This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

Towards Unconstrained Vocabulary Eavesdropping with mmWave Radar using GAN

Pengfei Hu, Wenhao Li, Yifan Ma, Panneer Selvam Santhalingam, Parth Pathak, Hong Li, Huanle Zhang, Guoming Zhang, Xiuzhen Cheng, *Fellow, IEEE*, Prasant Mohapatra, *Fellow, IEEE*

Abstract—As acoustic communication systems become increasingly common in our daily life, eavesdropping brings severe security and privacy risks. Current methods of acoustic eavesdropping either provide low resolution due to the use of sub-6 GHz frequencies, work only for limited words based on classification approaches, or cannot work through-wall because of the use of optical sensors. In this paper, we present MILLIEAR, a mmWave acoustic eavesdropping system that leverages the high-resolution of mmWave FMCW ranging and generative machine learning models to not only extract vibrations but to reconstruct the audio. MILLIEAR combines speaker vibration estimation with conditional generative adversarial networks to eavesdrop and recover high-quality audios (i.e., with no vocabulary constraints). We implement and evaluate MILLIEAR using off-the-shelf mmWave radars deployed in different scenarios and settings. Evaluation results clearly show that MILLIEAR can accurately reconstruct the audio even at different distances, angles, and through the wall with different insulator materials. In addition, our subjective and objective evaluations demonstrate that the reconstructed audio has a strong similarity with the original audio.

Index Terms—Acoustic Eavesdropping, mmWave Radar, Vibration Sensing, Generative Adversarial Network

1 Introduction

ACOUSTIC communication systems such as video conferencing, personal digital assistants, and home entertainment are becoming increasingly popular. While our digital communication (data transmission) over the Internet is protected through encryption techniques, the "last hop" of the acoustic communication systems, i.e., the voice emitting from speakers, is unencrypted. This unencrypted information coming from the speaker can reveal highly private or confidential information. Therefore, acoustic eavesdropping poses major security and privacy risks, considering the increasing prevalence of acoustic communication systems in homes and offices.

Acoustic eavesdropping attacks have been studied extensively where the core idea is to capture the vibrations generated by a speaker using different types of sensors. As an example of the "in-room" category of attacks, an IMU sensor can be used to listen to acoustic signals [1]–[4]. While these methods primarily operate by placing the sensor in the same room as the speaker or pre-installed on the victim's devices, "outside-room" attacks can remotely eavesdrop while being next door or farther away from the audio source. For example, high-speed cameras [5], lasers [6], photodiodes [7], or WiFi signals [8], [9] have been used for remotely discerning the spoken text through vibrations. Compared to "in-room" eavesdropping, "outside-room" eavesdropping is more difficult to prevent, and thus posees a higher risk.

In this paper, we target the "outside-room" scenario, where the attacker device has no near access to the victim. We propose to leveraging wireless communications as they can penetrate walls or soundproof windows. Specifically, we present MILLIEAR, a system that combines the high sensing resolution through mmWave signals and the regenerative capabilities provided by machine learning models to create a highly effective acoustic eavesdropping attack. It addresses many limitations of the prior attack systems including the following aspects:

1) *Higher resolution:* The sensing resolution is closely related to the wireless bandwidth. Therefore, compared to existing RF-based eavesdropping systems that operate at sub-6 GHz frequencies [8], [9], MILLIEAR uses a mmWave FMCW radar that can leverage the large available bandwidth at mmWave spectrum to provide better range resolution. As we show in this work, speaker vibrations of as low as tens of microns can be detected using a mmWave radar for accurate eavesdropping.
2) *Unconstrained vocabulary:* Majority of existing eavesdropping systems such as [1]–[4], [8]–[10] regard acoustic signal extraction as a classification problem by profiling a handful of words. That is, their systems classify each eavesdropped sound to one of the pre-defined limited words (e.g., good, happy, thanks). Generally, the number of words in their systems is within hundreds at most, as the model becomes untractable to classify more words. In practice, however, the content of human conversation is extremely diverse and thus predefining a small set of words does not work well for real eavesdropping attacks. In comparison, MILLIEAR

Pengfei Hu, Wenhao Li, Yifan Ma, Xiuzhen Cheng, Huanle Zhang, and Guoming Zhang are with the School of Computer Science and Technology at Shandong University, China. Email: {phu, dtczhang, guomingzhang, xzcheng}@sdu.edu.cn, {li_wenhao, mayifan}@mail.sdu.edu.cn
Panneer Selvam Santhalingam and Parth Pathak are with the Computer Science Department at George Mason University, USA. Email: {psanthal, phpathak}@gmu.edu
Hong Li is with the Institute of Information Engineering, Chinese Academy of Sciences, China. Email: lihong@iie.ac.cn
Prasant Mohapatra is with the Department of Computer Science, UC Davis, USA. Email: pmohapatra@ucdavis.edu
Huanle Zhang and Guoming Zhang are corresponding authors.

demonstrates the attack with unconstrained vocabulary as it does not require training for classifying words. Instead, it provides the reconstruction of entire conversational audio entirely from the mmWave vibrations.

3) *Remote, low-cost and smaller sensor footprint:* Unlike [11] and [12] eavesdropping systems which only work when spyware is pre-installed in the victim's systems or devices, MILLIEAR works even behind glass, wooden doors, and walls. Compared to [5], [6] and [7] which require expensive camera sensors, laser transducer or telescope, mmWave radars are low-cost and will become an integral component for next-generation smartphones (5G/6G communications). Furthermore, due to the much smaller wavelength of mmWave signals, the sensor footprint is significantly smaller compared to the large multi-antenna system setup required by sub-6 GHz frequencies.

However, building a high-quality mmWave-based eavesdropping system for the unconstrained vocabulary attack entails several challenges, including:

(1) *Speaker vibration extraction using mmWave radar signals in the presence of multi-path noise.* The signal received at mmWave radar sensor consists of both the signal reflected from the vibrating speaker as well other nearby objects. The multipath effect greatly affects the signal quality. To launch an eavesdropping attack in a real-world scenario, we should design an accurate vibration extraction scheme in the presence of multi-path noise. To address this problem, we measure the phase changes through *virtual* sub-chirps. Specifically, we firstly apply a sliding window on the raw mmWave data to generate sub-chirps. Then, we apply a range-FFT to the sub-chirps for deciding the candidate vibration bins and other bins (i.e., mmWave noise sources). Last, we apply a Doppler-FFT on the refined bins to help us extract the vibrations from the speaker.

(2) *Accurate reconstruction of the audio from mmWave vibrations with unconstrained vocabulary.* The audio captured through mmWave signals can contain any unknown words. This means that we need an machine learning model that can not only classify the existing words based on limited training, but can also learn to reconstruct the acoustic components of any word based on prior training. We address this problem by developing a conditional generative adversarial network (cGAN) that uses mel-spectrograms as images to enhance the mmWave vibration extraction. The cGAN is trained using spectrograms of original audio and their corresponding mmWave captured data, by learning to enhance the mmWave spectrogram to the ones similar to the original. Our cGAN model can remove noise and add representative acoustic components for accurate audio reconstruction.

We implement and evaluate MILLIEAR using off-the-shelf mmWave radars and deploy them in different scenarios and settings. The evaluation results show that MILLIEAR achieves a high similarity between the the reconstruction audio and the original audio. Our contributions can be summarized as follows:

- We present a mmWave acoustic eavesdropping system, named MILLIEAR, that uses off-the-shelf mmWave FMCW radar to accurately capture speaker vibrations. The captured speaker vibrations are then enhanced through a generative machine learning model that requires no prior knowledge of the words in the audio signals. Our model can recreate high-quality audio signal directly from the mmWave radar signals by leveraging cGANs.
- We perform an extensive evaluation of MILLIEAR. We use audio samples from 7 public personalities played through speakers and then captured by a mmWave radar. We use audio samples of more than 25000 words in training and testing, and our thorough evaluations show that MILLIEAR can accurately reconstruct the original audio with the average MCD (Mel-Cepstral Distortion) of 3.68 and the average likert user score of 6.83. In addition, we evaluate MILLIEAR in different scenarios with varying distances and angles between speaker and radar, different types of soundproofing material/wall between the speaker and radar, and different types of speakers. The valuation results clearly show the premium performance of MILLIEAR.

The paper is organized as follows. Section 2 provides the related work. Section 3 discusses mmWave radar and GAN preliminaries with a feasibility study, and Section 4 describes the system overview. Our vibration extraction methods and cGAN architectures are presented in Section 5 and Section 6, respectively. We implement MILLIEAR in Section 7 and evaluate MILLIEAR in Section 8. Last, we discuss MILLIEAR in Section 9 and conclude this paper in Section 10.

2 Related Work

In this section, we review and categorize related works focusing on audio eavesdropping. Table 1 summarizes these works and compares them with MILLIEAR.

Several studies have shown that an attacker can deploy an IMU sensor near the audio source to perform eavesdropping. They show that IMU-based audio sensing can classify words, small phrases, and the speaker gender [1]–[4]. [10] touches on the audio recovery with unconstrained vocabulary. Other similar forms of audio eavesdropping have also been proposed. For example, [12] implements a malware prototype which can turn the speaker into a microphone for the eavesdropping purpose; [11] recovers the audio using a vibration motor; and [13] uses a magnetic hard disk to recover songs and voices by measuring the offset between the read/write head and the track center of the disk. The main disadvantage of these eavesdropping methods is that they require to have physical access to the equipment/sensor in a close proximity of the victim, which reduces their applicability in practice. Also, given that some of the attacks require installing spyware on victim's device (referred as invasive approaches in Table 1), these attacks can be prohibited even if the victim only adopts simple defense strategies.

Wireless signals have also been used to eavesdrop audios. Two studies [8], [9] used WiFi signals to profile movements or vibrations and identify audio. Authors in [8] proposed a method to analyze the WiFi channel state

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

	Sensor	Capability		
		Unconstrained vocabulary	Non-invasive	Through-wall (opaque)
IMU	Gyroscope [1]	✗	✗	✗
	Accelerometer [2]–[4]	✗	✗	✗
	AccEar [14]	✓	✗	✗
	IMU fusion [10]	✗	✗	✗
Misc.	Vibration motor [11]	✓	✗	✗
	Speakers [12]	✓	✗	✗
	Magnetic hard drive [13]	✓	✗	N/A
Optical receiver	High speed camera [5]	✓	✓	✗
	Laser transceiver [6]	✓	✓	✗
	Photodiode [7]	✓	✓	✗
Radio receiver	WiFi-CSI [8]	✗	✓	✓
	WiFi-MIMO [9]	✗	✓	✓
	RFID tag [15]	✓	✗	✓
	UWB [16]	N/A	✓	N/A
	WaveEar [17]	✓	✓	N/A
	mmSpy [18]	✗	✓	N/A
	MILLIEAR	✓	✓	✓

TABLE 1: Eavesdropping approaches in literature and their comparison with MILLIEAR.

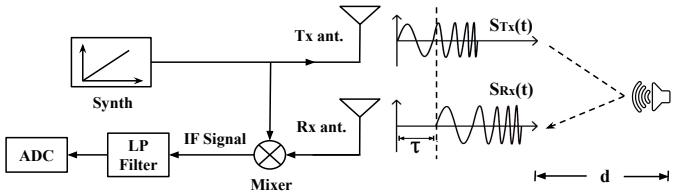


Fig. 1: Structure of an FMCW radar.

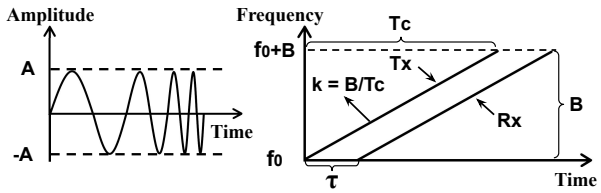


Fig. 2: The waveform definition of FMCW radar signal.

information (CSI) for classifying words. Similarly, in [9], authors analyzed the received signal strength (RSS) of the WiFi signals where the audio vibrations are considered as low-rate modulations of RF signals. Akin to WiFi works, RFID [15] and Doppler radar [19] have been leveraged for eavesdropping. In particular, [15] requires a pre-installed tag in the victim's room. Compared to our approach, these works relying on low resolution traffic data due to lower frequencies and packet rates. Also, they require a multi-antenna setup to localize victims and thus result in larger physical footprint compared to mmWave, making the attack more difficult to be carried out in practice. [16] presents an Impulse Radio Ultra-Wideband based system that is able to simultaneously recover and separate sounds from multiple sources. Using the same RF technology, [20] can recover audio below 400 Hz. However, their capability for recovering unconstrained vocabulary has not been studied. Besides, these works do not explicitly target complete audio reconstruction with unconstrained vocabulary.

Cameras and lasers have also been used for acoustic eavesdropping. Authors in [6] used a laser beam pointing to the sound source or an object near the sound source, to receive the reflected signal and convert it to audio signal. Similarly, [5] used a high-speed video camera to obtain the video of an object in the victim's room (such as a plastic bag, water, etc.) and analyze the response as sound waves impinge on the object to recognize audio. [7] proposed to use a remote electro-optical sensor to analyze the fluctuations to sound of the victim's light bulb. The main disadvantage of these methods is that, apart from the limited vocabulary, these attacks are difficult to carry out as they require expensive, special purpose hardware such as the high-speed cameras.

In other similar research, [21] and [22] use mmWave radar to recover audio below 1 kHz, but the performance of the human audio reconstruction has not been evaluated. mmSpy [18] can eavesdrop on phone calls by using mmWave and machine learning. However, mmSpy studies eavesdropping on constrained vocabulary (hot words and numbers). Authors in [17] used mmWave to acquire high-quality voice from user's vocal vibrations from near-throat region. [23] proposed a speech eavesdropping approach by leveraging the piezoelectric films and mmWave signals. [24] proposed a remote and through-wall screen attack that used mmWave to remotely collect information from LCD screens. [25] showed how mmWave radar can be used for micrometer-level vibration measurement in industrial environments. [26] presents a noise-resistant multi-modal speech recognition system by fusing mmWave radar and microphone. While similar, these works do not focus on acoustic eavesdropping and audio reconstruction.

3 PRELIMINARIES

In this section, we introduce the Frequency Modulated Continuous Wave (FMCW) radar based vibration measurement and the Generative Adversarial Networks (GAN) based signal enhancement.

3.1 Vibration Estimation

An FMCW radar transmits a signal called "chirp". A chirp is a sinusoid whose frequency increases linearly with time. FMCW radars can be used to accurately estimate the object distance and its relative velocity by comparing the transmitted and received signals. Figure 1 illustrates the structure of an FMCW radar, which includes a transmitter antenna and a receiver antenna. The distance d to the object (speaker) can be estimated by calculating the difference between the transmitted and the received signals. With the accurate estimation of distance changes from a mmWave radar, MILLIEAR can infer the vibrations of the speaker and reconstruct the audio.

Figure 2 illustrates the waveform of the FMCW radar signal, where A is the amplitude of signal, f_0 is the start frequency, B is bandwidth of radar, k is the slope of the frequency increase, T_c is the signal duration, T_x and R_x is the transmitted and received signal, respectively. Let $S_{Tx}(t)$ and $S_{Rx}(t)$ be the FMCW transmitted and received (reflected by target) signal represented as

$$S_{Tx}(t) = A_{Tx} \cdot \cos[2\pi \cdot f_{Tx}(t) \cdot t + \phi_{Tx}] \quad (1)$$

$$S_{Rx}(t) = A_{Rx} \cdot \cos[2\pi \cdot f_{Rx}(t) \cdot t + \phi_{Rx}] \quad (2)$$

where $f_{Tx}(t)$, ϕ_{Tx}, and A_{Tx} are the frequency, the phase, and the amplitude of the transmitted signal, respectively. Correspondingly, $f_{Rx}(t)$, ϕ_{Rx}, and A_{Rx} are receiver's signal features. We denote the round-trip delay between the transmitted and received signals as τ, so $f_{Rx}(t) = f_{Tx}(t - \tau)$ is the τ-delayed version of $f_{Tx}(t)$. After applying a mixer on

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

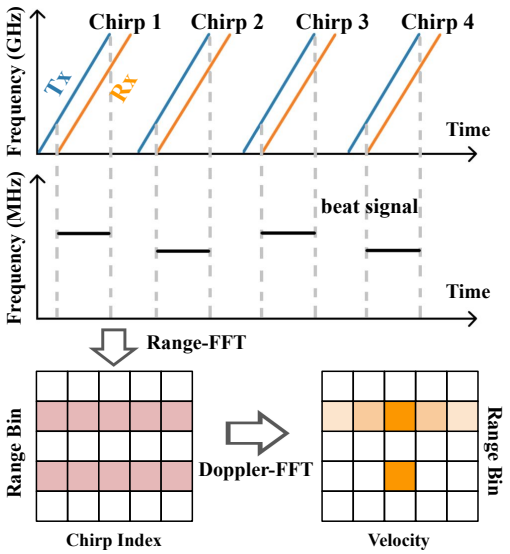


Fig. 3: Chirp generation and processing.

the transmitted and received signal, we can obtain the *beat frequency* signal as follows

$$\begin{aligned} S_b(t) =& S_{Tx}(t)S_{Rx}(t) \\ =& \frac{1}{2}A_{Tx}A_{Rx} \cdot \{\cos[2\pi \cdot f_b(t) \cdot t + \phi_b] + \\ & \cos[4\pi \cdot f_{Tx}(t) \cdot t - 2\pi \cdot f_b \cdot t + \phi_b]\} \end{aligned}$$

where $f_b(t) = f_{Tx}(t) - f_{Rx}(t)$ is the frequency change function of beat signal and $\phi_b = \phi_{Tx} - \phi_{Rx}$. Since the beat frequency (at MHz level) is much lower than the carrier frequency (at GHz level) [27], we can apply a low-pass filter to exclude the carrier. Then the beat frequency signal can be expressed as follows

$$S_b(t) = A_b \cdot \cos[2\pi k\tau t \pm 2\pi f_0 \tau - 2\pi k\tau^2 \mp \phi_b] \quad (3)$$

where $A_b = \frac{1}{2}A_{Tx}A_{Rx}$ is the synthesized amplitude of the transmitter and the receiver. In fact, due to the presence of reflected signals from objects at different distances in the original data, the frequency components in $S_b(t)$ are different. As shown in Figure 3, we perform Range-FFT on fast-time samples in a chirp. It maps the time domain signal to the frequency domain. Objects at different distances have a peak in the frequency domain. Then, we perform Doppler-FFT on the results of Range-FFT for our vibration source positioning task.

From Eq. 3, we can derive the phase of the intermediate frequency signal as follows,

$$\phi = 2\pi f_0 \tau - 2\pi k\tau^2 + \phi_b \quad (4)$$

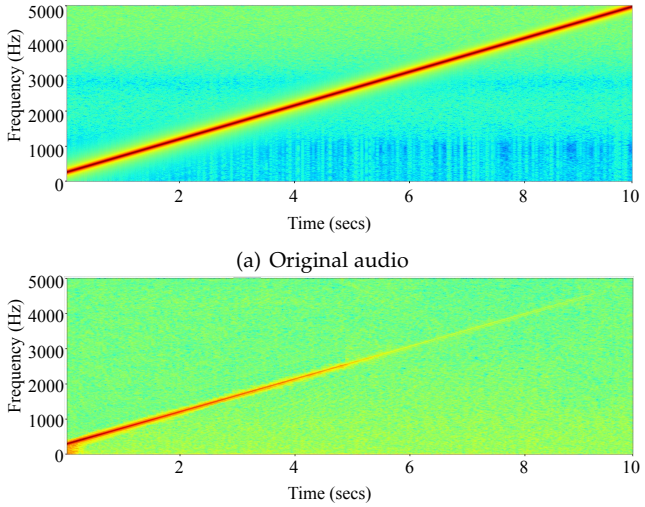
Since τ involves the speed of light c, the accuracy of its calculation will be rough. Therefore, we combine $\tau = 2d/c$ and Eq. 4 to eliminate τ,

$$\phi = 2\pi f_0 \cdot \frac{2d}{c} - 2\pi k \cdot \frac{4d^2}{c^2} + \phi_b \quad (5)$$

Simplifying Eq. 5, we can obtain:

$$8\pi k d^2 - 4\pi f_0 c d + (\phi - \phi_b)c^2 = 0 \quad (6)$$

from Eq. 6, we can derive an accurate distance measurement as:



(a) Original audio

(b) Reflected mmWave signal

Fig. 4: The spectrograms for (a) original audio and (b) reflected mmWave signal from the speaker.

$$d = \left(\frac{f_0}{k} + \sqrt{\frac{f_0^2}{k^2} - 2 \cdot \frac{\phi - \phi_b}{\pi k}} \right) \cdot \frac{c}{4} \quad (5)$$

We perform the linear parabolic interpolation in the phase spectrum from Range-FFT to obtain a wrapping phase. Combined with the phase calculated in Eq. 4, we can achieve an accurate phase estimation, which can be used in Eq. 5. Therefore, we can calculate the distance from the FMCW radar to the speaker by chirps. The vibration estimation can be obtained from the difference between successive distance measurements.

3.2 A Feasibility Study

In order to launch an eavesdropping attack, we verify the correlation between the received mmWave signal and the audio played through a speaker using a proof-of-concept experiment. In the experiment, we let the speaker play an test audio (as shown in Fig. 4(a)) while the mmWave radar is placed in front of the speaker at a $1m$ distance without any blockage. The frequency of the test audio is from $200Hz$ to $5kHz$ to measure the frequency response. Fig. 4 shows the played audio spectrogram and the corresponding mmWave spectrogram captured by the FMCW radar. We observe that the mmWave signal shows a high similarity with the audio signal. Due to the low sampling rate of the FMCW radar, the radar signals show poor similarity with the audio at high frequencies. Also, The FMCW radar suffer from white noise over the whole spectrum. To address these two issues, we enhance the mmWave radar signals reflected from the speaker using a generative machine learning model.

3.3 Generative Adversarial Networks

Generative adversarial networks (GANs) belong to the class of generative models [28]. The goal for GANs is to learn a function that can map between two distributions: the source and the target. The source is a random noise distribution $(p_z(z))$ and the target is the underlying distribution of the data (p_{data}). Once this mapping is learned, GANs can take

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

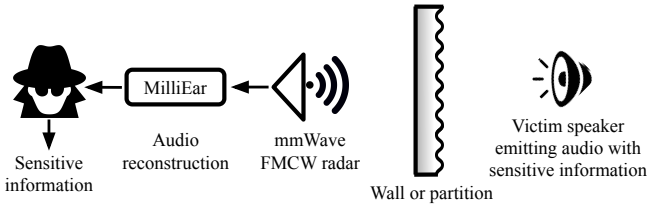


Fig. 5: Our attack scenario of mmWave-based audio eavesdropping.

a sample $z \in p_z$ and map it to sample $x \in p_{data}$. GANs implicitly learn this mapping function and have enabled a plentiful of novel applications [29]–[33]. GAN models are trained by emulating a min-max game between the two networks, one is the generator (G) and the other is the discriminator (D). The generator's objective is to fool the discriminator by generating samples from the noise distribution $p_{z(z)}$ which are similar to those sampled from p_{data}. The discriminator's job is to correctly label the data from the generator as fake and the data from p_{data} as real. The objective function $V(G, D)$ for this min-max game between the two networks can be written as

$$V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[log\ D(x)] + \mathbb{E}_{z\sim p_z(z)}[log\ (1 - D(G(z)))]$$

where the objective of the generator is to minimize $log\ (1 - D(G(z)))$ and the objective of the discriminator is to minimize $log\ D(x)$. An equilibrium is reached when the generator has successfully approximated p_{data} and the discriminator can no longer differentiate between real and fake data.

3.4 Attack Model

Previous approaches to preventing acoustic eavesdropping rely on the use of isolators, such as soundprooof glass, polyethylene foam, and plywood. In this work, we consider the eavesdropping threat which leverages the mmWave radar to reconstruct the sound of the speaker even with the existence of sound-proof isolators. As illustrated in Fig 5, a practical eavesdropping attack is expected to work in the following conditions: (i) there is an acoustic isolation between the attacker and the victim, i.e., the victim's sound cannot penetrate the sound-proof isolator; the attacker cannot deploy any equipment/sensor in the same room as the victim; (ii) the attacker has no prior information about the type of audio information emitting from the victim speaker. The attacker is required to not only able to classify a handful of audio signals (i.e., words or numbers), but to recreate any audio from the entire vocabulary including full sentences. (iii) the device to launch an attack is portable and affordable. The attack model considered in this work is more practical and challenging than existing works. In our work, the attacker can perform sound eavesdropping in this scenario with a low-cost commercial mmWave radar outside the soundproof space.

4 System Overview

Fig. 6 illustrates our mmWave voice eavesdropping system. MILLIEAR has a mmWave radar which can capture the minute vibration cause by the sound. First, the mmWave radar emits an FMCW chirp signal to the vibrating speaker. Then, the signal arrived at the speaker is reflected back to the radar. Last, through careful processing and enhancement of the received signal, MILLIEAR extracts the speaker vibrations. However, due to the background reflection and the multipath effect [34], [35], there are errors in the received signal, resulting in inaccurate estimation of vibrations. To solve this problem, MILLIEAR feeds the vibration data into our Generative Adversarial Network for enhancement and denoising, which achieves high-quality audio reconstruction. MILLIEAR is mainly composed of two modules:

(1) Spectrogram Generation (SG): SG consists of two phases, namely, target (speaker) localization and spectrogram extraction. In order to locate the position of the speaker, MILLIEAR takes several steps. First, it receives the raw data from the mmWave radar as an input. Second, it performs Range-FFT on the raw data to measure the distance to the target. Third, it conducts Doppler-FFT on the result of Range-FFT to find candidate range bins and identify the one that contains the desired vibration. In order to improve the resolution of the FFT, each chirp of a frame was split into multiple sub-chirps to provide multiple observation while extracting the displacement of vocal vibrations. Last, MILLIEAR performs Short-time Fourier Transform (STFT) to each chirp to obtain the time-frequency domain spectrogram. STFT is essentially a windowed Fourier Transform. The formula for STFT and the other details of SG will describe in Section 5.

(2) Audio Reconstruction (AR): The AR module uses a conditional GAN that is trained using two sources of spectrogram images - one from the mmWave radar and the other from original audio. Using the training data, the GAN learns how to enhance the mmWave spectrogram by enhancing representative frequency and amplitude components and reducing noise. The trained GAN model is then used to reconstruct audio directly from the captured mmWave spectrograms. Please note that the GAN training is agnostic to the spoken text and thus does not require any manual annotation during the training. We elaborate on AR in Section 6.

5 Spectrogram Generation

This section explains the spectrogram generation module that consists of the vibration detection component and the vibration extraction component.

5.1 Vibration Detection

To facilitate signal processing, we collect raw binary ADC data via a mmWave radar and convert it into a multidimensional IQ array. To avoid spectral leakage, we segment the acquired IF signal through a windowing process. In this process, we choose the Hanning window. Then, we perform a fast Fourier transform (FFT) to output the Range-FFT spectrum and the phase spectrum, which contain a single chirped frequency bin. The results of the Range-FFT can be used to distinguish multiple objects based on their intermediate frequencies. We identify peaks on the Range-FFT spectrum by applying a continuous wavelet transform (CWT) based peak detection algorithm [36].

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

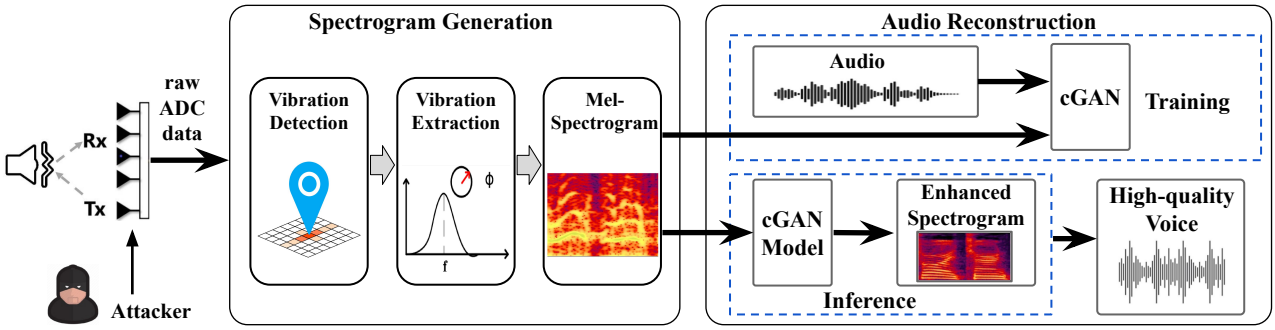


Fig. 6: The MILLIEAR system mainly consists of a mmWave radar, a Data-preprocessing module to extract the vocal spectrogram, and an Audio Reconstruction module to recover high-quality voice.

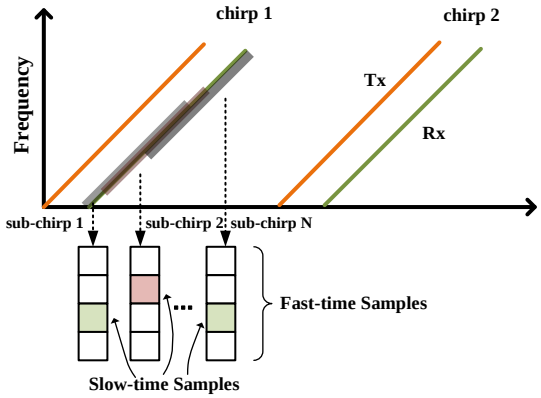


Fig. 7: Vibration extraction from FMCW chirps.

Each frequency peak corresponds to an object within the radar range. However, Range-FFT can only give us the distance of the target. To locate reverberating objects, we performed a Doppler-FFT test based on the results of the Range-FFT. Similar to the Range-FFT spectrum, we can identify objects within the radar vision.

In the Doppler-FFT spectrum, a vibrating object has a significant velocity magnitude on the velocity axis. For an object in a certain range bin, the higher magnitude on the velocity axis, the higher the probability of the vibrating object. Therefore, we select objects with high velocity and set the processing priority in descending order of velocity values. In this way, we can achieve vibrating object localization. Regarding the selection of vibrating objects, a high-pass threshold is set as a buffer in order to avoid the effects of weak object vibrations and other errors.

5.2 Vibration extraction

In order to restore the audio, the vibration displacement must be accurately extracted. We adopt the similar method in [25] for the vibration extraction.

The mmWave radar emits chirps at a fixed time interval and groups a bunch of chirps as one frame for Range-Doppler processing. Range-FFT typically takes all fast-time samples of one chirp as input and generates one slow-time sample. However, low-cost commercial mmWave radars cannot guarantee accurate phase extraction under low SNR based on a single chirp. To improve the phase extraction, we apply a sliding window on fast-time samples within one single chirp to generate more *virtual* sub-chirps as shown in Fig. 7. These sub-chirps are used for cross-referencing with each other. We then conduct Range-FFT on each sub-chirp to obtain multiple slow-time samples. Since the duration of slow-time samples (one frame) are much longer than fast-time samples (one chirp), the time variance of a group of sub-chirps within one chirp can be ignored, i.e., we can consider these sub-chirps being transmitted simultaneously. As shown in Fig. 7, the position of the voice bin detected by sub-chirp 2 (red bin) is different from that of other sub-chirps (green bins). Since we have multiple observations for cross-validation, the abnormal bin (red) can be identified and eliminated. By this approach, we can accurately recognize the correct voice bin.

With the accurate extraction of the voice bin, we perform Doppler-FFT on the slow-time samples to derive the phase. The vibration displacement is calculated according to Eq. 5 once the phase is available. Since the displacement at a specific time is the direct result of the amplitude of audio, we consolidate all the vibration displacements with a timestamp into a waveform as shown in Fig. 4. The maximum chirp rate of the mmWave sensor used in our work is $10kHz$ which is much smaller than the sampling rate of common audio $44.1kHz$. In order to recover audio from the under-sampled vibration waveform, we resort to GAN to enhance the vibration information with more details.

5.3 Mel-spectrogram generation

Our vibration waveform is a one-dimensional signal. However, the conditional generative adversary network (cGAN) in audio reconstruction requires image-like input with correlations among surrounding pixels. Hence, we first transform the waveform to mel-spectrograms. A mel-spectrogram [37] is a popular representation for audio signal which has been widely used in the speech synthesis, audio denoising, etc. We feed this image-like spectrogram into cGAN for enhancement. The enhanced spectogram is then converted back to audio, which leads to a little information loss.

In this work, we choose Short-time Fourier Transform (STFT) to get the time-frequency spectrogram. STFT can be calculated as follows,

$$STFT(t,f) = \int_{-\infty}^{+\infty} x(\tau)h(\tau - t)e^{-j2\pi f\tau}d\tau \qquad (6)$$

where $h(\tau - t)$ is the window function, τ is the half window size of time t, and x is the waveform. Since the magnitude of the generated spectrogram is relatively large, in order to

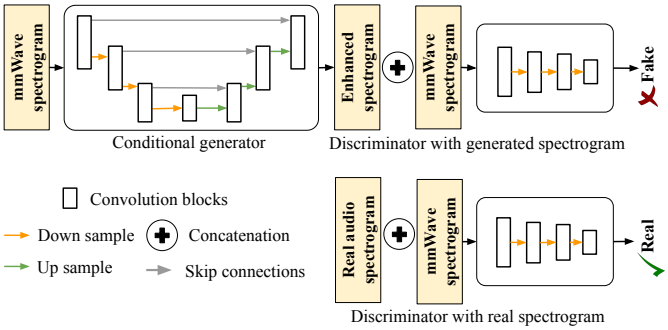


Fig. 8: MILLIEAR cGAN architecture.

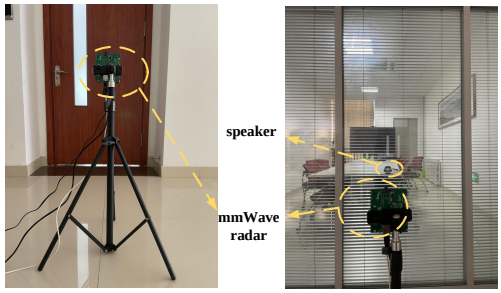


Fig. 9: Two examples of experiment setup of MILLIEAR. **Left**: Conference Room with a dense wood door; **Right**: Conference room with a double-panel glass wall.

obtain a sound feature of a suitable size, it is usually passed through a mel-scale filter bank to produce a mel spectrum. Studies have shown that humans do not perceive frequencies linearly [38]. Instead, humans are better at detecting differences in low frequencies than in high frequencies. For example, we can easily distinguish the difference between 500 Hz and 1000 Hz, but it is difficult for us to distinguish the difference between 10,000 Hz and 10,500 Hz. In order to capture this feature, we convert the spectrogram produced by STFT to mel-spectrogram [37]. The conversion process to calculate the mel-frequency $mel(f)$ follows the equation $mel(f) = 2595 * \log_{10}(1 + \frac{f}{700})$, where f is the frequency. The transformation is performed on both the vibration signal as well as the corresponding audio waveform for the cGAN training and only on the vibration signal during the testing.

6 Audio Reconstruction

This section describes our audio reconstruction module. It covers the GAN architecture and the reconstruction applied in MILLIEAR.

6.1 GAN Architecture

We adopt an image to image translation approach [39] for enhancing the mmWave vibration mel-spectrograms. We use the conditional version of GAN referred as cGAN. Unlike GANs which generate data from a random noise vector (as described in Sec. 3.3), cGANs additionally take a conditional variable, enabling control on the generated data [40]. The objectives of the generator and the discriminator are modified to include the conditional input y. The modified objective functions for the generator and the discriminator are $log(1 - D(y, G(z, y)))$ and $log(D(y, x))$ respectively. Fig. 8 shows our cGAN architecture. While training, the generator takes a mmWave vibration mel-spectrogram as a conditional input and enhances it. The enhanced mel-spectrogram is concatenated with mmWave mel-spectrogram and input to the discriminator. The discriminator is expected to classify this as fake. Additionally, when input with the mel-spectrogram from real audios concatenated with mmWave mel-spectrogram, the discriminator classifies it as real. Inputting the mmWave mel-spectrogram conditions the discriminator and forces the generator to generate the output corresponding to the input mmWave mel-spectrogram instead of any real looking mel-spectrogram. As the training progresses, the generator learns to enhance the input such that it becomes difficult for the discriminator to discriminate between the generator enhanced mel-spectrogram and the real mel-spectrograms obtained from real audio. During the testing, the generator is used to enhance the mmWave vibration mel-spectrogram, without the presence of a discriminator. It can be observed that the discriminator essentially helps the generator learn by indicating the errors in the generated data. After training with cGAN, the difference between the enhanced spectrum and the original spectrum is further minimized. In other words, the high-frequency part of the audio is complemented and the low-frequency part of the audio is enhanced.

For the generator network, we utilize the UNET [41] architecture with skip connections. UNET is an encoder-decoder based architecture proposed for biomedical image segmentation. Each convolutional block in the generator and discriminator is comprised of convolutional layers with square kernels of size 4×4 and stride value 2, followed by batch normalization and rectified linear units (ReLU) for non-linearity [42]. Batch normalization normalizes the activation of different units and accelerates the network converge [43]. A dropout value of 0.5 is used in the intermediate layers and the number of filters is set as multiples of 64 with the filter size decreasing linearly in the subsequent layers following the suggestions in [41]. For the discriminator, we use three convolutional blocks, followed by patch wise predictions of real or fake, with a patch size of 30×30. In contrast to having pixel wise or per image prediction, patch wise predictions take advantage of the independence in patches that are further apart. Additionally, as the captured mmWave data does not include the high-frequency components of the audio, the network's prediction on those patches can be independently improved. The generator and discriminator networks are trained alternatively following the approach delineated in [44]. We use the binary cross-entropy loss [45] between the predicted and ground truth patch labels along with L1 norm [46] over the generator network as the loss function. L1 norm provides regularization without blurry artifacts of the L2 norm. We empirically observe that a learning rate of 0.0002 generates faster convergence. We use the Adam [47] optimizer for optimizing the network. The network is trained for 200 epochs and the performance on a validation set is used to select the optimal training epoch.

6.2 Reconstruction from Enhanced Spectrograms

Once the cGAN enhances the mmWave mel-spectrogram with richer acoustic features, we use a vocoder to convert

Label	Person	# of words for testing	# of words for training	# of words overlapping
$User_1$	Barack Obama	1703	6812	51
$User_2$	Taylor Swift	1605	6421	48
$User_3$	Bill Gates	1594	6377	47
$User_4$	Anne Hathaway	1509	6037	45
$User_5$	Amitabh Bachchan	1143	25647*	34
$User_6$	Meryl Streep	1084		32
$User_7$	Hugh Jackman	1072		30

TABLE 2: Audio dataset used for evaluating MILLIEAR.

the mel-spectrogram to the audio. Specifically, we use the Griffin-Lim algorithm [48] to synthesize waveform from the generated spectrogram due to its efficiency and simplicity. Griffin-Lim uses the phase constraint between frames to achieve iterative convergence and can reconstruct the speech signal using the frequency spectrogram on the basis of the lack of original phase information. It is proposed to finding an approximate phase without destroying the adjacent amplitude spectrum and its own amplitude spectrum. Given that there is a large difference between the worst case and the best case phase, a more accurate phase is obtained through iteration. By this way, even without the original phase information, we can restore the audio waveform to a large extent using the Griffin-Lim algorithm. The reconstructed audio is expected to be as similar to the original human audio as possible.

7 IMPLEMENTATION

This section provides the implementation setup of MILLIEAR and the dataset used for the training and testing.

7.1 Experiment Setup

We implement MILLIEAR on TI IWR1642 BoosterPack which includes an evaluation board (IWR1642BOOST) and a real-time data-capture adapter (DCA1000EVM) [49]. IWR1642 has 2 transmitter (Tx) and 4 receivers (Rx) antennas with the working frequency range of 76-81 GHz. We use one Tx antenna to transmit the FMCW signal and all four Rx antennas to receive the reflected signal. The DCA1000EVM board is used to collect raw ADC data (fast-time samples). The pre-processing of the raw data was conducted on a laptop with an AMD Ryzen 7 4800H CPU and 16GB memory.

The sampling rate of all the audio samples used in our experiments is 44.1 KHz. We use a typical conference room setting with speaker volume set to 70dB and background noise of approximately 45dB (typical indoor office background noise [50]). Fig. 9 shows two typical conference room scenarios used in our experiments. MILLIEAR was evaluated under various settings to capture the influence of sensing distance and angle, materials of isolators, etc. For each setting, we collect at least 4500 audio samples and their corresponding raw mmWave data. The training was performed offline on a server with 10 GPUs (Nvidia RTX 3090). Training for a single user for 200 epochs takes about 2.5 hours and average testing time is 20s.

7.2 Dataset

Our dataset contains audios from 7 English-speaking public personalities as shown in Table 2. We refer to them as $User_1$ through $User_7$. For each user, we randomly select speech samples available online from websites such as YouTube. Table 2 also shows the length of speech audios used in number of words for training and testing for each user . Since our objective is to demonstrate the capability of our model to reconstruct unconstrained vocabulary, we organize the dataset such that there is only a small overlap (shown in Table 2) between words in speech used for training versus testing. The audio samples are played on a speaker in the conference room settings discussed before. The audio and mmWave data are split into 2 seconds segments for input to cGAN model. The total amount of mmWave data is 1.2TB. For $User_1$ through $User_4$, the cGAN model is trained using their own data (training and testing for the same user). For $User_5$ through $User_7$, the model is trained using the audio samples of $User_1$ through $User_4$ and tested on $User_5$ through $User_7$. This setting enables us to validate the performance of model in terms of how it generalizes across different users with cross-subject training.

8 EVALUATION

In this section, we analyze the results of our experiments in two parts: (i) the overall audio reconstruction performance of MILLIEAR and (ii) robustness of MILLIEAR in various scenarios and settings. We perform both subjective and objective evaluation of MILLIEAR, in terms of the following metrics:

- **Mel-Cepstral Distortion**. Mel-Cepstral Distortion (MCD) [51] is an objective measure used for speech quality assessment. It has been widely used in comparing the quality of synthesized speech to the original speech. A smaller MCD score indicates a closer similarity between the reconstructed audio and the original audio. It is believed that a reconstructed audio with MCD below 8 can be recognized by a typical speech recognition system [52].
- **Likert Score**. For subjective evaluation of the reconstructed audio, we recruit 20 volunteers to listen to the recovered audio. These participants include both native and non-native English speakers with ages from 20 to 30 years old. We ask them to listen to the reconstructed audio and the original audio one after the other and then rate the quality of restored audio on a likert scale of 0 to 10. A higher likert score indicates better quality of reconstructed audio. Score of 0 indicates the reconstructed audio is unintelligible while 10 means there is no difference between the reconstructed and original audios.

8.1 Overall audio reconstruction performance

We first evaluate MILLIEAR's ability to reconstruct audio signals in the conference room setting as shown in Fig. 9 (right). The mmWave sensor and the speaker are isolated by a double-panel glass wall with a distance of 1.5m. Fig. 10 illustrates the three types of spectrograms for $User_1$: original audio, directly generated from the mmWave radar without any enhancement, and audio reconstructed from the mmWave radar enhanced by our cGAN model. We observe that the original audio and reconstructed audio spectrograms show high similarity. This is because our

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

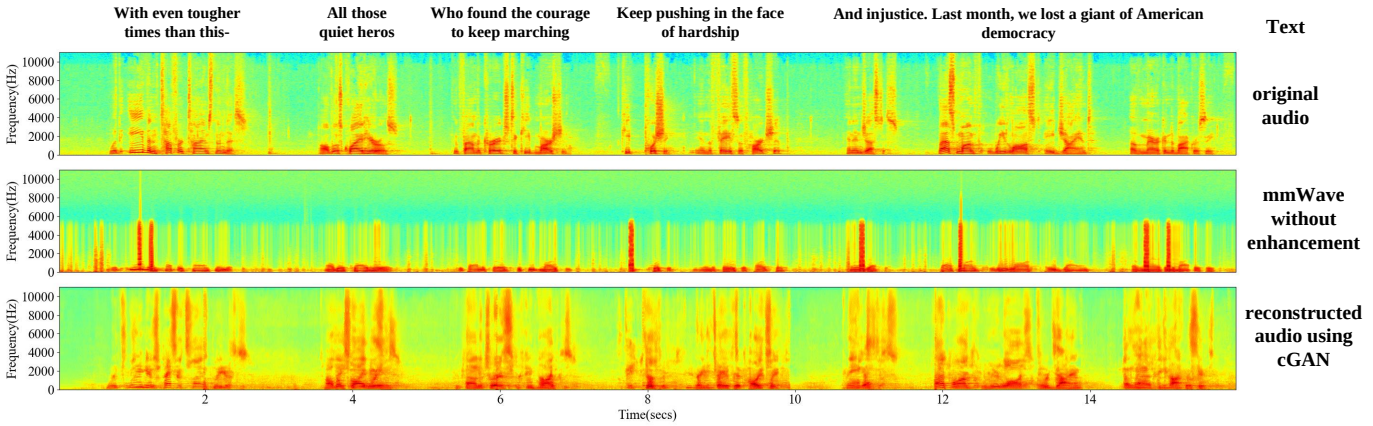


Fig. 10: User$_1$ speech spectrograms for (a) original audio, (b) directly generated from mmWave data without enhancement and (c) audio reconstructed from mmWave data using our cGAN.

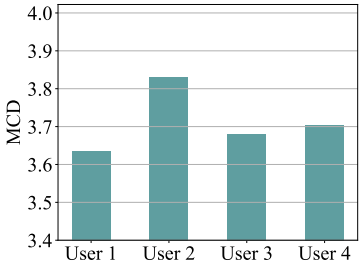
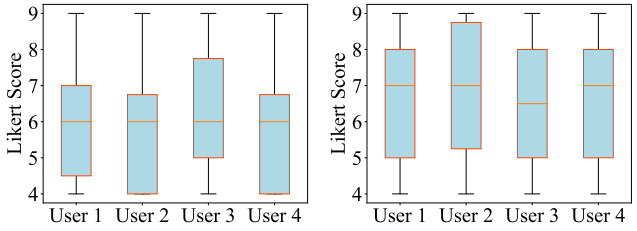
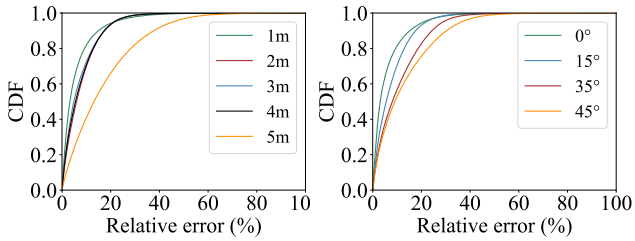


Fig. 11: Objective assessment based on MCD for the recovered audio.



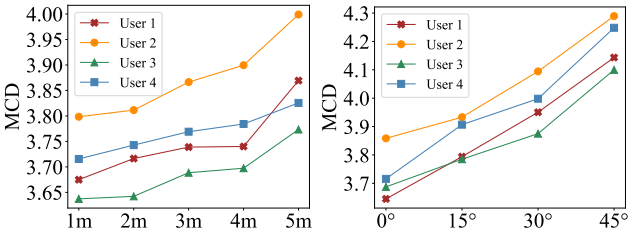
(a) Audio snippet 1 (60s duration) (b) Audio snippet 2 (58s)

Fig. 12: Subjective assessment by volunteers for the recovered audio.



(a) Varying distance (b) Varying angle

Fig. 13: Vibration extraction performance (relative amplitude error between mmWave vibration waveform and original audio) at different distances and angles.



(a) Distance (b) Angle

Fig. 14: Audio reconstruction performance at different (a) distances and (b) angles.

cGAN model is able to learn how to enhance the mmWave spectrograms by reducing noise in the mmWave data and adding specific acoustic components at different frequencies and their amplitude. Given that the overlap (in terms of words) in our training and testing data is small (Table 2), the accurate reconstruction clearly demonstrate our cGAN's ability to work with unconstrained vocabulary. Even in the example shown in Fig. 10 that only 10 words (mostly frequency used words such as *the*, *to*, and *of*) are part of the training speech, MILLIEAR still performs very well.

Fig. 11 shows the MCD for Users 1 through 4. The cGAN model is trained and tested separately for each user. We observe that the average MCD is less than 4 for all users. This implies that the reconstructed audio is not only human discernible but also shows high similarities with the original speech. We further evaluate this similarity using subjective evaluation. Fig. 12 shows the median Likert score from the 20 volunteers for the audio samples of 4 users (both original and reconstructed). As shown in Fig 12, the median score of each user on both two audio sample snippets is higher than 6 which indicates that MILLIEAR has the ability to reconstruct voice that is clearly human recognizable.

8.2 Impact of distance and direction

In real-world scenarios, an attacker may need to adjust the position of the mmWave sensor in order to carry out the eavesdropping. However, adjusting the position will change the distance and direction between the victim device and the mmWave radar. Therefore, we evaluate the robustness of MILLIEAR for different distances and directions. We vary the distance between the mmWave sensor and the speaker from 1m to 5m, and vary the angle from $0°$ to $45°$ in our

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3226690

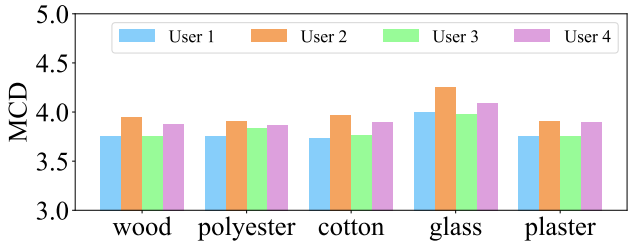


Fig. 15: Audio reconstruction performance with different insulation types.

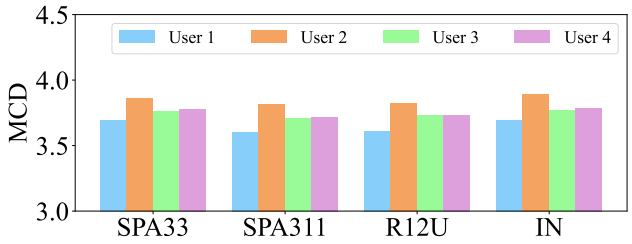


Fig. 16: Audio reconstruction performance with different types of speakers.

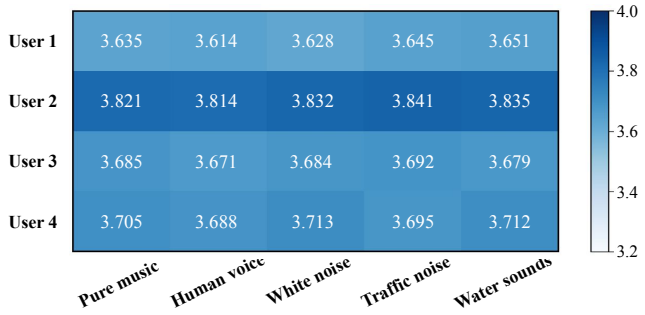


Fig. 17: MCD of User 1 through 4 at different background noise.

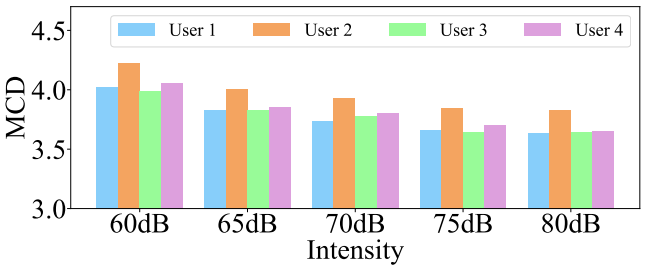


Fig. 18: The effect of different sound intensities on MCD of User 1 through 4.

experiments. These settings are evaluated for the 4 users' audio with individually trained models.

Fig. 13 shows the performance of our proposed vibration extraction. We use the relative error e_r to evaluate the accuracy of vibration extracted from the mmWave signals (without enhancement). Since the amplitudes are at different scales, we normalize them before calculating the relative error of different distance and angles. The relative error to the original audio is derived based on $e_r = \frac{|A_v - A_o|}{A_o}$, where A_v and A_o are the normalized amplitude of the vibration waveform and the original audio signal, respectively. We can see that MILLIEAR achieves 8.9% distance average relative error and 9.6% angle average relative error. The comparison shows the relative error of MILLIEAR between $1m$ and $5m$ is 10.2%, and the relative error between $0°$ and $45°$ is 8.8%. The results clearly shows that MILLIEAR's vibration extraction achieves a good accuracy in our experiments.

Fig. 14(a) shows the MCD for four users (User_1 to User_4) with varying test distances from 1m to 5m. We observe increased MCD scores, indicating gradual reduction in reconstruction quality. However, the overall degradation is not significant at least within the range of the radar. Fig. 20(b) shows that angle has a greater impact on the quality of the reconstructed audio compared to the distance. This is probable because the vibration detection of the speaker surface (i.e., the reciprocating motion) is increasingly difficult to capture through the radar when they are at a larger angle from each other. Nonetheless, MILLIEAR can accurately reconstruct the audio within $45°$. The above experiments show that MILLIEAR can carry out the eavesdropping even at different distances and directions.

8.3 Impact of different types of insulation materials and speakers

The soundproof isolators have been widely used to prevent eavesdropping in practical scenarios. Hence, we conduct experiments to test the robustness of MILLIEAR against different types of insulation materials. We choose 5 types of popular soundproof panels that are composed of dense wood, polyester, cotton, glass and soundproof plaster, respectively. As shown in Figure 15, except for glass, the performance of MILLIEAR does not change significantly with the observed MCD being within 4. Since glass is the strongest reflector of mmWave signals among the materials studied (based on permittivity and attenuation values found in [53], [54]), the sound reconstruction is deteriorated by a small margin. In general, we observe that MILLIEAR achieves a decent performance through penetrating most insulating and soundproofing materials, and thus MILLIEAR can carry out the eavesdropping in common indoor spaces such as homes and offices.

Given that speakers from different manufacturers have distinct features (shapes, material, etc.), we evaluate MILLIEAR with four different types of speakers. They are Philips SPA33, Philips SPA311, Edifier R12U, and Tmall IN. Note that there is no cover on the diaphragm of Philips SPA311 and Edifier R12U, while the diaphragm is covered in Philips SPA33 and Tmall IN speakers. Fig. 16 shows that can achieve better eavesdropping performance on Philips SPA311 and Edifier R12U than Philips SPA33 and Tmall IN, because the vibrating surfaces of the former two speakers are directly exposed to the mmWave sensor.

8.4 Impact of different types of background noise

In the real world, the sound source is usually surrounded by a variety of background noise. Consequently, to make the experiments more practical, we study the impact of background noise on MILLIEAR as follows. We select five different background noises, which are pure music, human voice, white noise, traffic noise[1], and water sounds[2]. Specif-

1. City Traffic Sounds, https://www.youtube.com/watch?v=fh3EdeGNKus
2. Water Sounds, https://www.youtube.com/watch?v=jkLRith2wcc

ically, we choose "Summer[3]" for pure music and "I Have a Dream[4]" by Martin Luther King Jr. for the human voice, and we create a white noise with an amplitude of 0.1. We use another speaker (volume set to 50dB) to play background noise at $5m$ from the radar. Figure 17 shows the MCD of User 1 through 4 under five types of background noises. It can be easily observed that the MCD of each User does not change significantly under different noises. The reason for this is that MILLIEAR reconstructs audio by extracting vibrations of the audio source. The speaker actively modulates the vibration, and thus the sound waves from the background noise have a weak effect on the diaphragm of the speaker. Therefore, the performance of MILLIEAR does not change significantly in the presence of background noise.

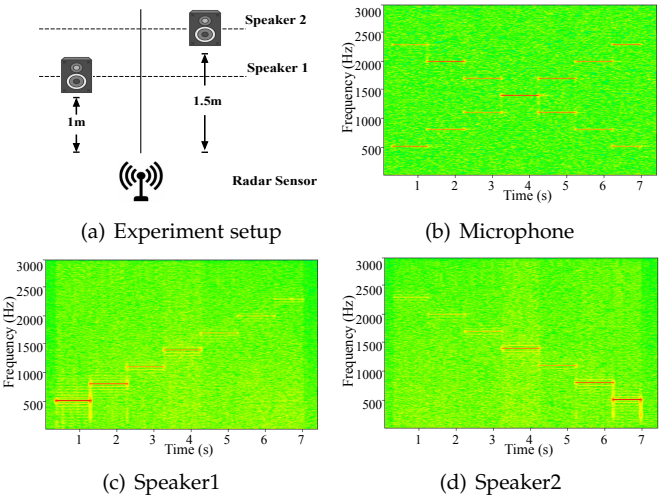
8.5 Impact of different sound intensities

In a real eavesdropping scenario, the sound intensity of the target is usually not constant. Therefore, in order to provide a more comprehensive experimental evaluation, we investigate the effect of the sound intensity of the speaker on the audio reconstruction. We place the speaker at $1m$ from the radar. We set the intensity from $60dB$ to $80dB$ and let the speaker play audio from User 1 through 4. We evaluate the 4 users' audio with individually trained models. Figure 18 shows that, as the sound intensity increases, the MCD decreases, which means that the spectral similarity becomes higher. This is because the sound intensity of the speaker is determined by the modulation of the amplitude, and the reduction of the sound intensity indicates the reduction of the vibration amplitude. Nevertheless, within the typical range of sound intensities, MILLIEAR can accurately recover the audios, indicating the effectiveness of the eavesdropping.

8.6 Multiple audio sources reconstruction

Conventional eavesdropping methods only attack one audio source in default. For instance, a microphone can record the overlapped audio from different sources, but it is difficult to separate them. In this experiment, we study the ability of MILLIEAR for multi-source audio reconstruction. We put two speakers in the mmWave sensor's field of view. Speaker1 and Speaker2 (both are Philips SPA33) are placed at $1m$ and $1.5m$ from the sensor respectively. Speaker1 plays audio1 and Speaker2 plays audio2. To prevent overlap of sound sources, the distance between each source is larger than the distance resolution[5] of the radar. In addition, we put a microphone next to the radar as a comparison.

As shown in Figure 19(b), multiple audio sources are super-imposed in the microphone spectrogram. This is because they are entangled in both the time and frequency domains, which makes it difficult to separate the audio sources. In contrast, by selecting different range bins, the mmWave signals from each audio source can be processed independently. As shown in Figure 19(c) and Figure 19(d),



(a) Experiment setup (b) Microphone

(c) Speaker1 (d) Speaker2

Fig. 19: Multiple audio source evaluation: (a) Experimental setup (b) Spectrogram of the audio recorded by microphone (c-d) Spectrograms of reconstructed audio from Speaker1 and Speaker2.

MILLIEAR can effectively separate multiple audio sources. Our experiment results show that MILLIEAR can reconstruct sounds from multiple audio sources.

8.7 Model generalization with cross-user training

To show that MILLIEAR has a good model generalization capability, we train and test the cGAN model for different users (cross-user training and testing). We train the model using User_1 data and then test it with Users 2, 3 and 4. Fig. 20(a) shows the MCD reduction when Users 2, 3 and 4's speeches are tested with their own individually trained model vs. the model trained using User_1's data. We find that while there is clearly a reduction in audio reconstruction performance, the overall performance is still good to carry out the attack. The reconstruction quality degrades because the voice characteristics of different people have different dominant frequency components that are not always accurately reconstructed during cross-user training.

To evaluate if adding more user's data to the training can further improve the cross-user performance, we train the model with data from Users 1 through 4, and test it on Users 5 through 7. Fig. 20(b) shows the resultant MCD. We find that when more users are used in the training, the model generalizes better by learning to capture more diverse set of acoustic features. For example, the MCDs of User_5 with model of User_1 through User_4 are all above 5.6, while model trained using multiple users' data generates a much lower MCD of 3.8. These cross-user training results show that an attacker can train the model offline with a large number of users' audio data and then carry out the eavesdropping attack on an unknown user's audio data.

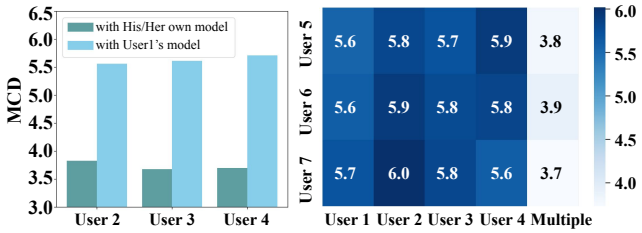
9 Discussion

MILLIEAR is a mmWave-based acoustic eavesdropping with unconstrained vocabulary, which achieves premium performance. This section discusses MILLIEAR in the following aspects.

3. Joe Hisaishi - Summer, https://www.youtube.com/watch?v=l0GN40EL1VU

4. I Have a Dream, https://www.youtube.com/watch?v=vP4iY1TtS3s

5. According to theoretical calculation, we know that the spatial resolution of radar is roughly $5cm$, which means that two objects less than $5cm$ apart will be overlapped into one object.



(a) own vs. others (b) cross-user test

Fig. 20: Model generalization: (a) test result based on own model vs. other's model; (b) cross-user test results.

Countermeasure to mmWave Eavesdropping. As we clearly show in this paper, mmWave signals can accurately eavesdrop sounds from speakers. There are several methods to prevent or mitigate eavesdropping, all with drawbacks. (1) A straightforward method it to wrap the room or the speaker with electromagnetic shields to block all wireless signals. However, considering the cost and inconvenience, it is very unlikely that electromagnetic shields will become widely available in daily lives. (2) Another method is to disrupt the mmWave frequency bands by broadcasting jamming mmWave signals, since eavesdropping with sub-6 GHz does not work well. However, mmWave is also used by high-speed short-distance data transmission (e.g., TV connection and 5G/6G communications), and thus jamming mmWave frequency bands interferes with legit applications. (3) Adding jitters to the speakers to create man-made vibration could mitigate the eavesdropping quality. However, the original sound quality might deteriorate as well. In addition, the attacker can simply apply filtering techniques to remove jitters. (4) Jamming the signal of a target radar using another radar. Nevertheless, an FMCW radar receiver expects to receive signals with a predefined frequency pattern and to filter signals from other frequency bands. (5) Moving audio sources. The mmWave radar in this paper has a chirp rate of 10k, i.e., the interval time between adjacent chirps is only 0.1 ms. Therefore, manually moving speakers do not create significant displacement within this short chirp interval. Moreover, the application of cGAN also eliminates the effect of such noise and thus our experimental results are valid even when people deliberately move the speakers as a countermeasure. As we can see, it is difficult to prevent mmWave eavesdropping, which calls for more research effort to design effective counter-measurements.

Performance Improvement. In our prototyping system, we adopt UNET for the GAN generator and a simple convolutional network for the GAN discriminator. It is expected that applying other more advanced models can further improve the accuracy of MILLIEAR. Since the focus of this paper it to build the eavesdropping system with no constrained vocabulary by using mmWave and GAN techniques, we leave it as future work to obtain the best performance. Nonetheless, the simple models used in this paper have already achieved exciting overall performance, strongly supporting that the workflow in MILLIEAR is general and effective.

Mobile Version of MILLIEAR. Since mmWave modules will be integrated in next-generation smartphones for emerging applications, we plan to design a mobile version of MILLIEAR. The corresponding eavesdropping attack from phone is more difficult to discover, which raises severe concerns about the privacy of human conversation over speakers. Meanwhile, the mobile version will poses new research challenges such as how to remove the interfering vibration from the attacker's phone in pockets.

10 Conclusion

In this work, we propose a mmWave eavesdropping system that combines the mmWave FMCW and generative machine learning networks to reconstruct the original audio. Our evaluation results show that MILLIEAR is highly effective in eavesdropping voices, achieving the average MCD of 3.68 and the average likert user score of 6.83. This paper sheds light on a mmWave-based eavesdropping system and could motivate more research along this direction, given the inspiring results.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2021YFB3100400), National Natural Science Foundation of China (Grant No. 62202276, 62202274, 61832012), Shandong Science Fund for Excellent Young Scholars (No. 2022HWYQ-038), and NSF grants CNS-1815945 and CNS-1730083.

References

[1] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *23rd USENIX Security Symposium (USENIX Security 14)*, 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 *NDSS*, 2020.

[3] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 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 *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 301–315.

[5] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," 2014.

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

[7] B. Nassi, Y. Pirutin, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Real-time passive sound recovery from light bulb vibrations." *IACR Cryptol. ePrint Arch.*, vol. 2020, p. 708, 2020.

[8] 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.

[9] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 130–141.

[10] J. Han, A. J. Chung, and P. Tague, "Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2017, pp. 181–192.

[11] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, 2016, pp. 57–69.

[12] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, "Speake (a) r: Turn speakers to microphones for fun and profit," in *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, 2017.

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

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

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

[16] 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 *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 1–14.

[17] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 14–26.

[18] S. Basak and M. Gowda, "mmspy: Spying phone calls using mmwave radars," in *2022 2022 IEEE Symposium on Security and Privacy (SP) (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2022, pp. 995–1012. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/SP46214.2022.00058

[19] 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.

[20] Y. Rong, S. Srinivas, A. Venkataramani, and D. W. Bliss, "Uwb radar vibrometry: An rf microphone," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1066–1070.

[21] 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.

[22] 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 *2021 International Conference on Microwave and Millimeter Wave Technology (ICMMT)*. IEEE, 2021, pp. 1–3.

[23] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu, and K. Ren, "mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 5 2022.

[24] 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 *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 217–232.

[25] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, "mmvib: micrometer-level vibration measurement with mmwave radar," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–13.

[26] 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 *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2021, pp. 97–110.

[27] "Iwr1642 single-chip 76- to 81-ghz mmwave sensor datasheet (rev. b)." [Online]. Available: https://www.ti.com/lit/ds/symlink/iwr1642.pdf?ts=1627443405952

[28] "Generative model," https://en.wikipedia.org/wiki/Generative_model.

[29] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9465–9474.

[30] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1857–1865. [Online]. Available: http://proceedings.mlr.press/v70/kim17a.html

[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4396–4405.

[32] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2089–2093.

[33] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[34] C. Zhang, F. Li, J. Luo, and Y. He, "iLocScan: Harnessing Multipath for Simultaneous Indoor Source Localization and Space Scanning," in *Proc. of the 12th ACM SenSys*, 2014, p. 91—104.

[35] Z. Chen, G. Zhu, S. Wang, Y. Xu, J. Xiong, J. Zhao, J. Luo, and X. Wang, "M^3: Multipath Assisted Wi-Fi Localization with a Single Access Point," *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 588–602, 2021.

[36] 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.

[37] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Very short time environmental sound classification based on spectrogram pattern matching," *Information Sciences*, vol. 243, pp. 57–74, 2013.

[38] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[40] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[42] A. F. Agarap, "Deep learning using rectified linear units (relu)," 2019.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[44] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[45] "Binary cross entropy," https://en.wikipedia.org/wiki/Cross_entropy.

[46] "L1 norm," https://en.wikipedia.org/wiki/Regularization_(mathematics).

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[48] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[49] [Online]. Available: https://www.ti.com/tool/IWR1642BOOST

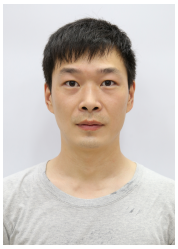
[50] "Common noise levels." [Online]. Available: https://noiseawareness.org/info-center/common-noise-levels/

[51] 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.

[52] 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.

[53] B. Langen, G. Lober, and W. Herzig, "Reflection and transmission behaviour of building materials at 60 ghz," in *5th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Wireless Networks - Catching the Mobile Future.*, vol. 2, 1994, pp. 505–509 vol.2.

[54] J. Lu, D. Steinbach, P. Cabrol, P. Pietraski, and R. V. Pragada, "Propagation characterization of an office building in the 60 ghz band," in *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, 2014, pp. 809–813.



Pengfei Hu is a professor in School of Computer Science and Technology at Shandong University, China. He received Ph.D. in Computer Science from UC Davis. His research interests are in the areas of IoT security, AI security, mobile computing. He has published over 30 papers in premier conferences and journals on these topics, e.g. IEEE S&P, ACM CCS, IEEE INFOCOM, IEEE TMC, etc. He also served as TPC for numerous prestigious conferences and associate editors for IEEE TWC and IoTJ.

Wenhao Li is currently working toward Ph.D from the School of Computer Science, Shandong University, China. He received BE degree in computer science and technology from the Beijing Information Science and Technology University, China. His current research interests include system security and RF sensing.

Yifan Ma is currently working toward M.D. from the School of Computer Science, Shandong University, China. He received BE degree in computer science and technology from Guangzhou University. His recent research has focused on mmWave sensing.

Panneer Selvam Santhalingam is a Ph.D. candidate in the department of Computer Science at George Mason University, advised by Dr. Parth Pathak. He did his Masters in Information Security and Assurance at George Mason University and his Bachelors of Engineering at SSN college of engineering, Chennai, India. He is interested in Wireless and Mobile Computing with a focus on high-frequency wireless, and wearable sensing with application in areas like: Human computer interaction (Gesture recognition and activity recognition), Accessibility (American Sign Language recognition), and Cyber physical systems.

Parth Pathak is an assistant professor of Computer Science at George Mason University. Before joining George Mason University, he was a postdoctoral scholar at University of California, Davis, and before that he received his M.S. and Ph.D. degrees in Computer Science from North Carolina State University in 2012. His research interests include mobile and ubiquitous computing, wireless networking, and energy-efficient computing and communication systems. His recent focus is on design of high-speed millimeter-wave wireless networks and their application in cyber-physical systems such as robotic manufacturing and autonomous platforms such as UAVs and UGVs. Current work also includes use of machine learning for enhancing wireless sensing capabilities and improving security of next generation wireless networks.

Hong Li was born in 1989. He received the B.S. degree in computer science from Xi'an Jiao Tong University in 2011 and the Ph.D. degree in cyber security from University of Chinese Academy of Sciences in 2017. Since 2017, he is an associate professor at the Institute of Information Engineering, Chinese Academy of Sciences. His current research interests include IoT security, ICS Security and Blockchain.

Huanle Zhang is an associate professor in the School of Computer Science and Technology at Shandong University, China. He received the PhD degree in computer science from the University of California, Davis (UC Davis), in 2020. He was employed as a postdoc at UC Davis from 2020 to 2022 and a project officer at Nanyang Technological University from 2014 to 2016. His research interests include data-centric AI, IoT, and mobile systems.

Guoming Zhang received his Ph.D. degree from the Department of Electrical Engineering of Zhejiang University, supervised by Prof. Wenyuan Xu and Donglian Qi. He received the Master degree in School of Mechanical Engineering of Beijing Institute of Technology, supervised by Prof. Jie Hu. He obtained his Bachelor degree in College of Transportation from Ludong University in 2013. His research interests include IoT security and acoustic communication. He won the best paper awards of ACM CCS 2017, Qshine 2019.

Xiuzhen Cheng received her M.S. and Ph.D. degrees in computer science from the University of Minnesota – Twin Cities in 2000 and 2002, respectively. She is a professor in the School of Computer Science and Technology, Shandong University. Her current research interests include wireless and mobile security, cyber physical systems, wireless and mobile computing, sensor networking, and algorithm design and analysis. She has served on the editorial boards of several technical journals and the technical program committees of various professional conferences/workshops. She also has chaired several international conferences. She worked as a program director for the US National Science Foundation (NSF) from April to October in 2006 (full time), and from April 2008 to May 2010 (part time). She received the NSF CAREER Award in 2004. She is Fellow of IEEE and a member of ACM.

Prasant Mohapatra is serving as the Vice Chancellor for Research at University of California, Davis. He is also a Professor in the Department of Computer Science and served as the Dean and Vice-Provost of Graduate Studies at University of California, Davis during 2016-18. He was the editor-in-chief of the IEEE Transactions on Mobile Computing. He has served on the editorial boards of the IEEE Transactions on Computers, the IEEE Transactions on Mobile Computing, the IEEE Transaction on Parallel and Distributed Systems, the ACM Journal on Wireless Networks, and Ad Hoc Networks. He is a fellow of the IEEE and a fellow of the AAAS