mmEavesdropper: Signal Augmentation-based Directional Eavesdropping with mmWave Radar

Yiwen Feng, Kai Zhang, Chuyu Wang, Lei Xie, Jingyi Ning, Shijia Chen
State Key Laboratory for Novel Software Technology, Nanjing University, China
Email: {ywfeng, mg1933091, ningjiy, MF20330008}@smail.nju.edu.cn, {chuyu, lxie}@nju.edu.cn

Abstract—With the popularity of online meetings equipped with speakers, voice privacy security has drawn increasing attention because eavesdropping on the speakers can quickly obtain sensitive information. In this paper, we propose mmEavesdropper, a mmWave based eavesdropping system, which focuses on augmenting the micro-vibration signal via theoretical models for voice recovery. Particularly, to augment the receiving signal of the target vibration, we propose to use beam-forming to facilitate the directional augmentation by suppressing other orientations and use Chirp-Z transform to facilitate the distance augmentation by increasing the range resolution compared with traditional FFT. To augment the vibration signal in the IQ plane, we build a theoretical model to analyze the distortion and propose a segmentation-based fitting method to calibrate the vibration signal. To augment the spectrum for sound recovery, we propose to combine multiple channels and leverage an encoder-decoder based neural network to reconstruct the spectrogram for voice recovery. We perform extensive experiments on mmEavesdropper and the results show that mmEavesdropper can reach the accuracy of 93% on digit and letter recognition. Moreover, mmEavesdropper can reconstruct the voice with an average SNR of 5dB and peak SNR of 17dB.

Index Terms—mmWave, Sound vibration, Signal augmentation, Eavesdropping

I. INTRODUCTION

Acoustic eavesdropping is regarded as one of the critical privacy leakage problems because the human voice is the easiest way to obtain sensitive information. On one hand, intelligent voice assistant systems are widely deployed in people’s homes, which extend the usage of speakers and bring new risks of being attacked. For example, the Remember function of Google home may replay passwords, private messages, and schedules, which may be leaked and cause privacy problems. On the other hand, with the growth of remote cooperation in international companies and the prevalence of working from home during the COVID-19 pandemic, online meetings have been widely used, where different kinds of confidential information may be mentioned during the meeting. The speaker plays these sensitive voices during the meetings and is probably leaked to the adversary. For instance, information including trade secrets and passwords may be leaked when someone discusses them in online meetings, which can bring unpredictable dangers of economic losses and even personal safety.

Traditionally, acoustic eavesdropping is based on an embedded wiretapping device, which can record the original sounds. However, it is affected by environmental noise and other sound. Researchers have recently proposed new approaches to sense the sound source’s vibration for eavesdropping directly.

Sensor-based approaches use the vibration effect caused by sound pressure to reconstruct the voice [1]–[3]. However, these approaches are limited by the contact sensing manner, and usually, they can only realize audio classification instead of voice reconstruction. Vision-based approaches [4] use high-speed cameras to capture the sound-related vibration of tiny objects for sound reconstruction. However, it is vulnerable to illumination and can be obscured. Wireless-based approaches have gained more interest in recent years. Researchers use the reflected signal pattern of RFID [5] [6] and mmWave [7] to sense the sound vibration. However, most work can only classify specific words instead of reconstructing the sound, which has obvious limitations. MILLIEAR [7] and mmPhone [8] realize audio reconstruction based on mmWave. However, they focus on directly using the original mmWave signal for reconstruction with a well-trained neural network, which usually depends on the specific sensing environment.

In this paper, we propose mmEavesdropper, a mmWave-based approach to eavesdrop on the speaker’s sound based on the received signal. Compared with the traditional wiretap, mmEavesdropper can leverage the multiple receiving antennas to perform beam-forming such that only the vibration of the target orientation can be captured. Specifically, we focus on augmenting the vibration caused by the sound. We model the sound vibration and optimize the Frequency Modulated Continuous Wave (FMCW) signal data processing algorithm with Commercial Off The Shelf (COTS) mmWave radar. We augment the vibration effect from the following three aspects. Firstly, to obtain the exact vibration signal from the mmWave signal, we augment the received signal by focusing on the orientation and distance of the target vibration membrane with
the beam-forming technique and Chirp-Z transform. Secondly, to extract the exact signal feature from the vibration signal, we augment the vibration signal by proposing an IQ-based distortion model and a segmentation-based fitting method to calibrate the vibration signal. Finally, to completely recover the frequency spectrum of the sound, we propose to fuse the vibration signal feature of multiple channels and train an encoder-decoder network to reconstruct the sound.

There are three key challenges to be addressed in this paper. The first challenge is to accurately extract the vibration signal from the raw received signal. Traditionally, the Fast Fourier Transform (FFT) method is used to determine the distance of the vibrating membrane. After that, we can extract the corresponding range bin signal. However, due to the limitation of sampling rate, the resolution of such method is only 4 cm, which is not accurate enough to extract the exact vibration signal. To address this challenge, we propose to apply the Chirp-Z transform (CZT), which can efficiently increase the resolution of the FFT algorithm without increasing the computation overhead significantly. Instead of getting the uniform spectrum over the whole frequency band, CZT can define the target frequency band within a specific range and increase the sampling precision of based on sparse point FFT. As a result, we can get higher resolution for the frequency spectrum of the target band, such that we can accurately determine the exact distance of the target with higher resolution.

The second challenge is to calibrate the signal distortion due to the multipath effect. Theoretically, the signal samples tend to rotate on an arc-like trace in the IQ plane. However, according to our empirical study, the vibration signals in the IQ plane are distorted due to the complicated multi-path effect and approximate treatment of peak frequency in CZT. Such distortion can import massive noise on both the phase and strength of the vibration signal, making the vibration signal’s spectrum different from the original human voice. To address this challenge, we build a theoretical model of the signal in the IQ plane, which can quantitatively derive the phase error due to the Fourier transform and strength error due to the variation of the multi-path effect. Based on this model, we propose a segmentation-based fitting method of phase augmentation, which can efficiently divide the signal into two parts according to the model. After that, each part is a complete arc consisting of the fundamental vibration, and we can leverage the phase of the arc to estimate the vibration.

The third challenge is to augment the frequency spectrum due to the frequency response issue of the speaker. Usually, low-frequency sound has a larger vibration amplitude for the same loudness than the high-frequency sound due to the specific frequency response curve of the speaker [9]. As a result, since the vibration signal measured from the mmWave represents the vibration amplitude, the high-frequency band has weaker power than the low-frequency band, although the high-frequency vibration is also captured. To address this challenge, we propose a channel fusion-based learning method to compensate for the frequency band. Although the vibration signal of the target bin is more accurate, we find that the vibration spectrum also appears in the vibration of the nearby range bins. Therefore, we propose to comprehensively consider the vibration signal of multiple channels, which all contain sound vibration. After that, we leverage an encoder-decoder network to learn the frequency features from these vibration signals and then reconstruct the human voice based on these features. Based on the above mechanism, we can efficiently recover the human voice from multiple channels of the vibration signal.

In this paper, we make the following contributions. First, we comprehensively investigate the capability of sound eavesdropping with mmWave radar and propose a novel system mmEavesdropper to sense the sound vibration based on the augmentation of mmWave signal. Second, we propose theoretical models to augment the receiving signal by focusing on the accurate orientation and distance of the vibration target, augment the vibration signal by calibrating the IQ signal distortion, and augment the frequency spectrum by fusing multiple channels with high-frequency compensating. The proposed models can provide essential insights to extract and study the micro-vibration based on the COTS mmWave radars. Third, we implement a real system prototype based on COTS mmWave radar and train an encoder-decoder based neural network to recover the human voice. According to extensive experiments, our system can efficiently eavesdrop on the human voice with 93% accuracy.

II. RELATED WORK

Sensor-based Approach. Sensor-based eavesdropping approaches mainly convert vibration displacement caused by audio signal to an electrical signal to reconstruct the audio signal. Researchers propose to use mobile phone built-in sensors such as gyroscope [10], accelerometer [11] and Inertial Measurement Unit (IMU) [2] [12]. These approaches utilize the vibration effect caused by sound pressure to recognize the sound. Kwong et al. [3] use disk drives to eavesdrop. However, these approaches require contact sensing and have low sampling rates. Additionally, they only classify audio signals rather than recovering the signals.

CV-based Approach. Approaches based on Computer vision (CV) mainly leverage high-speed cameras and lidars. Researchers use cameras to capture the vibration of tiny objects in the environment caused by sound waves and apply CV algorithms to videos or image sequences [4]. Some researchers utilize customized lidar sensors to sense the audible vibration [13] [14]. For example, Sami et al. [1] propose an eavesdropping scheme using lidar equipped on a cleaning robot. Specifically, thin and light objects such as clothes and plastic bags vibrate due to the audio source, and lidars can scan the vibration to restore sound. Unfortunately, these perception methods depend on the illumination condition, and the sensing target must be in a non-line-of-sight scene. Meanwhile, the sampling rate of cheap household lidars is often low, which reduces their availability.

Wireless-based Approach. Wireless signals provide inspiring solutions for sensing [15]–[26]. The mechanism of the
Wireless-based approach is that the vibration target modulates the reflected signals, and the signal processing algorithms can demodulate the original sound signal. There are many solutions based on centimeter-wave. Wei et al. [27] use WIFI beam-forming technology with an antenna array to sense speaker vibration and realize eavesdropping through walls. Some researchers use RFID signals to sense the vibration of objects [28] [29]. Li et al. [6] propose that RFID tags can use harmonic signals to sense single-frequency sound and improve the single-frequency vibration sensing ability of RFID to 2.5kHz. In recent years, the development and application of millimeter-wave have provided new possibilities for eavesdropping. Xu et al. [30] realize human voice reconstruction through a customized mmWave radar. Wavoice [31] use the commercial mmWave radar to capture information in the human voice and recognize the speech.

Unlike the previous work, mmEavesdropper is implemented on a portable COTS device and can realize directional sensing. It has no limit of illumination and can work in dark scenarios. Besides word classification, mmEavesdropper can recover sound with high SNR, which is human recognizable.

III. EMPIRICAL STUDY

In order to investigate the sensing ability of mmWave, we execute some empirical experiments to extract the sound vibration from the mmWave signal. As shown in Fig. 2(a), we place the mmWave radar in front of the speaker at 1m and try to leverage the radar to sense the sound vibration. Similar to the existing solutions [8] [7] [31], we first conduct range-FFT to determine the range bin of the target vibrating membrane and then extract the sound vibration from the phase of the signal belonging to the target bin.

A. Feasibility of sound sensing

Observation 1. The mmWave can clearly capture the monotone sound vibration from the frequency spectrum with much noise, but it is challenging to capture the clear human voice with a complicated frequency spectrum.

We first directly verify the sound-capturing ability of the mmWave by playing two monotone sounds of 300Hz and 500Hz with the speaker. We extract the target bin based on the strongest reflection signal power by applying an FFT and present the time-frequency spectrogram by applying Short-time Fourier transform (STFT) on the signal phase. The spectrograms are shown in Fig. 2(b) and Fig. 2(c). There are two predominant frequencies at the corresponding values, meaning the potential for audio sensing. However, there are also harmonic wave interference and noise interference, which can severely affect the sensing performance. Hence, it is necessary to filter the noise to obtain a stable vibration signal.

Moreover, we further try to sense the human voice with a complicated frequency spectrum played by the speaker. As shown in Fig. 2(d), we apply STFT to the received signal phase and present the time-frequency spectrogram. Compared to the original voice signal, as shown in Fig. 2(e), it is evident that the mmWave radar has captured the human voice. Notably, due to the constraint of the sampling rate, the captured frequency is below 2.5 kHz, which is still sufficient to cover the human voice. However, there is also a lot of noise interference on the spectrogram, which unfortunately submerges the vibration signal completely. As a result, the spectrogram of human voices is unclear, so removing such noise is necessary.

B. Interference related range bin estimation

Observation 2. The estimated bin of the target vibration based on the range-FFT can only determine the coarse position of the target, leading to the noise of the signal phase.

We firstly play a monotone sound on the speaker and apply FFT on the captured signal. Then we get the bin with the strongest reflection, corresponding to 1.23m. The Signal Noise Ratio (SNR) of the signal phase in this bin is 12.15dB. After that, we move the speaker 2cm away from the radar and repeat the experiment above. The result indicates that the strongest reflection still occurs at 1.23m, which is equal to the initial position. However, after a 2cm movement, the SNR of the signal in the target bin reduces to 9.35dB. The experiments show that range-FFT cannot catch this change in the frequency domain when we move the speaker. Furthermore, a coarse range bin also reduces the SNR of the phase frequency, which will significantly damage the performance of sound reconstruction.

However, the radar resolution is 4cm, which is relatively ambiguous. As a result, signals from that wide range bin can be mixed up with some irrelevant components of other objects, leading to interference to the target voice portion. Therefore, it is essential to find an efficient method to confirm the speaker’s location accurately.

C. Phase distortion

Observation 3. The extracted signal does not ideally rotate as an arc in the IQ plane due to the multi-path effect.

In addition to the accurate distance of the target vibrating membrane, we further investigate the signal pattern of the target range bin. According to the theoretical models [32] [30], as the speaker vibrates back and forth, the signal tends to rotate on an arc on the IQ plane ideally. However, our experiments find that the signal variation pattern is not always an arc, as shown in Fig. 2(f). Since the phase changes are inherently sensitive, especially for the slight vibration estimation, such distortion can dramatically destroy the sound sensing. Hence, it is essential to determine the reason for this distorted pattern and remove them.

D. Signal in multiple channels

Observation 4. Signals of different channels show associated patterns, which can be used to sense the micro-vibration comprehensively.

The mmWave radar in the experiment has multiple Receiving (RX) antennas. Usually, only one channel’s data is used, while the others are be discarded, which may lose much vibration information. As shown in Fig. 2(g) and Fig. 2(h), signals from two RX antennas have associated patterns, although they still have some differences. Therefore, it is necessary to take
advantage of all the RX antennas and propose a fusion method to combine all the valuable vibration information to sense the micro-vibration of the sound.

IV. MILLIMETER-WAVE EAVESDROPPING MODEL

This section introduces the theoretical model of vibration sensing based on the mmWave radar. Mainly, we introduce the CZT (Chirp-Z transform) to optimize the resolution of range-FFT and then analyze the IQ distortion from the view of FFT calculation and multi-path effect.

A. Basic Vibration Sensing Model

The mmWave radar leverages the FMCW (Frequency-Modulated Continuous Wave) technology, which transmits a sinusoid signal called “chirp” with linearly increasing frequency. A chirp is transmitted from the transmitting (TX) antenna, then reaches an object with distance $R(t)$ and bounces back to the receiving (RX) antenna. We compare the RX signal with the TX signal to estimate the distance. The TX signal $s_{TX}$ and RX signal $s_{RX}$ can be expressed as [33] [34]:

$$s_{TX}(t) = \exp \left[ j(2\pi f_c t + \pi K t^2) \right],$$
$$s_{RX}(t) = \alpha s_{TX}[t - 2R(t)/c],$$

(1)

where $f_c$ and $K$ are the starting frequency and the chirp slope of the FMCW signal, respectively, $\alpha$ is the path loss. Then a mixer is used to generate the Intermediate Frequency (IF) signal as [34]:

$$s_{IF}(t) = s_{TX}(t) \cdot \text{conj}[(s_{RX}(t))]$$

$$= \alpha \exp \left[ 2\pi j f_c R(t)/c + 2KR(t)/c \cdot t \right],$$

(2)

where conj$[(s_{RX}(t))]$ means conjugate of $s_{RX}(t)$. Therefore, we can use the Fourier Transform on the IF signal to get the target frequency [35] as:

$$S_{IF}(f) = F(s_{IF}(t)) = A\delta(f - 2KR(t)/c),$$
$$A = \alpha \exp \left[ j4\pi f_c R(t)/c \right],$$

(3)

where $\delta$ is Dirac Delta function [36]. It means that when $f = 2KR(t)/c$, there is a frequency peak. Then we substitute $f = 2KR(t)/c$ into Eq. (3) and get the signal of the corresponding frequency as $\alpha \exp \left[ j4\pi f_c R(t)/c \right]$. It indicates that for the IF signal, there is a linear connection between the signal frequency and the reflection distance $R(t)$. We can extract the signal of corresponding frequency to represent the signal of the distance.

If the reflected object is vibrating, i.e., $R(t) = R_0 + d(t)$, where $R_0$ is the static distance and $d(t)$ is the vibrating function related to time $t$. Therefore, for each chirp, we can set $f = 2KR(t)/c$ to get the signal of frequency peak, and the continuous chirp can form a continuous signal $x(t)$ related to the vibrations as:

$$x(t) = \alpha \exp \left[ j4\pi f_c R(t)/c \right].$$

(4)

According to Eq. (4), when $R = R_0 + d(t)$, the vibration $d(t)$ can lead to the phase change of $x(t)$, where the phase $\theta(t) = \delta \cdot d(t)$. Therefore, we can first estimate the frequency peak to determine $R_0$ and then extract $x(t)$ to perceive the sound vibration.

B. Optimization of Target Detection

Based on the above model, it is vital to accurately estimate the vibration distance, i.e., $R_0$. Conventional methods use
FFT to determine the target range bin where the vibration is located. However, the locating accuracy is limited because the resolution of FFT is determined by the sampling rate and data size of a chirp according to Nyquist sampling theory [37]. For the IF signal $s_{IF}(t)$ with $N$ samples, the formula of FFT can be presented as [38]:

$$S_{IF}(k) = \sum_{t=0}^{N-1} s_{IF}(t)e^{-j2\pi k/N},$$  \hspace{1cm} (5)

where $k$ is the frequency index. Supposing $F_s$ is the sampling rate, the FFT method divides the target frequency band $[-F_s/2, F_s/2]$ into $N$ bins, where the resolution is $F_s/N$. This bin is called a "range bin". Therefore, if the actual frequency of IF signal for an object is $f_{actual}$ and the corresponding range bin frequency estimated from the peak of FFT is $f_{FFT}$, it means:

$$f_{actual} \in [f_{FFT} - \frac{F_s}{2N}, f_{FFT} + \frac{F_s}{2N}).$$  \hspace{1cm} (6)

Therefore, it is inaccurate to directly use the estimated frequency $f_{FFT}$ to indicate the location. According to the empirical experiment result, inaccurate location will lead to noise in the signal.

To solve the problem, we import Chirp-Z transform (CZT) [39], which can sufficiently improve the resolution. CZT can zoom out the spectrum of a specific frequency band to improve the resolution. For the same IF signal $s_{IF}(t)$, we can calculate the CZT as:

$$S_{CZT}(k) = \sum_{t=0}^{N-1} s_{IF}(t)G^{-t}W^tk, \hspace{1cm} k = 0, \ldots, M - 1$$  \hspace{1cm} (7)

where $G = G_0e^{j2\pi \theta_0}$ and $W = W_0e^{j2\pi \phi_0}$. $\theta_0$ and $\phi_0$ are the unit circle’s start angle and angular spacing. Based on Eq. (7), CZT can be considered as M-points sampling of original FFT result, whose starting frequency and interval are $G = G_0e^{j2\pi \theta_0}$ and $W = W_0e^{j2\pi \phi_0}$, respectively. When $G = 1$ and $W = e^{-j2\pi k/N}$, CZT degenerates into FFT.

Compared to FFT, CZT only focuses on a specific frequency band instead of the whole frequency band, which guarantees precision and also reduces computational complexity.

In our scenario, we first determine the range bin according to the FFT peaks as $f_{FFT}$. In order to find the more precise peak in the fit range bin, we set the range of CZT as $[f_{FFT} - \frac{2F_s}{N}, f_{FFT} + \frac{2F_s}{N}]$. Now, we use M-points CZT to search for the peaks in the spectrum, which can improve the resolution $M$ times compared with the FFT. For example, in Fig. 3, the result of FFT indicates the speaker is at 1m. Nevertheless, the peak coverage of FFT is coarse, but CZT can find a more fine-grained result within an FFT range bin, which is 1.018m.

### C. Signal Distortion in the IQ Plane

After determining the range bin of the vibrating membrane, we can extract the corresponding signal related to the distance of the vibrating membrane for sound sensing. The basic idea is to extract the signal $x(t)$ based on the estimating range bin from the CZT method and then estimate the vibration based on $x(t)$ in the IQ plane. Usually, the received signal of $x(t)$ is a superposition of the vibration signal $s_{membrane}$ from the vibrating membrane and the signal of the other static reflections $s_{static}$ as shown in Fig. 4(a):

$$x(t) = s_{membrane} + s_{static}.$$  \hspace{1cm} (8)

Since only $s_{membrane}$ varies the phase due to the vibrating, the signal moves along an arc in the IQ plane as shown in Fig. 4(b). Here, the center of the arc is related to $s_{static}$ instead of the original signal, and the variation phase of $x(t)$ is related to the vibration displacement $d(t)$. Thus, the traditional method can leverage the fitting method to estimate the center and extract the vibration displacement accordingly.

However, due to reciprocating motion during vibrating, the position of the vibrating membrane is not static at the original range bin but may slightly move out of the range bin. Thus, the signal directly extracted from the range bin may be distorted, which cannot accurately reflect the vibration. Next, we analyze this distortion and propose to resolve the problem.

1) Phase Distortion: For the phase distortion, we find it is caused by approximate treatment in signal processing. For simplicity, we take the FFT as an example to analyze the distortion, which can be extended to the CZT.

Conventionally, before applying the FFT on the IF signal to get the frequency spectrum, we need to add a rectangular window on the IF signal. Therefore, instead of the theoretical Fourier Transform function of Eq. (3), the actual frequency spectrum of FFT is:

$$S_{IF}(f') = F(s_{IF}(t) \cdot W_T(t)) = F(s_{IF}(t)) \ast F(W_T(t)) = [A\delta(f - 2KR(t)/c) \ast T\text{sinc}(fT) \exp(-j\pi fT)],$$  \hspace{1cm} (9)

where $W_T(t) = \begin{cases} 1, & 0 \leq t < T \\ 0, & \text{else.} \end{cases}$

Here, $\ast$ means convolution. It represents multiplication in the time domain and convolution in the frequency domain. $W_T(t)$ is the rectangular window, and $\text{sinc}(f) = \frac{\sin(\pi f)}{\pi f}$ is sinc function for calculating the spectrum of $W_T(t)$ [40]. Moreover, since a convolution with Dirac Delta function $\delta$ is equal to graph translation on coordinates [41], we can write Eq.(9) as:

$$S_{IF}(f') = A\text{sinc}(f - 2KR(t)/c) \exp(-j\pi(f - 2KR(t)/c)).$$  \hspace{1cm} (10)

Here, if we can accurately calculate the value of $R(t)$ from $f$ regardless of the variation of $R(t)$ due to the vibration, i.e.,
f = 2KR(t)/c, the underline part in this equation is equal to 1 and Eq. (10) degenerates to Eq. (3). However, since we need to fix the range bin to investigate the signal change for vibration sensing, i.e., f = f_{static} = 2KR_0/c, the distance R(t) is continuously changing due to the vibration, which is not always matching the frequency. Then, by substituting f = f_{static} and R = R_0 + d(t) into Eq.(10) and we get:

\[ S_{dist}(f)' = \alpha' \exp \left[ \frac{4\pi R_0}{\lambda} + \frac{4\pi d(t)}{\lambda} + \frac{2\pi TK d(t)}{c} \right]. \]  

(11)

Here \( \alpha' \) is a generalization of coefficients, which varies in a small range around one related to d(t). Therefore, we can easily omit the variation of amplitude. Under this circumstance, the actual phase \( \theta_{real} = \theta_{ideal} + \frac{2\pi TK d(t)}{c} \). Compared to the ideal phase \( \theta_{ideal} \), the additional error is:

\[ \Delta \theta = \frac{2\pi TK d(t)}{c}. \]  

(12)

In our system, \( T \cdot K = 4\text{GHz} \) represents the radar band. For a speaker’s micro-vibration \( d(t) = 5 \text{mm} \), the phase error ranges from 0.1\pi \sim 0.12\pi. Therefore, the phase distortion can be up to 21.6^\circ, which is non-negligible.

2) Amplitude Distortion: When the speaker plays the sound, not only the vibrating membrane but also the body of the speaker is vibrating. As a result, the vibration signal is the superposition of all the vibration signals:

\[ x(t)' = s_{static} + s_{membrane} + s_{speaker}. \]  

(13)

Since the reflection coefficient of the speaker depends on the multi-path effect and the sound vibration, it may either enhance or reduce the strength of the raw vibration signal \( x(t) \), leading to amplitude distortion. As shown in Fig. 5, the extra interference of \( s_{speaker} \) will affect vibration signal \( x(t) \), and make both the amplitude and the rotation radius either increase in Fig. 5(a) or decrease in Fig. 5(b). Moreover, the amplitude of \( s_{speaker} \) is related to the vibrating membrane, which means that when the phase distortion is more significant, the amplitude distortion is also more considerable.

3) Analysis of Signal Distortion: To verify the above model, we have collected different sounds played by the speaker and summarized three kinds of signal distortion in the IQ plane as shown in Fig. 6. We find that instead of varying along an arc theoretically, the practical signal in the IQ place may show different patterns. To analyze that pattern, we try to use the above model to explain the distortion in Fig. 6. Particularly, we define \( P_{static} \) as the stationary point, when \( d(t) = 0 \). Based on the stationary point and the surrounding
interference from other orientations, we firstly exploit the beam-forming technology to focus on the target orientation. Beam-forming can help to direct the signal and improve the amplitude of the RX signals. According to the layout of the antennas and the rule of beam-forming, we can combine the multiple channels with different coefficient weights. For example, as shown in Fig. 7, the distance between the adjacent RX antennas is $d_a$ and the incidence angle of the receiving signal is $\beta$. Therefore, there is a distance difference $\Delta d = d_a \cos \beta$ during propagation between the adjacent antennas, which will cause a phase difference. Hence, beam-forming can manually reduce the phase difference with the specific coefficient and then combine the signal of multiple channels as:

$$s_{BF}(t) = \sum_{i=1}^{N} s_{IF,i}(t) \exp \left( -j \frac{d_a \sin \beta}{\lambda} \times (i-1)\pi \right), \quad (14)$$

where $N$ is the number of antennas, i.e., 4 in our system. By applying beam-forming, we can not only focus on the target orientation, but also combine the multiple channels to augment the IQ distortion based on the stationary point $P_{\text{static}}$. The basic deviation idea is that the angle from each point in $\mathbb{P}$ to $P_{\text{static}}$ is opposite when the points are belong to different parts. Therefore, we firstly move all the points in $\mathbb{P}$ to origin $O$ by subtracting $P_{\text{static}}$ as: $\mathbb{P}' = \mathbb{P} - P_{\text{static}}$. Then, we leverage a clustering method to divide these points based on the idea that two points $P_1$, $P_2$ belong to the same part if $\angle P_1OP_2 < \delta$ as shown in Fig. 8(a). Here, $\delta$ is an angle threshold, which is set to 45° in practice. Finally, we can get two point set $\mathbb{P}_p$ and $\mathbb{P}_n$, representing the positive and negative part, respectively.

After getting the two part, we apply the least square method on each part to fit the 2 arcs, which can get the circle center $O_p$ and $O_n$ as shown in Fig. 8(b). Then the vibration signal related to the sound is extracted as:

$$\theta_{\text{vib}}(t) = \begin{cases} 
\angle O_p OP, & P \in \mathbb{P}_p \\
-\angle O_n OP, & P \in \mathbb{P}_n 
\end{cases} \quad (15)$$

D. Channel Fusion-based Voice Augmentation

After calibrating the IQ signal, the phase sequence $\theta_{\text{vib}}(t)$ can indicate the vibration of the membrane, especially when the sound is monotone. However, such phase sequence is still insufficient to directly reveal the human voice played by the speaker due to the following reasons: 1) The vibration amplitude of different frequencies are different due to the frequency response pattern of the speaker. Thus, for the same loudness, the amplitude of high frequency sound is much smaller than the low frequency sound. Therefore, the spectrum of vibration extracted based on mmWave is concentrated on the low frequency band. 2) MmWave radar utilizes Voltage Controlled Oscillator to generate FMCW, which brings multiplicative noise distortion in TX signal [42]. As a result, there are always periodic frequency noise in the spectrum.

To solve the above problem, we propose a channel fusion-based solution to augment the phase sequence. The basic idea is that besides the target bin, the phase sequences of other nearby bins also contain the information of sound vibration according to our study. Because the vibration information of different bins are all generated by the same sound source, we can fuse multiple channels together to augment the vibration of the same source and mitigate the other noise.

To fuse the multiple channels and augment the vibration signal, we propose to leverage an encoder-decoder based
neural network to get the final recovered human voice as shown in Fig. 9. To better represent the feature of human voice, we conduct STFT on recovered signal, which transfer the time-series signal into spectrogram, which can efficiently express the features of human voice from both the time and frequency domain. Then we use 4 convolution layers for downsampling in the encoder and 4 deconvolution layers for upsampling in the decoder. Moreover, 3 skip connection layers are utilized to connect convolution layers and the corresponding deconvolution layer, which ensures the integration of both the deep and shallow features in the final output of deconvolution layer. After that, we use the last convolution layer to generate the output spectrogram, such that the size the same at the input. We use the spectrogram of the voice audio as the groundwork to train the network, such that the decoder tend to generate the spectrogram similar to the voice audio.

VI. PERFORMANCE EVALUATION

A. Experiment setting

We implement our system on COTS mmWave radar system, which includes an evaluation module IWR1843BOOST [43] and a real-time capture card DCA1000EVM. We enable 1 transmitter and 4 receiver antennas. The experiment setup is shown in Fig. 10. The speaker is placed on a desk and the mmWave radar is 1m in front of the speaker. The captured data are transmitted to a Lenovo ThinkPad laptop with Intel Core i7 CPU and 16GB memory. The raw data are preprocessed with MATLAB, and the further deep learning algorithm is implemented with Python.

The playing voice contains the digits $0 \sim 9$ and letters $a \sim z$, including male and female voices. We collect the voice signal in a total of 1000 minutes. Among them, 80% are taken as training sets, and the others 20% are testing sets.

B. Metrics

1) Signal-to-noise ratio (SNR) and Peak signal-to-noise ratio (PSNR): SNR can be represented as $SNR = 10\log\frac{E_s}{N}$, PSNR has a good performance in evaluating speech noise reduction [44]. Therefore, it is an objective index for speech articulation measurement. PSNR can be defined as:

$$PSNR = 10\log_{10}\frac{V_{max}^2}{MSE}$$

$$V_{max} = \max(\max(V_{en}(k)), \max(V_{origin}(k)))$$

$$MSE = \frac{1}{N} \sum_{k=1}^{N} [V_{origin}(k) - V_{en}(k)]^2,$$

where $V_{en}$ is the voice information captured by our system and $V_{origin}$ is the unambiguous primitive audio signal.

2) Recognition accuracy: In order to estimate the ability of voice restoration, we design a recognition network for voice spectrogram based on LeNet-5 [45]. The network classifies letters and numbers and gets recognition accuracy.

C. Spectrogram recovery

mmEavesdropper can recover the mmWave-captured spectrogram with high quality. Our system can recover audio signals of high quality while eavesdropping. Fig. 11 illustrates that the primitive mmWave signals are reconstructed into relatively pure signals by the augmentation network. Compared with the microphone-captured voice, the high-frequency parts of mmWave-captured signals are augmented, and the details of low-frequency parts are enriched after training. Moreover, the noise on the mmWave-captured spectrogram is eliminated. As a result, the reconstructed signals have little noise and are quite similar to the microphone-captured signals. The results intuitively show that mmEavesdropper has the ability to restore the sound spectrogram.

D. Audio reconstruction

mmEavesdropper can reconstruct the mmWave-captured voice with the average SNR and PSNR of 4.8dB and 17dB, respectively. As introduced in VI-B, we use SNR and PSNR to evaluate the quality of the reconstructed voice signal based on mmEavesdropper. As shown in Fig. 12, the average SNR of reconstructed voice for all digits and letters is 4.8dB. In general, SNR is always over 3dB. Furthermore, the maximum value of SNR is 8.2dB, i.e., the letter “e”. The average PSNR reaches 17dB. PSNR is always over 15dB, and the maximum PSNR is 22dB, which appears at the digit “6”. The results prove that the information components of the voice signal are always more dominant than the noise components, i.e., our system can realize effective human voice eavesdropping.

E. Recognition accuracy

mmEavesdropper can reach the totally average recognition accuracy of 93%. In order to visually show the restoration
effect of the system, we take the augmented spectrograms as the input and train the recognition network based on LeNet-5. The results of digit recognition are shown in Fig. 13. The average accuracy of digit recognition can reach 95%. The highest recognition accuracy is digit “5”, and the lowest is “3”. Fig. 14 shows the results of letter recognition. The average letter recognition accuracy is 93%. Letter “r” has the lowest accuracy. It is because the pronunciation characteristics of “r” are insufficient. In summary, *mmEavesdropper* can reach the average recognition accuracy of 93%, meaning the recovered voice is of high quality and can be distinguished.

F. Distance robustness

*mmEavesdropper* can recover the voice with high quality with 1 ~ 3m and achieve the recognition accuracy over 94%. Since the speaker can be at any location in a real scenario, *mmEavesdropper* is supposed to be able to recover the voice from speakers at different distances. Therefore, we execute the experiments in different ranges to evaluate the distance robustness. Fig. 15(a) shows the influence of distance on the system’s performance. When the radar is 1 ~ 3m from the speaker, SNR of the reconstructed voice always exceeds 5dB, and PSNR is always greater than 17dB. Additionally, SNR and PSNR do not get affected as the distance increases. Fig. 15(b) illustrates that recognition accuracy will slightly decline with the distance increasing but will consistently exceed 94%. The experiments prove that *mmEavesdropper* can work stably within the distance of 1 ~ 3m.

G. Volume robustness

*mmEavesdropper* can recover the voice with high quality and achieve the recognition accuracy over 93% as the volume changes from 60dB to 90dB. In practice, the speaker volume can be set to different levels, affecting the eavesdropping effect. To evaluate this impact on our system, we adjust the speakers’ volume within the common range of 60dB ~ 90dB. As shown in Fig. 15(c), SNR and PSNR are over 5dB and 17dB, respectively. As for recognition accuracy, it is always greater than 93% as shown in Fig 15(d). Furthermore, the results indicate that SNR, PSNR, and accuracy are not affected significantly according to the volume change. Therefore, *mmEavesdropper* can work at high performance as the volume changes from 60dB to 90dB.

VII. CONCLUSION

The popularity of virtual online meeting has stimulated the increasing usage of speakers, which brings new threats to human privacy due to different kinds of eavesdropping. This paper proposes *mmEavesdropper*, the mmWave-based solution to eavesdrop on the speaker by augmenting the micor-vibration signal from three aspects. Particularly, we augment the receiving signal by focusing on the target orientation and distance. Then, we augment the vibration signal by calibration the signal on the IQ plane. Finally, we augment the frequency spectrum by fusimg multiple channels. We have evaluated the system via extensive experiments, which show that *mmEavesdropper* can efficiently eavesdrop on the human voice with 93% accuracy.

VIII. ACKNOWLEDGEMENT

This work is supported in part by National Key R&D Program of China under Grant No. 2022YFB3303900; National Natural Science Foundation of China under Grant Nos. 61902175, 61872174, 61832008, 62272216; The Key R&D Program of Jiangsu Province under Grant BE2020001-3. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization. This work is partially supported by the Fundamental Research Funds for the Central Universities No. 2022300296 (020214380096). Chuyu Wang is the corresponding author.