Thru-the-wall Eavesdropping on Loudspeakers via RFID by Capturing Sub-mm Level Vibration

CHUYU WANG, State Key Laboratory for Novel Software Technology, Nanjing University, China
LEI XIE*, State Key Laboratory for Novel Software Technology, Nanjing University, China
YUANCAN LIN, State Key Laboratory for Novel Software Technology, Nanjing University, China
WEI WANG, State Key Laboratory for Novel Software Technology, Nanjing University, China
YINGYING CHEN, Electrical and Computer Engineering, Rutgers University, USA
YANLING BU, State Key Laboratory for Novel Software Technology, Nanjing University, China
KAI ZHANG, State Key Laboratory for Novel Software Technology, Nanjing University, China
SANGLU LU, State Key Laboratory for Novel Software Technology, Nanjing University, China

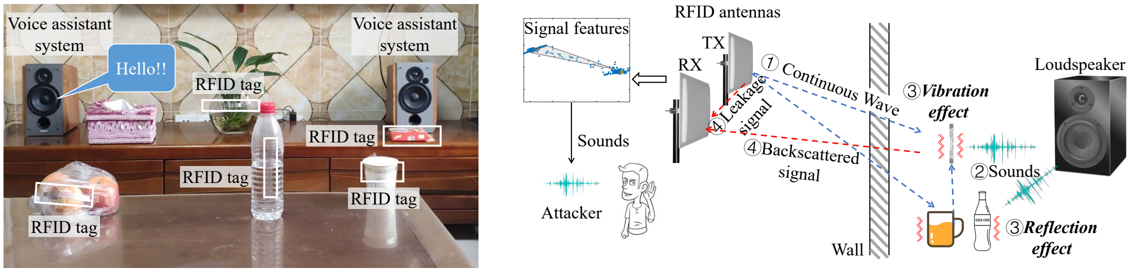
The unprecedented success of speech recognition methods has stimulated the wide usage of intelligent audio systems, which provides new attack opportunities for stealing the user privacy through eavesdropping on the loudspeakers. Effective eavesdropping methods employ a high-speed camera, relying on LOS to measure object vibrations, or utilize WiFi MIMO antenna array, requiring to eavesdrop in quiet environments. In this paper, we explore the possibility of eavesdropping on the loudspeaker based on COTS RFID tags, which are prevalently deployed in many corners of our daily lives. We propose *Tag-Bug* that focuses on the human voice with complex frequency bands and performs the thru-the-wall eavesdropping on the loudspeaker by capturing sub-mm level vibration. *Tag-Bug* extracts sound characteristics through two means: (1) *Vibration effect*, where a tag directly vibrates caused by sounds; (2) *Reflection effect*, where a tag does not vibrate but senses the reflection signals from nearby vibrating objects. To amplify the influence of vibration signals, we design a new signal feature referred as Modulated Signal Difference (MSD) to reconstruct the sound from RF-signals. To improve the quality of the reconstructed sound for human voice recognition, we apply a Conditional Generative Adversarial Network (CGAN) to recover the full-frequency band from the partial-frequency band of the reconstructed sound. Extensive experiments on the USRP platform show that *Tag-Bug* can successfully capture the monotone sound when the loudness is larger than 60dB. *Tag-Bug* can efficiently recognize the numbers of human voice with 95.3%, 85.3% and 87.5% precision in the free-space eavesdropping, thru-the-brick-wall eavesdropping and thru-the-insulating-glass eavesdropping, respectively. *Tag-Bug* can also accurately recognize the letters with 87% precision in the free-space eavesdropping.

CCS Concepts: • **Networks → Cyber-physical networks**; • **Security and privacy → Mobile and wireless security**.

*Lei Xie is the corresponding author.

Authors' addresses: Chuyu Wang, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, chuyu@nju.edu.cn; Lei Xie, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, lxie@nju.edu.cn; Yuancan Lin, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, yclin@smail.nju.edu.cn; Wei Wang, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, ww@nju.edu.cn; Yingying Chen, Electrical and Computer Engineering, Rutgers University, New Brunswick, USA, yingche@scarletmail.rutgers.edu; Yanling Bu, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, yanling@smail.nju.edu.cn; Kai Zhang, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, mg1933091@smail.nju.edu.cn; Sanglu Lu, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, sanglu@nju.edu.cn.

2474-9567/2021/0-ART0 $15.00
https://doi.org/00.0000/0000000.1122456



(a) Application scenario

(b) *Tag-Bug*: Acoustic thru-the-wall eavesdropping

Fig. 1. Thru-the-wall eavesdropping via RFID tags.

Additional Key Words and Phrases: Eavesdropping, RFID, Sub-mm Level Vibration

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 0 (2021), 25 pages. https://doi.org/00.0000/0000000.1122456

1 INTRODUCTION

Acoustic eavesdropping is one of the most significant security concerns, as the voice communication between people is an unencrypted transmission channel, making it easy to obtain the sensitive information. Traditional acoustic eavesdropping methods, which employ hidden or tampered microphones [8, 23], can be prevented by using soundproof insulation. Due to such insulation, the user may involuntarily neglect the acoustic eavesdropping in such scenario, making the loudspeaker a potential threat for eavesdropping. Particularly, benefiting from the unprecedented success of the advancement in speech recognition, the intelligent audio systems have been widely integrated into our daily life, which largely extends the usage of loudspeakers and brings new attack opportunities. For example, Google Home may replay the passwords, when the 'Remember' function is activated to record the private information by the user. Then, private information, *e.g.*, daily schedule, passwords and even life style, may be leaked. Another example is that online meetings during COVID-19 bring great convenience to many companies and employees when working from home. However, all these meetings involve the usage of loudspeakers heavily, which may lead to severe personal and corporate proprietary information leakage.

Due to its severe consequences, there have been active research efforts on eavesdropping of loudspeakers. Davis *et al.* leverage a high-speed camera to capture the vibrations of objects (*e.g.*, a glass of water or a potted plant) caused by the loudspeaker to perceive the sound [10], which relies on the existence of line-of-sight communication. Sensors such as gyroscopes embedded in a smartphone have also been exploited to capture the sound from the loudspeaker [26]. This approach works through the common medium with the loudspeaker and does not work for the thru-the-wall eavesdropping. It is also limited by the battery power of mobile devices. ART eavesdropper uses wireless signals to perceive the vibration of the loudspeaker diaphragm based on a specific MIMO antenna array [37]. This solution incurs hardware (*i.e.*, MIMO antenna array) with relatively high cost and works mostly in quiet environments. Any nearby vibrations, *e.g.*, a spinning fan, can affect the receiving signal. Some advanced work has shown that Ultra High Frequency (UHF) RFID tags can capture tiny vibrations. TagSound [20] perceives the mono-tone sound vibration by using harmonic signals, and others [40, 41] capture the ambient vibrations based on the phase variation by using the compressive sensing. However, the harmonic signals are too weak to perform the thru-the-wall eavesdropping, and the compressive sensing cannot be used to extract the human voice with none-sparse frequency bands.

In this paper, we explore the possibility of eavesdropping the human voice played by the loudspeaker based on the surrounding COTS RFID tags, which could be attached on many everyday objects as shown in Figure 1(a).

On one hand, many daily products from online purchasing, such as water bottles, delivery packages, hang tags, envelopes books, etc. , come with RFID tags. It greatly improves the chances of RFID tags appearing in our lives, and makes the tags easily overlooked. On the other hand, the adversary can even intentionally hide the battery-less and light-weighted RFID tags beside the loudspeaker, *e.g.*, under the table, which is hard to be detected and is able to eavesdrop in a long term. As shown in Figure 1(b), we develop *Tag-Bug*, an effective system to perform the thru-the-wall eavesdropping on the loudspeaker based on the received physical-layer signals. Similar to the previous attacks [6, 17, 26], we consider the loudspeaker as the sound source, which is widely used in a voice assistant system, *e.g.*, Google Home and Amazon Alexa, rather than the live human speech. The reason is that the live human speech mainly leads to the air flow from the mouth with small vibration of vocal cords, while the loudspeaker mainly leads to the diaphragm vibration. Thus, the human speech can be drowned by the vibration due to the air flow in the extracted sound. In particular, *Tag-Bug* can extract the sound from loudspeaker through two ways: (1) *Vibration effect*, the tag directly vibrates caused by sounds, *e.g.*, the tag vibrates directly due to the playing sounds when attached on the delivery package. (2) *Reflection effect*, the tag does not vibrate but senses the reflection signals from nearby vibrating objects due to the sound, *e.g.*, the tag captures the reflection signal from a cup of water, which vibrates due to the playing sounds. To extract the tiny vibration of the sound, we build a model to decompose the received signals and extract the Modulated Signal Difference (MSD) as the vibration indicator. Since the RFID tag is more sensitive to the low-frequency sound due to the larger sound energy, we leverage a Conditional Generative Adversarial Network (CGAN) to recover the high-frequency band by referring to the low-frequency band, so as to improve the quality of recovered human voice.

There are three main challenges in performing the eavesdropping via RFID tags. *The first challenge is to detect the sub-mm level vibration caused by the sound.* Traditionally, the vibration of the loudspeaker diaphragm is usually smaller than 1mm [16]. However, such tiny vibration results in the phase change below 0.04 radians, which is close to the noise level [39]. To address this challenge, we build a transmitting model and extract amplified vibration features from the received signal. Particularly, we extract the *Modulated Signal Difference (MSD)* as the difference of signals between the ON and OFF modulation states. The phase change of MSD indicates the tag displacement due to the vibration. Furthermore, we propose the amplified MSD by subtracting the average signal of OFF states in a time window. The amplified MSD can extract the sound from either the *vibration effect* or the *reflection effect*. In this way, *Tag-Bug* can extract the sub-mm level vibration, when either the tag itself or the nearby object vibrates caused by the sound wave.

The second challenge is to reduce the interference of the periodic commands sent by the RFID reader. In RFID systems, the periodic reader signal, *e.g.*, the `QUERY` and `ACK` commands, is much stronger than the backscattered signal from the tag. Even if the reader signal does not overlap with the tag signal in the time domain, the periodic reader signal will lead to the large noise in the frequency band when received by the antenna. To address this challenge, we randomize the tag response mechanism based on the C1G2 protocol. In particular, we randomly set the frame-size of each query cycle and let the tag randomly retransmit the `EPC` command. Then, the noise due to the periodic commands can be significantly reduced.

The third challenge is to refine the recovered human voice extracted from the amplified MSD. Human voice is the main target of the concerns during the eavesdropping. However, limited by the inherent material characteristics of RFID tags, the signals of high-frequency bands are very weak in the extracted sound from the amplified MSD, so the recovered sound is unclear for recognition. To address this challenge, we investigate the correlation of signals with different frequencies, and find that high-frequency signals are usually harmonic of low-frequency signals. To efficiently capture the correlation among different frequency bands, we develop a CGAN to recover the full-frequency band by referring to multiple low-frequencies. In this way, the refined sound has more comprehensive frequency band, and could be recognized more accurately.

This paper makes three contributions. First, we show the possibility of using low-cost and easily-overlooked RFID tags to effectively perform the thru-the-wall eavesdropping, pushing the limit of RFID sensing capability to the sub-mm level. Particularly, *Tag-Bug* can extract the sound vibration either from the *vibration effect* or the *reflection effect*, improving the applicability of our system. Second, we build a signal transmitting model to extract the vibration from the amplified *Modulated Signal Difference (MSD)* by removing the strong interference. A CGAN based method is designed to improve the quality of the recovered human voice. Third, we implemented our system *Tag-Bug* on the USRP platform. Real-world experiments show that *Tag-Bug* can successfully capture the monotone sound when the loudness is larger than 60dB. *Tag-Bug* can efficiently recognize the numbers of human voice with 95.3%, 85.3% and 87.5% precision in the free-space eavesdropping, thru-the-brick-wall eavesdropping and thru-the-insulating-glass eavesdropping, respectively. *Tag-Bug* can also accurately recognize the letters with 87% precision in the free-space eavesdropping.

2 PROBLEM FORMULATION

In this paper, we consider the novel problem of launching the side-channel eavesdropping on the *loudspeaker* by leveraging the vibration of ambient RFID tag due to the sound. Our attack mainly focuses on the sound played by the loudspeaker, rather than the voice of live human speech, because the live human speech mainly leads to the air flow instead of the air vibration due to the sound. As a result, the vibration extracted from the tag signal is related to the air flow, instead of the human voice. In this paper, we use the USRP platform to extract the sound due to the convenient access to the physical-layer signal.

2.1 Attack Model

We assume a victim user with a loudspeaker and some surrounding objects, which are attached with the passive RFID tags. Since the RFID tags are widely used to identify the objects in either the online shopping or the unmanned supermarket, any tagged object can be a potential threat to the user privacy. For example, the labeling tags on the delivery packages or the hang tags of the clothes from the online market may all open up a window of opportunity for eavesdropping. Besides, the adversary can even intentionally hide the battery-less and light-weighted RFID tags beside the loudspeaker, *e.g.*, under the table, which is hard to be detected and is able to eavesdrop in a long term. In this paper, we mainly focus on the private information, which are made up of number or letters, *e.g.*, social security number, a password, a credit card number, etc.

The adversary leverages an RFID system that can interrogate the RFID tags, which can work even in thru-the-wall scenario, and further extract the sound from the RF-signal and deduce the private information. Once any tag is placed beside the loudspeaker, the RF-signal backscattered by the tag can capture the sound vibration. Particularly, the tag can be directly vibrated by the sound due to the *vibration effect*, or affected by a nearby vibrating object due to the *reflection effect*. The adversary continuously collects the RF-signals and extracts the sound information when the loudspeaker is playing an audio sound, *e.g.*, a conversation during an online meeting. By analyzing the spectrogram energy distribution, the adversary can extract the sound from the RF-signals to deduce the private information, even if the adversary is outside the room of the victim.

2.2 Eavesdropping Scenarios

The side channel attack described in this paper can be launched via three different means: *medium-based*, *aerial-based* and *reflection-based* eavesdropping. *Medium-based* eavesdropping means the tag is directly attached on the vibration medium, *e.g.*, the loudspeaker. Hence, the sound transmission can lead to the tiny vibration of the medium and the tag. *Aerial-based* eavesdropping means that the tag is vibrated due to the aerial sound played by the loudspeaker, where recent work[6, 17, 26] has already shown its feasibility of capturing aerial sound using motion sensors. Both *Medium-based* and *Aerial-based* eavesdropping methods are leveraging the vibration effect to extract the sound information. *Reflection-based* eavesdropping means that a tag does not vibrate itself, but instead it is affected by the vibration of a nearby object, *e.g.*, a cup of water.

Online meeting. One possible attack scenario is that the victim is using a loudspeaker to discuss in the online meeting, which is frequently used during the COVID-19 period. The adversary can leverage the surrounding RFID tags to eavesdrop the sound played by the loudspeaker. As a result, the sensitive information talked during the online meeting can be obtained by the adversary, which may threat the personal life and property safety.

Voice assistant system. With the success of AI technique in the speech recognition, intelligent voice assistant systems, *e.g.*, Google Home, Amazon Echo Dot, are widely used due to their convenience. The voice assistants may replay the messages which includes some private information, *e.g.*, Google Home can remember the passwords or social security number with the 'Remember' function and replays them when needed. Such replayed sounds from the loudspeaker open up the possibility of the adversary eavesdropping on the private information.

3 FEASIBILITY STUDY

In this section, we use several experiments to study the feasibility of extracting the sound vibrations via RFID tags. Particularly, we focus on the mono-tone sound vibration to study the sensitivity of the RFID tags, which can be extended to the human voice.

3.1 COTS RFID Reader V.S. USRP Reader

We first compare the COTS RFID reader with the USRP reader in terms of sensing the tag vibration. We place the tag in front of the loudspeaker, as shown in Figure 2(a). We study the impact of mono-tone sounds with frequencies of 100Hz and 300Hz. By default, for the COTS ImpinJ Speedway $R420$ RFID reader [3], we have the sampling rate of 228Hz; for the USRP reader based on open project [4], we have the sampling rate of 2MHz.

Observation 1: *The USRP reader with higher sampling rate is more suitable for eavesdropping than the COTS RFID reader.*

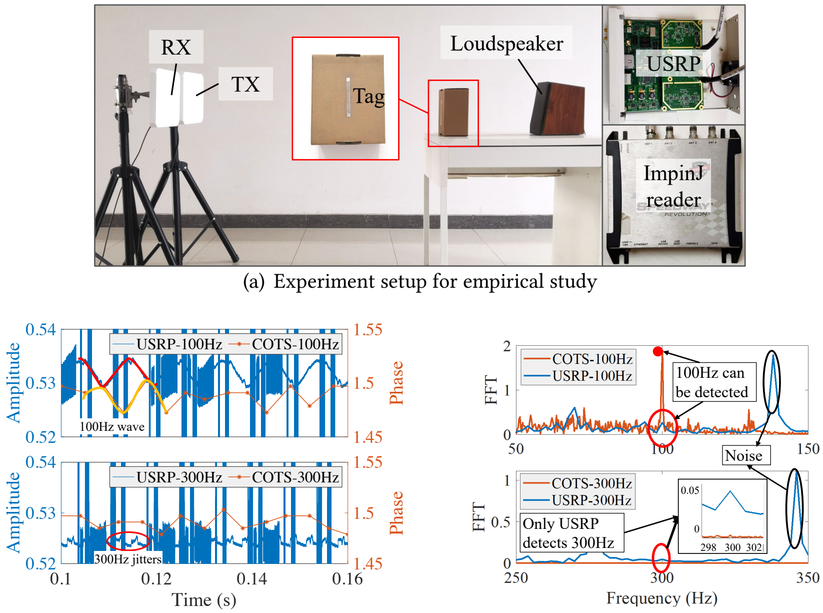
For the COTS RFID reader, we can only detect the 100Hz sound from both the frequency domain and time domain, *i.e.*, the orange wave in Figure 2(b) and the orange peak in Figure 2(c). According to the Shannon's law [31], over 600Hz sampling rate is required to capture the 300Hz sound. Even if the compressive reading [40, 41] can solve the mechanical vibration, it cannot sense the human voice, which has complicated frequency bands. Therefore, we do not consider the compressive sensing and use the traditional FFT to measure the frequency bands. For the USRP reader, even if the reader signal is much stronger than the tag signal, leading to the huge signal noise, we can still observe the weak tag signals of 100Hz and 300Hz in the time domain and frequency domain, *i.e.*, the 100Hz red wave and 300Hz jitters in Figure 2(b), and the corresponding blue peaks in Figure 2(c). Thus, when we focus on the human voice with complicated frequency bands, the USRP platform is more suitable to capture the human voice than the COTS RFID readers.

3.2 Tag Movement V.S. Tag Vibration

Since the tag vibration can be regarded as a small tag movement, we next investigate how the physical-layer signal changes with the tag movement by pushing the tag close to the antenna.

Observation 2: *The tag movement leads to the wavy change in the time domain, and the rotation of signal vector in the IQ plane.*

As shown in Figure 3(a), when we push the tag close to the antennas from 1.5m to 1.3m, the signal amplitude is changing as the cosine function. As shown in Figure 3(b), when we push the tag close to the antenna, the signal rotates in the IQ plane, and the rotation center is not at the origin. It means that the received signal does not change with the tag-antenna distance linearly. Moreover, two main circles are formed in this figure. In the enlarged signal in the time domain of Figure 3(a), we can clearly see the `QUERY` and `ACK` commands from the reader, as well as the `RN16` response and `EPC` response from the tag. Comparing Figure 3(b) with Figure 3(a), two circles in Figure 3(b) are caused by the changes of `CW` signals and tag backscattered signals, which correspond to the OFF and ON states of tag modulation [13]. Note that when we push the tag about $20cm$, which is about 1.23×



(a) Experiment setup for empirical study

(b) Signal in time domain

(c) Signal in frequency domain

Fig. 2. Signal analysis of USRP signal for vibration sensing.

of the half wave length of CW signals, the signal rotates about 1.23× circles. Since the tag vibration is a small tag movement, the tag vibration leads to the small wavy change in the time domain, and small rotation of signal vector in the IQ plane, which are used to build the model in Section 4.

3.3 Tag Vibration V.S. Diaphragm Vibration

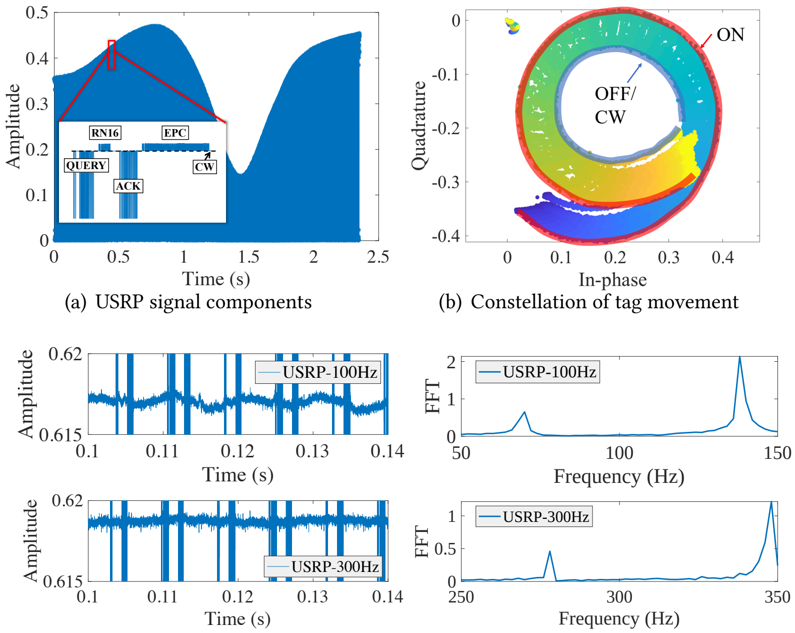
Since both the tag and the diaphragm may vibrate due to the sound pressure, we conduct experiments to study the different influences. Particularly, we remove the tag in front of the loudspeaker as shown in Figure 2(a) to capture the diaphragm vibration from the CW signal.

Observation 3: *The tag vibration captured by backscattered signals is much larger than the loudspeaker diaphragm vibration captured by CW signals.*

Comparing Figure 2(b) with Figure 3(c), when we remove the tag from the loudspeaker, the periodic patterns without tags are distinctly reduced. Particularly, for the 100Hz sound, we can still observe the weak periodic pattern in Figure 3(c), but the amplitude is much weaker than Figure 2(b). For the 300Hz sound, no periodic pattern can be found in Figure 3(c) and Figure 3(d). The reason is that the metallic tag can backscatter more RF-signals than the papery diaphragm. Thus, the attached tag can amplify the interference of the loudspeaker through backscattering.

4 SYSTEM DESIGN

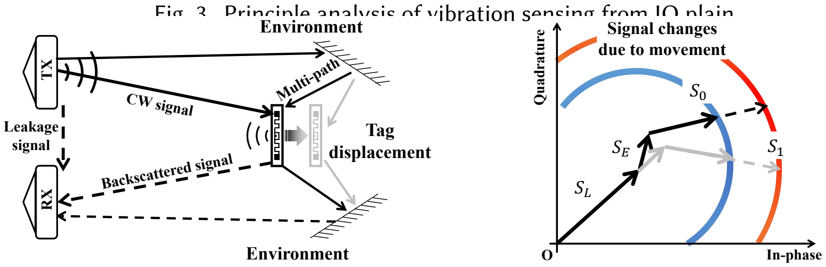
In this section, we introduce the principle of *Tag-Bug* by extracting the vibration of tag based on the signal model. In particular, we propose to extract the sound from either the *vibration effect* or the *reflection effect* of the tag. According to the sound extraction model, we design a new tag response mechanism, which can randomize the tag responses and improve the sound quality.



(a) USRP signal components (b) Constellation of tag movement

(c) Amplitude of USRP signal (d) Frequency analysis of raw signal

Fig. 3. Principle analysis of vibration sensing from IQ plain



(a) Signal components in RFID (b) Signal in the IQ plane

Fig. 4. Transmission model in RFID system.

4.1 Transmitting Model

Uplink. In RFID systems, the transmitting antenna TX sends the CW signal to activate the tag as shown in Figure 4(a). Due to the interference of multi-path effect, the signal reflected from the environment also arrives at the tag together with the CW signal:

$$S_{tag} = S_{TX}(h_d + h_E). \tag{1}$$

Here, S_{tag} indicates the signal received by the tag, S_{TX} is the CW signal sent by the TX antenna, h_d is the signal attenuation due to the transmitting distance and h_E is the signal attenuation due to the multi-path effect of the environment. Particularly, in an ideal channel model [13], h_d can be calculated as $h_d = \frac{1}{d}e^{\mathbf{j}\theta_d}$, where d is the distance between the TX antenna and the tag, $\mathbf{j}$ is the imaginary number. θ_d is the phase calculated from distance d and wave length λ, as:

$$\theta_d = 2\pi\frac{d}{\lambda} \mod 2\pi. \tag{2}$$

h_E is related to distance d and the transmitting environment in principle.

Downlink. After the tag receives the signal, the tag backscatters the signal with FM0 or Miller modulations, which encodes the binary bits with ON and OFF states [13]. For the OFF state, the tag backscatters all the CW signal, which has a small amplitude. Therefore, the signal received by the reader is the combination of the backscattered signal from tag $S_{tag}(h_{d'} + h_{E'})$ and the leakage signal from reader $S_{TX}h_L$:

$$S_{RX,0} = S_{TX}h_L + S_{tag}(h_{d'} + h_{E'}) = S_{TX}(h_L + h_d h_{d'} + h_{E,d}), \tag{3}$$

where $h_{d'}$ is the signal attenuation due to the downlink transmitting distance, $h_{E'}$ indicates the environment influence in the backscattered channel. For simplicity, we use $h_{E,d}$ to represent the overall signal attenuation due to the environment, which is also related to the distance d.

For the ON state, the tag backscatters a large amplitude signal by changing the state of tag antenna. Thus, the received signal is:

$$S_{RX,1} = S_{TX}h_L + S_{tag}(h_{d'} + h_{E'})h_1 = S_{TX}(h_L + h_1 h_d h_{d'} + h'_{E,d}), \tag{4}$$

where h_1 is the modulation gain of the tag, and $h'_{E,d}$ is the overall signal attenuation due to the environment for the ON state. In RFID systems, the tag changes the antenna capacitance to modulate the CW signal during the backscattering, so that h_1 is usually regarded as the signal enhancement. Particularly, because the multi-path effect from the environment is relative small, we thus omit the influence of h_1 and regard $h'_{E,d}$ approximates to $h_{E,d}$. As a result, the signal received by the reader can be divided into three parts: *the leakage signal* S_L, *the multi-path signal* S_E and *the backscattered signal* S_0 or S_1, where

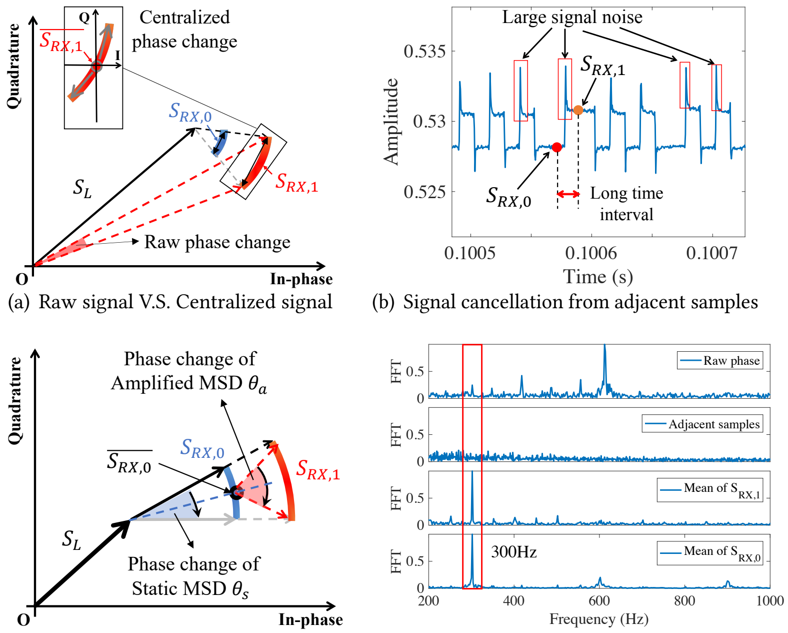
$$\begin{cases} S_L = S_{TX}h_L, \\ S_E = S_{TX}h_{E,d}, \\ S_0 = S_{TX}h_d h_{d'},\ S_1 = S_{TX}h_d h_{d'} h_1. \end{cases} \tag{5}$$

When the TX antenna and RX antenna are placed close to each other and the tag is relatively far from the two antennas, we regard $d' \approx d$. Thus, both S_0 and S_1 are proportional to $h_d h_{d'} = h_d^2 = \frac{1}{d^2}e^{\mathbf{j}2\theta_d}$, indicating that the phase change is $2\pi\frac{2d}{\lambda}$. Such phase change is compatible with the results in Figure 3(b), where 20cm movement leads to 2.45π radians phase change.

IQ plane analysis. Figure 4(b) presents the signal model in the IQ plane. The transmitting distance $d\backslash d'$ changes with the tag movement, leading to the change of both the multi-path signal S_E and the backscattered signal $S_0\backslash S_1$. Thus, the phases of S_E and $S_0\backslash S_1$ get changed, resulting in the rotation of the corresponding signals. The phase change of $S_0\backslash S_1$ is caused by the signal attenuation h_d^2, whose phase change is $2\pi\frac{2d}{\lambda}$. Therefore, both S_0 and S_1 rotate with the transmitting distance d, which leads to two arcs in the IQ plane. Since S_E is usually static, we omit it for simplicity. Such results exactly explain the signal change in Figure 3(b).

4.2 Sound Extraction from Vibration Effect

Theoretically, the *vibration effect* of the tag due to the sound can lead to the variation of the transmitting distance as $d = d_0 + f(t, d_v)$. Here, d_0 is the average tag-antenna distance, and $f(t, d_v)$ is the distance variation related to time t and vibration amplitude d_v. For the mono-tone sound with the frequency ϕ, $f(t, d_v) = d_v\cos(2\pi\phi t)$, which can be extended to any complicated sound with multiple tones. For simplicity, we introduce the algorithm with mono tone sound. In an ideal model, such tag vibration can be directly captured by the received signals S_0 and S_1. However, since the leakage signal S_L is much stronger than the backscattered signal S_0 and S_1, the small changes of S_0 and S_1 will not remarkably affect the received signal $S_{RX,0}$ and $S_{RX,1}$. Figure 5(a) plots the vibration-based signal change by omitting S_E. Both $S_{RX,0}$ and $S_{RX,1}$ slightly rotate, and the raw phase change is much small due to the strong leakage signal. Moreover, sub-mm level vibration of the tag due to the sound can be easily drowned by the ambient noise. Thus, *we need to amplify the vibration effect by removing the strong interference.*



(a) Raw signal V.S. Centralized signal (b) Signal cancellation from adjacent samples

(c) Amplified MSD V.S. Static MSD (d) Vibration extraction results of different cancellation methods

Fig. 5. Vibration extraction mechanisms.

Naïve Normalization. The direct way is to centralize $S_{RX,1}$ by subtracting the average value $\overline{S_{RX,1}}$, as shown in Figure 5(a). The phase variance range can be amplified to $[0, 2\pi]$. However, in the real system, $S_{RX,1}$ contains the large ambient noise, and such subtracting can import the additional noise signal. Thus, both the vibration effect and the signal noise are amplified.

To efficiently amplify the *vibration effect*, our basic idea is to extract the backscattered signals, which are related to the tag displacement. If we can obtain the backscattered signal S_0 or S_1, the corresponding phase change can indicate the tag displacement. However, it is difficult to measure the leakage signal S_L and the environment signal S_E, thus, we cannot individually get either S_0 or S_1 by referring to $S_{RX,0}$ and $S_{RX,1}$. Fortunately, since both S_L and S_E are static in most scenarios, by regarding $h'_{E,d}$ approximates to $h_{E,d}$, we can remove S_L and S_E from Eq. (4) and Eq. (3) as:

$$\Delta S_{RX} = S_{RX,1} - S_{RX,0} \approx S_{TX}(h_1 - 1)h_d^2. \tag{6}$$

We call it *Modulated Signal Difference (MSD)*. Here, only h_d changes with the tag vibration in principle, meaning that the vibration can be extracted from the MSD phase.

However, in any snapshot, only one of $\underline{S_{RX,0}}$ and $\underline{S_{RX,1}}$ can be received. Therefore, we cannot get the MSD ΔS_{RX} in reality. For a static tag, we can use $\overline{S_{RX,0}}$ and $\overline{S_{RX,1}}$ to calculate the MSD ΔS_{RX}, which is called *Static MSD*. But for a vibrating tag, both $S_{RX,0}$ and $S_{RX,1}$ get changed even during one tag response. Therefore, we cannot simply calculate the MSD from the average value. To address the problem, two kinds of cancellation solutions are considered to extract the MSD efficiently.

Instantaneous MSD. The first solution is the cancellation based on adjacent samples. As shown in Figure 5(b), since $S_{RX,0}$ and $S_{RX,1}$ cannot be collected in one snapshot, we use adjacent samples to approximate uncollected samples, which is called instantaneous MSD. It is similar to the standard RFID channel estimation, but the traditional estimation targets on a relatively stable tag while we focus on a vibrating tag. As shown in Figure 5(b),

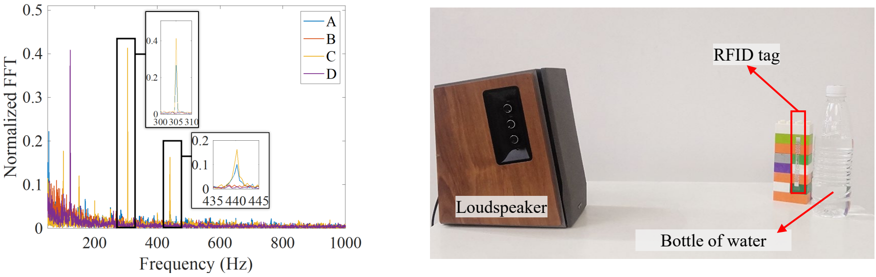


Fig. 6. Influence of enhanced multi-path effect.

due to the large signal noise around the signal edge, adjacent samples should be selected from the stable part of the square wave, and thus there is a longer time interval between adjacent samples. Such interval may not affect a stable tag, but can introduce the large noise for the high-frequency vibration.

Amplified MSD. The second solution is the cancellation based on $\overline{S_{RX,0}}$ within a small time window. The basic idea is to use $\overline{S_{RX,0}}$ to replace $S_{RX,0}$ for cancellation. Due to the vibration influence, $S_{RX,0}$ and $S_{RX,1}$ change with the time due to the vibration. Since the tag is vibrating at a fixed position during the time window, $\overline{S_{RX,0}}$ can be roughly regarded as $S_{RX,0}$ when the tag is static at its original position. As $\overline{S_{RX,0}}$ is an average value, there is no time interval between $S_{RX,1}$ and $\overline{S_{RX,0}}$, and the vibration feature is extracted as:

$$\Delta S'_{RX} = S_{RX,1} - \overline{S_{RX,0}} \approx S_{TX}(h_1 h_d^2 - \overline{h_d^2}). \tag{7}$$

Here, the time variation of $\overline{S_{RX,0}}$ is removed while $S_{RX,1}$ still contains the time variation signal due to the sound. Thus, Eq. (7) can be used to derive the vibration.

Compared with the static MSD, Eq. (7) omits the variation of S_0 and focuses on the variation of S_1 to extract the vibration. It is related to both the transmitting distance d and the modulation attenuation h_1. Since h_1 is the modulation factor caused by the impedance change of the tag antenna, the amplitude of h_1 is greater than 1. Hence, Eq. (7) can amplify the MSD phase by using $\overline{S_{RX,0}}$, which is called the *amplified MSD*. Figure 5(c) illustrates the amplification principle. By connecting the end of S_L and $\overline{S_{RX,0}}$, according to the exterior angle theorem of a triangle, the phase change of the amplified MSD θ_a, *i.e.*, summation of the two exterior angles, is larger than the phase change of the static MSD θ_s, *i.e.*, the summation of the two remote interior angles.

We use the 300Hz mono-tone sound to test the performance of different solutions as shown in Figure 5(d). For the raw phase of received signals, the 300Hz sound is buried by the 610Hz noise, which is caused by the measurement noise of the hardware. For the cancellation based on adjacent samples, the noise spreads over the frequency band, due to the large interval between adjacent samples. For the cancellation based on $\overline{S_{RX,1}}$, although we can get the clear peak at 300Hz, the ambient noise is also amplified, leading to several noise peaks. For the cancellation based on $\overline{S_{RX,0}}$, we detect frequency peaks at 300Hz, 600Hz and 900Hz, caused by harmonic signals. Thus, the amplified MSD is better for the vibration extraction, which is also suitable for the extraction of the sound with multiple tones.

4.3 Sound Extraction from Reflection Effect

Next we demonstrate how to extract the sound vibration considering the reflection effect. Currently, the sound can be extracted from the *vibration effect* by attaching the tag on the vibrating objects in some situations. When the tag vibrates on materials such as liquid, the tag signal could be seriously absorbed and the vibration signal is drowned by the ambient noise. Nevertheless, these materials are more easily vibrated by the sound than the tags. Therefore, by deploying the RFID tag close to the vibration-sensitive materials, even if the tag itself cannot be effectively vibrated, it can still perceive the vibration via the CW signal reflected from the surrounding vibrating materials. We call it *reflection effect*.

Usually, received signals contain the leakage signal S_L, the tag backscattered signal S_{tag} and the environment noise S_E:

$$S_{RX} = S_L + S_{tag} + S_E, S_{tag} = S_0 \text{ or } S_1. \tag{8}$$

Particularly, the environment noise can be further divided into several reflection signals $S_{B,i}$ and the thermal noise S_N:

$$S_E = S_N + \sum_{i=1}^{\mathcal{M}} S_{B,i}, \tag{9}$$

where $\mathcal{M}$ is the number of reflection signals. Normally, the reflection signals $S_{B,i}$ are usually static and weaker than the tag signal S_{tag}. However, if the reflection object resonates together by the sound, the vibration of the object will enhance the reflection signal.

We can divide the environment noise S_E into the static part $S_{E,s}$ and the vibration part $S_{E,v}$. According to Eq. (7), we can cancel the static part and have:

$$\Delta S''_{RX} = S_{RX,1} - \overline{S_{RX,0}} \approx S_{TX}(h_1 h_d^2 - \overline{h_d^2}) + (S_{E,v} - \overline{S_{E,v}}). \tag{10}$$

Comparing with Eq. (7), we add another time variation part $S_{E,v}$ caused by the reflection object, and $\Delta S''_{RX}$ can be used to sense the sound. Basically, the variation of $S_{E,v}$ is captured by the nearby tag, such that, the received signals $S_{RX,0}$ and $S_{RX,1}$ contain the reflection effect besides the vibration effect. Thus, Eq. (10) is the general case of Eq. (7).

We validate the model via experiments. We consider 4 kinds of setup: A. a single tag, B. a single bottle of water, C. a tag close to a bottle of water, D. a tag attached on a bottle of water. Instead of using mono-tone sound, we use the multi-tone sound with frequencies of 305Hz and 440Hz. As shown in Figure 6, setup A can achieve the FFT values of 0.27 and 0.1, respectively. Moreover, setup C achieves larger FFT values, *i.e.*, 0.41 and 0.15, which benefits from the reflection effect of the water vibration. But the FFT values of setup B and D are all below 0.01, meaning that they cannot sense the vibration. Thus, it is sometimes difficult to perceive the sound directly from the vibrating object itself. But the reflecting signal of vibrating object can be captured by the receiver through the tag backscattering. Hence, the reflection effect improves the capability of the vibration detection.

4.4 Tag Response Mechanisms

We next arrange tag responses based on the EPC C1G2 standard [2] to reduce the influence of the reader signal. In RFID systems, the reader uses the Framed-Slotted-ALOHA (FSA) anti-collision protocol to communicate with the RFID tags. Thus, the reader signal is alternately received with the tag signal by the receiving antenna RX as shown in Figure 3(a). According to the vibration model, since only the backscattered signal can be used for sensing, the reader signal should be removed from the time domain. Traditionally, the removed reader signal can be well interpolated and the reader signal does not affect the perceiving. But if each frame has the same size, the QUERY command appears periodically, leading to the large noise around the corresponding periods even the removed reader signal is interpolated.

Therefore, *it is essential to reduce the interference of the reader signal.* The basic idea is to import random factors for the tag response so that reader commands are randomly distributed. To break the periodicity, we first propose *Random Retransmission*, inspired by literatures [5, 20]. Different from continuously retransmitting EPC commands, we require the tag to randomly retransmit EPC commands. Thus, the length of each query cycle is randomly distributed according to retransmission numbers, and the periodicity problem is resolved. The second mechanism is *Random Frame-size*, which sets the frame size randomly. Based on the EPC C1G2 standard, we can modify the frame-size by setting Q parameter. Therefore, by randomly setting the frame-size, we can also solve the periodicity problem of reader commands.

Both *Random Retransmission* and *Random Frame-size* can break the periodicity problem of reader commands in principle, but they focus on randomizing different signals. *Random Retransmission* mechanism retransmits the EPC repeatedly, meaning that both $S_{RX,0}$ and $S_{RX,1}$ are collected. But there are large modulation noises around the signal edges between $S_{RX,0}$ and $S_{RX,1}$. *Random Frame-size* mechanism retransmits the CW repeatedly, meaning that more continuous CW samples, *i.e.*, $S_{RX,0}$, are collected. But they cannot be used to calculate the MSD and thus reduce the time resolution for the backscattered signal. Since *Random Retransmission* provides $S_{RX,0}$ and $S_{RX,1}$, and *Random Frame-size* provides continuous $S_{RX,0}$ to capture the vibration, a hybrid method of the two mechanisms is used to improve the perceiving performance.

5 SYSTEM IMPLEMENTATION

Tag-Bug perceives the sound of the loudspeaker based on the changing backscattered RF-signal of the tag. Specifically, the sound from the loudspeaker vibrates the tag or nearby objects, and such tiny vibration can be captured via either the *vibration effect* and the *reflection effect*. The implementation of *Tag-Bug* consists of five main components: 1) *Sampling randomization* first breaks the periodic pattern of reader signals by randomizing the tag retransmission and the frame size. 2) *Vibration amplification* amplifies the vibration influence on the received signal by making the vibration signal orthogonal to the noise. 3) *Vibration feature extraction* then separates the tag response from the received signal and extracts the amplified MSD as the vibration feature from the tag response. 4) *Signal filter* filters the signal noise based on the frequency analysis. 5) *CGAN based sound refinement* recovers the high-frequency band by referring to the low-frequency band to improve the quality of the recovered human voice.

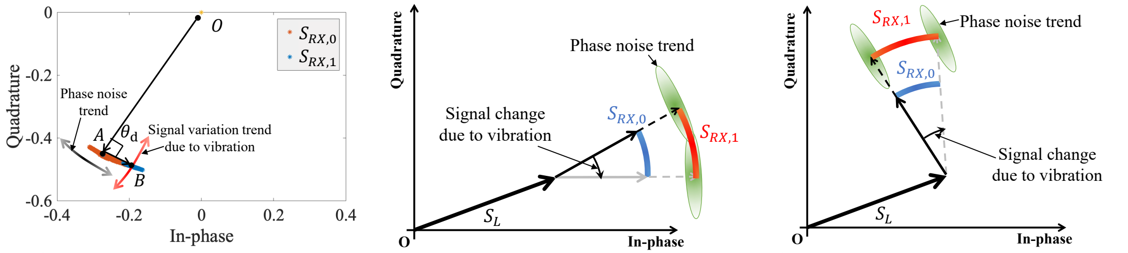
5.1 Sampling Randomization

Based on Section 4.4, we show the hybrid method of tag response mechanism. The basic idea is that the tag randomly retransmits the EPC command and the reader randomly sets the frame size. For the *Random Retransmission*, we let the tag retransmit its EPC command with 50% probability. For the *Random Frame-size*, we randomly set the frame-size to 1 or 2 with 50% probability. The probability settings guarantee that we import the randomization, and meanwhile avoid long-term EPC retransmission and the continuous empty slots. Thus, we randomize the tag response in both the query-level and slot-level to solve the period problem of reader signal.

5.2 Vibration Amplification

Even though the vibration is captured by the backscattered signal, the environment noise can usually overwhelm the tiny signal changes due to the sound wave. Thus, it is essential to derive the noise distribution and amplify the influence of the vibration caused by the sound on the received signal.

Based on our empirical study, the phase variance of the USRP platform is much larger than the amplitude variance, due to the hardware noise. Figure 7(a) shows the signal constellation of a static tag. The arc-shaped distribution of signal samples is caused by the phase variance, while the actual distribution should be one point in principle. Figure 7(b) shows the interference of the phase noise on the signal changes due to vibration, where the signal changes are almost drowned by the phase noise. To reduce the influence of the phase noise, our basic idea is *to make the vibration-based signal change orthogonal to the signal noise by adjusting the transmitting distance as shown in Figure 7(b)*. According to Figure 3(b), the signal changes due to the tag movement form a circle, and the radial direction is indicated by the line between $S_{RX,0}$ and $S_{RX,1}$. In Figure 7(a), the signal change due to the vibration is orthogonal to the radial direction AB, where A is the center of $S_{RX,0}$ samples, B is the center of $S_{RX,1}$ samples. Besides, the signal noise due to the phase variance is orthogonal to OA. Therefore, if AB is orthogonal to OA, the vibration-based signal change is orthogonal to the signal noise and the influence of the phase variance is reduced. According to literature [33], $\angle OAB$ is related to the transmitting distance, which is the same as θ_d in Eq. (2). Therefore, we can vary the transmitting distance such that $\angle OAB$, *i.e.*, θ_d, is close to $\pi/2$. In the eavesdropping



(a) Signal distribution of a static tag (b) Interference of phase noise on signals (c) Elimination of phase noise on signals

Fig. 7. Reducing interference of phase variance.

scenario, we can use a mobile-framework or manually to move reader antennas towards the tag to adjust θ_d. As a result, the vibration-based signal change due to the sound is orthogonal to the phase noise, and the influence of the phase noise is thus reduced.

5.3 Vibration Feature Extraction

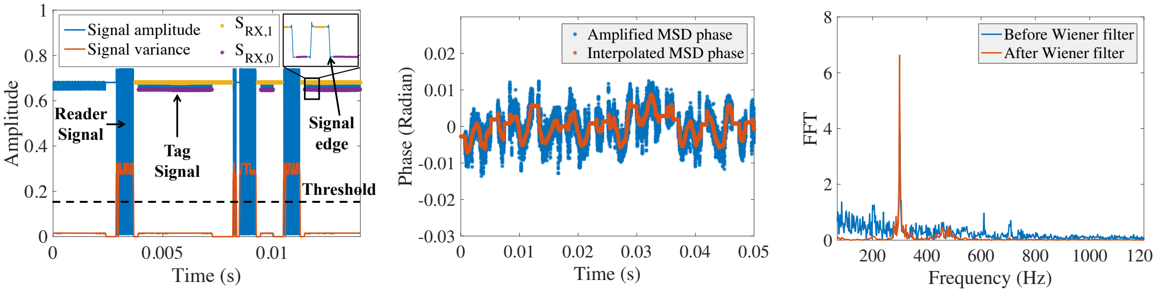
We next extract the vibration features from the received signals. We first segment the reader signal and the tag signal, and separate $S_{RX,0}$ and $S_{RX,1}$ from the tag signal. Then, we can extract the MSD feature $\Delta S''_{RX}$ based on Eq. (10). For the signal separation, the basic idea is that the reader signal has the larger energy variance compared with the tag signal. Therefore, we leverage a sliding-window to calculate the variance of the received signal. As shown in Figure 8(a), the variance of tag signal is very small because it is modulated on the CW signal, while the reader signal has much larger variance. Then, we can separate the tag signal based on the small signal variance.

After obtaining the tag signal, we further extract $S_{RX,0}$ and $S_{RX,1}$ from the backscattered signal. The basic idea is to detect signal edges and separate the signal according to the signal edges. When the tag modulates the signal by changing its state between ON and OFF, the received signals vary between $S_{RX,0}$ and $S_{RX,1}$. Therefore, by leveraging the signal edges, we can detect $S_{RX,0}$ and $S_{RX,1}$. Finally, we can calculate the vibration features MSD based on Eq. (10).

5.4 Signal Filter

The basic idea to recover the sound from the vibration is to analyze the signal extracted from the amplified MSD in the frequency domain, and further filter the ambient noise. Since the MSD features extracted based on either the *vibration effect* or the *reflection effect* are based on the signal segments of tag backscattering and the signal segments of reader command are useless in the sound extraction, we need to firstly handle the useless reader signal. In RFID system, the reader signal is relatively shorter compared with the tag signal, *e.g.*, RN16 command from the tag is 4.5 times longer than EPC command from the reader. Thus, we can remove the reader signal and the sound will not be distorted significantly. Here, the MSD features are not uniformly sampled due to the removed reader signal. Then, we further resample all the features with the spline interpolation, which can fill the gap of the reader signal based on the trend of MSD features. As shown in Figure 8(b), after the interpolation, the periodic feature can be roughly detected from the signal wave. After the interpolation, we further filter the signal via the frequency analysis. Typically, the main frequency bands of human voice are between 100Hz and 1000Hz [32]. Hence, we use a Butterworth filter to remove the out-of-band signal. In addition to the high frequency noise, it also removes the low frequency noise caused by ambient human movements.

For the noise signal inside the frequency band of the human speech, we design a Wiener filter [29] to further remove it. The basic idea is that the noise signal is usually stable during a short period, thus we can get the noise signal when the loudspeaker is not playing, and then remove the noise when the loudspeaker is playing. We design the Wiener filter based on the known signal, *i.e.*, the ambient noise, and then estimate the vibration



(a) Separating tag signal (b) Amplified MSD interpolation (c) Frequncy with Wiener filter

Fig. 8. Signal pre-processing and filtering.

signal, by removing the known signal from the received signal. As shown in Figure 8(c), after we remove the ambient environmental noise based on the Wiener Filter, the 300Hz signal due to vibration is much clearer in the frequency domain. Therefore, it indicates that the Wiener Filter can efficiently reduce the ambient noise.

5.5 CGAN based Sound Refinement

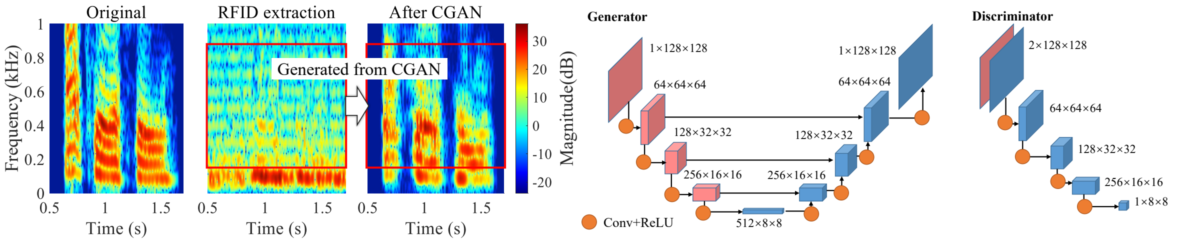
Even if we extract the sound based on either the *vibration effect* or the *reflection effect*, the frequency band is incomplete due to the inherent physical characteristics of RFID tags. Particularly, since the low-frequency sound has larger energy than the high-frequency sound to vibrate the tags, the sound extracted from the RFID tags have clear low-frequency band and relatively vague high-frequency band. Figure 9(a) compares the frequency band between the original sound and the extracted sound from RFID tags in the spectrogram. For the human voice, the low-frequency band is the fundamental frequency, while the high-frequency band is the harmonic frequency. Thus, we need to recover harmonic frequency band based on the fundamental frequency.

To solve the problem, we leverage a Conditional Generative Adversarial Network (CGAN) [27] to generate the whole frequency coefficients Y conditional on the fundamental frequency coefficient X. As shown in Figure 9(b), we take X as the input and generate the spectrogram from the fundamental frequency of X. Particularly, we divide the sound signals into separate word of 0.5s window length, and then calculate the spectrogram figure with Short-Time Fourier Transform (STFT). For the generator, we use a U-Net [28] to recover the whole frequency band. It firstly downsamples the original spectrogram with 4 convolution layers, and then upsamples the features to the original size with another 4 convolution layers. In the upsample part, we use skip connections to concatenate the input of upsample layer with the corresponding cropped features, which can combine both the high-level and low-level features for generation. For the discriminator, we use a traditional CNN model with 3 hidden convolution layers to judge the real spectrogram. We take the filtered sound extracted from RF-signals as the input of CGAN, and use the original sound recorded as the real spectrogram. We have generated more than 3000 minutes of sound files, where 80% of the sound files are used to train the CGAN model, and the rest data is used to test the model. By training the CGAN model, we can automatically generate the whole frequency band and obtain the full-frequency band from the extracted sound as shown in Figure 9(a). Finally, we use inverse STFT to transform the spectrogram to achieve better human voice. For the refined sounds, traditional methods such as the neural networks can be further used to recognize them. We omit this part and train our own LeNet-5 [18] to recognize the human voice, which are later introduced in Section 7.

6 PERFORMANCE EVALUATION OF MONO-TONE SOUNDS

6.1 Experiment Settings

To evaluate the sound quality of the extracted sound, we first quantitatively analyze the Signal-Noise-Ratio (SNR) when eavesdropping the mono-tone sound. We implement *Tag-Bug* based on the USRP N210 platform with two Laird *S*9028*PCL* directional antennas and an SBX Daughter-board according to the open source project [1], which



(a) Spectrogram comparison

(b) CGAN framework

Fig. 9. Sound refinement based on CGAN.

works at a center frequency of 920MHz. During the whole experiments, the sensing tag is vertically deployed on the Express Package with a default tag-antenna distance of 1m. The Express Package is filled with textile fabrics, which has a total weight of 0.65kg. We vary different kinds of parameters to evaluate the performance, including transmitting environment, sound characteristic and hardware, which are shown in Table 1.

Transmission environment. We examine the influence of different transmission environment by varying the tag-antenna transmitting distance, the tag-loudspeaker distance. For the transmitting distance, we vary the distance from 1m to 4m. For the tag-loudspeaker distance, we vary the distance from 10cm to 200cm. By default, the transmitting distance is set to 1m and the tag-loudspeaker distance is set to 20cm.

Sound characteristics. We vary the frequency and the loudness of the sound to evaluate the accuracy of the extracted vibration from RFID. Particularly, we vary the frequency of the mono-tone sound ranging from 80Hz to 1000Hz, which can cover the main frequency band of the human speech. For the loudness, we deploy a decibel meter beside the RX antenna to measure it. To evaluate the sensitivity of transient response, we vary the duration of each monotone sound from 10ms to 200ms. By default, we play the 261Hz mono-tone sound, with 75dB loudness.

Hardwares. We evaluate the performance by varying loudspeakers, RFID tags and everyday objects attached with RFID tags. Particularly, we consider 5 different loudspeakers to play the sound, *i.e.*, Edifier $R1700BT$ loudspeaker (Edifier), JBL $Clip3$ mini-speaker (JBL), Thinkpad laptop (Thinkpad), Anker speaker (Anker) and Huawei bluetooth speaker (HW). Besides, we consider 4 different RFID tags, including AZ-9640, AZ-$U73$, $E51$ and $ER62$. Moreover, we attach the tags on different objects, *i.e.*, package of Playing Card (PC), Plastic Bag (PB),

Table 1. Evaluation parameters

Parameter name	**Parameter range**	**Default settings**
Tag-antenna transmitting distance	1 - 4 m	1 m
Tag-loudspeaker distance	10 - 200 cm	20 cm
Mono-tone frequency	80 - 1000 Hz	440 Hz
Sound loudness	60 - 84 dB	80 dB
Mono-tone duration	10 - 200 ms	200 ms
Loudspeakers	Edifier, JBL, Think pad, Anker, Huawei	Edifier
RFID tags	E51, ER62, AZ-9640, AZ-U73	E51
Objects	Playing Card, Plastic Bag, Express Package, Airtight food Container	Express Package

Express Package (EP) and Airtight food Container (AC). By default, we use the loudspeaker Edifier with the tag *AZ*-9640, attached on the express package.

Metrics. For the quantitative analysis of mono-tone sound, we use the Signal-Noise-Ratio (SNR) to evaluate the eavesdropping performance as $SNR = 10\log\frac{E_S}{E_N}$, where E_S is the power of the sound signal and E_N is the power of the noise. Since mono-tone sound has one typical frequency, given $P(f)$ as the frequency, we modify the frequency as:

$$\begin{cases} P(i) = P(i)/5, if\ P(i) < \max(P(f))/2, \\ P(i) = P(i), if\ P(i) \geq \max(P(f))/2. \end{cases} \tag{11}$$

Finally, we calculate the improved SNR accordingly.

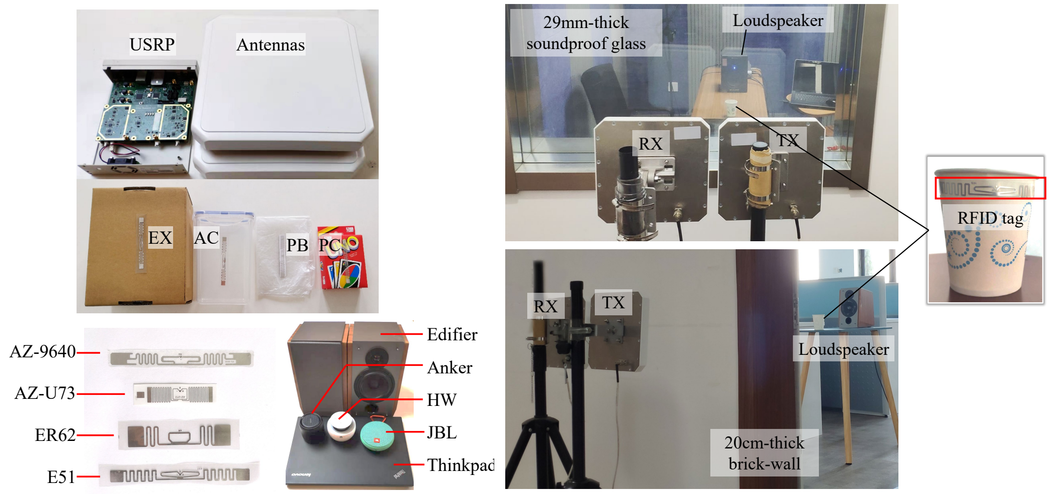
6.2 Impact of Transmission Environment

Tag-Bug can achieve over 4*dB SNR when the transmitting distance is over* 2*m in the Line-Of-Sight (LOS) transmission.* As shown in Figure 11(a), for the transmitting distance, the SNR decreases with the transmitting distance, because the propagation of the RF-signal may import the larger noise and the vibration signal also fades away during the propagation. As a result, when the transmitting distance increases, the SNR value decreases. Anyway, when the distance is larger than 2m, we can still achieve 4dB SNR.

Tag-Bug can achieve 2.6*dB SNR when the tag-loudspeaker distance is* 100*cm.* For the tag-loudspeaker distance, since it is difficult to sense the sound from the vibration effect when the tag-loudspeaker distance is larger than 50cm, we evaluate the impact of the tag-loudspeaker distance based on the reflection effect, *i.e.*, sensing the sound by deploying the tag beside a bottle of water. As shown in Figure 11(b), the SNR slightly decreases with the increasing tag-loudspeaker distance, since the sound power reduces along with the large tag-loudspeaker distance. Even the tag-loudspeaker distance is as long as 100cm, *Tag-Bug* can still achieve 2.6dB SNR. Moreover, *Tag-Bug* can still perceive the sound vibration when the tag-loudspeaker distance is as far as 2*m*.

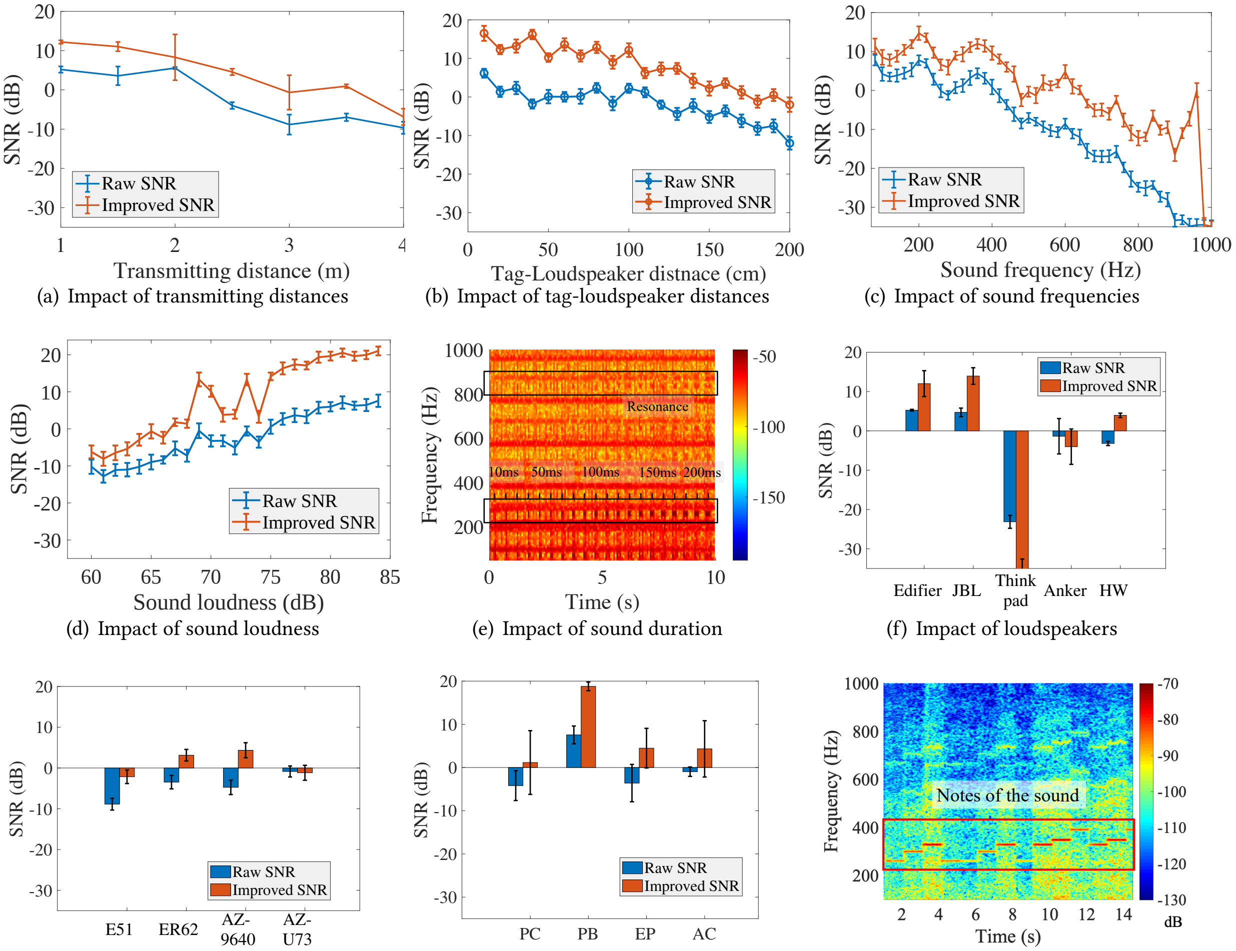
6.3 Impact of Sound Characteristics

Tag-Bug can achieve 2.3*dB average SNR when eavesdropping mono-tone sounds with different frequencies, and the highest SNR is up to* 14.7*dB.* We first evaluate the impact of different mono-tone frequencies. We focus on the mono-tone sound with the frequency ranging from 80Hz to 1000Hz, which can usually cover the frequency of



(a) Experiment devices (b) Experimental deployment in a soundproof room

Fig. 10. Experiment setups.



(a) Impact of transmitting distances (b) Impact of tag-loudspeaker distances (c) Impact of sound frequencies

(d) Impact of sound loudness (e) Impact of sound duration (f) Impact of loudspeakers

(g) Impact of tags (h) Impact of objects (i) Nursery rhyme Frère Jacques

Fig. 11. Performance of different transmitting environments.

the human speech. As shown in Figure 11(c), the lower frequency sound has the larger SNR value, while the higher frequency sound has the smaller SNR. Particularly, for the frequency lower than 200Hz, the SNR is larger than 0dB, while the SNR decreases to −10dB when the frequency is larger than 600Hz. It is mainly caused by two folds. First, the loudspeaker has different gains to the sounds with different frequencies, due to the physical imperfection. Second, the low-frequency sounds suffer from the less attenuation than the high-frequency sounds during the propagation, leading to the larger power to vibrate the tags and the nearby objects..

Tag-Bug can still achieve 1.8*dB SNR when the loudness of sound reduces below* 67*dB.* Next, we manually vary the loudspeaker volume and use a decibel meter to record the loudness. It is as expected that the louder sound leads to the larger SNR value in Figure 11(d). Particularly, we can still perceive the sound vibration even the loudness is 60dB. It indicates the possibility to eavesdrop in different scenarios. Moreover, the improved SNR is 8dB larger than the raw SNR, indicating that we can efficiently eavesdrop the sound.

Tag-Bug *can efficiently capture the sound vibration of monotone if the sound duration is longer than* 50*ms.* To evaluate the transient response on the sound vibration, we further vary the duration of each 261Hz monotone sound from 10ms to 200ms, and extract the sound in the default scenario. Figure 11(e) shows the spectrogram of all the 20 kinds of sounds. We find *Tag-Bug* can always detect the transient sound pulse, and achieves a stable extraction performance when the duration is longer than 50ms. Besides, we also find the resonance signal of the transient sound, especially when the duration is longer than 100ms. However, they all achieve an average raw SNR of over 3.21dB for the 261Hz monotone sound. It indicates that *Tag-Bug* can efficiently capture the sound vibration although it is relatively short.

6.4 Impact of Hardware

Tag-Bug achieves more than 3*dB SNR for the eavesdropping even on mini speakers.* As shown in Figure 11(f), we compare the SNR values among five different loudspeakers. The Edifier and JBL loudspeakers are the easiest to eavesdrop, due to the large gain of the loudspeaker. As for the laptop, since the embedded loudspeakers are designed for the personal usage, the gain is usually designed to be small. For the Huawei and Anker mini loudspeakers, they have more than 3dB SNR, indicating the efficiency of *Tag-Bug* in eavesdropping different kinds of loudspeakers.

Tag-Bug can leverage different tags for the eavesdropping. As shown in Figure 11(g), the four types of tags have similar SNR values. Particularly, AZ-9640 tag has the best performance, because the tag antenna is larger and the backscattered signals have less noises compared with other tags.

Tag-Bug can achieve up to 18*dB SNR when the tag is attached on the plastic bag.* As shown in Figure 11(h), we compare the eavesdropping performance by sticking the tag on four everyday objects, *i.e.*, package of Playing Card (PC), Plastic Bag (PB), Express Package (EP) and Airtight food Container (AC). We find that the plastic bag has the best performance, because it is lighter than the other three objects and thus easy to be vibrated by the sound. Since the plastic bag can be any package bag from the market, which is easily to be overlooked, it also indicates the high threat of the eavesdropping. The other three objects have similar performances, because they are all rigid-bodies, and only partial surface is vibrated by the sound. Therefore, the sound cannot lead to the smaller vibration of the objects compared with the plastic bag, leading to lower SNR.

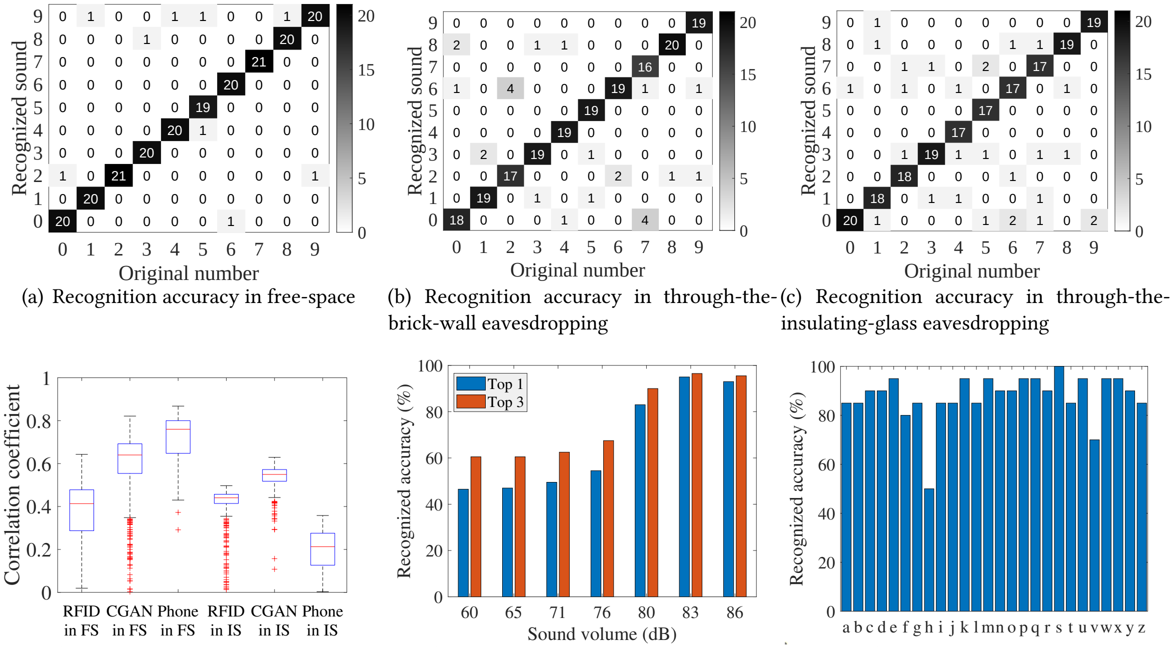
6.5 Impact of Continuous Sounds

Tag-Bug can efficiently recover the continuous sounds in time-series. We evaluate the performance of *Tag-Bug* on eavesdropping the continuous sounds. The loudspeaker plays a nursery rhyme Frère Jacques, which consists of a sequence of mono-tone sounds with different frequencies. As shown in Figure 11(i), we can clearly see each note corresponding to "do-re-mi-do-...". Since we reduce the amplitude to 1/2 from 5s to 9s, we can also observe such volume change in Figure 11(i). It indicates that *Tag-Bug* can derive both the frequency and the relative amplitude during the eavesdropping.

7 PERFORMANCE EVALUATION OF HUMAN VOICE

We further use our system to eavesdrop the human voice, which is played by the loudspeaker with the default setting. In comparison, we also eavesdrop the human voice in both the through-the-wall eavesdropping and through-the-insulating-glass eavesdropping, which is compared with the free-space eavesdropping. In our scenarios, we use either a 20*cm-thick brick-wall* or a 29*mm-thick insulating-glass with three layers of glass* to separate the loudspeaker and the RFID reader as shown in Figure 10(b). Particularly, we focus on the numbers and letters, which can be the main component of some private information, *e.g.*, social security number or passwords, *etc.* We invited 10 volunteers (8 males and 2 females) with IRB approval to record the human voice of 10 numbers and 26 letters. Then we use the Edifier loudspeaker to play all the human voice and use *Tag-Bug* to eavesdrop the sound in different scenarios.

We have generated 630 sound records for each number and letter to evaluate the recognition performance. We propose two ways to evaluate the performance: 1) We calculate the correlation coefficient between the original



(a) Recognition accuracy in free-space

(b) Recognition accuracy in through-the-brick-wall eavesdropping

(c) Recognition accuracy in through-the-insulating-glass eavesdropping

(d) Similarity of extracted sound

(e) Recognition of different volumes

(f) Recognition of letters

Fig. 12. Performance of recognizing human voice.

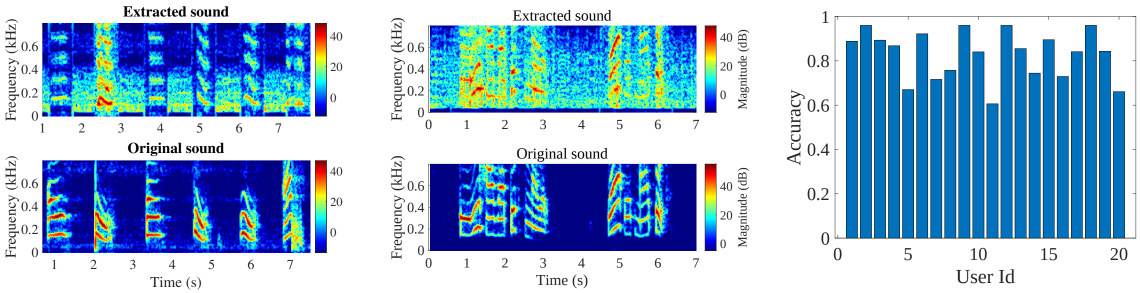
sound and the eavesdropping sound to evaluate the similarity from the time domain. 2) We train two basic LeNet-5 [18] to recognize the sound of numbers and letters based on the spectrogram, respectively.

In our experiment, we do not use a commercial speech recognition system, because the spectrogram captured by Tag-Bug is different from the traditional sound records. Therefore, we train our own LeNet-5 for speech recognition. We mix the records of all the volunteers together and separate the data according to the eavesdropping scenarios. Then we use 9/10 of the data to train the network and use the rest to test it. For a real adversary, other efficient methods such as transfer learning can be used to improve the recognition accuracy, which is beyond the scope of this paper.

7.1 Impact of Insulation between Reader and Loudspeaker

Tag-Bug can recognize the number with a macro averaged precision of 95.5% *in the free-space and with a macro averaged precision of* 86.3% *of through-the-brick-wall scenario and through-the-insulating-glass scenario.* We further use the class-wise averaged precision of spoken number recognition to evaluate the performance of the extracted sound. As shown in Figure 12(a), *Tag-Bug* can accurately recognize the numbers with over 95% precision in the free-space. Only 1 or 2 instances are incorrectly recognized out of all the 21 instances. Particularly, *Tag-Bug* can accurately recognize number '2' and '7' with 100%. When we eavesdrop the sound through a 20*cm-thick brick-wall*, we can still achieve an average precision of 85.2% in Figure 12(b). The recognition class-wise averaged precision of all numbers are over 75%.

Besides, we further move to the soundproof room and eavesdrop the sound through the *29mm-thick insulating-glass*. As shown in Figure 12(c), we can still achieve an average precision of 87.5%. All the numbers can be accurately recognized with over 80%. It indicates that the adversary can eavesdrop the user privacy through the different kinds of insulation. Moreover, since we train a comprehensive LeNet-5 for all the users for speech recognition, the speech related voice features can be efficiently captured by *Tag-Bug* for recognition, while the



(a) Spectrogram of male voice (b) Spectrogram of female voice (c) Recognition accuracy

Fig. 13. Case study of eavesdropping daily conversation.

individual characteristics such as voice fingerprint are ignored by the network. To improve the performance of speech recognition, advanced deep learning methods such as transfer learning can be used to make the recognition network suitable for different scenarios.

To explain the reason for the high accuracy, we further examine the similarity between the original sound played by the loudspeaker and the sound extracted by our system. The similarity is defined as the cross-correlation between the extracted sounds and the original sounds. We calculate the similarity of both the sound directly from the RFID tag and refined by the CGAN model. In comparison, we also use a smartphone to record the sound and calculate the similarity. We examine the similarity in both the free-space (FS) and the insulation scenarios (IS), including through the wall and the insulating glass. As shown in Figure 12(d), the CGAN can obviously improve the similarity with over 50% and 10% for the FS and IS, respectively. As for the smartphone, it is only 17% better than *Tag-Bug* for the FS. However, the smartphone is worse that *Tag-Bug* , whose similarity is only 38% of *Tag-Bug* due to the insulation of the sound. Thus, *Tag-Bug* can achieve good performance in eavesdropping, which guarantees the high accuracy in recognition.

7.2 Impact of Sound Volume

Tag-Bug *can achieve over* 60% *recognition accuracy when the volume is larger than* 60*dB.* We next linearly change the volume of the player and use a decibel meter to measure the absolute volume. To comprehensively examine the effectiveness, we use both top1 and top3 accuracy to evaluate the recognition of numbers. As shown in Figure 12(e), there is a large gap when the volume increases from 76 to 80dB. It indicates that once the sound is louder that 76dB, *Tag-Bug* can accurately recognize the sound with about 90% accuracy. Nevertheless, even when the volume decreases to 60dB, *Tag-Bug* can still achieve top3 accuracy of about 60%. It still gives the chance to speculate the privacy with high probabilities.

7.3 Recognition Accuracy of Letters

Tag-Bug can accurately recognize the letters of human voice with about 87% *class-wise averaged precision.* Next, we further use another LeNet-5 to recognize the spoken letters played by the loudspeaker with the same deployment in the free-space. As shown in Figure 12(f), *Tag-Bug* can accurately recognize all the 26 letters, where 23 letters achieve over 80% precision. Here, 'h' achieves only 50% precision, because the rest of 'h' is recognized as 'f' due to the similar syllables. Nevertheless, the adversary can still obtain the private information according to the similar syllables. It is *Tag-Bug* is able to eavesdrop the human voice of both the numbers and letters.

8 CASE STUDY

Experiment Settings: As shown in Figure 1(a), we deploy our system in a living room to eavesdrop the daily conversations. We let the loudspeaker play different kinds of human voice in a real living environment with ambient noise. For the sound source, we ask 2 male users and 2 female users to record 20 sentences with different pitches, *e.g.*, "do you have a dictionary", "would you please hand me the book", *etc.*, which have totally 84 words. Several RFID tags along with some reflection objects are placed in front of the loudspeaker, with different distances relative to the loudspeaker. We deploy *Tag-Bug* to eavesdrop the sound played by the loudspeaker, where the antennas are 3m away from the loudspeaker. We recover the sound and compare with the origin sound from both the time and frequency domain. Moreover, we ask 20 volunteers to manually recognize the extracted human voice based on the candidates of all the 20short sentences. The accuracy is defined as the number of correctly recognized sentences over all the extracted human voice for recognition.

Performance Evaluation: *Our system can efficiently eavesdrop the complicated human voice even in the noisy environments.* Figure 13(a) and Figure 13(b) use the spectrograms to show the similarity between the RFID sound and the origin sound, spoken by a male and a female, respectively. From the time domain, each word is correctly detected for both of the two users; from the frequency domain, the frequency band larger than 400Hz can be efficiently recovered by our CGAN. Here, our CGAN can efficiently recover most of these high-frequency bands for either the male with low pitch or female with high pitch. Although the spectrogram is not exactly the same as the original sound, CGAN provides an efficient way to recover the harmonic frequency bands for the human voice with multiple tones. Besides, we further ask 20 volunteers to manually recognize the extracted human speech and plot the accuracy as shown in Figure 13(c). The recovered human speech can be recognized with an average accuracy of 81.5% and a maximum accuracy of 94%. It indicates the high opportunity to eavesdrop the human voice.

9 POSSIBLE DEFENSES

To defend the eavesdropping via RFID tags, the possible ways are to physically block the transmission of either the sound from the loudspeaker to the tag or the RF-signals from possible tags to the reader.

The first way is to manually reduce the influence of the sound on the RFID tags. For example, by removing the potential tagged objects around the loudspeaker, users can avoid both the vibration effect and the reflection effect due to the loudspeaker, and thus defend the eavesdropping from the sensing source. Besides, the user can also decrease the volume of the loudspeaker or use an earphone, when receiving the private information. Since the power energy of the sound is much weaker, which can hardly vibrate the tags in these scenarios.

The second way is to manually increase the energy fading when the RF-signals are transmitting from the tag to the reader. Since the performance of *Tag-Bug* is mainly determined by the transmission environment, *e.g.*, the transmitting distance, the user can reduce the possibility of eavesdropping by increasing the propagation path. For example, keeping the loudspeaker away from the surrounding walls can lead to the larger transmitting distance and efficiently avoid thru-the-wall eavesdropping. Besides, deploying the shielding material in the wall can prevent the reader from interrogating the tags outside the room. Moreover, the user can deploy an RF jammer inside the meeting room, which can significantly affect the signal transmitting. It may block or affect the transmitting of the weak signal backscattered from the tag, and thus avoid thru-the-wall eavesdropping.

10 LIMITATION & DISCUSSION

Eavesdropping on live human speech. In this paper, we mainly focus on the loudspeaker. Nevertheless, we also try to eavesdrop on the live human speech by letting the user shouting at the tag. We find the tag is actually vibrated by the human speech, but the extracted sounds are very noisy. We speculate that the tag vibration is mainly caused by the air flow during human speech (*e.g.*, speaking number '2'), which leads to large wind noise. Eavesdropping on the live human speech can be a potential future work.

Tag-loudspeaker Distance. Since the COTS RFID tag is still relatively thick compared with the diaphragm of the microphone, the tag-loudspeaker distance should be less than 150cm to achieve the eavesdropping on human voice. Adversary can intentionally hide the tags closer to the loudspeaker to improve the eavesdropping or use a tag with ultra-light material.

Reader-Tag Distance. In our current system, we use USRP to collect the physical-layer signal for convenience, and the evaluated reader-tag distance is below $2m$ due to the power limitation of USRP. For a powerful reader, it is easy to increase the interrogation distance by increasing the transmitting power. Nevertheless, the eavesdropping principle is the same as *Tag-Bug* and our system can be easily extended to the longer distance.

Types of Materials. In a real eavesdropping scenario, the materials of surrounding objects can undoubtedly affect the eavesdropping. For example, when the tag is attached to different objects as shown in Figure 11(h), the eavesdropping performance is totally different due to the different resonance effects of these materials. Nevertheless, since the widely used objects, *e.g.*, plastic bag, express package or paper cup, usually have relatively good resonance effect, they can still pose a potential threat for the sound leakage. Besides, the materials of the separating walls can also affect the eavesdropping, due to the signal absorption of these materials. For example, walls with chicken wire in earthquake zones can block transmission of RF-signals and prevent the eavesdropping. However, our methods can usually work for the general walls, *e.g.*, wood walls or wet walls, since most RF-signal can penetrate such walls. As for the vibration of the walls due to the resonance, it is usually quite weak and can be ignored for sound eavesdropping.

Multiple Sensing Tags. In a real eavesdropping scenario, more than one tagged objects may exist together, leading to the uncertainty for sound eavesdropping. To solve this problem, the attacker can iteratively try to eavesdrop with each tag and choose the tag with the best sound performance for continuous eavesdropping. Since the vibration of each tag can be extracted via either the vibration effect or the reflection effect, the attacker need to try with both of the two mechanisms and choose the better one for eavesdropping. In this paper, we focus on the fundamental principle and the feasibility of RFID-based eavesdropping, and thus omit part of the implementation description.

11 RELATED WORK

RFID-based vibration detection. Recently, RFID has been widely used in indoor localization [24, 25, 38, 39, 42], trajectory tracking [15, 21, 30, 34] and activity recognition [11, 12, 22, 36]. But only a few works apply RFID to detect the vibration. TagBeat [40] makes the first attempt to measure the mechanical vibration based on one RFID tag. TagTwins [14] further improves TagBeat by leveraging the dual RFID tags to eliminate the interference of ambient noises. RF-Ear [41] extends device-based sensing scenarios to device-free scenarios. Even if these methods leverage the compressive reading to estimate the period, they are all limited by the COTS RFID system, and thus cannot sense the complicated human voice. TagSound [20] proposes the possibility of sensing the sound with the RFID tag by leveraging the harmonic backscattered signal. Since the harmonic signal is weaker than the raw UHF signal, the harmonic signal is difficult to achieve thru-the-wall eavesdropping.

Remote vibration detection. Due to the mature technique of radar system, the radar principle has been used to achieve the remote vibration detection for a long time. LADAR [9] first investigates the possibility of using radar signals for verifying the rotational speed of mechanical vibration systems. Davis *et al.* [10] further leverage a high-speed camera to estimate the vibration of everyday objects, *e.g.*, the plant leaves, based on the principle of LADAR. But the vision-based attacks always suffer from the line-of-sight communication problem. Inspired by LADAR, Wei *et al.* [37] further propose ART, which estimates the sound spectrum based on the WiFi signal, which requires a MIMO antenna array. Moreover, the WiFi signal is easily affected by other surrounding vibration objects. Kwong *et al.* [17] use the magnetic hard disk to eavesdrop, which needs to gain the access to the target hard disk. In comparison, UHF RFID systems work at 920MHz with battery-less RFID tags, which can penetrate the wall with less energy loss compared with the WiFi signals.

Acoustic eavesdropping. Recently, there have been active research efforts on sensing human activities, especially the keystroke, via the acoustic eavesdropping. In particular, Asonov *et al.* [7] leverage a supervised learning method to recognize the keystroke. Ubik [35] locates the keystrokes on the solid surface by leveraging the multi-path fading with the machine learning method. Liu *et al.* [23] leverage the dual microphones on the smartphone to locate the keystrokes with the TDoA methods. Gyrophone [26] leverages the signal changes of the gyroscope to infer the coarse information of the speaker. Li *et al.* [19] use mmWave to capture the vibration of throat for user authentication. Instead of sensing human activities from the sound, we directly perform eavesdropping via RFID tags.

12 CONCLUSION

In this paper, we explore the possibility of eavesdropping the human voice from COTS RFID tags and present *Tag-Bug*, a battery-less approach for the thru-the-wall eavesdropping attack via COTS RFID tags. The key challenge and principle lie in extracting and amplifying the vibration features due to the sound. We investigate a novel feature, *i.e.*, *Modulated Signal Difference (MSD)* from the signal model, which can amplify the influence of the vibration on the received signal from either the *vibration effect* or the *reflection effect*. To achieve the full-frequency band human voice, we propose a Conditional Generative Adversarial Network to recover the high-frequency band by referring to the low-frequency band. Real-world experiments show that *Tag-Bug* can successfully capture the monotone sound when the loudness is larger than 60dB. *Tag-Bug* can efficiently recognize the numbers of human voice with 95.3%, 85.3% and 87.5% precision in the free-space eavesdropping, thru-the-brick-wall eavesdropping and thru-the-insulating-glass eavesdropping, respectively. *Tag-Bug* can also accurately recognize the letters with 87% precision in the free-space eavesdropping.

ACKNOWLEDGMENTS

This work is supported in part by National Natural Science Foundation of China under Grant Nos. 61902175, 61832008, 61872174, 61872173, 61802169; JiangSu Natural Science Foundation under Grant Nos. BK20190293, BK20180325 and the Key K&D Program of Jiangsu Province under Grant BE2020001-3. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] 2018. USRP SDR reader. https://github.com/nkargas/Gen2-UHF-RFID-Reader.
[2] 2019. EPC Gen2, EPCglobal. https://www.gs1.org/epcglobal.
[3] 2019. Impinj,Inc. http://www.impinj.com/.
[4] 2019. USRP reader. https://github.com/nkargas/Gen2-UHF-RFID-Reader.
[5] Zhenlin An, Qiongzheng Lin, and Lei Yang. 2018. Cross-Frequency Communication: Near-Field Identification of UHF RFIDs with WiFi!. In *Proc. of ACM Mobicom*.
[6] S Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1000–1017.
[7] Dmitri Asonov and Rakesh Agrawal. 2004. Keyboard acoustic emanations. In *Proceedings of IEEE Symposium on Security and Privacy*. IEEE, 3–11.
[8] Michael Backes, Markus Dürmuth, Sebastian Gerling, Manfred Pinkal, and Caroline Sporleder. 2010. Acoustic Side-Channel Attacks on Printers.. In *USENIX Security symposium*. 307–322.
[9] P Castellini, M Martarelli, and EP Tomasini. 2006. Laser Doppler Vibrometry: Development of advanced solutions answering to technology's needs. *Mechanical systems and signal processing* 20, 6 (2006), 1265–1285.
[10] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Frédo Durand, and William T Freeman. 2014. The visual microphone: passive recovery of sound from video. (2014).
[11] Han Ding, Chen Qian, Jinsong Han, Ge Wang, Zhiping Jiang, Jizhong Zhao, and Wei Xi. 2016. Device-free detection of approach and departure behaviors using backscatter communication. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 167–177.
[12] Han Ding, Longfei Shangguan, Zheng Yang, Jinsong Han, Zimu Zhou, Panlong Yang, Wei Xi, and Jizhong Zhao. 2015. Femo: A platform for free-weight exercise monitoring with rfids. In *Proc. of ACM SenSys*.
[13] Daniel M Dobkin. 2012. *The rf in RFID: uhf RFID in practice*. Newnes.

[14] Chunhui Duan, Lei Yang, Huanyu Jia, Qiongzheng Lin, Yunhao Liu, and Lei Xie. 2018. Robust spinning sensing with dual-rfid-tags in noisy settings. *Proc. of IEEE INFOCOM.*

[15] Chengkun Jiang, Yuan He, Xiaolong Zheng, and Yunhao Liu. 2018. Orientation-aware RFID tracking with centimeter-level accuracy. In *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks.* IEEE Press, 290–301.

[16] Wolfgang Klippel and Joachim Schlechter. 2006. Measurement and visualization of loudspeaker cone vibration. In *Audio Engineering Society Convention 121.* Audio Engineering Society.

[17] Andrew Kwong, Wenyuan Xu, and Kevin Fu. 2019. Hard Drive of Hearing: Disks that Eavesdrop with a Synthesized Microphone. In *2019 IEEE Symposium on Security and Privacy (SP).* IEEE, 905–919.

[18] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[19] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems.* 312–325.

[20] Ping Li, Zhenlin An, Lei Yang, and Panlong Yang. 2019. Towards Physical-Layer Vibration Sensing with RFIDs. In *Proc. of IEEE INFOCOM.*

[21] Jia Liu, Min Chen, Shigang Chen, Qingfeng Pan, and Lijun Chen. 2017. Tag-compass: Determining the spatial direction of an object with small dimensions. In *Proceedings of IEEE Conference on Computer Communications (INFOCOM).* 1–9.

[22] Jia Liu, Xingyu Chen, Shigang Chen, Xiulong Liu, Yanyan Wang, and Lijun Chen. 2019. TagSheet: Sleeping posture recognition with an unobtrusive passive tag matrix. In *Proceedings of IEEE Conference on Computer Communications (INFOCOM).* 874–882.

[23] Jian Liu, Yan Wang, Gorkem Kar, Yingying Chen, Jie Yang, and Marco Gruteser. 2015. Snooping keystrokes with mm-level audio ranging on a single phone. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking.* ACM, 142–154.

[24] Jia Liu, Feng Zhu, Yanyan Wang, Xia Wang, Qingfeng Pan, and Lijun Chen. 2017. RF-scanner: Shelf scanning with robot-assisted RFID systems. In *Proceedings of IEEE Conference on Computer Communications (INFOCOM).* 1–9.

[25] Yunfei Ma, Nicholas Selby, and Fadel Adib. 2017. Minding the billions: Ultra-wideband localization for deployed rfid tags. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking.* ACM, 248–260.

[26] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing speech from gyroscope signals. In *Proc. of 23rd USENIX Security Symposium (USENIX Security 14).* 1053–1067.

[27] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv 2014. *arXiv preprint arXiv:1411.1784* (2014).

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention.* Springer, 234–241.

[29] Pascal Scalart et al. 1996. Speech enhancement based on a priori signal to noise estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing Conference*, Vol. 2. IEEE, 629–632.

[30] Longfei Shangguan and Kyle Jamieson. 2016. Leveraging Electromagnetic Polarization in a Two-Antenna Whiteboard in the Air. In *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies.* ACM, 443–456.

[31] Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.

[32] Michael Vorländer. 2007. *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality.* Springer Science & Business Media.

[33] Chuyu Wang, Lei Xie, Wei Wang, Tao Xue, and Sanglu Lu. 2016. Moving Tag Detection via Physical Layer Analysis for Large-Scale RFID Systems. In *Proc. of IEEE INFOCOM.*

[34] Jue Wang, Deepak Vasisht, and Dina Katabi. 2015. RF-IDraw: virtual touch screen in the air using RF signals. In *Proc. of ACM SIGCOMM.*

[35] Junjue Wang, Kaichen Zhao, Xinyu Zhang, and Chunyi Peng. 2014. Ubiquitous keyboard for small mobile devices: harnessing multipath fading for fine-grained keystroke localization. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services.* ACM, 14–27.

[36] Zhongqin Wang, Fu Xiao, Ning Ye, Ruchuan Wang, and Panlong Yang. 2018. A see-through-wall system for device-free human motion sensing based on battery-free RFID. *ACM Transactions on Embedded Computing Systems (TECS)* 17, 1 (2018), 6.

[37] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proc. of ACM Mobicom.* ACM, 130–141.

[38] Teng Wei and Xinyu Zhang. 2016. Gyro in the air: tracking 3D orientation of batteryless internet-of-things. In *Proc. of ACM Mobicom.* 55–68.

[39] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices. In *Proc. of ACM MobiCom.*

[40] 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 *Proc. of ACM Mobicom.* ACM, 16–28.

[41] 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 *Proc. of IEEE INFOCOM.*

[42] Yilun Zheng, Yuan He, Meng Jin, Xiaolong Zheng, and Yunhao Liu. 2018. RED: RFID-based eccentricity detection for high-speed rotating machinery. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1565–1573.