MILLIEAR: Millimeter-wave Acoustic Eavesdropping with Unconstrained Vocabulary

***Abstract*—As acoustic communication systems become more common in homes and offices, eavesdropping brings significant security and privacy risks. Current approaches of acoustic eavesdropping either provide low resolution due to the use of sub-6 GHz frequencies, work only for limited words using classification, or cannot work through-wall due to 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 with unconstrained vocabulary. We implement and evaluate MILLIEAR using off-the-shelf mmWave radar deployed in different scenarios and settings. We find that it can accurately reconstruct the audio even at different distances, angles and through the wall with different insulator materials. Our subjective and objective evaluations show that the reconstructed audio has a strong similarity with the original audio.**

I. INTRODUCTION

With a large portion of the global workforce working remotely, acoustic communication systems such as video conferencing, personal digital assistants, and home entertainment systems are becoming more popular than ever. While our digital communication over the Internet is protected through strong encryption, 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 information about the users. With the increasing prevalence of video conferencing systems in homes and offices, acoustic eavesdropping poses major security and privacy risk.

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 [1]–[5] can be used to listen to acoustic signals. 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 source. A high-speed camera [6], lasers [7] or photodiodes [8] have been used for remotely discerning the spoken text through vibrations. Authors in [9], [10] proposed using WiFi signals to extract speaker vibrations.

In this paper, 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. MILLIEAR addresses many limitations of the prior attack systems:

(1) Higher resolution: Compared to existing RF-based eavesdropping systems that operate at sub-6 GHz frequencies [9], [10], MILLIEAR uses mmWave FMCW radar that can exploit the large available bandwidth at mmWave spectrum to provide better range resolution. As we show in this work, speaker vibrations of up to tens of microns can be detected using mmWave radar for accurate eavesdropping.

(2) Unconstrained vocabulary: Majority of existing eavesdropping systems such as [1]–[5], [9], [10] treat acoustic signal extraction as a classification problem through profiling of a handful of words. In comparison, MILLIEAR demonstrates the attack with unconstrained vocabulary as it does not require training for classifying specific words. Instead, it provides the reconstruction of entire conversational audio purely 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 system, MILLIEAR works even behind glass, wooden doors and walls. Compared to [6], [7] and [8] which require expensive camera sensors, laser transducer or telescope, mmWave radar is low-cost. Furthermore, due to the smaller wavelength of mmWave signals, the sensor footprint is smaller as well compared to the large multi-antenna system setup needed at sub-6 GHz frequencies.

Building a mmWave eavesdropping system with unconstrained vocabulary requires us to address two challenging questions:

(1) *How do we extract the speaker vibrations using mmWave radar signals in 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. 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. We address this problem by measuring the phase change through *virtual* sub-chirps. We firstly apply a sliding window on the raw mmWave data to generate sub-chirps. A range-FFT is then applied on the sub-chirps to candidate vibration bins and other bins (i.e., mmWave noise sources). A Doppler-FFT applied on the refined bins can then help us extract the vibrations related to the speaker as measured by the mmWave radar sensor.

(2) *How do we accurately reconstruct the audio from mmWave vibrations assuming unconstrained vocabulary in the audio?* The audio captured through mmWave signals can contain any words unknown in advance to MILLIEAR. This means that we need a learning model that can not just classify the existing words based on limited training, but can 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 to learn 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 reconstruction for any audio.

Our contributions can be summarized as follows:

(1) We present a mmWave acoustic eavesdropping system 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 presented model can recreate high-quality audio signal directly from the mmWave radar signals using cGANs.

(2) We perform an extensive evaluation of MILLIEAR. We demonstrate the attack for audio from 7 public personalities played through speakers and captured through a mmWave radar. With audio samples with over 25000 words used in training and testing, our objective and subjective 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. We evaluate MILLIEAR in different scenarios with varying distances and angles between speaker and radar, different types of sound-proofing material/wall between the two, and different types of speakers. Lastly, our results also show that MILLIEAR has a strong regenerative and generalizability capacity where cGAN models trained using audio of users (cross-user training) other than the victim can also perform very well in eavesdropping for the victim audio.

The remaining paper is organized as follows. Section II discusses the related work. Section III discusses mmWave radar and GAN preliminaries with a feasibility study, and Section IV describes the system overview. Our vibration extraction methods and cGAN architectures are discussed in Sections V and VI-A, respectively. We evaluate MILLIEAR in Section VII and conclude in Section VIII.

II. RELATED WORK

Table I provides a comparison of related works with our system. Several studies have shown that deploying a IMU sensor near the audio source can enable an attacker to perform eavesdropping. Authors in [1]–[4] show that IMU-based audio sensing can classify words and small phrases and even speaker gender. [5] can recover the audio with unconstrained vocabulary. [12] implements a malware prototype which can turn the speaker (headphones, earphones) connected to the computer into a microphone for the eavesdropping purpose. Authors in [11] recovered the audio using a vibration motor. [13] uses a magnetic hard disk to recover audio where measuring the offset between the read/write head and the track center of the disk can be used to recover songs and voices. The main disadvantage of these eavesdropping methods is that they need 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 I), the attacks can be restricted based on victim's active defense strategies.

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

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

Two studies [9], [10] have used WiFi signals to profile movements or vibrations and identify audio. Authors in [9] proposed a method to analyze the WiFi channel state information (CSI) to classify words. In [10], authors analyzed the received signal strength of the WiFi signals where the audio vibrations are considered as low-rate modulations of RF signals. [14] presents an Impulse Radio Ultra-Wideband based system which is able to simultaneously recover and separate sounds from multiple sources. However, its capability for recovering unconstrained vocabulary has not been studied. Compared to our approach, these works relying on WiFi traffic offers lower resolution due to lower frequency and packet rate. Also, victim localization requires a multi-antenna setup that has larger physical footprint at lower frequencies compared to mmWave, making the attack more difficult to be carried out in practice. Finally, these works do not explicitly target complete audio reconstruction with unconstrained vocabulary as shown in our work.

Cameras and lasers have also been used for acoustic eavesdropping. Authors in [7] 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, [6] used a high-speed video camera to obtain the video of an object in victim's room (such as a plastic bag, water, etc.) and analyze the response as sound waves impinge on the object to recognize audio. [8] 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 apart from limited vocabulary, these attacks are difficult to carry out as they require expensive, special purpose hardware such as the high-speed camera.

In other similar research, authors in [15] used mmWave to acquire high-quality voice from user's vocal vibrations from near-throat region. [16] proposed a remote and through-wall screen attack that used mmWave to remotely collect information from LCD screens. [17] showed how mmWave radar can be used for micrometer-level vibration measurement in industrial environments. While similar, these works do

not focus on acoustic eavesdropping and audio reconstruction which are the focus of this work.

III. Preliminaries

In this section, we briefly present the general idea of the Frequency Modulated Continuous Wave (FMCW) radar based vibration measurement and the principles of Generative Adversarial Networks for signal enhancement.

A. Vibration Estimation

An FMCW radar transmits a signal called "chirp". A chirp is a sinusoid whose frequency increases linearly with time. An FMCW signal can be used to estimate the displacement (Δd) of an object using the phase difference of reflected signal as

$$\Delta d = \frac{\lambda \Delta \phi}{2\pi} \tag{1}$$

where $\Delta\phi$ is the phase change and λ is the wavelength of wireless signal. For example, based on Eq. 1, the phase change of 1 degree will result in the displacement of about 10 microns for a 77 GHz FMCW radar used in our experiments. The vibration amplitude of a speaker is normally in the order of hundred microns level [17]. Hence, mmWave is capable of capturing minute vibrations of a speaker.

Based on Eq. 1, the vibration displacement Δd is directly related with the phase change $\Delta\phi$. Once we extract the accurate phase change from mmWave signal, it will be possible to derive the vibration displacement. Let $S_{Tx}(t)$ and $S_{Rx}(t)$ be the FMCW transmitted and received (reflected by target) signal represented as

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

where $f_{Tx}(t)$ and $f_{Rx}(t)$ are the frequency of transmitted signal and received signal at time t, respectively, ϕ_{Tx} and ϕ_{Rx} are the phase of transmitted signal and received signal, respectively, A_{Tx} and A_{Rx} are the amplitude of the transmitted and received signal. After applying a mixer on 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) [18], we can apply a low-pass filter to exclude the carrier. Then the beat frequency signal can be expressed as follows

$$S_b = A_b \cdot \cos[2\pi \cdot f_b(t) \cdot t + \phi_b] \tag{4}$$

where $A_b = \frac{1}{2} A_{Tx} A_{Rx}$ is the synthesized amplitude of transmitter and receiver.

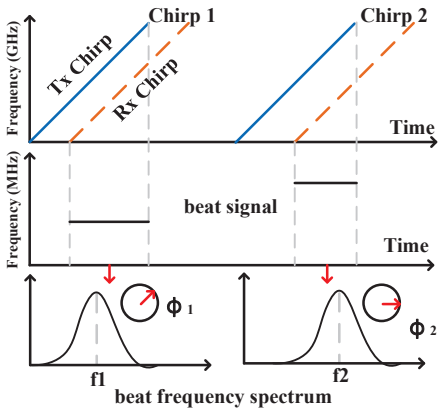
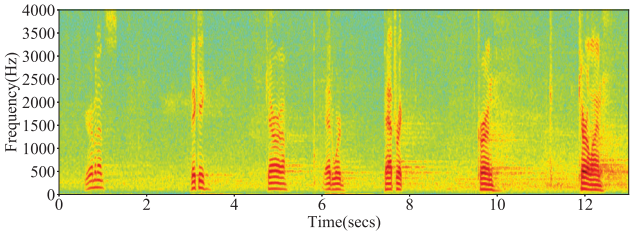
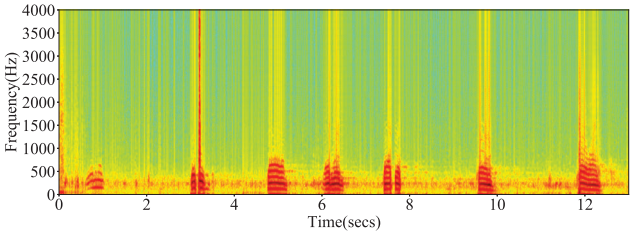


Fig. 1: Phase extraction from FMCW chirps.



(a) Original audio



(b) Reflected mmWave signal

Fig. 2: The spectrograms for (a) original audio and (b) reflected mmWave signal from the speaker.

Therefore, the final beat signal is only related to f_b and ϕ_b. 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 1, 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 will have a peak in the frequency domain. For our vibration source positioning task, we will only focus on the range bin of the corresponding distance. Then, as shown in Figure 1, we can further perform Doppler-FFT on the results of Range-FFT to derive the phase change by $\phi_1 - \phi_2$. For the same range bin in two chirps, performing the Doppler-FFT operation will extract the phase of the corresponding position. This provides us the capability to derive the time-variant phase caused by the speaker vibration.

B. A Feasibility Study

In order to launch an eavesdropping attack, we verify the correlation between the received millimeter wave signal and an audio played through a speaker using a proof-of-concept experiment. In the experiment, we let the speaker play a

speech while the mmWave radar is placed in front of the speaker at $1m$ distance without any blockage. Fig. 2 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 low sampling rate of the FMCW radar, the radar signals show poor similarity with the audio at high frequencies. Also, FMCW radar suffer from noise at lower frequencies. To address these two issues, mmWave radar signals reflected from the speaker can be enhanced using a generative machine learning model to reconstruct the original audio.

C. Generative Adversarial Networks

Generative adversarial networks (GANs) belong to the class of generative models [19]. 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 a sample $z \in p_z$ and map it to sample $x \in p_{data}$. GANs implicitly learn this mapping function and have enabled several novel applications [20]–[24]. 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 $log\ D(x)$. As the two models play this game, an equilibrium is reached when the generator has successfully approximated p_{data} and the discriminator can no longer differentiate between real and fake data.

D. Attack Model

Previous approaches to prevent acoustic eavesdropping depends on the usage of isolators, such as soundprooof glass, polyethylene foam, plywood, etc. In this work, we consider the eavesdropping threat which leverages the mmWave radar to reconstruct the sound of speaker even with the existence of sound-proof isolators. As shown in Fig 3, we assume the following about the attacker: (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 not only able to classify a handful of audio signals (i.e., words or numbers), but can recreate any audio from the entire vocabulary including full sentences. (iii) the device to launch an attack is portable and affordable. The attacker can perform sound eavesdropping in this scenario with a low-cost commercial mmWave radar outside the soundproof space.

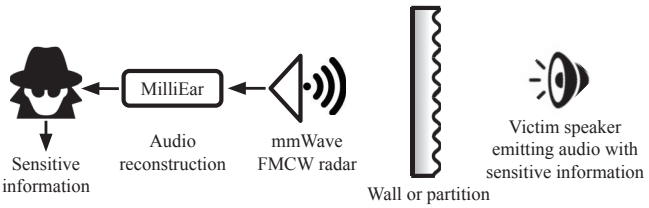


Fig. 3: Our attack scenario of mmWave-based audio eavesdropping.

IV. System Overview

We design and implement our mmWave voice eavesdropping system MILLIEAR as shown in Fig. 4. It has a mmWave radar which can capture the minute vibration cause by the sound. The mmWave radar will emit an FMCW chirp signal to the vibrating speaker at first. The signal arriving at the speaker will be reflected back to the radar. Through careful processing and enhancement of the received signal, MILLIEAR can extract the speaker vibrations. However, due to background reflection and multipath effects, there may be errors in the received signal, resulting in inaccurate vibration estimation. The vibration data will then be fed into our Generative Adversarial Network for enhancement and denoising to finally achieve high-quality audio reconstruction. MILLIEAR consists 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 speaker, MILLIEAR takes raw samples from mmWave radar as input. We perform Range-FFT on the raw data to measure the distance to the target. Then we conduct 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. We then perform STFT to each chirp to get the time-frequency domain spectrogram.

(2) Audio Reconstruction (AR): AR uses a conditional GAN that is trained using two 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. We note that the GAN training is agnostic to the spoken text and does not require any manual annotation during training.

V. Spectrogram Generation

A. Vibration extraction

In order to extract the vibration displacement, we must accurately locate the vibration target at first. We directly follow

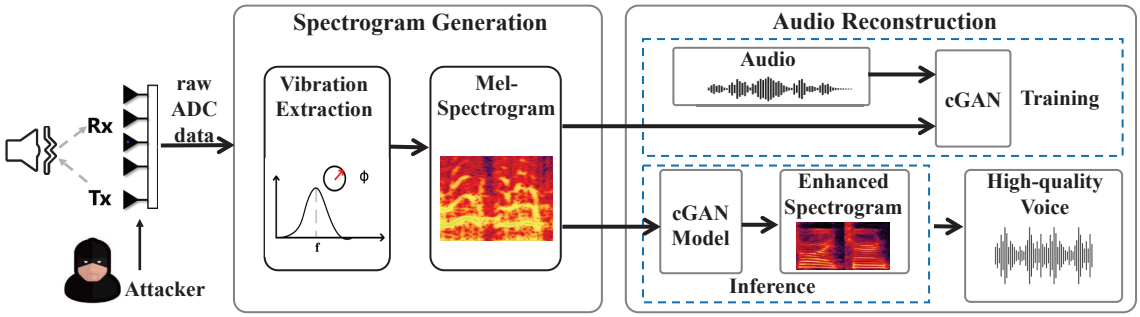


Fig. 4: The MILLIEAR system mainly consists of a mmWave radar (TI IWR1642boost and TI DCA1000EVM), Data-preprocessing module to extract the vocal spectrogram and Audio Reconstruction module to recover high-quality voice.

the solutions proposed by [17] for the target localization.

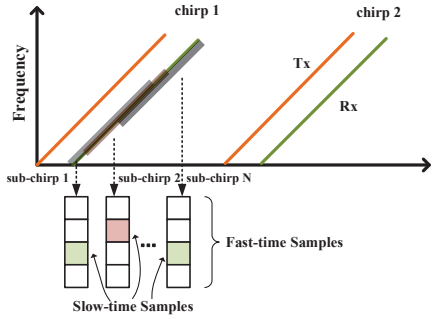


Fig. 5: Vibration extraction from FMCW chirps.

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 can apply a sliding window on fast-time samples within one single chirp to generate more *virtual* sub-chirps as shown in Fig. 5. These sub-chirps will be 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. 5, the position of the voice bin detected by sub-chirp 2 (red bin) is different from that of other sub-chirps (green bin). Since we have multiple observations for cross-validation, the abnormal bin (red) can be identified and eliminated. Through this approach, we could 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 change. The vibration displacement could be calculated according to Equ. 1 once the phase change 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. 2. The maximum chirp rate of the mmWave sensor used in our work is $10kHz$ which is far less 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 it with more details.

B. 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 [25] is a popular representation for audio signal which has been widely used in the speech synthesis, audio denoising, etc. We can directly feed this image-like spectrogram into cGAN for enhancement. The enhanced spectogram is then converted back to audio with little information loss.

In this work, we choose Short-time Fourier transform (STFT) to get the time-frequency spectrogram. STFT is essentially a windowed Fourier Transform, which has been defined as follows,

$$STFT(t,f) = \int_{-\infty}^{+\infty} x(\tau)h(\tau - t)e^{-j2\pi f\tau}d\tau \qquad (5)$$

where $h(\tau - t)$ is the window function and 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 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 [26]. 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 [25]. 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 cGAN training and only on the vibration signal during the testing.

VI. Audio Reconstruction

We now describe our audio reconstruction methodology.

A. GAN Architecture

We adopt an image to image translation approach [27] 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. III-C), cGANs additionally take a conditional variable, enabling control on the generated data [28]. 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. 6 shows our cGAN architecture. While training, the generator takes as conditional input a mmWave vibration mel-spectrogram and enhances it. The enhanced mel-spectrogram is concatenated with mmWave mel-spectrogram and input to the discriminator. The discriminator classifies this as fake. Additionally, when input with the mel-spectrogram for real audio 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. After this, for testing, the generator is independently used to enhance the mmWave vibration mel-spectrogram, without the presence of a discriminator. It can be observed that the discriminator is essentially helping the generator learn by indicating the errors in the generated data.

For the generator network, we utilize the UNET [29] 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 [30]. Batch normalization normalizes the activation of different units and accelerates the network converge [31]. 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 [29]. 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 [32]. We use the binary cross-entropy

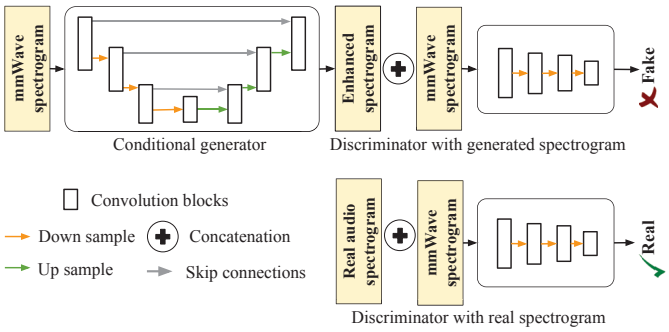


Fig. 6: MILLIEAR cGAN architecture.

loss [33] between the predicted and ground truth patch labels along with L1 norm [34] 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 yields faster convergence. Adam [35] optimizer is used for optimizing the network. The network is trained for 200 epochs and performance on a validation set is used to pick the best training epoch.

B. Reconstruction from Enhanced Spectrograms

Once the cGAN enhances the mmWave mel-spectrogram with richer acoustic features, we can use a vocoder to covert the acoustic parameters into speech waveforms. In this paper, we use the Griffin-Lim algorithm [36] 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 it can reconstruct the speech signal using the frequency spectrogram on the basis of the lack of original phase information. It solves how to find 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. This way, even without the original phase information, we can restore the audio waveform to a large extent using the Griffin-Lim algorithm.

VII. Evaluation

A. Implementation and Experiment Setup

We implement MILLIEAR on TI IWR1642 BoosterPack which includes an evaluation board (IWR1642BOOST) and a real-time data-capture adapter (DCA1000EVM) [37]. 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

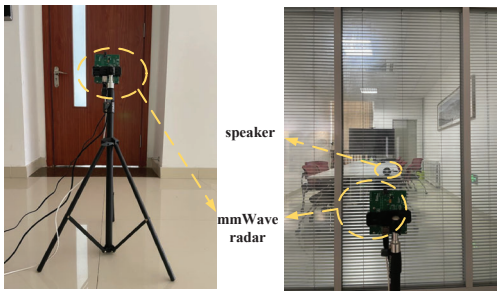


Fig. 7: Two examples of experiment setup of MILLIEAR. **Left**: Conference Room with a dense wood door; **Right**: Conference room with a double-panel glass wall.

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 II: Audio dataset used for evaluating MILLIEAR (*Model for $User_5$ through $User_7$ trained using data from $User_1$ through $User_4$ for generalizability testing with cross-user training).

[38]). Fig. 7 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.

B. Dataset

Our dataset contains audios from 7 English-speaking public personalities as shown in Table II. We refer to them as $User_1$ through $User_7$. For each user, we randomly pick speech samples available online from websites such as YouTube. Table II shows the length of speech audios used in number of words for training and testing for each user[1]. Since our objective is to demonstrate the capability of our model to reconstruct unlimited vocabulary, we organize the dataset such that there is only a small overlap (shown in Table II) 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 enables us to validate the performance of model in terms of how it generalizes across different users with cross-subject training.

[1]Authors will publish the entire dataset of original audios, mmWave data and reconstructed audios with the final version of the paper.

C. Evaluation Metrics

We perform both subjective and objective evaluation of MILLIEAR.

Mel-Cepstral Distortion. Mel-Cepstral Distortion (MCD) [39] is an objective measure used for speech quality assessment. It has been widely used in comparing the quality of synthesized speech to the original/natural speech. While a detailed explanation of MCD is outside the scope of the paper, 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 [40].

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. Here, higher likert score indicates better quality of reconstructed audio. Score of 0 indicates the reconstructed audio is unintelligible while 10 means there is little to no difference between the reconstructed and original audios.

D. Numerical Results

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.

Audio reconstruction performance. We first evaluate MILLIEAR's ability to reconstruct speech signals in the conference room setting as shown in Fig. 7 (right). Here, the mmWave sensor and the speaker are isolated by a double-panel glass wall with a distance of 1.5m. Fig. 8 shows the three types of spectrograms for $User_1$: original audio, directly generated from mmWave without any enhancement, and audio reconstructed from mmWave using our cGAN model. We observe that the original audio and reconstructed audio spectrograms show high similarity. This is due to the fact that our 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 II), the accurate reconstruction points to our cGAN's ability to work with unconstrained vocabulary. Even in the example shown in Fig. 8, only 10 words (mostly frequency used words such as *the, to, of*, etc.) of the shown text were part of the training speech.

Fig. 9 shows the MCD for Users 1 through 4. Here, 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

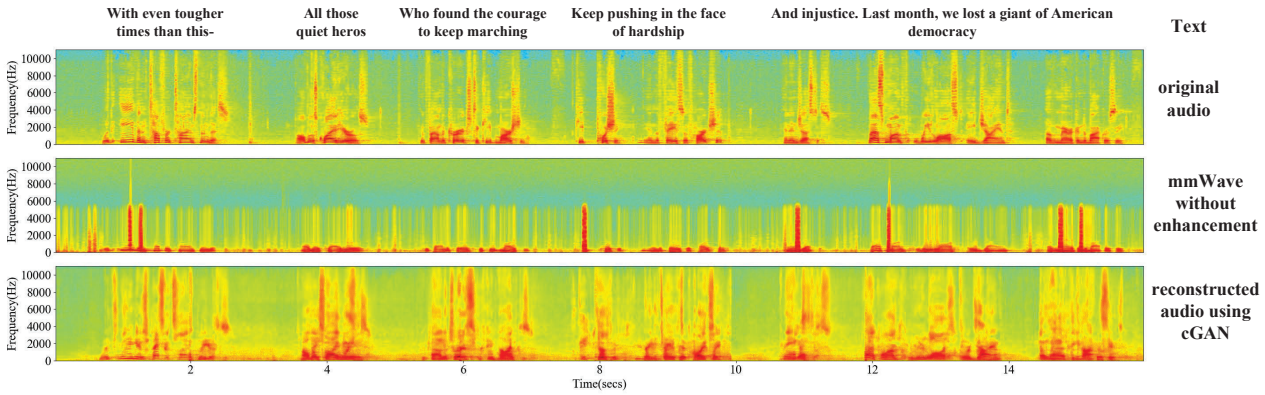


Fig. 8: User$_1$ speech spectrograms for (a) original audio, (b) directly generated from mmWave data without enhancement and (c) audio reconstructed from mmWave data using our cGAN.

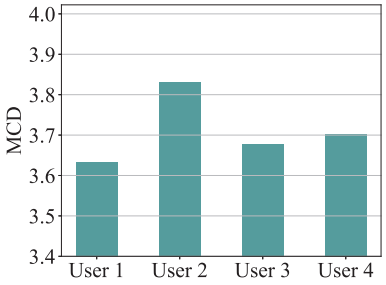
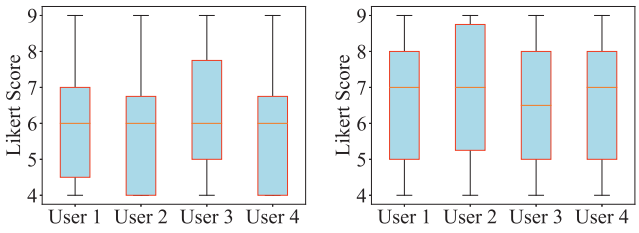
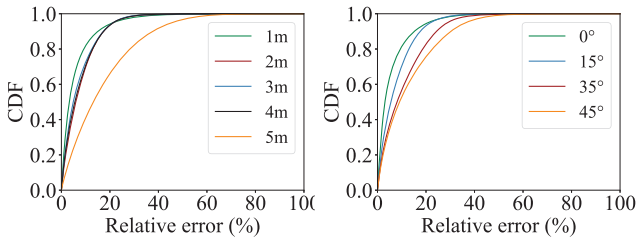


Fig. 9: Objective assessment based on MCD for the recovered audio.



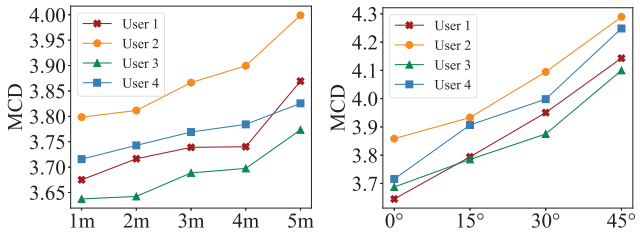
(a) Audio snippet 1 (60s duration) (b) Audio snippet 2 (58s duration)

Fig. 10: Subjective assessment by volunteers for the recovered audio.



(a) Varying distance (b) Varying angle

Fig. 11: Vibration extraction performance (relative amplitude error between mmWave vibration waveform and original audio) at different distances and angles.



(a) Distance (b) Angle

Fig. 12: Audio reconstruction performance at different (a) distances and (b) angles.

only human discernible but shows strong similarities with the original speech. We further evaluate this similarity using subjective evaluation. Fig. 10 shows median Likert score of 20 volunteers for the audio samples of 4 users (both original and reconstructed). As shown in Fig 10, 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.

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 must evaluate the robustness of MILLIEAR for different distances and directions. We vary the distance between the mmWave sensor and speaker from 1m to 5m, and vary the angle from $0°$ to $45°$ in our experiments. These settings are evaluated for the 4 users' audio with individually trained models.

Fig. 11 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. MILLIEAR achieved 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%. This shows that MILLIEAR's vibration extraction achieves a reasonable accuracy in our experiments.

Fig. 12(a) shows the MCD for four users (User$_1$ through User$_4$) with varying test distance from 1m to 5m. We observe

that the MCD score increases, indicating gradual reduction in reconstruction quality. However, the overall change is not observed to be significant at least within the range of the radar. Fig. 15(b) shows that angle has a greater impact on the quality of the reconstructed audio compared to the distance. This can be attributed to the fact that the vibration of the speaker surface (i.e., the reciprocating motion) is increasingly difficult to capture through the radar when they are at an angle from each other. We find that MILLIEAR can reconstruct the audio reasonably accurately within $45°$. This shows that our proposed system can be used by an attacker to carry out the eavesdropping even at different distances and directions.

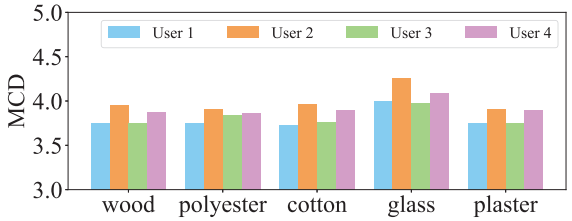


Fig. 13: Audio reconstruction performance with different insulation types.

Impact of different type 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 kinds of insulation materials. We chose 5 kinds of popular soundproof panels which are composed of dense wood, polyester, cotton, glass and soundproof plaster respectively. As shown in Figure 13, 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 [41], [42]), the sound reconstruction is deteriorated by a small margin. In general, we observe that MILLIEAR can achieve a decent performance through penetrating most insulating and soundproofing materials, making it possible to carry out the eavesdropping in common indoor spaces such as offices.

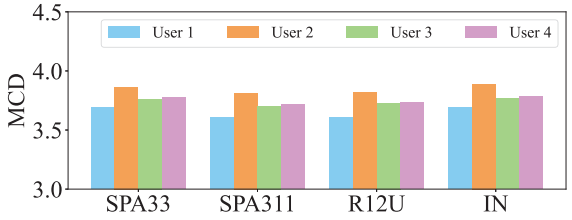
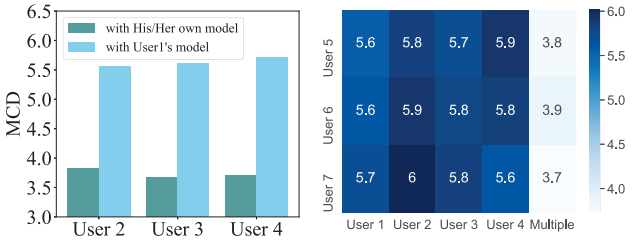


Fig. 14: Audio reconstruction performance with different types of speakers.

Given that speakers from different manufacturers have distinct features (spapes, material, etc.), we evaluate MILLIEAR with four different types of speakers. They are Philips SPA33, Philips SPA311, Edifier R12U, 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. 14 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 are directly exposed to mmWave sensor.

Model Generalization with cross-user training. In order to prove that generalizability of the MILLIEAR's model, we train and test the cGAN model for different users (cross-user training and testing). First, we train the model using User_1 data and test it with Users 2, 3 and 4. Fig. 15(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 provides reasonable to carry out the attack. The reduction can be attributed to the fact that the voice characteristics of different people have different dominant frequency components that are not always accurately reconstructed during cross-user training.



(a) (b)

Fig. 15: Model generalization: comparing MCD for cGAN model trained and tested for different users.

To evaluate if adding more user's data to training will 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. 15(b) shows the MCD. We find that when data from more users are considered 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 yields 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 can then carry out the eavesdropping attack on an unknown user's audio data, making our proposed eavesdropping attack even more harmful in practice.

VIII. 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 results and evaluations show that the attack can be highly effective in a range of practical constraints such as different angles and partitions. With increasing popularity of video conferencing systems and low-cost availability of mmWave radars, there is a need to protect against the proposed attack. Various defense strategies such as use of thicker insulation materials in walls, use of earphones/headphones for eliminating the speaker diaphragm exposure, or using active jamming signals for FMCW radars could be employed to prevent against the attack.

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] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[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] 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.

[15] 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.

[16] 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.

[17] 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.

[18] "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

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

[20] 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.

[21] 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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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

[32] 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.

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

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

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

[36] 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.

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

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

[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] 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.

[42] 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.