mmEcho: A mmWave-based Acoustic Eavesdropping Method

Pengfei Hu, Wenhao Li, Riccardo Spolaor*, Xiuzhen Cheng
School of Computer Science and Technology, Shandong University, Qingdao, China
Email: {phu, rspolaor, xzcheng}@sdu.edu.cn, li_wenhao@mail.sdu.edu.cn

***Abstract*—Acoustic eavesdropping targeting private or confidential spaces is one of the most severe privacy threats. Soundproof rooms may reduce such risks, but they cannot prevent sophisticated eavesdropping, which has been an emerging research trend in recent years. Researchers have investigated such acoustic eavesdropping attacks via sensor-enabled side-channels. However, such attacks either make unrealistic assumptions or have considerable constraints. This paper introduces mmEcho, an acoustic eavesdropping system that uses a millimeter-wave radio signal to accurately measure the micrometer-level vibration of an object induced by sound waves. Compared with previous works, our eavesdropping method is highly accurate and requires no prior knowledge about the victim. We evaluate the performance of mmEcho under extensive real-world settings and scenarios. Our results show that mmEcho can accurately reconstruct audio from moving sources at various distances, orientations, reverberating objects, sound insulators, spoken languages, and sound levels.**

***Index Terms*—mmWave radar, Audio eavesdropping, Vibrations, Signal processing**

1. Introduction

The sound eavesdropping attack is an age-old threat to users' private or confidential information. Along with the widespread usage of soundproof materials in modern buildings, non-effort sound eavesdropping has been alleviated. However, sophisticated eavesdropping techniques have grown steadily over the years, which pose a continuous threat to the user's information security.

Researchers have investigated acoustic eavesdropping attacks with various techniques to reconstruct audio information. We can classify these attacks into two main categories based on the deployment schemes: invasive and non-invasive acoustic eavesdropping. In an invasive scenario, several works in [1]–[9] have used the motion sensor (e.g., accelerometer, gyroscope) to infer the audio signal. These attacks assume having access to data from sensors on the victim's device or placed in the same room as the audio source. In contrast, non-invasive eavesdropping attacks rely on sensors placed in an adjacent or more distant locations. To carry out this type of attack, researchers have used sensors, such as laser transceiver [10], [11], high-speed cameras [12] and photodiode [13], to reconstruct speech through vibrations induced by sound waves. In addition, WiFi signals [14], [15] and millimeter wave [16]–[18] have also been used to extract vibrations and thus eavesdrop on audio information. However, these methods either recover a narrow range of sound frequencies, cannot carry out real-time eavesdropping (i.e., offline analysis), or make strong assumptions on prior knowledge (e.g., a large dataset of the victim's speech for training, the pre-installation of a malicious software on the victim's device).

In this paper, we present mmEcho, a system that leverages millimeter wave (mmWave) radar with signal processing to achieve efficient and accurate non-invasive acoustic eavesdropping. mmEcho can address the aforementioned limitations while achieving good performance.

(1) No machine learning and no prior knowledge required: Many eavesdropping attacks rely on machine learning techniques to achieve reasonable performance [16]–[18]. However, machine learning needs large labeled datasets for training, which requires high labor and computational cost. In addition, machine learning is highly data-dependent. It may require the target user's data to reach good performance. However, such data is hard or impossible to obtain in most cases. Compared to [16]–[18], mmEcho applies signal processing techniques instead of machine learning, hence it can reconstruct anyone's voice without any prior knowledge. Moreover, mmEcho can perform eavesdropping without any burdensome model training or inference processes.

(2) Unconstrained vocabulary and high frequency response: Most of the existing methods attain eavesdropping as a classification task by segmenting the audio to recognize individual phonemes or a small set of hot words [1]–[5], [14], [15], [18]. By only relying on signal processing techniques, mmEcho directly reconstructs the audio from the millimeter-wave signal, hence it can perform eavesdropping on an unconstrained vocabulary. Moreover, the majority of existing work can reconstruct only a limited range of audio frequency (e.g., below 1kHz in [12], [13]), which cannot fully cover the spectrum of human speech. With the intra-chirp method, mmEcho can reconstruct audio frequencies up to 5kHz with a radar's chirp rate of 10kHz.

(3) Low cost, portable, and high resolution: Unlike expensive or unwieldy sensors (e.g., laser transceiver [10],

* *Corresponding author.*

[11], high-speed cameras [12], telescopes [13]), mmEcho relies on a low-cost, portable, and off-the-shelf mmWave radar. Furthermore, the short-length wavelength of millimeter wave allows such radar to provide a better vibration resolution than other Radio Frequency (RF) based audio eavesdropping systems [14], [15]. The high resolution is necessary since the sound waves induced vibration on surrounding objects (that we define as *reverberating objects*) is at the order of a few micrometers.

(4) Simultaneous eavesdropping of multiple audio sources: Most of the state-of-the-art eavesdropping methods reconstruct audio by targeting a single audio source. For instance, the IMU-based eavesdropping methods in [1]–[5] can only extract sound vibrations produced by the mobile device's built-in loudspeaker. In contrast, mmEcho can reconstruct the original audio from the vibration of a wide range of reverberating objects. Instead of targeting directly at a single audio source [16], mmEcho targets a reverberating object that allows to simultaneously eavesdrop on multiple and mobile audio sources (e.g., humans, mobile devices).

To perform a non-invasive acoustic eavesdropping attack via a mmWave radar, we need to address two major challenges:

(1) *How do we perform intra-chirp distance measurement to cover the human speech spectrum?* The mmWave radar relies on the "chirp" to perform distance measurement. In an optimistic scenario, each chirp could provide a distance measurement. Hence, a 10k chirp rate radar can reconstruct audio up to 5kHz. However, the conventional single-chirp scheme cannot offer micrometer-level resolution in a realistic scenario. It is possible to exploit multiple chirps (i.e., inter-chirp) to measure the phase and derive better resolution. However, it will significantly limit the frequency of reconstructed audio, i.e., the capability to cover the human speech spectrum. We solve this problem by estimating the phase within a single chirp (i.e., intra-chirp). It can provide high-resolution distance estimation without sacrificing the chirp rate.

(2) *How do we accurately measure the micron-level vibration?* In this paper, we reconstruct the audio entirely via signal processing techniques to achieve machine learning-independent acoustic eavesdropping. To attain this goal, we need highly accurate vibration measurements at a micrometer-level distance by calibrating the signal phase via frequency and phase interpolation (see Section 5.2). Based on the high-precision distance measurement, we combine the distance information of all chirps to reconstruct the audio amplitude, which exploits the advantage of the high chirp rate of mmWave radar.

Contributions. In this paper, we provide the following scientific contributions:

- We present mmEcho, a mmWave-based non-invasive acoustic eavesdropping system that recreates sound information from micron-level vibrations on reverberating objects. By entirely relying on signal processing techniques, mmEcho does not require any training dataset or prior knowledge of the audio signal.
- We propose an intra-chirp scheme that provides high-resolution distance estimation without sacrificing the chirp rate. Hence, it can eavesdrop audio with unconstrained vocabulary and cover the full spectrum of human speech.
- mmEcho can achieve effective eavesdropping from reverberating objects made of a wide variety of everyday materials via penetrating various sound-insulating materials. mmEcho can eavesdrop on multiple and non-stationary audio sources of different nature (e.g., human, loudspeaker, smartphone).
- We perform an extensive evaluation of mmEcho under various settings such as distance, orientation, materials, sound volume, languages, and audio source mobility with subjective and objective metrics. Our experimental results demonstrate that mmEcho can accurately reconstruct the audio with the average MCD (Mel-Cepstral Distortion) of 3.36 (the lower, the better), the average MOS (Mean Opinion Score) of 4.09 (the higher, the better), and the average WER (Word Error Rate) of 18.10% (the lower, the better).

Organization. The rest of the paper is organized as follows. We discuss the related work on acoustic eavesdropping in Section 2. In Section 3, we define our attack scenario. Section 4 provides an overview of FMCW radar, the principles of vibration measurement, and the feasibility analysis for our attack. We describe the mmEcho design in Section 5 and we experimentally evaluate it in Section 6. We discuss the potential applications, insights, and limitations of mmEcho in Section 7. Finally, we draw some conclusions in Section 8.

2. Related Work

In this section, we provide an overview of the state-of-the-art work related to acoustic eavesdropping, and we compare them with our proposed system. We summarize the work related to acoustic eavesdropping in Table 1.

Motion Sensor-based acoustic eavesdropping. Researchers have demonstrated the feasibility of eavesdropping by using motion sensors [1]–[4] to reconstruct words, phrases, and even to identify the gender of the speaker. [5] uses sensor fusion (e.g., geophone, gyroscope, accelerometer) to eavesdrop on sound within a room. [19] reconstructs with unconstrained vocabulary the audio played by a mobile device's loudspeaker via the built-in accelerometer. [6] and [7] use the magnetic hard drive and vibrating motor to reconstruct audio, respectively. [9] achieves eavesdropping by transforming the speakers connected to the computer into microphones. [8] eavesdrops on audio by using the vibration sensors in the nasal pads of glasses. The major disadvantage of these motion sensors-based methods is that they are intrusive attacks, i.e., they require close access to the victim's device, which makes these attacks easy to prevent in practice.

Optical sensor-based acoustic eavesdropping. Cameras, lasers, and telescopes have also been used for acoustic eavesdropping. [12] uses a high-speed camera to record the vibrations of an object caused by sound waves and then

TABLE 1. ACOUSTIC EAVESDROPPING ATTACKS IN THE LITERATURE COMPARED WITH MMECHO.

Sensor Type	Acoustic Eavesdropping Attack	Competence				
		Non-Invasive	Through Opaque Insulator	Uncontrained Vocabulary	Unaided by ML	Mobile Audio Source
Motion Sensor	AccelWord [4]	✗	✗	✗	✗	✓
	PitchIn [5]	✗	✗	✓	✓	-
	AccelEve [2]	✗	✗	✗	✗	-
	Accear [19]	✗	✗	✓	✗	-
	Speechless [3]	✗	✗	✗	✗	-
	Gyrophone [1]	✗	✗	✗	✗	-
	HDD [6]	✗	-	✓	✓	-
	VibraPhone [7]	✗	✗	✓	✗	-
	V-Speech [8]	✗	-	✓	✗	-
	SPEAKE(a)R [9]	✗	✗	✓	✓	-
Optical Sensor	Visual Microphone [12]	✓	✗	✓	✓	-
	LidarPhone [11]	✗	✗	✓	✗	-
	Lamphone [13]	✓	✗	✓	✓	-
Radio Receiver	WiHear [15]	✓	✓	✗	✗	✗
	ART [14]	✓	✓	✗	✓	-
	Tag-Bug [21]	✗	✓	✓	✗	-
	Uwhear [22]	✓	✓	-	✓	✗
	WaveEar [16]	✓	-	✓	✗	✗
	MILLIEAR [17]	✓	✓	✓	✗	✗
	mmSpy [18]	✓	-	✗	✗	✗
	mmEcho (this work)	✓	✓	✓	✓	✓

reconstructs the audio. Both [10] and [11] use laser sensors to implement eavesdropping. [11] controls a robot vacuum cleaner to point its laser sensor at a audio source or other vibrating object and converts the vibrations into audio by analyzing the received signal. [13] uses a remote electro-optical sensor to analyze the vibration of light bulbs due to sound to reconstruct the audio. The main disadvantages of these attacks are constrained vocabulary, lower frequency response, and easy to prevent by obstructing the line-of-sight channel (e.g. using a curtain). In addition, the recently published SoK paper in [20] demonstrates that none of the work in [1], [2], [4]–[6], [8], [11]–[13] can effectively eavesdrop on a live human speech in a real-world scenario.

RF-based acoustic eavesdropping. The Great Seal bug [23] is one of the first acoustic eavesdropping devices to use passive RF techniques to transmit an audio signal. However, it requires a pre-installed sensor in the room. In recent years, researchers have proposed eavesdropping attacks based on RF technologies, such as WiFi, RFID, and mmWave. [14] and [15] can recognize specific words by analyzing WiFi Received Signal Strength (RSS) and Channel State Information (CSI), respectively. By using Impulse Radio Ultra-Wideband, [22] can separate multiple audio sources (household objects) via vibration sensing. Using the same RF technology, [24] can recover audio only below 400 Hz. Due to low audio frequency response, these works have not considered the recovery of human speech in their performance evaluations. [21] relies on RFID tags in combination with cGAN for acoustic eavesdropping. However, it needs to pre-install RFID tags in the victim's proximity, which reduces its practicality. In summary, these WiFi signal- and impulse radio-based methods [14], [15], [21], [22], [24] suffer a insufficient vibration resolution due to long wavelength and low packet rate. Moreover, compared to portable mmWave radars, they require large antenna setups, which lead to an amplified physical footprint of the attacker and increase the operational difficulty in a real-world scenario.

In the area of mmWave-based eavesdropping, MILLIEAR [17] relies on inter-chirp-based coarse-grained phase estimates from mmWave signal to extract large amplitude vibrations (from 0.1 to 1mm). Nevertheless, all inter-chirp-based methods have limitations in terms of low-frequency response and inaccuracies, thus the authors address these challenges with cGAN. mmSpy [18] can use mmWaves to eavesdrop on phone calls. Since mmSpy considers eavesdropping as a classification problem, it can only identify specific keywords or digits (i.e., constrained vocabulary). Since both works [17], [18] leverage machine learning techniques, they require large labeled datasets, which are hard to obtain in an eavesdropping attack.

Other RF-based methods unsuitable for eavesdropping. Among other methods that use RF-based technologies, [25] exploits Doppler radar to recognize a limited set of keywords and the frequency response is limited below 200Hz [26] presents a wall-permeable attack that infers the content visualized on an LCD screen via a mmWave radar. In an industrial environment, mmVib [27] achieves micron-level vibration measurements (below 500Hz). Authors in [28] use mmWave radar to sense vibration below 1kHz, but the radar sensor needs to be quite close to the speaker (≤ 5cm) for eavesdropping. [29] presents a system for user verification with mmWave radar, which does not focus on speech reconstruction and has a frequency response below 200Hz. [30] uses FMCW radar to reconstruct audio at frequency below 1kHz and inadequate experimental evaluation is provided. [16] can reconstruct a high-quality voice from the user's throat by using mmWave radar, but it requires the subject to stay still and at a short distance from the radar probe (less than 2m). Authors in [31] use a customized mmWave radar to achieve vibration monitoring. [32] applies speech enhancement algorithms to improve the speech signal captured by customized mmWave radar. In [33], authors provide a noise-resistant multi-modal speech recognition system that combines mmWave and microphone by using machine learning.

However, the work in [25], [27], [28], [30], [33] can achieve a frequency response that cannot fully covers the human speech spectrum (300Hz to 3.4kHz) [34]. In addition, since [16], [25], [26], [29], [33] use machine learning, they require prior knowledge and a large dataset for model training, thus they cannot achieve unconstrained-vocabulary eavesdropping. The works in [31], [32] rely on customized hardware and perform line-of-sight audio reconstruction. Hence, these works are not cost-effective and cannot achieve eavesdropping when the target is non-line-of-sight. Therefore, none of the aforementioned work fulfills the requirements for acoustic eavesdropping in a real-world scenario.

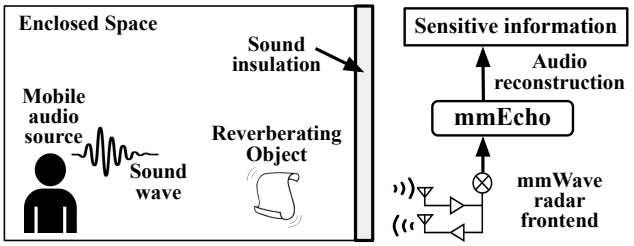


Figure 1. Our attack model for mmWave-based acoustic eavesdropping.

3. Attack Model

The use of soundproof materials is a traditional method to prevent eavesdropping. Such materials include wood, glass, acoustic wool, plasterboard, etc. In this paper, we consider the threat model depicted in Figure 1, where the victim stays in a soundproof room. The pressure of the sound wave produced by an audio source (e.g. human, loudspeaker, etc.) can induce vibrations in the surrounding objects (defined as *reverberating object*). An attacker can measure such vibrations with a mmWave radar outside the room to further recover the original sound.

Under these settings, the attacker's main objective is to eavesdrop on any sound in the victim's room. In particular, we assume the following scenario and attacker capabilities:

- the attacker can only place the mmWave radar outside the victim's room, and it cannot deploy any equipment or sensor (e.g., camera, microphone) inside the victim's room;
- the victim's room has a partial aperture composed of sound insulation material which is penetrable by mmWave;
- everyday objects (i.e., reverberating objects, such as chip bags, carton boxes, etc.) are in the victim's room. Sound waves can induce minute vibrations on such objects. However, the attacker has no prior knowledge on the location of such objects within the victim's room;
- the attacker has no prior information on the sound in the victim's room, however, the attacker has to recover audio which should be human-comprehensible;
- the equipment used by the attacker has to be portable and cost-efficient.

4. Preliminaries

In this section, we first introduce the basic principles of using mmWave radar to perform displacement measurement. We then assess the feasibility for our eavesdropping attack.

4.1. Frequency Modulated Continuous Wave

The Frequency-Modulated Continuous Wave (FMCW) radar is a special type of radar sensor that transmits a signal called "chirp". A chirp is a sinusoid whose frequency increases linearly with time. FMCW radar can be used to accurately estimate the object distance and its relative velocity. Figure 2 shows the simplified system architecture of FMCW radar.

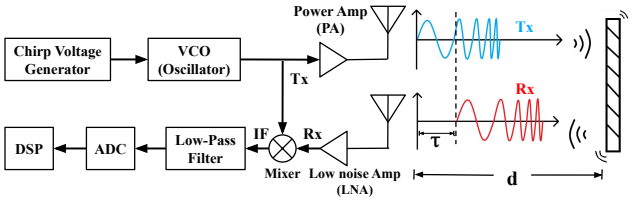


Figure 2. The simplified system architecture of FMCW radar.

We denote the round-trip delay between the transmitted and received signals as τ,

$$\tau = \frac{2d}{c} \tag{1}$$

where d is the distance between the radar and the target object, and c is the speed of the millimeter-wave signal.

We can define the transmitted and received signals of a mmWave radar as follow:

$$s_{Tx}(t) = A_{Tx} \cdot \cos[2\pi \cdot f_{Tx}(t) \cdot t + \phi_{Tx}] \tag{2}$$

$$s_{Rx}(t) = A_{Rx} \cdot \cos[2\pi \cdot f_{Tx}(t-\tau) \cdot (t-\tau) + \phi_{Rx}] \tag{3}$$

where $f_{Tx}(t)$ is frequency of transmitted signal, ϕ_{Tx} and ϕ_{Rx} are the phase of transmitted signal and received signal, A_{Tx} and A_{Rx} are the amplitude of the transmitted and received signal. The transmitted signal $f_{Tx}(t) = f_0 + kt$, where f_0 is the start frequency, $k = B/T_c$ is the slope of chirp signal, where B is the bandwidth of radar and T_c is the duration of a chirp as shown in Figure 3.

As shown in Figure 2, a mmWave radar transmits a chirp signal and receives a reflected signal. The transmitted signal and received signal are combined by a mixer. As a result of this process, we obtain an Intermediate Frequency (IF) signal as follow,

$$\begin{aligned} s_{IF}(t) =& s_{Tx}(t) \cdot s_{Rx}(t) \\ =& \frac{1}{2} A_{Tx} A_{Rx} \{\cos[2\pi \cdot f_{Tx}(t) \cdot t + \phi_{Tx} \\ & - 2\pi \cdot f_{Tx}(t-\tau) \cdot (t-\tau) - \phi_{Rx}] \\ & + \cos[2\pi \cdot f_{Tx}(t) \cdot t + \phi_{Tx} \\ & + 2\pi \cdot f_{Tx}(t-\tau) \cdot (t-\tau) - \phi_{Rx}]\} \end{aligned}$$

Since the frequency of the first cosine signal (frequency difference of two carrier signals, which is at MHz level) is much lower than the second one (frequency sum of two carrier signals, which is at GHz level) [35], we can apply a low-pass filter to exclude the second one. Therefore, we obtain:

$$\begin{aligned} s_{IF}(t) =& \frac{1}{2} A_{Tx} A_{Rx} \cdot \cos[2\pi \cdot f_{Tx}(t) \cdot t + \phi_{Tx} \\ & - 2\pi \cdot f_{Tx}(t-\tau) \cdot (t-\tau) - \phi_{Rx}] \end{aligned} \tag{4}$$

After the substitution $f_{Tx}(t)$ for $f_0 + kt$ in Eq. (4), we derive the intermediate frequency signal as follows,

$$s_{IF}(t) = A_{IF} \cdot \cos\left(4\pi k\tau t + 2\pi f_0 \tau - 2\pi k\tau^2 + \phi_d\right) \tag{5}$$

where $A_{IF} = \frac{1}{2} A_{Tx} A_{Rx}$ is the IF signal amplitude and $\phi_d = \phi_{Tx} - \phi_{Rx}$ is the phase difference between the

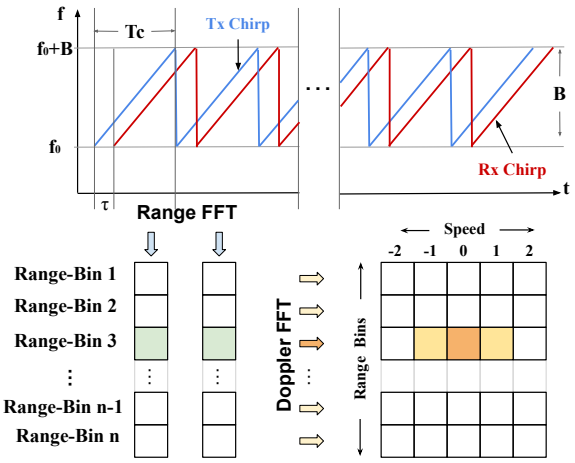


Figure 3. Generation of chirps

transmitted and received signal. Figure 3 demonstrates the generation of the chirp sequence, in which the transmitted chirp is in blue, and the received chirp is in red.

4.2. FMCW-based Vibration Measurement

Since the sound waves is a kind of mechanical wave, it can cause minute vibration on other objects. Our eavesdropping approach aims to measure the tiny vibration on reverberating objects to infer the sound that causes such vibrations. Once the vibration can be measured accurately, we can indirectly reconstruct the original sound, which is similar to the mechanism of the eardrum for sound perception.

The conventional FMCW based distance calculation formula is as follows,

$$d = \frac{f_{IF}}{k} \cdot \frac{c}{2} \tag{6}$$

However, it cannot provide micrometer-level resolution for vibration measurement.

Highly accurate distance estimation is a prerequisite for vibration extraction. We use the intra-chirp method to estimate the distance between the target object and radar, i.e., using a single chirp to estimate the distance. Our intra-chirp method combined with the radar's high chirp rate greatly improves the performance of the system, which can now acquire a larger number of measurements per second. For example, with a chirp rate of $10k$ chirps/second, the system can obtain up to ten thousand individual distance measurements each second. Therefore, the variation of two successive measurements provides displacement of vibration. From Eq. (5), we can derive the phase of intermediate frequency signal as follows,

$$\phi = 2\pi f_0 \tau - 2\pi k \tau^2 + \phi_d \tag{7}$$

With Eq. (1) and Eq. (7), we can derive the accurate distance measurement:

$$d = \left(\frac{f_0}{k} + \sqrt{\frac{f_0^2}{k^2} - 2 \cdot \frac{\phi - \phi_d}{\pi k}} \right) \cdot \frac{c}{4} \tag{8}$$

Furthermore, we use a phase calibration method which is described in Section 5 to obtain the signal phase for more accurate distance measurement. Then, we extract the vibration information from the displacement between successive chirps.

4.3. Feasibility Analysis

To evaluate the feasibility of our eavesdropping attack, we choose several materials as a reverberator for a proof-of-concept experiment, i.e., tinfoil, chip bag, projector screen, carton box, and paper. We use a loudspeaker as the audio source and frosted glass as a sound insulator in front of the mmWave radar, and we point the radar probe towards the reverberating object.

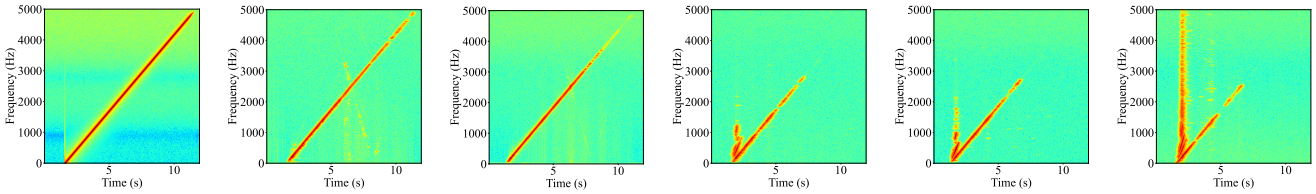
To assess the frequency response of our system, we play a sweep-tone ranging from 10Hz to 5kHz on the loudspeaker. This frequency range fully covers the spectrum of human speech (300Hz to 3.4kHz) [34], [36]. We report the results of our feasibility analysis in Figure 4. We can notice that the frequency response of tinfoil and chip bag is below 4kHz (in figures 4(b) and 4(c), respectively). The frequency response of carton box, projector screen and paper is below 3, 2.8, and 2.5kHz, respectively (in figures 4(d), 4(e), and 4(f)). The reason for the variation of frequency response on different materials is due to their different elastic deformation capabilities and reflectivity. The distortion at beginning of sweep-tone is mainly caused by fact that the actual transmitted sinusoidal signal is modulated by the step function when the excitation signal is applied. This distortion does not occur with human-speech, hence, it has a negligible impact on human-speech reconstruction.

The result of the feasibility analysis indicates that mmWave radar can capture the vibration spectrum induced by the human voice. Hence, we can leverage an FMCW radar to eavesdrop on sound from a reverberating object.

5. mmEcho System

In this section, we present the system architecture of mmEcho. Figure 5 provides an overview of the mmEcho system and its modules. In what follows, we introduce the three modules that compose mmEcho and their roles in reconstructing high-quality audio from the mmWave radar's raw ADC data.

1) Reverberating Objects Detection (ROD) analyzes the data received by the radar to locate the reverberating objects within the victim's room. In particular, this module identifies the objects with adequate sound-induced vibrations suitable for processing in the subsequent modules by filtering out the unsuitable ones. This module can also guide directional adjustments in pointing the radar towards the selected reverberating object to attain a high signal-to-noise ratio (SNR). To achieve these goals, the ROD module pre-processes the intermediate frequency signal and performs a Range-FFT to obtain the range bins as shown in Figure 3. After that, ROD applies a Doppler-FFT (i.e., 2D-FFT) on such range bins



(a) Original audio (b) Tinfoil (c) Chip Bag (d) Carton Box (e) Projector Screen (f) Paper

Figure 4. The spectrograms for original and reconstructed audio from different reverberating materials.

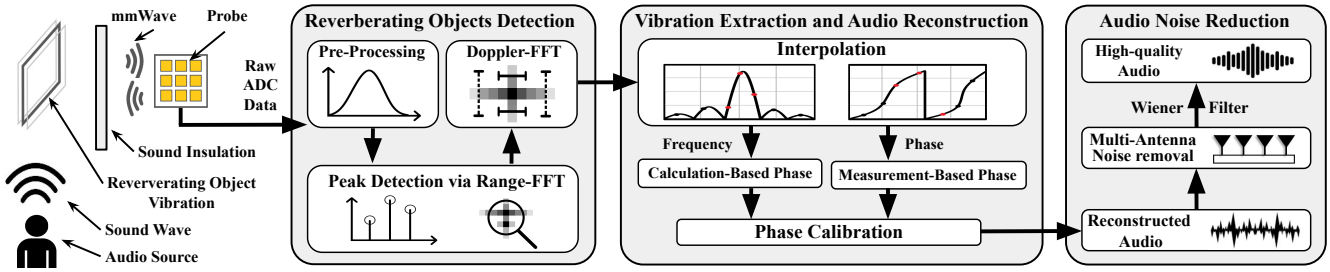


Figure 5. The mmEcho system mainly consists of a mmWave radar (TI IWR1642boost and TI DCA1000EVM), a Data-preprocessing module to extract the vocal spectrogram, and an Audio Reconstruction module to recover high-quality voice.

and sets the range gates of the reverberating objects in the Doppler-FFT spectrum to detect the vibration in each object within the radar vision.

2) Vibration extraction and Audio reconstruction (VA) is the core of our method since it reconstructs the audio from the vibrations induced by the pressure of sound waves on a reverberating object. Hence, VA aims to measure such vibrations as radar-to-object distance (i.e., displacement of the object caused by vibrations). According to Eq. (8), distance calculation requires an accurate phase estimation. To measure the distance with micron-level precision, VA accurately estimates the phase of the received radar signal by combining the frequency- and measurement-based phase estimations. Since we apply an intra-chirp method, VA measures the accurate radar-to-object distance for every chirp at 10k samples/second. Thus, VA derives the amplitude of the original sound waves from the displacement with consecutive chirps. By transposing such amplitude in the time domain, VA provides a preliminary version of the reconstructed audio.

3) Audio Noise Reduction (ANR) aims to improve the quality of the audio reconstructed by the VA module. Measuring the micron-level displacement on a reverberating object requires accurate measurements from the radar. However, such measurements may contain anomalies due to RF interference from the surrounding environment. Such anomalies result in audible "clicks" in the reconstructed audio. Since a radar provides data from multiple independent antennas, ANR identifies and removes most anomalies by applying a sliding window on such data. In addition, ANR further improves the quality of the resulting audio by filtering the white noise via a Wiener filter.

In what follows, we describe in detail the aforementioned modules.

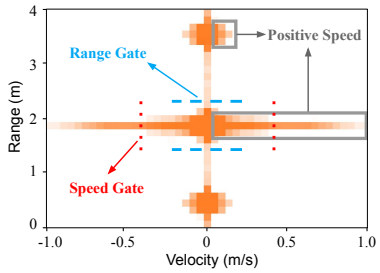
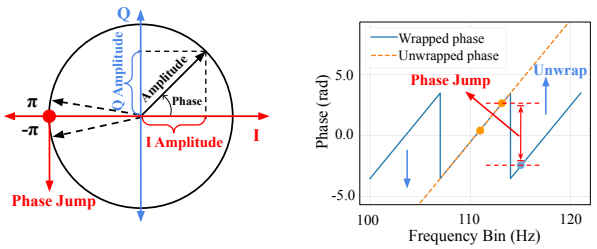


Figure 6. The spectrum of Doppler-FFT

5.1. Reverberating Object Detection

To facilitate signal processing, we collect raw binary ADC data by mmWave radar and convert it to multi-dimensional IQ arrays. We partition the collected intermediate frequency signal via a windowing process to avoid spectrum leakage. For this process, we select several commonly used window functions: Boxcar, Triang, Blackman, Nuttall, and Hanning. We test the impact of such window functions on audio reconstruction in Section 6.4.1. Then, we perform a Fast Fourier Transform (FFT) to output the Range-FFT frequency spectrum $Spec_{IF}(f)$ and phase spectrum $Spec_{\phi}(f)$, which contains the frequency bins of a single chirp. The result of Range-FFT can be used to differentiate multiple objects based on the intermediate frequency. We identify the peaks on the Range-FFT spectrum by applying a Continuous Wavelet Transform (CWT)-based peak detection algorithm [37].

Each frequency peak corresponds to an object within the radar range. However, Range-FFT can only provide us with the target's distance. To locate the reverberating object, we perform Doppler-FFT based on the results of Range-FFT. In Figure 6, we provide an example of the Doppler-FFT spectrum. Similar to the Range-FFT spectrum, we can identify the objects within the radar's range. We can observe that a vibrating object produces a significant lateral velocity



(a) Phase principle of IQ data. (b) Wrapping and unwrapping

Figure 7. Phase wrapping

in the Doppler-FFT spectrum. Therefore, we separate one object from other objects on the range axis (i.e., *range gate*) and determine whether an object vibrates or not by setting a threshold on the velocity axis (i.e., *speed gate*). By detecting vibrating objects on the Doppler-FFT spectrum, we locate potential reverberating objects, which we can use for our eavesdropping attack.

5.2. Vibration Extraction and Audio Reconstruction

We run the CWT-based peak detection algorithm [37] in the frequency spectrum. Each peak in the Range-FFT corresponds to a distinct object and its distance. However, the resulting peak frequency f_{in} is in low resolution due to the limited number of sampling points in a single chirp (e.g., maximum 400 sampling points per chirp on our mmWave sensor). Therefore, we apply the parabolic interpolation algorithm to obtain precise peak frequencies. In particular, we perform the interpolation by selecting three, five, or seven points near a frequency peak. We choose the three-point parabolic interpolation since it provides accurate results and has a low computational complexity. We refer to a peak frequency obtained from the interpolation process as $\hat{f}_{in}$.

With the intermediate frequency, we can measure the distance of an object at a millimeter-level accuracy. However, such a low accuracy cannot extract the micrometer-level vibration on reverberating objects. To address the challenge, we resort to the phase of the intermediate frequency signal to obtain a micrometer-level distance measurement. Since we use the intra-chirp method, the precise estimation of phase is essential for minute displacement estimation. Unfortunately, the phase measurement is easily affected by the phase wrapping phenomenon, as shown in Figure 7. In the actual measurement, we obtain the IQ data consisting of In-phase (I) and quadrature (Q) components. As depicted in Figure 7(a), the phase is within the range $[-\pi, \pi]$. When the actual phase is outside this range, the phase value is added or deducted by multiple 2π to enforce the phase value within $[-\pi, \pi]$.

To the best of our knowledge, the number of phase jumps on a reverberating object is unknown even if we unwrap the phase. Therefore, we propose a phase calibration algorithm based on a combination of phase calculation and phase measurement.

(1) Calculation-Based Phase (CBP). We use the previously estimated frequencies $\hat{f}_{in}$ in Eq. (6) to obtain a frequency-based distance estimation and the Round-trip delay τ. According to Eq. (7), we can calculate the phase ϕ_{CBP} from frequency. The frequency-based phase calculation contains information on the number of wrapping. However, the phase value based on calculation is not accurate due to the defects of the frequency-based method.

(2) Measurement-Based Phase (MBP). We can also use the exact peak frequency $\hat{f}_{in}$ to extract an accurate phase value based on measurement. To obtain such a value, we need to interpolate in the phase spectrum. Unfortunately, directly applying such interpolation produces errors due to the phase wrapping phenomenon. Therefore, in the phase spectrum $Spec_{\phi}(f)$, we perform phase unwrapping around the $\hat{f}_{in}$ as follow:

$$\phi(\lfloor \hat{f}_{in} \rfloor) = \begin{cases} \phi(\lfloor \hat{f}_{in} \rfloor) + 2\pi, & \phi(\lceil \hat{f}_{in} \rceil) - \phi(\lfloor \hat{f}_{in} \rfloor) > \pi \\ \phi(\lfloor \hat{f}_{in} \rfloor) - 2\pi, & \phi(\lfloor \hat{f}_{in} \rfloor) - \phi(\lceil \hat{f}_{in} \rceil) > \pi \\ \phi(\lfloor \hat{f}_{in} \rfloor), & otherwise \end{cases}$$

where $\phi(\lfloor \hat{f}_{in} \rfloor)$ and $\phi(\lceil \hat{f}_{in} \rceil)$ are the phases of the nearest integer to the left and right of $\hat{f}_{in}$ in $Spec_{\phi}(f)$, respectively. Then, we apply linear interpolation to the phase spectrum. We obtain ϕ_{MBP} as a result of our phase interpolation algorithm. However, the number of phase jumps cannot be calculated only with the measurement-based phase.

(3) Phase Calibration. It is worth noticing that the ϕ_{CBP} contains the phase wrapping information while the ϕ_{MBP} provides a more accurate phase value. Therefore, we can perform the phase calibration by combining the ϕ_{CBP} and ϕ_{MBP} as follows:

$$\hat{\phi} = \phi_{MBP} + 2\pi \cdot round\left(\left|\frac{\phi_{CBP} - \phi_{MBP}}{2\pi}\right|\right) \quad (9)$$

Although ϕ_{CBP} is a rough phase estimation, it contains the number of phase wrapping. In the best case, $\phi_{CBP} = \phi_{MBP} + 2\pi \cdot n$, where n is the number of phase wrapping. Hence, we use ϕ_{CBP} to extract n and combine ϕ_{MBP} to get more precise phase estimate. In the real world, there are inevitable errors between theoretical calculations and actual measurements, but the errors cannot exceed one phase period. Therefore, we use $(\phi_{CBP} - \phi_{MBP})/2\pi$ to calculate the number of wrapping. Due to the insufficient accuracy of the frequency-based method, we use the ϕ_{MBP} for calibration. Hence, $\hat{\phi}$ includes both the accuracy of the measurement-based phase and the wrapping information contained in the calculation-based phase.

(4) Audio Reconstruction. With the intra-chirp based method in sections 5.1 and 5.2, we can derive a micrometer-level distance for each chirp according to Eq. (8). With the precise measurement of distance, we can easily calculate the displacement (i.e., vibration) of the reverberating object between successive chirps. From the measured vibration on a reverberating object, mmEcho aims to generate an audio signal of the original sound as a time series of amplitudes. As shown in Figure 8, each chirp c can provide an accurate distance d_c, thus we obtain the displacement from consecutive chirps $\Delta d = d_{c+1} - d_c$, i.e., the vibration amplitude.

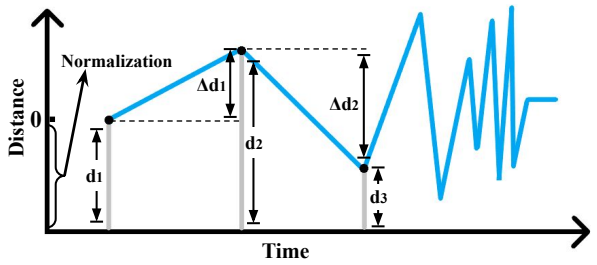


Figure 8. Extract vibration information from the distance in the time domain.

After calculating the displacement of all chirps, we obtain the vibration amplitude information on the time domain. Hence, the chirps include information about the vibrations on a reverberating object. Note that the chirp rate (i.e., in chirp/second) sets a limitation on the maximum frequency of reconstructed audio. For instance, we can reconstruct audio up to 5kHz with a chirp rate of 10kHz. We then normalize and amplify the amplitude of such vibrations to obtain an intermediate signal.

We convert the intermediate signal to a single-channel audio signal (i.e., monophonic sound) where a 16-bit depth is used to represent each sample. As a final step, we re-scale the intermediate signal within the interval $[-2^{15}+1, 2^{15}-1]$ to obtain reconstructed diatonic audio.

5.3. Audio Noise Reduction

Since radio frequency signal is prone to surrounding interference, abnormal data may be present during the acquisition process. In particular, such anomalies can result in an audible "click" in the reconstructed audio signal, which affects the comprehensibility.

A typical mmWave radar normally equips multiple receiving antennas (i.e., four receiving antennas in our case) for signal enhancement purpose. Since each receiver antenna is independent, it is possible to remove clicks by combining multiple acquired data from different antennas as shown in Figure 9.

We process the data from the multiple receiver antennas simultaneously in the following way. Since different receiver antennas have different path lengths for signal propagation, it can result in a minor difference in arrival time among antennas. Hence, we cannot compare the whole temporal amplitude series among multiple antennas. We resort to sliding windows to address this issue as in most cases two consecutive clicks do not occur in a short period of time. A sliding window is set for the data acquired by each antenna, and all the windows move synchronously. Then we perform decision-making for peaks in sliding windows. If the peak at a certain point in $window_n$ also exists in the other windows, it is identified as a "valid peak"; otherwise, it is identified as a "click" and dropped. Sliding windows partially solve the arrival time delay problem. However, it fails if there are multiple peaks within the same window. Therefore, the selection of a proper window length is essential. We assess the effect of window length and step size on the audio

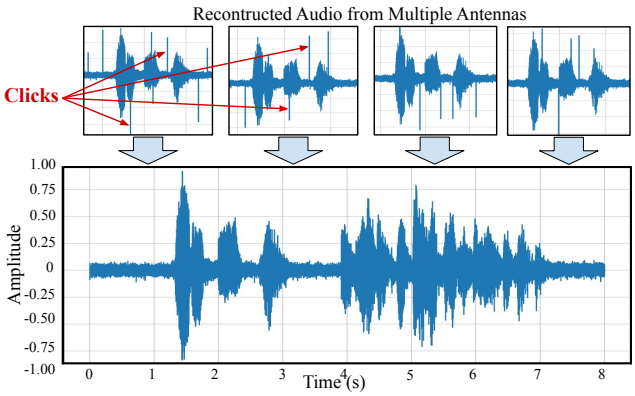


Figure 9. Click removal from multi-antenna reconstructed audio.

reconstruction results in Section 6.4.1. Finally, we apply a Wiener filter [38] to remove the white noise from the reconstructed audio.

6. Evaluation

In this section, we report the results of the experimental evaluation of mmEcho. In particular, we present the detailed setup of our experiments, the dataset, and the performance under extensive scenarios and settings.

6.1. Experiment Setup

Hardware. We implement mmEcho based on the mmWave sensor IWR1642BOOST and the data capture adapter DCA1000EVM (both from Texas Instruments). The IWR1642 BOOST works in the 76-81GHz band and is equipped with two transmitter antennas and four receiver antennas. The DCA1000EVM provides real-time data acquisition and streaming for two- and four-lane LVDS traffic from IWR1642BOOST. The maximum chirp rate of the mmWave sensor is 10kHz. The mmWave radar is connected to the laptop via an Ethernet RJ45 interface. We perform data processing and analysis on a laptop (Lenovo Legend Y7000P) equipped with an Intel Core i7-10750H CPU @ 2.60GHz and 16GB of RAM. We use a loudspeaker (Philips SPA311) as audio source. The sampling rate of all the audio samples in our experiments is 44.1kHz.

Software. We use mmWave SDK[1] to configure mmWave sensor modules and acquire the data from Analog-to-Digital Converter (ADC). The mmWave SDK also provides basic post-processing and visualization of ADC data. We provide more details for our algorithm implementation and its computational overhead in Appendix A.

Material. Our eavesdropping method measures the vibrations on a reverberating object caused by sound waves. Since the intensity of vibration and the reflected signal directly depends on the properties of the material of a reverberating object (henceforth referred to as *reverberators*), we analyze such properties on several widely used materials in daily life, which are tinfoil, chip bag, plastic, poly bag, carton,

1. mmWave SDK Ver. 03.05.00.04, www.ti.com/tool/MMWAVE-SDK

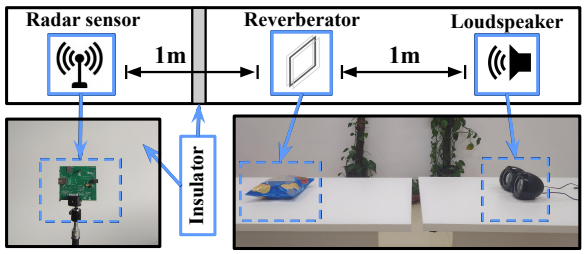


Figure 10. Experimental setup of an attack scenario.

projector cloth, leather, laptop lid (i.e., a MacBook pro 2021 13 inches), and paper.

In our attack scenario in Section 3, we consider the presence of a sound insulator between the millimeter-wave probe and a *reverberator* in the victim's room. We choose several common acoustic insulation materials (referred to as *insulators*) in modern buildings, which are drywall, wood, glass, frosted glass, cotton, and polyester.

Attack Setup. In our experimental evaluation, we place the mmWave radar sensor outside the victim's room. We position a reverberator composed of different materials (e.g., tinfoil, chip bag, and plastic) on a table within the room. We also position the loudspeaker as an audio source at different distances from the reverberator. The sound waves resulting from the audio played by the loudspeaker induce minute vibrations on the reverberator. By pointing the mmWave radar towards the reverberating object, we can measure such sound-induced vibrations and reconstruct the original audio played by the loudspeaker. In Figure 10, we depict the meeting room scenario used for our experiments.

6.2. Dataset

We collect audio data of five commonly spoken languages in the world, which are English, Chinese, Italian, Spanish, and French. In Table 2, we report the detailed information of our dataset. The English audio is from the IELTS listening test exam[2,3], including two male ($User_1$ and $User_2$) and two female ($User_3$ and $User_4$). The Chinese audio is from the training audio for Chinese pronunciation[4]. For the other languages, we select instructional audio for learning Italian[5], Spanish[6], and French[7].

6.3. Metrics

We assess the quality of the audio reconstructed by mmEcho using the following subjective and objective evaluation metrics.

Mel-Cepstral Distortion (MCD) [39] is an objective measure used for speech quality assessment. It has been widely used to compare the quality of synthesized speech and original/natural speech. A smaller MCD value indicates a closer similarity between the reconstructed audio and the original audio. Typically, reconstructed audio with MCD below 8 can be recognized by speech recognition systems [40].

Word Error Rate (WER) [41] is an objective metric for evaluating speech comprehensibility in terms of recognized words. We input the original audio and the reconstructed audio into the Google speech recognition system[8] and calculate its accuracy as $WER = (D + I + S)/N$, where D, I, and S represent the number of word deletions, insertions, and substitutions, respectively, while N is the total number of words in the original audio. A lower WER indicates a higher word recognition accuracy, i.e., a better reconstruction performance of mmEcho.

Mean Opinion Score (MOS) is a metric used for the subjective assessment of audio quality. MOS is expressed by a single discrete number, typically in the range from 1 to 5, where 1 indicates the lowest intelligible quality and 5 indicates the highest intelligible quality. Table 3 shows the rating scale for the MOS metric. In the user study for this subjective evaluation, we recruited 30 participants (i.e., 15 males and 15 females) among university students and faculties via group chats and mailing lists. These participants include both native and non-native speakers with ages from 20 to 30 years old. Participants are volunteers (i.e., no reward) and with no conflict of interest. We provided the participants with the experiment instructions in advance. The participants could choose a time and place at their convenience and they could take breaks, interrupt, or decide to quit the experiment at any time. The overall duration of the experiment is around 10 minutes.

TABLE 2. AUDIO DATASET USED FOR EVALUATION.

Label	Language	# of words	Duration (s)	Gender
$User_1$	English	1034	6124	Male
$User_2$	English	1107	6245	Female
$User_3$	English	1152	6308	Male
$User_4$	English	1075	6276	Female
$User_5$	Italian	1134	6417	Male
$User_6$	Chinese	1178	6482	Male
$User_7$	French	1009	6112	Male
$User_8$	Spanish	1071	6294	Male

TABLE 3. RATING SCALE FOR MOS

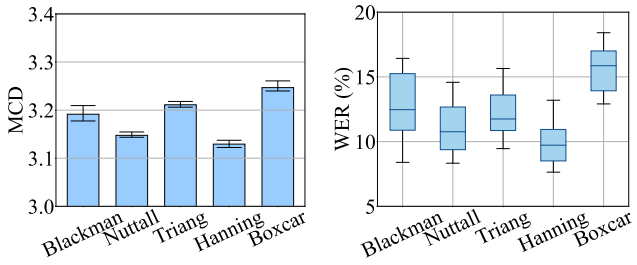
Score	Label	Description
5	Excellent	All the original speech is recovered
4	Good	Most of the original speech is recovered
3	Fair	Half of the original speech is recovered
2	Poor	Little of the original speech is recovered
1	Bad	None of the original speech is recovered

6.4. Experimental Results

In this section, we analyze the performance of mmEcho under various settings.

6.4.1. Impact of Window Function and Size. This experiment aims to assess which window function and which

2. IELTS Listening 21.8, www.youtube.com/watch?v=OWAjEPFqmVY
3. IELTS Listening 21.9, www.youtube.com/watch?v=5cG3VcPRYhM
4. Spoken Mandarin training, https://music.163.com/song?id=29808867
5. Italian Children's Stories, www.theitalianexperiment.com/stories
6. One Hour Spanish Mini, www.youtube.com/watch?v=gBJMt1_xjTM
7. Super easy French, www.youtube.com/watch?v=fq_4V-Ia1z0
8. Google Speech-To-Text, https://cloud.google.com/speech-to-text



(a) MCD (b) WER

Figure 11. Impact of different window functions on the performance.

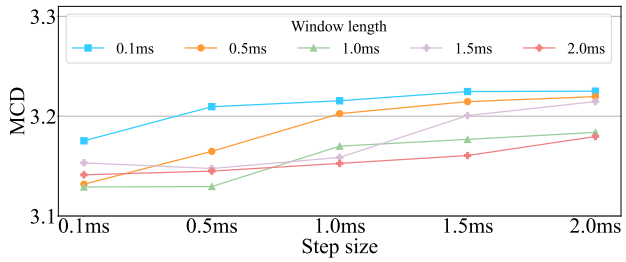


Figure 12. Effect of different lengths and step sizes of sliding window on audio reconstruction performance of mmEcho.

sliding window configuration are the most suitable for our requirements. We apply different window functions on the received signal to reconstruct the audio.

In Figure 11, we compare the performance of the considered window functions in terms of MCD and WER. As we can notice, the Hanning window function achieves the lowest MCD and WER. Hence, we apply the Hanning window function in all the afterward experiments.

In Figure 12, we show the effect of using different sliding window lengths and step sizes on the audio reconstruction results. For the same window length, the processing time is longer for shorter steps. As an optimal trade-off between processing time and MCD value, we apply 1ms window length and 0.5ms step size in all experiments reported in this section.

6.4.2. Overall Performance in a Realistic Scenario. We evaluate the audio reconstruction performance of mmEcho in a real-world meeting room scenario as depicted in Figure 10, where the drywall serves as insulator[9], a chip bag as reverberator, and a loudspeaker as the audio source at a distance of one meter from the reverberator.

In Figure 13, we report the spectrograms and the textual content of (a) the original audio (used for reference), and (b) the audio reconstructed by mmEcho from the signal reflected by the reverberator. We can clearly observe the high similarity between the reconstructed audio and the original audio in the low-frequency range (below 3.5kHz). The high-frequency range of the reconstructed audio has a lower intensity than the low-frequency part, which is due to the weak elastic deformation capacity of the reverberating material, i.e., the sound waves of the audio source does not induce high-frequency vibrations on the reverberator. Nonetheless, this has limited influence on the comprehensibility of the reconstructed audio since its spectrum encompasses most of the frequency spectrum of human speech [34], [42], [43].

9. The insulator in the realistic scenario is constructed with a wood framing covered with a single-layer drywall on each side. The drywall is made of calcium sulfate dihydrate, and each layer is 2cm thick.

6.4.3. Impact of Distance and Direction. In real-world attack scenarios, the attacker needs to adjust the position of the mmWave radar to target a reverberator. Hence, it will affect the distance and direction of the mmWave radar from such a reverberator (*radar-to-reverberator*). In this experiment, we assess the robustness of mmEcho across various distance and direction angles between the mmWave radar and a reverberator made of tinfoil. In particular, we adjust the distance from 0.5 to 5m at a fixed angle of 0° and the direction angle from 0° to 60° at a fixed distance of 1m. We use frosted glass as an acoustic insulator and a loudspeaker as an audio source. For each distance and direction setting, we play the English audio from different users, measure the Signal-to-Noise Ratio (SNR) of the received signal, and assess the reconstructed audio quality.

We report the results of these experiments in terms of MCD and WER metrics in Figure 14. mmEcho still achieves good performance at the distance of 5m. WER is between 4.17% at 0.5m to 35.51% at 5m. As reported in Figures 14(a) and 14(b), the increasing distance and angle cause the attenuation of millimeter-wave signal in terms of SNR, which also decreases the reconstruction performance. Compared with the distance, the performance degrades rapidly along with the increase of angle. Nonetheless, mmEcho still achieves a good performance up to the angle of 45° (i.e., 3.5 and 26.75% for average MCD and WER, respectively) and reasonable performance at even at a large angle of 60° (i.e., 4.13 and 45.61%).

We also assess the performance of mmEcho under various settings of the distance between the reverberator and the audio source (*reverberator-to-source*). In this experiment, we fix the radar-to-reverberator distance at 1m and vary the reverberator-to-source from 0.5 to 5m. We keep using the frosted glass as the acoustic insulator. For each distance setting, we play English audios from different users. Figure 15 reports the results of this experiment in terms of MCD and WER. As we move the loudspeaker away from the reverberator, the MCD and WER of the reconstructed audio increase accordingly. We can explain these results with the physical properties of sound. The relationship between the amplitude A of sound waves and the distance d follows the proportion $A \propto 1/d^2$ (i.e., an inverse-square law). Hence, as the amplitude of sound waves decreases with the distance, the intensity of the resulting vibrations on a reverberator also decreases. Even under the worst distance setting, mmEcho can still achieve a reasonable performance at an average MCD of 4.16 and WER of 45.18%.

In summary, the results of our experiment underline that the MCD does not deteriorate significantly, and the WER is within human comprehensibility under a large distance

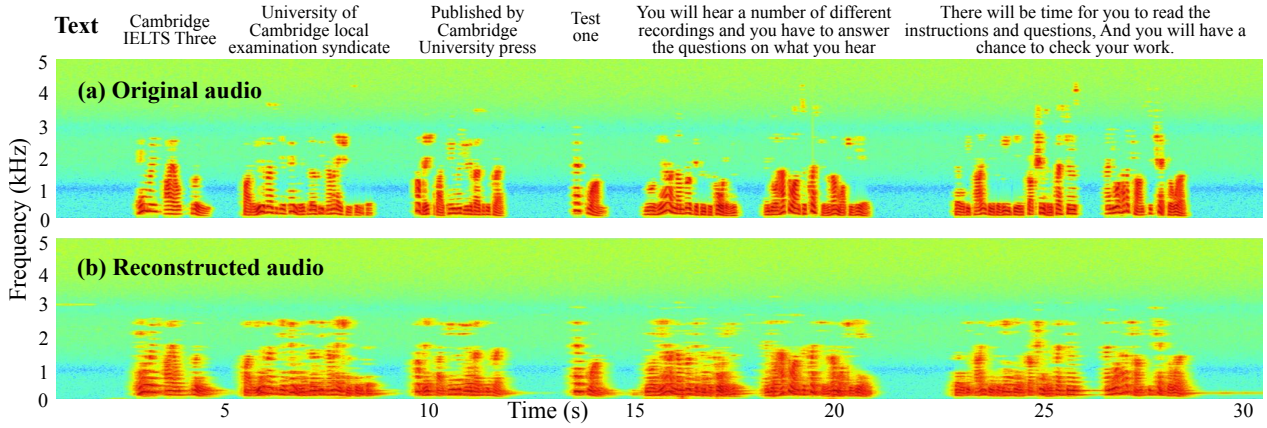
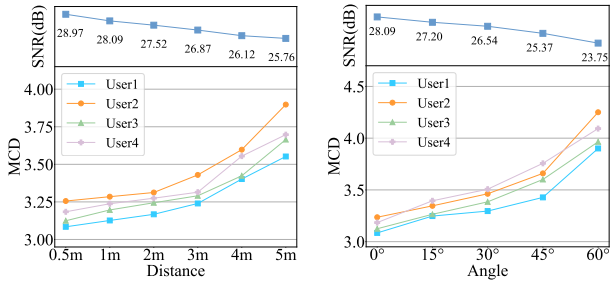
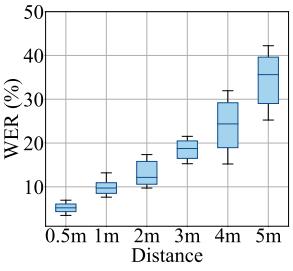


Figure 13. The spectrograms for (a) original audio and (b) reconstructed audio by mmEcho

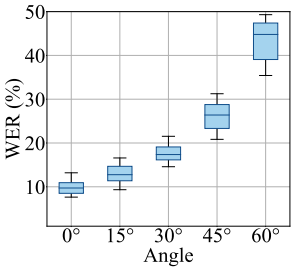


(a) MCD of various distance

(b) MCD of various angle

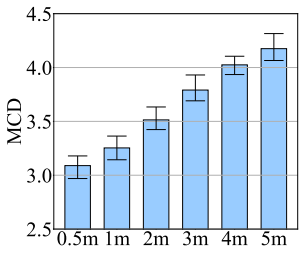


(c) WER of various distance

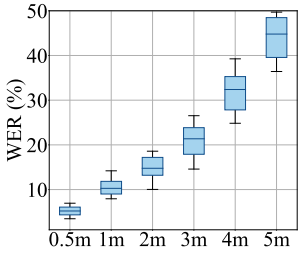


(d) WER of various angle

Figure 14. MCD and WER of different (a) distance and (b) angle between radar and reverberator.



(a) MCD at different distances



(b) WER at different distances

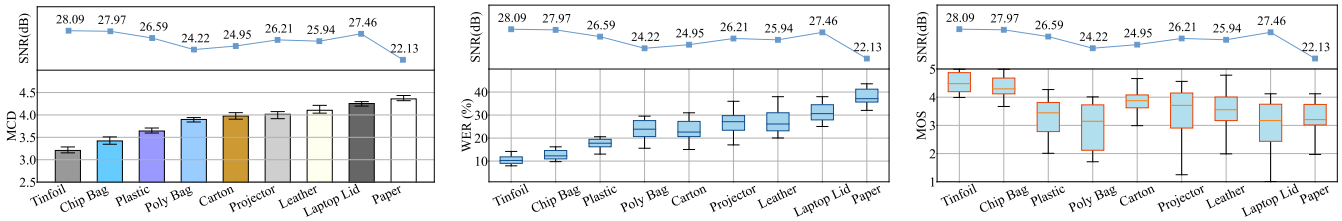
Figure 15. The impact of the distance between the audio source and the reverberator on the system performance.

and angle setting. It indicates mmEcho can launch effective eavesdropping.

6.4.4. Impact of Different Reverberating Materials. The same sound wave can activate various intensities of vibration on different reverberating materials. In this experiment, we evaluate the influence of different materials on the performance of mmEcho. In this experiment, we fix both the radar-to-reverberator and the reverberator-to-source distances at 1m, use frosted glass as insulator, and plays English audio from the loudspeaker.

Figure 16 shows the performance results in terms of MCD, WER, and MOS on different reverberating materials. Among the considered materials, tinfoil and chip bag achieve the lowest MCD of 3.25 and 3.4, respectively. Besides, the average WER of tinfoil and chip bag is also the best, which are 10.60% and 12.72%, respectively. Sound waves produce particularly strong vibrations on these two materials because of their low stiffness, high elasticity, and slimness. In addition to that, they have strong reflectivity for millimeter wave. This allows the antennas of mmWave radar to receive a high signal-to-noise signal. The high MCD and WER for cardboard box (i.e., carton) and paper is due to the weak reflectivity since most of the millimeter-wave signal from the radar penetrates it, and only a slight amount of signal has been reflected back. The MCD and WER of plastic materials (i.e., plastic and polybag) are higher than tinfoil since plastic has higher stiffness and lower reflectivity. The front side of a projector screen cloth is white plastic glass fiber (to scatter the light from a projector), but its backside is more reflective. In our attack scenario, the mmWave radar is outside a room. Hence, it would likely point at the backside of a projector screen. From the backside of the projector screen cloth, mmEcho can reconstruct the audio with an MCD of 4.02 and WER of 26.67%. The MCD and WER of leather are slightly better than that of the projector screen due to the leather's smoother surface, which allows a limited improvement in terms of signal reflectivity. In Figure 16, we can notice that the SNR is not directly related to the audio reconstruction performance on different reverberators. In particular, although some materials (i.e., projector, leather, and laptop lid) have a higher SNR than the other materials, they have worse MCD and WER because of their high stiffness and poor elasticity.

The results of this experiment show that the average MCDs and WERs of the considered reverberators are all below 4.36 and 36.53%, which indicates a significant similarity between the original and reconstructed audio, and high comprehensibility.



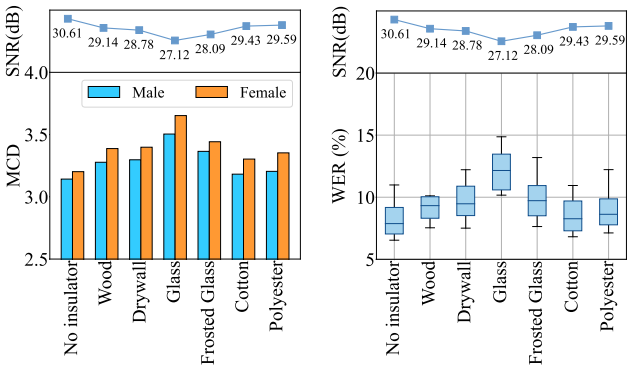
(a) MCD of various reverberator materials (b) WER of various reverberator materials (c) MOS of various reverberator materials

Figure 16. Performance of audio reconstruction on various reverberator materials

We further evaluate the performance of mmEcho with MOS. Figure 16(c) shows the MOS for the above-mentioned reverberators based on the rating of 30 volunteers on nine reconstructed audio samples. We further discuss the MOS results in Appendix B to show the relationship between MOS and reconstructed words. We can observe that each kind of reverberator has a median MOS value higher than 3 on all audio samples. These results further indicate that mmEcho can reconstruct human intelligible audio.

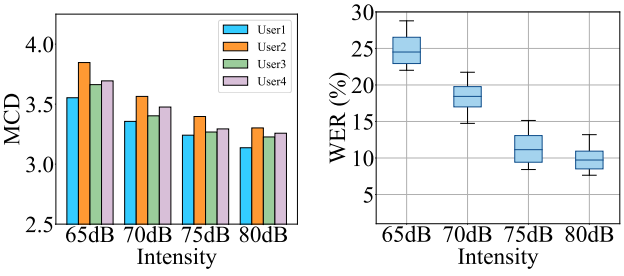
6.4.5. Impact of Different Acoustic Insulators. In our attack model, we assume that the victim uses a soundproof insulator (e.g., acoustic insulation glass, acoustic foam, wood panels) to prevent eavesdropping. In this experiment, we evaluate the influence of different insulation materials on the performance of mmEcho. We place a reverberator made of tinfoil at a distance of 1m from the insulator and the insulator closely in front of the radar. A loudspeaker is placed at a distance of 1m from the reverberator, and we play English audios from different users. We select five common acoustic insulation materials in the modern building for evaluation, which are dense wood, glass, frosted glass, cotton, and polyester.

In Figure 17, we report the results in terms of MCD and WER of the reconstructed audio considering different insulators. On the top side of figures 17(a) and 17(b), we also show the measured SNR of the radar transceiver signals for reference. Among the considered insulators, normal glass has the worst MCD and WER due to its strong reflectivity [44], [45]. Nevertheless, the altered surface of frosted glass reduces its reflectivity, which makes the MCD and WER for frosted glass better than the one for normal glass. We can also observe that we achieve the lowest MCD and WER values with insulators made of cotton and polyester, which is due to their low reflectivity for millimeter wave. In general, the audio reconstructed from male speech achieves a better MCD and WER than the one from female speech since a male voice typically has a lower frequency range than a female voice [46]. In Figure 17, we can observe that an insulator's physical properties affect the SNR of the radar signal and, in turn, the reconstruction accuracy. Nonetheless, mmEcho achieves an MCD lower than 3.65 and a WER lower than 13.25% for all the considered settings. Therefore, we can conclude that mmEcho can effectively reconstruct audio by penetrating an acoustic insulator, which makes eavesdropping possible in common indoor spaces.



(a) MCD of different insulators (b) WER of different insulators

Figure 17. The impact of various acoustic insulators (for each insulator, SNR is provided).

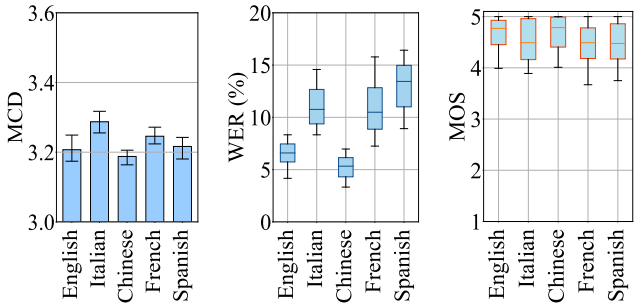


(a) MCD at various volume (b) WER at various volume

Figure 18. The impact of the various sound volume.

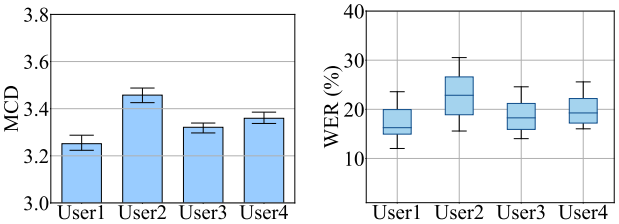
6.4.6. Impact of Sound Volume. The sound volume of the audio source has a direct influence on the intensity of vibration on a reverberator. In this experiment, we assess the impact of the sound level on the performance of mmEcho. We keep the same setting of our previous experiment in Section 6.4.5 and use frosted glass as an insulator and tinfoil as a reverberator. We place the loudspeaker at a fixed distance (1m) from the reverberator and vary the sound level from 65 to 80dB. Figure 18 shows the performance of mmEcho at different sound levels. As expected, the MCD and WER of the reconstructed audio increase with the decrease of the sound volume. At the voice level of a normal conversation (around 65dB) [42], [47], [48], the average MCD and WER are below 4 and 30% respectively, which indicates mmEcho can effectively reconstruct audio to eavesdrop on the speech.

6.4.7. Impact of Different Languages. The spoken language of the original audio may affect the quality of reconstruction due to the different combinations of phonemes



(a) MCD (b) WER (c) MOS

Figure 19. Performance of audio reconstruction in different languages



(a) MCD (b) WER

Figure 20. The impact of mobile audio source on mmEcho

used [43], [49] (i.e., phonotactics). For this reason, we evaluate the performance of mmEcho on different spoken languages: English, Chinese, Italian, Spanish, and French. For English, we use the audio of User_1. And for Italian, Chinese, French, and Spanish, we use the audio from User_5 to User_8 respectively. In this experiment, we use the same experimental settings as the previous experiment in Section 6.4.6. In Figure 19(b), we report the MCD, WER, and MOS scores for the reconstructed audio with different languages. MCD for all considered languages is below 3.28. For WER, we can observe there is a small variance, and this is because different languages have different phonology and morphology [49], which leads to different accuracy when the reconstructed audio is processed by a speech recognition system. As shown in Figure 19(b), all the researched languages have a WER below 15%. To exclude errors due to the speech recognition system, we further evaluate the impact of language via the subjective metric MOS. Figure 19(c) shows the evaluation from 30 volunteers, and we can observe that the median MOS of all five subject languages is above 4.45, which indicates that mmEcho has the ability to reconstruct different languages.

6.4.8. Impact of a Moving Audio Source. In a real-world scenario, the audio source can move around within the room. For example, a user may have a phone call while walking or sitting. Therefore, we assess the influence of a mobile audio source on the performance of mmEcho. In this experiment, we use a reverberator made of tinfoil and fix the radar-to-reverberator distance at 1m and direction angle at $0°$. During the experiment, we keep moving the loudspeaker with a speed of 0.5m/s in a reciprocating motion, varying the reverberator-to-source distance between 1 and 3m. The loudspeaker plays the English audio from User_1 to User_4. We report the results in terms of MCD and WER in Figure 20. Despite a slight increase of MCD compared to a stationary audio source (see Figure 14(a)), the average MCD is below 3.5 for all users. Moreover, the average WER for the four users is 18.72%, with a maximum value of 30.5%. Our results show that mmEcho can reconstruct the original audio even when the audio source is in motion.

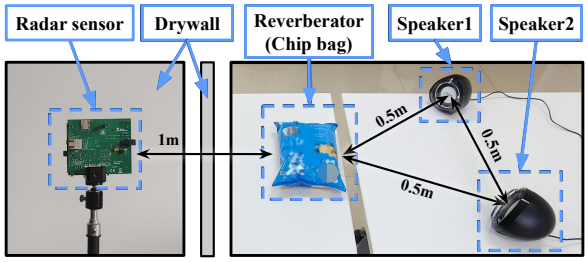
6.4.9. Multi-Source Reconstruction. In real-world scenarios, multiple audio sources can be present within the same room, e.g., people engaging in a conversation in person or over the phone. mmEcho enables eavesdropping on multiple audio sources simultaneously. We provide a proof-of-concept experiment to demonstrate this capability. As depicted in Figure 21(a), we place a reverberator (i.e., chip bag) at 1m distance from the radar and separate them with an insulator (i.e., drywall). We position the two loudspeakers (i.e., *Speaker1* and *Speaker2*) at the same distance (0.5m) from the reverberator and space them 0.5m apart from each other. We use Google speech synthesis system[10] to generate a conversation between two individuals, i.e., *Source1* and *Source2*, and play their audio on *Speaker1* and *Speaker2*, respectively. We report the resulting spectrogram for this experiment in Figure 21(b). The reverberator captures the sound from multiple audio sources. Hence, mmEcho can reconstruct the audio content of the whole conversation.

To further investigate the superposition of sound waves on the reverberator, we perform an additional experiment using the same setup in Figure 21(a) where *Speaker1* and *Speaker2* simultaneously play an incremental sweep-tone from 10Hz to 3kHz (*Source1*) and a decremental sweep-tone from 3kHz to 10Hz (*Source2*), respectively. From the resulting spectrogram in Figure 21(c), we can observe that mmEcho correctly reconstructs the sweep-tones. Therefore, mmEcho enables simultaneous eavesdropping from multiple audio sources.

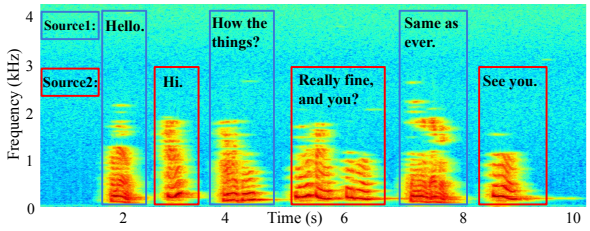
6.4.10. Comparison with Other mmWave-Based Eavesdropping Methods. In this section, we compare the audio reconstruction performance of mmEcho with two recent machine learning based works: mmSpy [18] and MILLIEAR [17]. We report the comparative results in terms of the evaluation metrics used in these two works. For the sake of fairness of comparison, we apply mmEcho to the attack scenarios defined by mmSpy and MILLIEAR for performance evaluation.

Applying mmEcho in the Attack Model of mmSpy. In this comparative analysis, we assess the reconstruction accuracy of mmEcho under the scenario and experimental settings considered by mmSpy in [18]. During a phone call, mmSpy aims to recover the remote caller's speech by measuring the vibrations produced by the phone's earpiece. According to the attack model of mmSpy, we point our mmWave radar towards the backside of the phone under test to measure the vibrations induced by its earpiece while play-

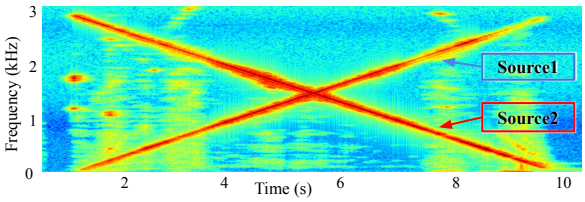
10. Google Text-To-Speech, https://cloud.google.com/text-to-speech



(a) Experimental setup of multi-source reconstruction.



(b) Conversation reconstruction performance where *Speaker1* plays the audio of *Source1* and *Speaker2* plays the audio of *Source2*.



(c) Superimposed audio reconstruction performance.

Figure 21. Multi-source reconstruction setup and results.

ing audio from the same dataset (i.e., AudioMNIST [50]). We compute the accuracy of the audio reconstructed by mmEcho via the Google Speech-to-Text system since, unlike mmSpy, mmEcho does not classify each word individually. Figure 22 shows the reconstruction accuracy of mmSpy reported in [18] and our mmEcho for the considered phones varying the radar-to-phone distance from 1 to 6ft (i.e., 30.48 to 182.88cm). The average accuracy of mmSpy is 61.89%. In comparison, the average accuracy of mmEcho is 86.27% (i.e., WER 13.73%), corresponding to a 24.38% improvement. Along with the increase in distance, the performance of mmSpy degrades rapidly while our scheme maintains a reasonable performance.

Applying mmEcho in the Attack Model of MILLIEAR. MILLIEAR aims to reconstruct the audio played by a loudspeaker by measuring the vibrations on the speaker's drive via a mmWave radar. In this analysis, we evaluate the reconstruction accuracy of mmEcho on the attack model considered by MILLIEAR in [17]. In particular, we position the mmWave radar and the loudspeaker (Philips SPA311) at a radar-to-speaker distance of 1.5m and separated by a glass insulator. We point the radar directly at the loudspeaker's drive. From the audio dataset used by MILLIEAR in [17], we play the audio of four users on the loudspeaker, i.e., $User_1$ from Barack Obama, $User_2$ from Taylor Swift, $User_3$ from Bill Gates, and $User_4$ from Anne Hathaway. Figure 23 shows the reconstruction performance in terms of MCD

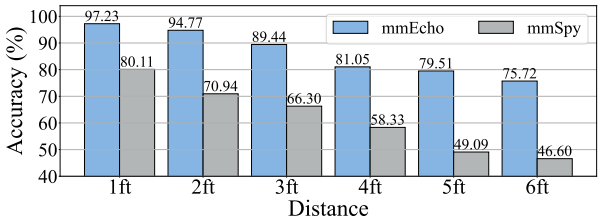


Figure 22. Comparison of reconstruction results by mmEcho and mmSpy [18]

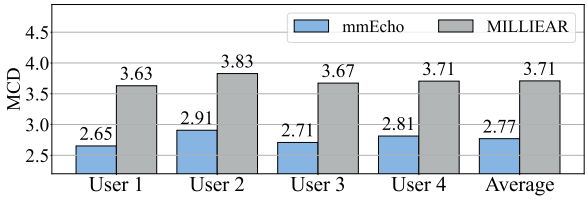


Figure 23. Comparison of reconstruction results by mmEcho and MILLIEAR [17]

of mmEcho and MILLIEAR. The average value of MCD for MILLIEAR is 3.71 while the one of mmEcho is 2.77, corresponding to a 25.34% performance improvement.

7. DISCUSSION

In this section, we discuss several insights from our experiments, the advantages and limitations of our approach, and possible future work.

Key Factors Influencing the Audio Reconstruction. As a result of the evaluation in Section 6, the audio reconstruction performance of our system is affected by five key factors: radar-to-reverberator angle and distance, reverberator-to-source distance, reverberator's material, and insulator's material. The reberberator's distance from the audio source and inherent properties of reverberator's material (e.g., elasticity and reflectivity) influence the amplitude of the sound-induced vibrations on the reverberator. The other factors influence the SNR of the radar signal due to signal attenuation [51]. Nonetheless, we could improve the SNR by increasing mmWave radar transmission power or relying on multi-antenna techniques, such as beamforming and MIMO [52].

Multi-object Simultaneous Eavesdropping. We can carry out an eavesdropping attack with the vibration of multiple reverberating objects to improve audio reconstruction. This is possible since distinct objects fall in different range bins in the Range-FFT spectrum. While we currently set a range gate on a single target object to exclude other objects, we can also measure the vibrations on another object (or more of them) by considering its range gate. The sound wave propagation in the air induce different vibration intensity on different objects (due to their distance and material). Hence, we can combine the variation of vibration measurements on multiple objects to further improve the audio reconstruction.

Eavesdropping from Audio Sources vs. Reverberating Objects. Several of the research work carry out eavesdropping by directly targeting loudspeakers as audio sources [9]. However, audio sources have many other forms, such as human beings and mobile devices (e.g., smartphones and tables). Different from the diaphragm in a loudspeaker, such audio sources do not expose a suitable surface to reflect the millimeter-wave signal. For example, the built-in loudspeaker on a mobile device presents a tiny surface while the signal reflected by a human being is negligible (i.e., the millimeter-wave easily penetrates organic material). In the case of a loudspeaker with a sufficiently large reflective surface (i.e., loudspeaker drive), a mmWave radar may not effectively capture vibrations due to the loudspeaker's orientation. Moreover, an audio source may change its position, which makes it unfeasible for the mmWave radar to point precisely and directly at such an audio source. By indirectly reconstructing audio from the vibration on reverberating objects, mmEcho can provide a novel and reliable eavesdropping approach even under those unfavorable settings to other state-of-the-art works. mmEcho only needs a stationary reverberating object and within the mmWave radar range, even behind an insulator among the ones in Section 6.4.5. In addition, the loudspeaker's diaphragm itself can be used as reverberator to reconstruct the human voice. As a future work, we will investigate the potential of mmEcho to eavesdrop from moving reverberating objects. Since adjacent chirps occur within 0.1ms and they propagate much faster than a moving object, measuring the sound-induced vibrations on such an object should have a negligible effect on the reconstructed audio.

Frequency of the Reconstructed Audio Limited by Chirp Rate. We utilize the intra-chirp method, i.e., we obtain the exact distance information from a single chirp and then extract the vibration information in the time domain. Therefore, the chirp rate also determines the sampling rate of the reconstructed audio thus, according to Nyquist's theorem [53], [54], we can reconstruct frequencies up to half of the chirp rate. In our work, the mmWave radar we use has a chirp rate of 10kHz. Hence, we can reconstruct audio up to 5kHz. Consequently, our audio reconstruction performance is limited by the radar parameters. Therefore, we believe that we could achieve even better performance by relying on customized radar with a higher chirp rate.

Signal Processing Unaided by Machine Learning. We implemented mmEcho using signal processing techniques without resorting to machine learning. Compared to machine learning, which is dataset-dependent and requires extensive training, the sole application of signal processing allows mmEcho to achieve practical and efficient eavesdropping while avoiding any dependence on data. In other words, our system can reconstruct the complete audio without requiring any information about the target (e.g., voice samples of the user, audio source's location), which enables a more practical eavesdropping, i.e., unconstrained vocabulary and no prior knowledge required.

Defenses against mmWave-Based Eavesdropping. As practical countermeasures against our attack, we propose RF shielding and signal jamming. The use of RF shielding methods is a possible countermeasure since they aim to thwart RF signals. The ideal yet unrealistic solution is to enclose the entire victim's room in a Faraday cage [55] to block every external electromagnetic field, including mmWaves. Working on a similar principle, applying an RF shielding mesh is a more practical solution to attenuate RF signals from outside the victim's room [56]. Therefore, we experimentally evaluate the effect of such a solution on the audio reconstruction results of mmEcho. Considering the experimental setup in Section 6.4.2, we apply a metal mesh (stainless steel, 5mm aperture, and 0.5mm wire diameter) on the insulator as an RF shield. The SNR of the received radar signal decreases from 28.78 (without) to 19.46dB (with RF shielding mesh). Likewise, the MCD of the reconstructed audio increases from 3.26 to 11.39. Therefore, this countermeasure is effective since the MCD of the reconstructed audio is higher than 8, i.e., speech recognition systems cannot recognize the content of such audio [40].

Due to cost constraints, commercial mmWave radars are typically not robust against signal interference. Hence, a viable defense is to jam the signal of a target radar (*eavesdropper*) using another radar (*jammer*). However, an FMCW radar receiver expects to receive signals with a predefined frequency pattern and filters signals from other frequency bands. To address this issue, we can first obtain the *eavesdropper*'s frequency band by using a spectrum analyzer and then use the *jammer* to continuously transmit RF signals in the same frequency band to degrade the *eavesdropper*'s performance.

8. Conclusion

In this work, we propose a mmWave-based non-invasive acoustic eavesdropping system that reconstructs the original audio via signal processing techniques and without the aid of machine learning or prior knowledge. The results of our extensive evaluation show the effectiveness of our attack in real-world scenarios and under different conditions, such as distance, direction, reverberating material, sound insulator, sound volume, and spoken language. Our method can also eavesdrop on sound from multiple and moving audio sources, which significantly improves the success rate of our attack.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work is supported by the National Key Research and Development Program of China (Grant No. 2021YFB3100400), National Natural Science Foundation of China (Grant No. 62202276, 61832012), Shandong Science Fund for Excellent Young Scholars (Grant No. 2022HWYQ-038), and Future Young Scholars Project of Shandong University.

References

[1] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *Proc. of USENIX Security Symposium*, 2014, pp. 1053–1067.

[2] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer," in *Proc. of Network and Distributed System Security (NDSS)*, 2020.

[3] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in *Proc. of IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 1000–1017.

[4] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *Proc. of ACM SIGMobile Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2015, pp. 301–315.

[5] J. Han, A. J. Chung, and P. Tague, "Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in *Proc. of ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2017, pp. 181–192.

[6] A. Kwong, W. Xu, and K. Fu, "Hard drive of hearing: Disks that eavesdrop with a synthesized microphone," in *Proc. of IEEE symposium on security and privacy (SP)*, 2019, pp. 905–919.

[7] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *Proc. of ACM SIGMobile Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2016, pp. 57–69.

[8] H. A. C. Maruri, P. Lopez-Meyer, J. Huang, W. M. Beltman, L. Nachman, and H. Lu, "V-speech: Noise-robust speech capturing glasses using vibration sensors," in *Proc. of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 2, no. 4, 2018, pp. 1–23.

[9] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, "Speake (a) r: Turn speakers to microphones for fun and profit," in *Proc. of USENIX Workshop on Offensive Technologies (WOOT)*, 2017.

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

[11] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: eavesdropping via lidar sensors," in *Proc. of ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2020, pp. 354–367.

[12] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics*, vol. 33, no. 4, 2014.

[13] B. Nassi, Y. Pirutin, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Real-time passive sound recovery from light bulb vibrations," in *Proc. of USENIX Security Symposium*, 2022.

[14] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proc. of ACM SIGMobile Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2015, pp. 130–141.

[15] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2907–2920, 2016.

[16] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proc. of ACM SIGMobile Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2019, pp. 14–26.

[17] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, 2021, pp. 11–20.

[18] S. Basak and M. Gowda, "mmspy: Spying phone calls using mmwave radars," in *Proc. of IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 995–1012.

[19] P. Hu, H. Zhuang, P. S. Santhalingam, R. Spolaor, P. Pathak, G. Zhang, and X. Cheng, "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *Proc. of IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 1530–1530.

[20] P. Walker and N. Saxena, "Sok: assessing the threat potential of vibration-based attacks against live speech using mobile sensors," in *Proc. of ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 2021, pp. 273–287.

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

[22] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "Uwhear: through-wall extraction and separation of audio vibrations using wireless signals," in *Proc. of ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2020, pp. 1–14.

[23] G. Brooker and J. Gomez, "Lev termen's great seal bug analyzed," *IEEE Aerospace and Electronic Systems Magazine*, vol. 28, no. 11, pp. 4–11, 2013.

[24] Y. Rong, S. Srinivas, A. Venkataramani, and D. W. Bliss, "Uwb radar vibrometry: An rf microphone," in *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1066–1070.

[25] R. Khanna, D. Oh, and Y. Kim, "Through-wall remote human voice recognition using doppler radar with transfer learning," *IEEE Sensors Journal*, vol. 19, no. 12, pp. 4571–4576, 2019.

[26] Z. Li, F. Ma, A. S. Rathore, Z. Yang, B. Chen, L. Su, and W. Xu, "Wavespy: Remote and through-wall screen attack via mmwave sensing," in *Proc. of IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 217–232.

[27] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, "mmvib: micrometer-level vibration measurement with mmwave radar," in *Proc. of ACM SIGMobile Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2020, pp. 1–13.

[28] L. Wen, Y. Li, Y. Ye, C. Gu, and J.-F. Mao, "Audio recovery via noncontact vibration detection with 120 ghz millimeter-wave radar sensing," in *Proc. of International Conference on Microwave and Millimeter Wave Technology (ICMMT)*. IEEE, 2021, pp. 1–3.

[29] Y. Dong and Y.-D. Yao, "Secure mmwave-radar-based speaker verification for iot smart home," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3500–3511, 2020.

[30] E. Guerrero, J. Brugués, J. Verdú, and P. de Paco, "Microwave microphone using a general purpose 24-ghz fmcw radar," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1–4, 2020.

[31] L. Piotrowsky, J. Siska, C. Schweer, and N. Pohl, "Using fmcw radar for spatially resolved intra-chirp vibrometry in the audio range," in *Proc. of IEEE/MTT-S International Microwave Symposium (IMS)*, 2020, pp. 791–794.

[32] F. Chen, S. Li, C. Li, M. Liu, Z. Li, H. Xue, X. Jing, and J. Wang, "A novel method for speech acquisition and enhancement by 94 ghz millimeter-wave sensor," *Sensors*, vol. 16, no. 1, p. 50, 2015.

[33] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren, "Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals," in *Proc. of ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2021, pp. 97–110.

[34] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[35] T. I. Inc., "Iwr1642 single-chip 76- to 81-ghz mmwave sensor datasheet (rev. b)," 2018. [Online]. Available: https://www.ti.com/lit/ds/symlink/iwr1642.pdf

[36] M. Chounlakone and J. Alverio, "The laser microphone," 2002.

[37] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[38] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E.-S. M. El-Rabaie, W. Al-Nuaimy, S. A. Alshebeili, and F. E. Abd El-samie, "Speech enhancement with an adaptive wiener filter," *International Journal of Speech Technology*, vol. 17, no. 1, pp. 53–64, 2014.

[39] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[40] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang, and W. Xu, "The feasibility of injecting inaudible voice commands to voice assistants," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[41] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 577–582.

[42] S. Rosen and A. Fourcin, "Frequency selectivity and the perception of speech," *Frequency selectivity in hearing*, vol. 373487, 1986.

[43] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. of National Academy of Sciences*, vol. 102, no. 7, pp. 2293–2298, 2005.

[44] C. Hoffman and R. Driggers, *Encyclopedia of Optical and Photonic Engineering (Print)-Five Volume Set*. CRC Press, 2015.

[45] J. Kokkoniemi, J. Lehtomäki, and M. Juntti, "Reflection coefficients for common indoor materials in the terahertz band," in *Proc. of ACM International Conference on Nanoscale Computing and Communication (NanoCom)*, 2018, pp. 1–6.

[46] T.-W. Lee and G.-J. Jang, "The statistical structures of male and female speech signals," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2001, pp. 105–108.

[47] R. T. Sataloff, "The human voice," *Scientific American*, vol. 267, no. 6, pp. 108–115, 1992.

[48] C. J. Chen and D. A. Miller, "Pitch-synchronous analysis of human voice," *Journal of voice*, vol. 34, no. 4, pp. 494–502, 2020.

[49] R. Jakobson and M. Halle, *Fundamentals of language*. De Gruyter Mouton, 2020.

[50] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR abs/1807.03418*, 2018.

[51] M. Marcus and B. Pattan, "Millimeter wave propagation: spectrum management implications," *IEEE Microwave Magazine*, vol. 6, no. 2, pp. 54–62, 2005.

[52] D. R. Fuhrmann and G. San Antonio, "Transmit beamforming for mimo radar systems using signal cross-correlation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 1, pp. 171–186, 2008.

[53] H. Landau, "Sampling, data transmission, and the nyquist rate," *Proc. of the IEEE*, vol. 55, no. 10, pp. 1701–1706, 1967.

[54] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh, "A universal sampling method for reconstructing signals with simple fourier transforms," in *Proc. of ACM SIGACT Symposium on Theory of Computing (STOC)*, 2019, pp. 1051–1063.

[55] S. J. Chapman, D. P. Hewett, and L. N. Trefethen, "Mathematics of the faraday cage," *SIAM Review*, vol. 57, no. 3, pp. 398–417, 2015.

[56] N. Amralah, "Shielding effectiveness of metal mesh and radio frequency shielding film for optical applications," in *Proc. of IEEE Antenna Measurement Techniques Association Symposium (AMTA)*, 2021, pp. 1–5.

Appendix A. Computational and Time Overheads.

We implement the signal processing and audio reconstruction in Python v3.10 using Cython v0.29.32, Scipy v1.9.0, and Numpy v1.23 libraries. We collect the signal from the radar sensor, buffering data batches of a duration of $d = 4$ second. The processing time and amount of memory required to store a batch is proportional to the duration d. Regarding the memory requirements, processing of a signal data batch with $d = 4$ requires at most 400MB of RAM. At our current implementation, the overall time for signal processing and audio reconstruction of an individual batch using a single CPU core is around 1.8s on our laptop (see Section 6). Since the overall processing time is lower than a batch's duration, we can attain a continuous audio stream with a delay of around 6s (i.e., batch data collection and processing time) from the start of signal acquisition.

Appendix B. Intelligibility of the Reconstructed Audio

Mean Opinion Score (MOS) is a metric that provides a subjective evaluation on the intelligibility of the reconstructed audio as a numerical value. In Table 3, we shows the rating scale for the MOS metric in a range from "Bad" and "Excellent" that correspond to 1 and 5, respectively. We also report examples of the relationship between the MOS value given by the participants and the related sentences in Table 4. We can notice that the MOS is associated with unrecognized words' length and significance in understanding the meaning of an overall sentence. Moreover, participants' failure to recognize uncommon (e.g., "drowsiness") or multiple consecutive words (e.g., "questions, and you") influences them to give a low MOS. When MOS is below 2, participants are unable to understand the audio content. When MOS is in a range between 3 to 5, participants can grasp useful information from the audio even if they do not recognize a few individual words.

TABLE 4. MOS AND THE CORRESPONDING RECOGNIZED WORDS. WE ENCLOSE IN PARENTHESES THE WORDS OR PART OF WORDS THAT USERS FAIL TO RECOGNIZE.

MOS	Participant's understanding from the reconstructed audio
4∼5	...guide giving instructions to (*a*) group of international students in Canada, preparing for (*a*) (*whale*) watching trip, before you hear the talk, you have some time to look at question...
3∼4	...(*At*) the end of the test, you will be given 10 minutes to transfer your answers to (*an*) answer (*sheet*). Now (*turn*) to section one, first you (*have*) some time to look at question one to six...
2∼3	...and (*questions*), (*and*) (*you*) will have a chance to check your (*word*), (*all*) the recordings will be played (*once*) (*only*), the test (*is*) in four sections...
1∼2	...if you think you might get (*seasick*) take one of these (*patches*) and (*put*) (*it*) on your (*arm*), it works on (*pressure points*) of the body and will (*release*) (*seasick*)ness without the (*drowsiness*) (*you*) can get (*from*) pills. (*Are*) there (*any*) other questions...