*. The final experimental results are the average value of cross validation.

Evaluation metrics. To evaluate the performance of *RFSpy*, we define several metrics as follows.

- *Mel-Cepstral Distortion (MCD)*, which represents the similarity of reconstructed sound spectrograms and corresponding recorded sound spectrograms. A smaller/larger MCD represents a higher/lower similarity between the reconstructed sound and the recorded sound.
- *Word Error Rate (WER)*, which represents sound recognizability by sound recognition API. $WER = \frac{S+D+I}{N}$, where S, D, and I are word numbers substituted, deleted, and inserted, respectively, and N is word numbers in ground truth. A smaller/larger WER implies a higher/lower sound recognition accuracy.

8.2 Overall Performance

We evaluate the performance of *RFSpy* in reconstructing sound spectrograms. Figure 14 shows examples of reconstructed sound spectrograms and recorded sound spectrograms (i.e., ground truth) when a dynamic speaker produces sound "which instrument do you play", and a dynamic microphone and capacitive microphone receive sound "I play piano since childhood", respectively. The involved sound content contains not only trained words (i.e., which, do, you, play, I, since, childhood) but also OOV words (i.e., instrument, piano). It can be observed from the figure that the reconstructed and recorded sound spectrograms (i.e., ground truth) exhibit high similarity. MCD between the reconstructed and recorded sound spectrograms for the dynamic speaker, dynamic microphone, and capacitive microphone are 5.62, 6.64, and 6.82, respectively. Usually, a sound with a MCD below 8 is well-recognized by sound recognition API [54], and a smaller MCD represents a higher similarity between the reconstructed sound and recorded sound.

To validate the effectiveness of *RFSpy* for eavesdropping on conversations with OOV words, we evaluate the performance for the trained and untrained words, respectively. The top of figure 15 shows the average MCD of *RFSpy* for trained and untrained words. We can observe from the figure that the average MCD of dynamic speaker, dynamic microphone, and capacitive microphone for trained words are 5.06, 6.1, and 6.4, respectively, and that for untrained words are 6.28, 7.06, and 7.32, respectively. In addition, the bottom of figure 15 shows the WER of *RFSpy* for trained words and untrained words. From the figure, we can observe that the WER of dynamic speaker, dynamic microphone, and capacitive microphone for trained words are 15.8%, 16.84%, and 18.5%, respectively, and that for untrained words are 17.6%, 18.98%, and 19.6%, respectively. The performance differences between trained and untrained words are not obvious.

Moreover, the overall WER of *RFSpy* for both trained and untrained words of dynamic speaker, dynamic microphone, and capacitive microphone are 16.7%, 17.91%, and 19.05%, respectively, and the overall MCD for both trained and untrained words of dynamic speaker, dynamic microphone, and capacitive microphone are 5.67, 6.58, and 6.86, respectively. The above results show that *RFSpy* can eavesdrop on online conversations effectively.

8.3 Impact of Environments and Blocking Materials

We evaluate the sound eavesdropping performance of *RFSpy* in four different environments with various blocking materials, and the results are shown in figure 16. It can be observed from the figure that the performance difference of *RFSpy* in hall, meeting room, and lab with different blocking materials are not obvious. For example, the maximum differences of MCD between hall, meeting room, and lab for dynamic speaker, dynamic microphone, and capacitive microphone are only 0.96, 0.52, and 0.46, respectively, and the maximum differences of WER between hall, meeting room, and lab for dynamic speaker, dynamic microphone, and capacitive microphone are only 2.42%, 2.62%, and 2.88%, respectively. In contrast, *RFSpy* obtains a worse performance in the office, this is because RF signals undergo significant attenuation in brick wall. However, even in this environment, MCD are all lower than 7.5 and WER are all lower than 28% for both speaker and microphones. Hence, the above results show the ability of *RFSpy* to penetrate common blocking materials and cause threats to in-room conversations.

Usually, the hall and lab are accompanied with passers-by, so we explore the impact of passers-by on the performance of *RFSpy*. Figure 17 shows the sound eavesdropping performance of *RFSpy* in these two environments with/without passers-by. We can observe from the figure that the performance differences of *RFSpy* in both hall and lab with/without passers-by are not obvious. For

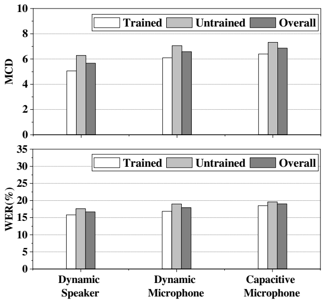


Figure 15: MCD and WER of *RFSpy* for trained and untrained words.

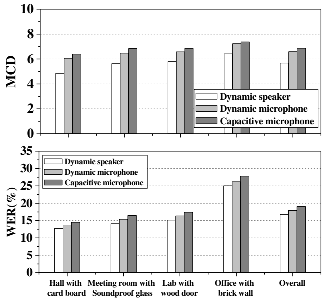


Figure 16: MCD and WER of *RFSpy* in different environments and blocking materials.

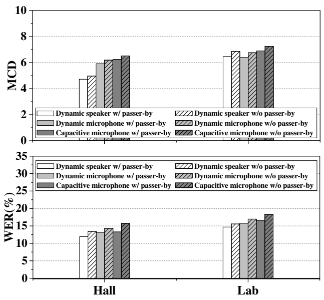


Figure 17: MCD and WER of *RFSpy* under environments with/without passers-by.

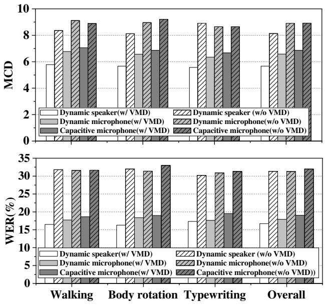


Figure 18: MCD and WER of *RFSpy* with/without VMD under different body motions.

example, MCD differences in hall with and without passer-by for dynamic speaker, dynamic microphone, and capacitive microphone are only 0.25, 0.27, and 0.26, respectively, and WER differences in hall with and without passer-by for dynamic speaker, dynamic microphone, and capacitive microphone are only 1.53%, 1.2%, and 2.6%, respectively. This is because *RFSpy* removes static environment and dynamic passers-by interference using spectral subtraction algorithm and VMD algorithm, respectively.

8.4 Comparison with Existing Works

We compare the sound eavesdropping performance of *RFSpy* with three state-of-the-art (SOTA) works, i.e., *RF-Mic* [7], *MagEar* [31], and *mmEve* [45]. Table 1 shows the comparison results of *RFSpy* with these three methods. From the table, we can observe that these three recent works enable eavesdropping on either human-speaking sound [7] or speaker-produced sound [31, 45] due to the limitations of employed sensors. In contrast, *RFSpy* is capable of eavesdropping on not only human-speaking sound but also speaker-produced sound. For eavesdropping on human-speaking sound, the WER of *RFSpy* is comparable to that of *RF-Mic* [7]. For eavesdropping on speaker-produced sound, *RFSpy* exhibits significantly lower WER compared to *Magear* [31] and similar WER as that of *mmEve* [45]. Therefore, *RFSpy* exhibits distinct advantages compared with existing works.

8.5 Impact of Body Motion

Usually, human body motions interfere with RF signals, so *RFSpy* utilizes VMD to remove the interference on RF signals. Figure 18 shows MCD and WER of *RFSpy* with/without VMD under different body motions, i.e., walking, body rotation on chair, and typewriting. We can observe from the figure that MCD and WER of *RFSpy* without VMD are larger than 8 and 30%, respectively, for both speaker and microphones, which is difficult to recover conversation content. In contrast, MCD and WER of *RFSpy* with VMD are smaller than 7.1 and 20%, respectively, for both speaker and microphones, which can recover conversation content effectively.

Table 1: A Comparison of *RFSpy* with Existing Works

Methods	Speaker-produced sound	WER of Speaker-produced	Human-speaking sound	WER of Human-speaking
RFMic[7]	✗	✗	✓	12.41%
MagEar[31]	✓	25.77%	✗	✗
mmEve[45]	✓	15.25 %	✗	✗
RFSpy	✓	17.48 %	✓	16.7 %

8.6 Impact of Distance

RF signals attenuate as propagating, so we explore the impact of distance between RFID antennas and tags on the perfomance of *RFSpy*. We select data when users are static, and face towards the RFID antenna under distance from $0.3m \sim 4.5m$ to evaluation the perfomance of *RFSpy*. Figure 19 shows the MCD and WER of *RFSpy* under different distances between RFID antennas and tags. We can observe from the figure that MCD and WER increases as the distance increases for both speaker and microphones. This is because the signal strength of received RF signals decreases as the distance increases. However, MCD and WER of *RFSpy* for the dynamic speaker, dynamic microphone, and capacitive microphone are all lower than 8 and 25% when the distance is within the range of $3.5m$. In general, *RFSpy* can be used for natural sound eavesdropping in various scenarios, such as the meeting room, offices, etc., within such a distance range.

8.7 Impact of User Orientation

To explore the impact of user orientation on the performance of *RFSpy*, we define an orientation as 0° when a user face toward the RFID antennas. As the user turn clockwise, his/her orientation is positive and increases sequentially. We conduct experiments when users are at a fixed position of $2m$ distance and orientation from 0° to 360° with 45° steps. Figure 20 shows the MCD and WER of *RFSpy* when users face toward different orientations. It can be observed from the figure that when users orientation is 90°, both the MCD and WER obtain the minimum value, this is because the RFID tag attached on headsets is towards the RFID antennas in this orientation. Moreover, both the MCD and WER obtain the maximum value under the orientation of 270°, this is because the side of headsets with an attached RFID tag back to the RFID antennas in this orientation, which results in RF signals being blocked by the human head. However, even under this orientation, the MCD are all lower than 8 and the WER are all lower than 25%,

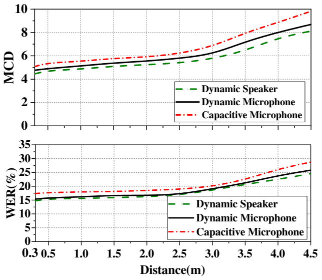


Figure 19: MCD and WER of *RFSpy* under different distances.

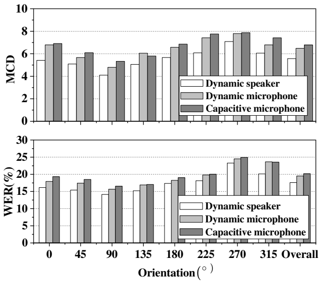


Figure 20: MCD and WER of *RFSpy* under different orientations.

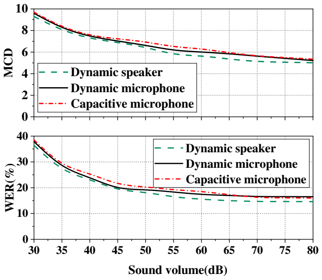


Figure 21: MCD and WER of *RFSpy* under different sound volumes.

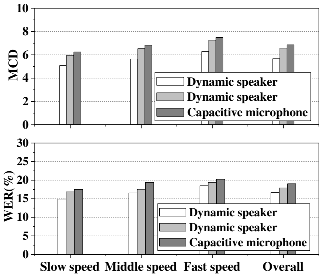


Figure 22: MCD and WER of *RFSpy* for different speech rates.

which demonstrates that *RFSpy* is applicable for most conversation eavesdropping scenarios.

8.8 Impact of Sound Volume and Speech Rate

Sound Volume. In general, a higher sound volume brings a larger magnitude of the metal coil vibration of speakers/microphones, so we evaluate the performance of *RFSpy* under different sound volumes. We record the microphone-received and speaker-produced sound volumes using a decibel meter, which is inserted under the ear cushion of headsets. Figure 21 shows the MCD and WER of *RFSpy* under different sound volumes. It can be seen from the figure that higher sound volumes lead to a lower MCD and WER. When the sound volume is large than 40dB, the MCD and WER of dynamic speaker, dynamic microphone, and capacitive microphone are all lower than 8 and 25%, respectively. Usually, speaker-produced/microphone-received sound volumes are larger than 45dB during use of headsets, so *RFSpy* is suitable for most of the application scenarios.

Speech Rate. In our experiments, the recruited volunteers speak at their natural speech rates. We divide users' speech rates into three categories, i.e., slow speed (<100 words/min), middle speed (100-150 words/min), and fast speed (>150 words/min) according to [18], and explore the impact of speech rates on the performance of *RFSpy*. Figure 22 shows the MCD and WER of *RFSpy* for different speech rates. It can be observed from the figure that the MCD and WER increases slightly as the speech rates increase for both speaker and microphones. However, the MCD and WER of *RFSpy* for the dynamic speaker, dynamic microphone, and capacitive microphone are all lower than 8 and 21%, respectively. Hence, *RFSpy* is rubust to various human speech rates.

8.9 Impact of Headset Category and Tag Position

Headset Category. We explore the impact of three categories of headsets, i.e., Sony, Edifier, and Philips, on the performance of *RFSpy*, and the results are shown in figure 23. From the figure, we can observe that the performance difference for different headset categories is not obvious. For example, the maximum differences of MCD between Sony, Edifier, and Philips for speaker and microphone are only 0.69 and 1.13, respectively, and the maximum differences of WER between Sony, Edifier, and Philips for speaker and microphone are only 1.83% and 3.09%, respectively.

Tag Position. In our experiments, RFID tags are attached at random positions under headset ear cushion. To explore the impact of tag positions on *RFSpy*, we group tag positions into three areas based on relative positions under ear cushion (i.e., left area, middle area, right area) and evaluate the performance of *RFSpy* under various tag positions. Figure 23 shows the sound eavesdropping performance under different tag positions. It can be seen from the figure that the performance difference of *RFSpy* for different tag positions are not obvious. For example, the maximum differences of MCD between left, middle, and right areas for speaker of Sony is only 0.36, and the maximum differences of WER between left, middle, and right areas for speaker of Sony is only 0.64%.

8.10 Impact of AER Algorithm

To validate the effectiveness of AER algorithm in removing echo interference, we compare the sound eavesdropping performance of *RFSpy* with/without AER algorithm, and the results are shown in Figure 24. It can be observed from the figure that MCD of dynamic speaker, dynamic microphone, and capacitive microphone with AER algorithm are all smaller than 7 while that without AER algorithm are all larger than 8. Usually, it is hard for sound recognition APIs to recognize a sound when MCD of the sound is larger than 8. Moreover, the WER of dynamic speaker, dynamic microphone, and capacitive microphone are all decreased from exceeding 24% to below 20% after using AER algorithm. The above results demonstrate AER algorithm's efficacy in removing interference and improving sound eavesdropping performance.

9 COUNTERMEASURE AND LIMITATION

Countermeasures. One potential way to defend *RFSpy* is to use RF signals shielding material in the headset shell. The RF signal shielding material usually consists of high-permeability metals, such as iron, silicon steel, or permalloy, etc. These materials suppress the propagation of RF signals, which can be used to defend *RFSpy*. Moreover, electromagnetic interference devices, such as RF devices working at the same frequency band as *RFSpy*, could cause RF signal distortion. Hence, in places with high security requirements, such as military missions, trade negotiations, etc., electromagnetic interference devices can be placed around to defend *RFSpy*.

Limitations. *(1) Micro-electro mechanical system (MEMS) microphone.* In this paper, we discuss two types of microphones, i.e., dynamic microphones and capacitive microphones. Besides, there

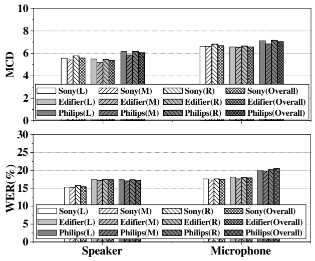


Figure 23: MCD and WER of *RFSpy* for different headset categories and tag positions.

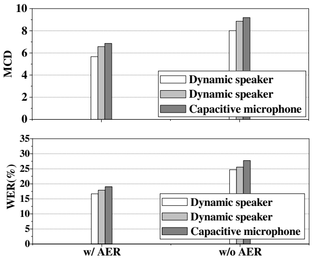


Figure 24: MCD and WER of *RFSpy* with/without AER algorithm.

also exists a type of MEMS microphones. The size of MEMS microphones is very small (about a few millimeters), so it is difficult for *RFSpy* to sense the metal coil vibration of MEMS microphones for eavesdropping. Fortunately, the dynamic microphone and capacitive microphone make up main headset microphone market.

(2) Small earphone eavesdropping. Since current mainstream RFID tags are about ten millimeters in size, our work only explores eavesdropping on online conversations through larger headsets. However, benefiting from the rapid development of RFID technologies, the size of RFID tags is becoming smaller and smaller, e.g., RFID tags based on the latest technologies are only 2 × 3mm in size[36]. Hence, our work also enables online conversation eavesdropping for smaller earphones by using the small RFID tags.

(3) Near-end user and far-end user speak simultaneously. RFSpy uses only one RFID tag to alternately sense the metal coil vibration of headset speakers and microphones for online conversation eavesdropping. When near-end user and far-end user speak simultaneously, the metal coil of headset speakers and microphones vibrate at the same time, and *RFSpy* cannot distinguish RF signals corresponding to the speaker or microphone.

(4) Word connection pronunciation and Different languages. In English, word connections is a common pronunciation habit. Since there are English connection reading rules, *RFSpy* can collect more samples with word connection for constructing a robust conversation eavesdropping model. Besides, we design *RFSpy* system based on standard English. Except for English, there exist many other languages, such as Chinese, Japanese, French, etc. Since the pronunciation and phoneme composition of different languages are different, we regard exploring eavesdropping on other languages as our future works.

10 RELATED WORK

In this section, we review works related to *RFSpy*.

Sensing sound through non-acoustic sensors. With the development of Internet of Things (IoT) technology, studies about sound sensing through non-acoustic sensors receive lots of attention. On one hand, some researchers use various sensors, such as accelerator [1, 6, 24, 39, 40, 62, 63], gyroscope [34], lidar [38], vibration sensor[33], and customized magnetic sensor [31], to capture smart device vibration for sensing speaker-produced sound [1, 24, 31, 34, 38, 62, 63] or various facial dynamics to sense human-speaking sound [6, 33, 39, 40, 64]. On the other hand, some works use RF signals, such as Wi-Fi [48], mmWave [3, 16, 22, 23, 43, 44, 51, 59], and RFID [7, 30, 46], to capture surrounding object vibration for sensing speaker-produced sound [3, 22, 23, 30, 44–46, 48] or capture body motions for sensing human-speaking sound [7, 16, 43, 51]. However, due to the limitations of employed sensors, although these methods enable sensing either speaker-produced sound or human-speaking sound, they fail to eavesdrop on both speaker-produced and human-speaking sound simultaneously.

Sensing sound through earphones/headsets. Recently, privacy security issues have been widely explored [6, 27, 31, 52, 60]. Some works explore achieving sound sensing through earphones/headsets equipped with various sensors[6, 31]. For example, *MagEar*[31] uses a customized coil to capture magnetic signals leaked by a microspeaker to sense speaker-produced sound. *EarSpy*[6] uses accelerometer on earphones/headsets to sense human mouth motions and vocal cord vibration during speaking for eavesdropping on human-speaking sound. However, liking non-acoustic sensor-based sound sensing methods, works [6, 31] still fail to eavesdrop on both speaker-produced and human-speaking sound through only one sensor. Moreover, works [6, 31] implement sound eavesdropping only for trained words, but cannot recover online conversations with OOV words.

RFID-based Sensing. The rapid development of Internet of Things (IoT) technology has led to many applications [2, 8, 17, 28, 29, 47, 49, 50, 53, 55, 56, 58, 65]. Moreover, due to the advantages of low cost and deployment flexibility, RFID tags have been widely utilized in various wireless sensing applications, such as localization [49], user authentication [8, 17], and vital signs monitoring [47], etc. Meanwhile, due to the excellent sensing ability of RF signals, recent years have witnessed the rapid development of RFID-based vibration sensing [29, 50, 55, 56, 58], which are generally utilized to sense mechanical vibrations. However, no existing works are available to realize online conversation eavesdropping through RFID sensing.

11 CONCLUSION

In this paper, we propose an online conversation eavesdropping system, *RFSpy*, which utilizes common headsets with an attached RFID tag to eavesdrop on both speaker-produced and microphone-received sound. *RFSpy* first leverages a VMD algorithm and an AER algorithm to remove human body motion interference and echo interference from the received RF signals, respectively. Then, we use a designed SSR network to train a mapping relationship model between RF signal spectrograms and sound spectrograms, and further construct a phoneme-based pixel-column mapping relationship to reconstruct sound spectrograms for not only trained words but also untrained (OOV) words. Afterward, *RFSpy* employs Griffin-Lim algorithm to convert the sound spectrograms to time-domain sound waveforms, which can be recovered by sound recognition API to sound content. Experiments in real environments demonstrate the effectiveness of *RFSpy* for eavesdropping on online conversations with OOV words.

ACKNOWLEDGMENTS

This research was sponsored by National Natural Science Foundation of China (No. 62172277).

REFERENCES

[1] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer. In *Proceedings of NDSS'20*. San Diego, USA.

[2] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. 2020. Acoustic-based sensing and applications: A survey. *Computer Networks* 181 (2020), 107447.

[3] Suryoday Basak and Mahanth Gowda. 2022. mmspy: Spying phone calls using mmwave radars. In *Proceedings of IEEE S&P'22*. San Francisco, USA.

[4] Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* 27, 2 (1979), 113–120.

[5] BUGHUNTER. 2023. Microphone jammer BugHunter DAudio bda-3 Voices with 7 ultrasonic transducers. [Online]. Available: https://www.amazon.com/BDA-3-Control-Microphone-Suppressor-Recording/dp/B089WH2B9F.

[6] Yetong Cao, Fan Li, Huijie Chen, Xiaochen Liu, Chunhui Duan, and Yu Wang. 2023. I Can Hear You Without a Microphone: Live Speech Eavesdropping From Earphone Motion Sensors. In *Proceedings of IEEE INFOCOM'23*. New York, USA.

[7] Yunzhong Chen, Jiadi Yu, Linghe Kong, Hao Kong, Yanmin Zhu, and Yi-Chao Chen. 2023. RF-Mic: Live Voice Eavesdropping via Capturing Subtle Facial Speech Dynamics Leveraging RFID. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–25.

[8] Yunzhong Chen, Jiadi Yu, Linghe Kong, Yanmin Zhu, and Feilong Tang. 2022. RFPass: Towards environment-independent gait-based user authentication leveraging RFID. In *Proceedings of IEEE SECON'22*. Virtual Conference.

[9] Google cloud. 2023. Cloud speech-to-text documentation. [Online]. Available: https://cloud.google.com/speech-to-text.

[10] Antonie Cohen. 1965. *The phonemes of English*. Springer.

[11] Israel Cohen, Yiteng Huang, Jingdong Chen, and Jacob Benesty. 2009. Pearson correlation coefficient. *Noise reduction in speech processing* (2009), 1–4.

[12] M Dobhn Daniel et al. 2008. The rf in rfid passive uhf rfid in practice. In *Elsevier*.

[13] R Dauman. 2013. Bone conduction: an explanation for this phenomenon comprising complex mechanisms. *European annals of otorhinolaryngology, head and neck diseases* 130, 4 (2013), 209–213.

[14] Kevin D'hoe, Van Nieuwenhuyse, et al. 2009. Influence of different types of metal plates on a high frequency RFID loop antenna: study and design. *Advances in electrical and computer engineering* 9, 2 (2009), 3–8.

[15] Konstantin Dragomiretskiy and Dominique Zosso. 2013. Variational mode decomposition. *IEEE transactions on signal processing* 62, 3 (2013), 531–544.

[16] Long Fan, Lei Xie, Xinran Lu, Yi Li, Chuyu Wang, and Sanglu Lu. 2023. mmMIC: Multi-modal Speech Recognition based on mmWave Radar. In *Proceedings of IEEE INFOCOM'23*. New York, USA.

[17] Chao Feng, Jie Xiong, Liqiong Chang, Fuwei Wang, Ju Wang, and Dingyi Fang. 2021. Rf-identity: Non-intrusive person identification based on commodity rfid devices. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5. 1–23.

[18] Thomas Gay. 1981. Mechanisms in the control of speech rate. *Phonetica* 38, 1-3 (1981), 148–158.

[19] Dedre Gentner and Arthur B Markman. 2006. Defining structural similarity. *The Journal of Cognitive Science* 6, 1 (2006), 1–20.

[20] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing* 32, 2 (1984), 236–243.

[21] Yejun He and Zhengzheng Pan. 2013. Design of UHF RFID broadband anti-metal tag antenna applied on surface of metallic objects. In *Proceedings of IEEE WCNC'13*. Glasgow, UK.

[22] Pengfei Hu, Wenhao Li, Riccardo Spolaor, and Xiuzhen Cheng. 2023. mmecho: A mmwave-based acoustic eavesdropping method. In *Proceedings of IEEE S&P'22*. Delft, Netherlands.

[23] Pengfei Hu, Yifan Ma, Panneer Selvam Santhalingam, Parth H Pathak, and Xiuzhen Cheng. 2022. Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary. In *Proceedings of IEEE INFOCOM'22*. Virtual Conference.

[24] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth Pathak, Guoming Zhang, and Xiuzhen Cheng. 2022. Accelerometer acoustic eavesdropping with unconstrained vocabulary. In *Proceedings of IEEE S&P*. San Francisco, USA.

[25] F. Mavromatis N. Kargas and A. Bletsas. 2019. USRP reader. [Online]. Available: https://github.com/nkargas/Gen2-UHF-RFID-Reader.

[26] Jung Ho Kim, Jim Tai Kim, Jin O Kim, and Jin Ki Min. 1997. Acoustic characteristics of a loudspeaker obtained by vibration and acoustic analyses. *WIT Transactions on The Built Environment* 28 (1997), 181–190.

[27] Hao Kong, Li Lu, Jiadi Yu, Yingying Chen, and Feilong Tang. 2020. Continuous authentication through finger gesture interaction for smart homes using WiFi. *IEEE Transactions on Mobile Computing* 20, 11 (2020), 3148–3162.

[28] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3track: mmWave-based Multi-user 3D Posture Tracking. In *Proc. of ACM MobiSys*. Portland, OR, USA.

[29] Ping Li, Zhenlin An, Lei Yang, and Panlong Yang. 2019. Towards physical-layer vibration sensing with rfids. In *Proceedings of IEEE INFOCOM'19*. Paris, France.

[30] Ping Li, Zhenlin An, Lei Yang, Panlong Yang, and QiongZheng Lin. 2019. RFID harmonic for vibration sensing. *IEEE Transactions on Mobile Computing* 20, 4 (2019), 1614–1626.

[31] Qianru Liao, Yongzhi Huang, Yandao Huang, Yuheng Zhong, Huitong Jin, and Kaishun Wu. 2022. MagEar: eavesdropping via audio recovery using magnetic side channel. In *Proceedings of ACM MobiSys'22*. Portland, Oregon.

[32] Weifeng Liu, Puskal P Pokharel, and Jose C Principe. 2008. The kernel least-mean-square algorithm. *IEEE Transactions on signal processing* 56, 2 (2008), 543–554.

[33] Héctor A Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Marco Beltman, Lama Nachman, and Hong Lu. 2018. V-Speech: noise-robust speech capturing glasses using vibration sensors. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2. 1–23.

[34] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing speech from gyroscope signals. In *Proceedings of USENIX'14*. San Diego, USA.

[35] Philip McCord Morse, Acoustical Society of America, and American Institute of Physics. 1948. *Vibration and sound*. McGraw-Hill New York.

[36] rfidhy. 2022. The Smallest RFID Tag as Thin as Sand. [Online]. Available: https://www.rfidhy.com/the-smallest-rfid-tag-as-thin-as-sand/.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE CVPR'22*. New Orleans, USA.

[38] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with your robot vacuum cleaner: eavesdropping via lidar sensors. In *Proceedings of ACM SenSys'20*. Yokohama, Japan.

[39] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. Face-Mic: inferring live speech and speaker identity via subtle facial dynamics captured by AR/VR motion sensors. In *Proceedings of ACM MobiCom '21*. New Orleans, USA.

[40] Weigao Su, Daibo Liu, Taiyuan Zhang, and Hongbo Jiang. 2021. Towards device independent eavesdropping on telephone conversations with built-in accelerometer. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5. 1–29.

[41] Richard EA van Emmerik and Erwin EH van Wegen. 2000. On variability and stability in human movement. *Journal of Applied Biomechanics* 16, 4 (2000), 394–406.

[42] Gianfranco Andía Vera, Yvan Duroc, and Smail Tedjini. 2013. Analysis of harmonics in UHF RFID signals. *IEEE Transactions on Microwave Theory and Techniques* 61, 6 (2013), 2481–2490.

[43] Chao Wang, Feng Lin, Zhongjie Ba, Fan Zhang, Wenyao Xu, and Kui Ren. 2022. Wavesdropper: Through-wall Word Detection of Human Speech via Commercial mmWave Devices. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 6. 1–26.

[44] Chao Wang, Feng Lin, Tiantian Liu, Ziwei Liu, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. 2022. mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect. In *Proceedings of IEEE INFOCOM'22*. Virtual Conference.

[45] Chao Wang, Feng Lin, Tiantian Liu, Kaidi Zheng, Zhibo Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Kui Ren. 2022. mmEve: eavesdropping on smartphone's earpiece via COTS mmWave device. In *Proceedings of ACM MobiCom'22*. Sydney, Australia.

[46] Chuyu Wang, Lei Xie, Yuancan Lin, Wei Wang, Yingying Chen, Yanling Bu, Kai Zhang, and Sanglu Lu. 2021. Thru-the-wall eavesdropping on loudspeakers via RFID by capturing sub-mm level vibration. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 5. 1–25.

[47] Chuyu Wang, Lei Xie, Wei Wang, Yingying Chen, Yanling Bu, and Sanglu Lu. 2018. Rf-ecg: Heart rate variability assessment based on cots rfid tag array. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2. 1–26.

[48] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of ACM MobiCom'15*. Paris, France.

[49] Fu Xiao, Zhongqin Wang, Ning Ye, Ruchuan Wang, and Xiang-Yang Li. 2017. One more tag enables fine-grained RFID localization and tracking. *IEEE/ACM Transactions on Networking* 26, 1 (2017), 161–174.

[50] Binbin Xie, Jie Xiong, Xiaojiang Chen, and Dingyi Fang. 2020. Exploring commodity rfid for contactless sub-millimeter vibration sensing. In *Proceedings of ACM SenSys'20*. Yokohama, Japan.

[51] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of ACM MobiSys'19*. Seoul, South Korea.

[52] Xiangyu Xu, Jiadi Yu, Yingying Chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: Towards Behavior-irrelevant On-touch User Authentication on Smartphones Leveraging Vibrations. In *Proc. of ACM MobiCom*. London, UK.

[53] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. BreathListener: Fine-grained Breathing Monitoring in Driving Environments Utilizing Acoustic Signals. In *Proc. of ACM MobiSys*. Seoul, South Korea.

[54] Chen Yan, Guoming Zhang, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2019. The feasibility of injecting inaudible voice commands to voice assistants. *IEEE Transactions on Dependable and Secure Computing* 18, 3 (2019), 1108–1124.

[55] Lei Yang, Yao Li, Qiongzheng Lin, Huanyu Jia, Xiang-Yang Li, and Yunhao Liu. 2017. Tagbeat: Sensing mechanical vibration period with cots rfid systems. *IEEE/ACM transactions on networking* 25, 6 (2017), 3823–3835.

[56] Lei Yang, Yao Li, Qiongzheng Lin, Xiang-Yang Li, and Yunhao Liu. 2016. Making sense of mechanical vibration period with sub-millisecond accuracy using backscatter signals. In *Proceedings of ACM MobiCom'16*. New York, USA.

[57] Lei Yang, Qiongzheng Lin, Xiangyang Li, Tianci Liu, and Yunhao Liu. 2015. See through walls with COTS RFID system. In *Proceedings of ACM Mobicom'15*. Paris, France.

[58] Panlong Yang, Yuanhao Feng, Jie Xiong, Ziyang Chen, and Xiang-Yang Li. 2020. Rf-ear: Contactless multi-device vibration sensing and identification using cots rfid. In *Proceedings of IEEE INFOCOM'20*. Virtual Conference.

[59] Chuyu Wang Lei Xie Jingyi Ning Yiwen Feng, Kai Zhang and Shijia Chen. 2023. mmEavesdropper: Signal Augmentation-based Directional Eavesdropping with mmWave Radar. In *Proceedings of IEEE INFOCOM'23*. New York, USA.

[60] Jiadi Yu, Li Lu, Yingying Chen, Yanmin Zhu, and Linghe Kong. 2019. An indirect eavesdropping attack of keystrokes on touch screen through acoustic sensing. *IEEE Transactions on Mobile Computing* 20, 2 (2019), 337–351.

[61] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proceedings of IEEE ISCA'19*. ISCA, Graz, Austria.

[62] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of ACM MobiSys'15*. Florence, Italy.

[63] Shijia Zhang, Yilin Liu, and Mahanth Gowda. 2023. I Spy You: Eavesdropping Continuous Speech on Smartphones via Motion Sensors. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 6. 1–31.

[64] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.

[65] Yanmin Zhu, Ruobing Jiang, Jiadi Yu, Zhi Li, and Minglu Li. 2014. Geographic routing based on predictive locations in vehicular ad hoc networks. *EURASIP Journal on Wireless Communications and Networking* 2014 (2014), 1–9.