Privacy Leakage via Speech-induced Vibrations on Room Objects through Remote Sensing based on Phased-MIMO

Cong Shi
Rutgers University
cs1421@scarletmail.
rutgers.edu

Tianfang Zhang
Rutgers University
tz203@scarletmail.
rutgers.edu

Zhaoyi Xu
Rutgers University
zx111@soe.
rutgers.edu

Shuping Li
Rutgers University
sl1567@scarletmail.
rutgers.edu

Donglin Gao
Rutgers University
dg921@soe.
rutgers.edu

Changming Li
Rutgers University
cl1361@scarletmail.
rutgers.edu

Athina Petropulu
Rutgers University
athinap@soe.
rutgers.edu

Chung-Tse
Michael Wu
Rutgers University
ctm.wu@rutgers.edu

Yingying Chen
Rutgers University
yingche@scarletmail.
rutgers.edu

ABSTRACT

Speech eavesdropping has long been an important threat to the privacy of individuals and enterprises. Recent research has shown the possibility of deriving private speech information from sound-induced vibrations. Acoustic signals transmitted through a solid medium or air may induce vibrations upon solid surfaces, which can be picked up by various sensors (e.g., motion sensors, high-speed cameras and lasers), without using a microphone. To date, these threats are limited to scenarios where the sensor is in contact with the vibration surface or at least in the visual line-of-sight.

In this paper, we revisit this important line of research and show that a remote, long-distance, and even thru-the-wall speech eavesdropping attack is possible. We discover a new form of speech eavesdropping attack that remotely elicits speech from minute surface vibrations upon common room objects (e.g., paper bags, plastic storage bin) via mmWave sensing, signal processing, and advanced deep learning techniques. While mmWave signals have high sensitivity for vibrations, they have limited sensing distance and normally do not penetrate through walls. We overcome this key challenge through designing and implementing a high-resolution software-defined phased-MIMO radar that integrates transmit beamforming, virtual array, and receive beamforming. The proposed system enhances sensing directivity by focusing all the mmWave beams toward a target room object, allowing mmWave signals to pick up minute speech-induced vibrations from a long distance and even through walls. To realize the attack, we design an object identification technique that scans objects in a room and identifies a prominent object that is most sensitive to speech vibrations for vibration feature extraction. We successfully demonstrate speech privacy leakage using speech-induced vibrations via the development of a deep learning framework. Our framework can leverage domain adaptation techniques to infer speech content based only on the unlabeled vibration data of a victim. We validate the proof-of-concept attack on digit recognition through extensive experiments, involving 40 speakers, five common room objects, and attack scenarios with mmWave devices inside and outside the room. Our phased-MIMO-based attack can achieve success rates of 88% ~ 98% and 64% ~ 86% with and without using speech labels for training. The success rates are 81% ~ 94% and 58% ~ 74% for thru-the-wall attacks. Furthermore, we discuss possible defense methods to mitigate this unprecedented security threat.

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**.

KEYWORDS

Speech privacy attack; mmWave sensing; phased-MIMO

1 INTRODUCTION

The unencrypted nature of voice makes speech eavesdropping always a lucrative attack as well as a core topic in computer security. A user's private information or an enterprise's financial/intellectual properties can be compromised if an adversary can listen onto the voice communication channel. Lessons learned from numerous cyber attacks triggered by voice leakage encourage people to speak in soundproof environments, such as rooms deployed with double-glazing glasses and sound absorption sheets.

Nonetheless, research studies reveal that voice communication can still be compromised by leveraging vibrations produced by speech. Motion sensors of the victim's smartphones can be compromised and exploited to sense speech played by loudspeakers [1, 3, 5, 31]. When speech is produced, vibrations can propagate through a solid medium (e.g., a desk or the body of the smartphone) and reach the motion sensors. Recent studies have shown the feasibility of sensing such vibrations on the shell of loudspeakers/smartphones through wireless sensing techniques, such as those based on lasers [51],
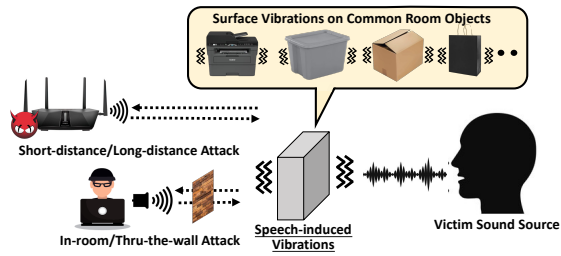
Figure 1: Illustration of possible attack scenarios through the proposed eavesdropping attack based on mmWave sensing upon speech-induced vibrations from room objects.

WiFi [53, 56] and ultra-wide-band (UWB) radars [54]. These studies show promising results, but rely on unrealistic assumptions, where the sensor is in direct contact or visual line-of-sight between the sensor and the sound source (or vibration source). A natural question is whether we can capture airborne speech in scenarios in which there is *no direct contact* or *not even the visual line-of-sight*. Our work here suggests that this attack is indeed possible.

New eavesdropping attack via high-resolution phased-MIMO radar sensing. In this work, we consider a new line of attacks that targets minute *speech-induced vibrations* from common room objects (e.g., paper bags, plastic storage bins, cardboard boxes). When human speech interacts with surrounding objects, it is reflected, refracted, and absorbed, inducing minute vibrations upon the objects' surfaces. Based on this phenomenon, we discover a new and stealthy speech eavesdropping attack that remotely captures such minute vibrations leveraging commercial mmWave devices as illustrated in Figure 1. The millimeter-level wavelength enables mmWave signals to capture vibrations with orders of higher sensitivity compared to traditional radio frequency techniques (e.g., WiFi, RFID). In addition, the integration of mmWave hardware onto mobile and IoT devices (e.g., 5G routers, smart home sensors) makes mmWave increasingly accessible and desirable for adversaries. A recent study shows the success of eavesdropping speeches replayed by smartphone earpieces [6] via mmWave under direct line-of-sight scenarios. However, this kind of attack with a short distance is easy to expose. Our attack capturing vibrations upon in-room objects to achieve eavesdropping is very stealthy. A recent initial work [35] demonstrates its possibility with a short distance, however, the short wavelength of mmWave signals incurs high signal propagation loss, making practical in-room (or even thru-the-wall) attacks difficult in reality. To tackle this inherent challenge, we develop a software-defined high-resolution radar sensing scheme, phased-MIMO radar, a technique that automatically adjusts signals applied to the transmitter and receiver antenna pairs to steer multiple mmWave beams towards the vibrating object, aiming to retain effective speech sensing even under long distances (e.g., >5m) and even occlusion. The designed attack is feasible on commercial mmWave off-the-shelf sensors without any hardware modifications, thus it posts high privacy concerns as the 5G era featured by mmWave communication/sensing has approached.

Differences with existing attacks. The proposed attack shows significant advantages over prior attacks relying on remote sensing:

Audio. Compared to traditional attacks via microphones, which directly sense sounds, our attack uses mmWave sensing to remotely turn room objects into acoustic sensors. It enables our attack to bypass sound insulation, which is designed to lock sounds (mechanical waves) instead of mmWave signals (electromagnetic waves).

Vision. Research studies show the potential of capturing speech-induced vibrations using vision sensors, such as lasers [48], high-speed cameras [11], and lidars [41]. These attacks may leave visual clues (e.g., laser dots, bulky cameras), and they rely on a visual line-of-sight between the laser/camera and the vibrating object. Differently, our attack does not have such limitations (an attack through a non-opaque wall is demonstrated in Section 8.1).

Radio frequency (RF). Our attack based on mmWave has over 25× and 65× higher resolution to minute displacements compared to existing RF sensing in lower frequency bands, such as WiFi [53, 56] and RFID [52]. Compared to eavesdropping based on conductive vibrations generated by loudspeaker or earpiece [6], we target speech vibrations upon room objects induced by airborne sounds, which is far more challenging yet more practical.

Challenges addressed in eliciting speech via speech-induced vibrations. We face several technical challenges to realize such an attack in practice: *1) Unknown angle of target vibrating object:* Effective vibration sensing with the phased-MIMO radar requires mmWave beams of all transmitter and receiver antenna pairs steered towards the vibrating object. Thus, we need to precisely detect the angle of interest where a target object with strong vibrations is located. *2) Unclear response to speech vibrations:* The mmWave signals capture speech-induced vibrations in terms of phase changes, which are sensitive but susceptible to hardware and environment noises. Therefore, effective algorithms for phase denoising and feature extraction are important for successful attacks. *3) Unavailability of labeled training data from the victim:* Successful speech inference relies on well-trained machine learning models, while the labeled training data, especially those from the victim, may not be available to the adversary in practice.

Overcoming the challenges. We design an attack system with advanced radar sensing, signal processing, and deep learning techniques to address the aforementioned challenges. *1) Beamforming based on object identification:* Our system scans through room objects and identifies a target object with the strongest vibration responses for speech extraction. By integrating multiple advanced radar sensing techniques, including transmit beamforming, virtual array, and receive beamforming, our phased-MIMO radar enhances sensing directivity by focusing all the mmWave beams towards the target object, making speech sensing feasible from a long distance and even through walls. *2) Phase calibration and feature extraction:* Our system then applies a series of signal processing techniques to denoise phase values and extracts time-frequency features carrying speech information. *3) Deep learning-based speech content inference:* Given the extracted features, our system performs speech recognition by designing a deep learning framework. Our framework leverages domain adaptation to infer speech content based only on the unlabeled vibration data of a victim, without requiring the victim's any ground-truth speech labels for training. We believe our work makes the following key contributions:

- **A new form of speech eavesdropping:** We discover a new line of practical attacks that elicit speech from minute speech-induced vibrations upon common room objects through mmWave sensing.

It is the first work that shows a *remote, long-distance,* and *thru-the-wall* eavesdropping attack based on micro-meter vibrations induced by airborne speech is feasible in reality.

- **A new attack system:** We design a high-resolution software-defined mmWave sensing scheme, *phased-MIMO radar,* which significantly enhances the signal-to-noise ratio on sensing speech-induced vibrations. We further design an attack system that incorporates a series of signal processing, feature extraction, and deep learning techniques to infer sensitive contents.
- **A new attack analysis:** We validate the proof-of-concept attack by conducting extensive experiments involving speeches of 50 speakers (i.e, 10 live speakers and 40 speakers from a public audio dataset) and 5 common room objects under various attack settings. Our attack can achieve success rates of 88% ∼ 98% and 58% ∼ 74% for recognizing 10 digits with and without using victims' speech labels for training.
- **A new defense study:** We discuss and study a set of passive and active defense methods to protect users' speech privacy from this unprecedented threat.

2 PRELIMINARY STUDY

2.1 Object Surface Vibration Model

During speech production, the sound source modulates the air into sinusoidal harmonic waves (acoustic signals), which propagate through the air in an omnidirectional manner and can be perceived by human ears. When the signals interact with surrounding objects (e.g., printers, plastic storage bins, paper bags), they are reflected, refracted, and absorbed as illustrated in Figure 2(a), causing the surfaces of those objects to vibrate. We refer to the vibrations upon the object's surface as *speech-induced vibrations*. On a closer look, the speech-induced vibrations are caused by the pressure upon the object's surface by the incident acoustic signals. Considering an incident angle of θ, the pressure on the object surface caused by the incident acoustic signal can be described as:

$$p_i(t) = A(t)\cos\theta \cdot e^{j\phi(t)}, \tag{1}$$

where $A(t)$ and $\phi(t)$ are the time-series amplitude and phase of the acoustic signal encoded with speech information. In addition, the acoustic signal reflected by the object's surface also excites the object and applies additional pressure [10]. The overall acoustic pressure applied to the object surface is represented as:

$$\begin{aligned} p(t) &= p_i(t) + p_r(t) \\ &= (1+\Gamma)A(t)\cos\theta \cdot e^{j\phi(t)}, \end{aligned} \tag{2}$$

where Γ is a reflection rate determined by the material of the object. We can find that the overall pressure $p(t)$ is directly related to the amplitude $A(t)$ and phase $\phi(t)$ of the acoustic signals. Note that a small proportion of the acoustic signal is refracted at the object's surface and propagates through the object. These signals do not generate vibrations on the object's surface [23].

The acoustic pressure $p(t)$ causes the object surface to have minute displacements, which are at the order of micrometer [35]. Under the same intensity of acoustic pressure, the magnitude of surface displacement is determined by the material and thickness of the object. We mathematically model the displacement of an object using a spring-mass system as illustrated in Figure 2(b). We denote Young's modulus [7] of the object as E, which reflects the stiffness
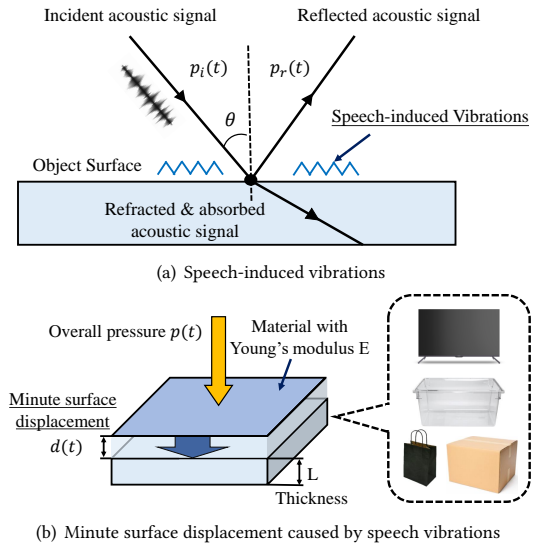

(a) Speech-induced vibrations

(b) Minute surface displacement caused by speech vibrations

Figure 2: Illustration of capturing speech-induced vibrations in terms of surface displacement of room objects.

of the object, while the thickness of the object is represented by L. When an overall acoustic pressure $p(t)$ is applied, we have:

$$d(t) = \frac{p(t)}{E} \cdot L. \tag{3}$$

The displacement of the object surface $d(t)$ has a linear relationship with $p(t)$, which indicates that speech information carried in the sound pressure is encoded into the surface displacements $d(t)$. It is noted that the displacement is also infected by the surface area of the object, which is merged with the pressure load and represented as $p(t)$ in Equation (3). The equation indicates that objects with smaller Young's modulus are more sensitive to acoustic pressure. We showcase Young's modulus of common room objects in Table 1.

2.2 Sensing Speech-induced Vibration via mmWave

We utilize mmWave sensing based on Frequency-Modulated Continuous Wave (FMCW) [47] to extract speech-induced vibrations.

Object-to-radar distance detection. To sense the vibrations of an object, we first need to determine the distance between the radar and a target object picking up the airborne acoustic signals. Particularly, we utilize a mmWave sensor that transmits and receives sequences of chirps with linearly increasing frequency in a fixed slope. The distance between the radar and the object can be calculated based on the frequency difference between the transmitted and the received chirps. Considering the slope of the chirp as β, the distance between the radar and the object can be calculated by: $r(t) = \frac{\Delta f(t)}{\beta}$, where $\Delta f(t)$ denotes the frequency difference. The distance detection procedure can be realized by applying dechirp and range-FFT operations on the received mmWave signals [38]. As we are interested in deriving vibrations upon the object's surface, we decompose the estimated distance into two parts, $r(t) = r + d(t)$, including the static distance between the object and the radar r and the surface displacements (vibrations) of the object $d(t)$.

(a) Experimental Setup

(b) Spectrogram of "Zero"

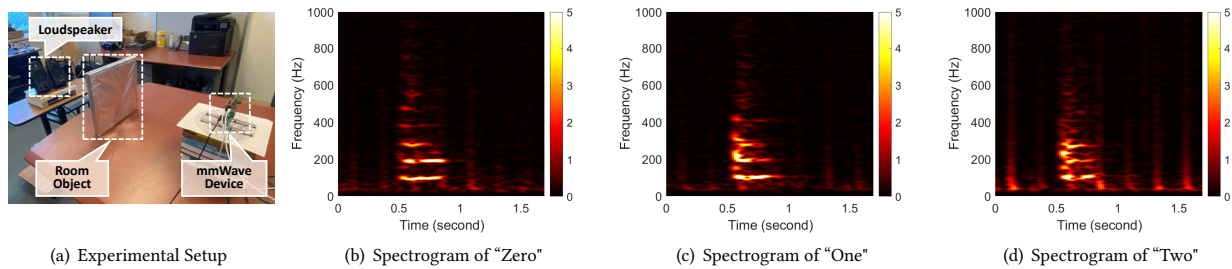(c) Spectrogram of "One"

(d) Spectrogram of "Two"

Figure 3: Illustration of capturing speech-induced vibrations using a tinfoil through mmWave sensing. The surface vibrations of the tinfoil can be captured in terms of phase changes of the received mmWave signals.

Capturing vibrations via phase extraction. The resolution of detected distance $r(t)$ is determined by the bandwidth. Commercial mmWave sensors normally have a $4GHz$ bandwidth [38], and the distance resolution is around $3.75cm$, which is not sufficient for capturing micrometer-level vibrations. Instead of using $r(t)$, our attack leverages the phase changes of the echoed signals at the distance r, which can measure the displacement in a scale even smaller than the wavelength of the mmWave signals:

$$\Delta\phi(t) = \frac{2\pi d(t)}{\lambda} + \Delta\phi_n, \qquad (4)$$

where λ is the wavelength of the mmWave signal and $\Delta\phi_n(t)$ denotes the phase noise due to the signal propagation and hardware. The phase changes $\Delta\phi$ are linearly related to the surface displacements $d(t)$ that are encoded with speech information. Given a $77GHz$ mmWave radar with a short wavelength of $3.89mm$, mmWave sensing can achieve a displacement resolution of $0.59um$ when the phase noise $\Delta\phi_n$ is 9.76×10^{-4} rads, i.e., any displacement larger than that will be detected and considered as the observed vibrations.

2.3 Proof of Concept

We conduct a preliminary experiment to validate the feasibility of using a commercial mmWave sensor to capture speech-induced vibrations of digits. The experimental setup is shown in Figure 3(a). Specifically, we use a pair of loudspeakers (i.e., Logitech Z623 loudspeakers) to play acoustic signals of three digits (i.e., "zero", "one", and "two") at a sound pressure level of $70dB$. A $77GHz$ mmWave radar (i.e., TI AWR2243) is used to sense the surface vibrations. We place a tinfoil to capture the speech-induced vibrations. The loudspeaker-to-object and the object-to-radar distances are all $0.5m$. We apply short-time Fourier transform (STFT) to the phase of mmWave data after performing dechirp and range-FFT operations. The spectrograms of the phases of the three digits are shown in Figure 3(b), 3(c), and 3(d). We can observe that the tinfoil can respond to speech vibrations of the three digits with different time-frequency patterns, indicating that speech-induced vibrations can be captured.

3 THREAT MODEL AND ATTACK OVERVIEW

3.1 Advantages of Exploiting Speech-induced Vibrations via mmWave

No direct contact. Our attack does not require any direct contact between the sound source and the object capturing speech vibrations. In contrast, prior attacks via motion sensors [1, 3, 5, 31] focus

Table 1: Approximate Young's modulus for various objects

Object	Major Material	Young's modulus E (GPa)
Tinfoil	Tin	68 [27]
Glass window	Float glass	47.7 [25]
Plastic storage bin	Polypropylene	1.68 [28]
Cardboard box	Corrugated cardboard	0.644 [2]
Digital TV	Multiple materials	22.68 [18]
Drone	Carbon-fiber-reinforced plastic	183 [36]
Steel cabinet	Steel	200 [26]

on the scenarios of using built-in motion sensors of smartphones to pick up conductive vibrations from a shared solid medium.

Hard to be noticed. With the proposed phased-MIMO radar, our attack can be launched inside common room sizes without exposing the eavesdropping device. Prior attacks based on lasers/lidars [12, 41] will leave visual clues on the object (e.g., a laser dot). The bulky high-speed cameras [11] are also easy to expose.

No visual line-of-sight. Our attack can leverage the penetration property of mmWave to capture speech-induced vibration through opaque obstacles. Prior works explore using laser vibrometers to capture surface vibrations caused by speeches [12, 51]. But these attacks rely on a visual line-of-sight between the laser/camera and the surface, which precludes scenarios with visual occlusion. Differently, our attack does not have such a limitation. We confirmed this via through-wooden(opaque)-wall attacks in Section 8.1.

3.2 Attack Scenarios

We consider scenarios where an adversary aims to eavesdrop on speeches produced by either a playback device of a victim or the victim himself/herself. The adversary can compromise a 5G/mmWave-enabled IoT device to launch the attack. There has been a growing trend of deploying mmWave modules on in-room IoT devices (e.g., 802.11ad WiFi routers) to support high-precision sensing and high-throughput wireless communication. For example, mmWave sensors have been used in intrusion detection [15], robot navigation [42], and hands-free appliance control [4]. These devices and sensors are interconnected and linked to the Internet, and they normally lack security protection [44]. An adversary may compromise the devices by exploiting software vulnerabilities on a large scale (e.g., injecting malicious code [55], attacking access control [50], and exploiting vulnerable software [30]). For example, the adversary can spread malicious scripts over a local network of a company to compromise some of the unsecured IoT devices (e.g., using the devices' factory default login information) and re-program them

for our attack. In addition, the adversary can even launch more profound attacks by using a commercial mmWave device to perform eavesdropping attacks in a different location/room. The adversary can leverage the penetration properties of mmWave signals to sense speech-induced vibrations inside a soundproof room, such as those installed with soundproof barriers, which prevents traditional audio eavesdropping. The acoustic signal will be absorbed if the barriers are soundproof, while our attack can be launched outside the room to sense surface vibrations of in-room objects, which is much more practical compared to traditional eavesdropping attacks via microphones, cameras, lasers, etc.

3.3   Adversary's Capability

Supervised training with victim's labeled data. The adversary will collect the victim's audio samples for generating speech vibrations and obtaining speech labels for training. For instance, the adversary can obtain audio samples from the victim's publicly exposed speech (e.g., YouTube, Zoom Webinar). He/she then uses a loudspeaker to replay the audio samples to generate vibrations on similar room objects, which can be collected using the adversary's mmWave device. As the audio samples are available, the adversary can easily generate speech labels for the collected mmWave data to train a speech recognition model.

Unsupervised training via domain adaptation. The adversary will leverage the victim's mmWave data collected during the attack phase to perform domain adaptation, without requiring audio samples or speech labels from the victim. Particularly, the adversary has a pre-trained speech recognition model built on other people's labeled data. The collected unlabeled data of the victim (mmWave data) during the attack phase is utilized to update the parameters of the pre-trained model in an unsupervised fashion, making the model better fit the victim's feature space.

3.4   Challenges of Adversary

Low signal-to-noise ratio (SNR) for remote vibration sensing. The resolution of displacement sensing is affected by the level of phase noises (i.e., $\Delta\phi_n(t)$ in Eq. (4)). The phase noises can be induced by mmWave signal propagation over the air. The noises will be particularly significant when sensing in-room objects through a room barrier (e.g., a wall or a window), which greatly decays the mmWave signals. We need to design techniques to enhance the SNR for capturing speech vibrations.

Interference of non-vibration-sensitive objects. As discussed in Section 2.1, room objects with relatively smaller Young's modulus normally have stronger responses to airborne acoustic signals. To achieve effective speech-induced vibration extraction, we need to identify/localize these objects and decouple the corresponding mmWave reflections from other room objects, such as those with large Young's modulus and weaker vibrations.

Unreliable speech characteristics from passive speech vibrations. The airborne acoustic signal undergoes complex transformations before being captured in mmWave data. The distortions render low-fidelity speech characteristics in passive speech vibrations. For example, over-the-air propagation attenuates acoustic signals at high frequencies [34]. In addition, successful speech recognition relies on well-trained deep learning models, while the
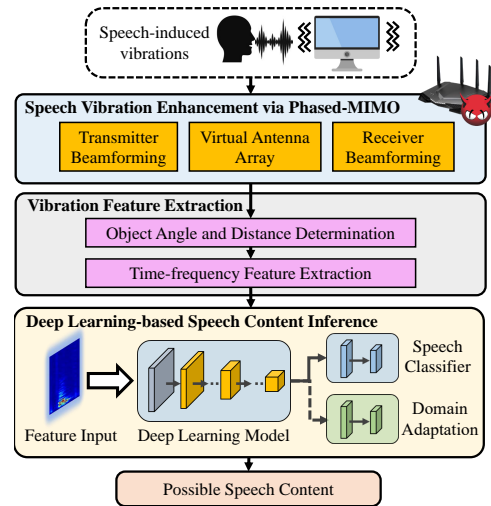

Figure 4: Overview of the proposed attack system.

labeled mmWave data encoding speech-induced vibrations, especially those from the victim, may not be available in practice. We need to derive reliable speech representations for the victim.

3.5   Overview of Attack System

We design an attack system to address the three aforementioned challenges, with the architecture shown in Figure 4.

Speech vibration enhancement via TDM phased-MIMO. We design a high-resolution time-division multiplexing (TDM) phased-MIMO radar based on advanced radar sensing techniques, including transmit beamforming, virtual array, and receive beamforming. Particularly, our scheme steers all the mmWave beams (i.e., the angle where the energy of mmWave signals is confined) towards the vibrating object to enhance the strength of its signal reflections, making the mmWave sensing retains its effectiveness under long distances and through walls. Our scheme further utilizes MIMO techniques to constructively integrate mmWave data from all transmitting and receiving pairs to boost the SNR.

Vibration feature extraction. Our system examines the phases of mmWave data across all distances and angles to identify an object with the strongest vibrations. By combining object identification with the phased-MIMO scheme, our system makes the mmWave sensing focus only on the desired vibrating object in both angle and distance dimensions. Our system then applies a series of signal processing techniques for denoising and extracts time-frequency features that encode speech information.

Deep learning-based speech content inference. We design a deep learning-based framework to recognize the encoded speech content. If the victim's mmWave data and speech labels can be obtained, our framework correlates the time-frequency features of the mmWave data with the speech labels to train a deep learning model with a speech classifier (supervised training) for speech recognition. Otherwise, our framework utilizes only the mmWave data of the victim to adapt a pre-trained model built on other people's data through domain adaptation (unsupervised training). The adapted model better fits the victim's feature space, and it can be used to infer the victim's speech content.
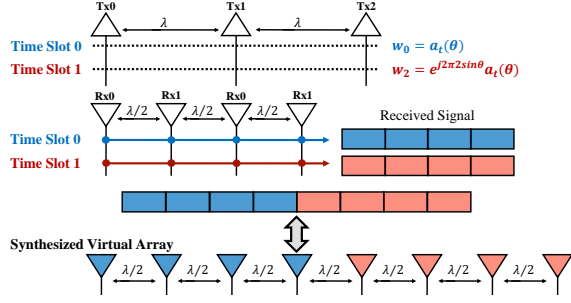
Figure 5: Implementation of phased-MIMO scheme based on a TI AWR2243 mmWave sensor with 3 TX and 4 RX, which gives rise to a receiving virtual array with 8 RX antennas. Two time slots are used to avoid overlapping elements.

4 SPEECH VIBRATION ENHANCEMENT VIA PHASED-MIMO RADAR

The proposed attack needs to reliably capture the minute speech-induced vibrations upon object surfaces (i.e., $d(t)$ in Equation (3)), which are on the order of a micrometer. Prior works [6, 35, 59] show the potential of recovering sounds or speeches using mmWave, but the small wavelength of mmWave signals renders significant propagation loss and low penetration capability. This characteristic results in low SNR of remote sensing on speech-induced vibrations, especially under long-distance and thru-the-wall attack scenarios.

To address this challenge, we design a high-resolution mmWave sensing scheme. For a sequence of frequency-modulated mmWave chirps to transmit, our scheme splits the chirps into several groups and leverages a subset of transmitter antennas (i.e., a phased array) to send each group of chirps. By applying additional phases to the transmitter antennas, the beam of mmWave signals is tuned to focus on a specific direction where the vibrating object is located. In addition, we extend the size of the receiver array by synthesizing a virtual receiver array, which significantly improves the displacement sensing resolution (i.e., aperture [33]) of mmWave. Our scheme then applies a time-division multiplexing (TDM)-MIMO operation to combine received signals from all receiver antennas.

4.1 Analog Transmit Beamforming

As a combination of MIMO radar and phased array, our software-defined phased-MIMO radar integrates the advantages of both techniques in the sense that the radar transmits orthogonal signals, each feeding a phased array structure. The orthogonal signals enable the construction of the virtual array which enjoys a longer aperture, while the phased array structure enables analog beamforming which focuses the transmit power on the desired target. Here, the TDM strategy is deployed to achieve orthogonality in the time domain. In each time slot, the weighted version of the same chirp signal will be transmitted by the phased array, where the weight will be different between slots. During different slots, the transmitted signals can be separated at the receiver. The orthogonality allows one to formulate a virtual array with an increased aperture. The independent observations of the object obtained at the virtual array enable improved estimation of the vibrations.

We consider a transmitter that has a uniform linear array (ULA) with N_t transmit antennas (TXs) spaced apart by d_t, and a receiver

(a) Setup for speech vibration sensing with in-room devices

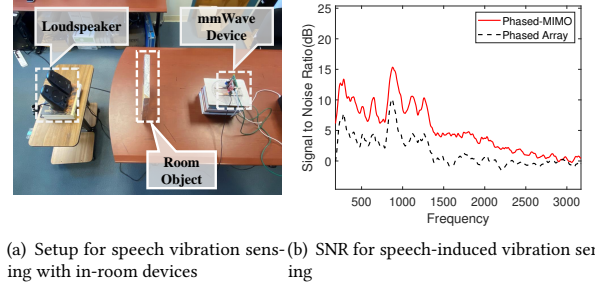(b) SNR for speech-induced vibration sensing

Figure 6: Comparing the vibration sensing sensitivity (SNR) of the proposed phased-MIMO radar and traditional single-channel phased array with an in-room mmWave device.

that has a ULA with N_r receive antennas (RXs) spaced apart by d_r. The array transmits FMCW chirp signals in a time-slotted fashion. In each slot, each TX transmits a weighted version of baseband waveform $x(t)$, using different weights between slots. The weights are chosen so that the transmissions of all antennas add up coherently in the direction of the object. By using different weights in each slot. we create different channels that provide diversity in observing the same object and thus can lead to improved SNR. The weights for the p-th time slot are represented as:

$$\mathbf{w}_p(\theta) = e^{j2\pi p\alpha(\theta)}\left[1, e^{-j2\pi\alpha(\theta)}, \ldots, e^{-j2\pi(N_t-1)\alpha(\theta)}\right]^T$$
$$= e^{j2\pi p\alpha(\theta)}\mathbf{a}_t(\theta), \tag{5}$$

where $\alpha(\theta) = d_t \frac{\sin(\theta)}{\lambda}$ is the normalized propagation delay between antennas on direction θ, λ is the wavelength of signal and $\mathbf{a}_t(\theta)$ is the transmit steering vector.

Considering θ_0 as the direction of the object, the signal transmitted in the p-th slot towards direction θ is:

$$z_p(t, \theta) = \mathbf{a}_t^H(\theta)\mathbf{w}_p(\theta_0)x(t) = b_p(\theta)x(t),$$
$$s.t. \ b_p(\theta) = e^{j2\pi p\alpha(\theta)} \sum_{n=0}^{N_t-1} e^{j2\pi n[\alpha(\theta)-\alpha(\theta_0)]}, \tag{6}$$

where $\{\cdot\}^H$ denotes the conjugate transpose operation. The power of the transmitted signal at direction θ from the p-th slot is:

$$Q(\theta) = E\{z_p(t, \theta)z_p^H(t, \theta)\} = |b_p(\theta)|^2 Q_x \tag{7}$$

where Q_x is the baseband signal power. One can see that $b_p(\theta_0) = N_t$, and the transmitted power is maximized at direction θ_0. The focused power facilitates the estimation of the displacement on the surface $r(t)$ and the corresponding phase change $\Delta\phi(t)$, making the estimation more robust to the noise $\Delta\phi_n(t)$ (see Eq.(4)). Also, the signal transmitted towards direction θ_0 in each slot is the same as that of a TDM-MIMO radar using the same array, except that it is amplified by the number of antennas. The boosted SNR makes it more feasible to capture the minute surface displacements induced by airborne acoustic signals.

4.2 Virtual Array and Receive Beamforming

Our attack synthesizes a virtual array to increase the sensing resolution. At a receiver side with N_r receiving antennas, after mixing with the conjugate of the transmitted signals and stacking the received signals from P slots, we can formulate a virtual array of

(a) Experiment setup
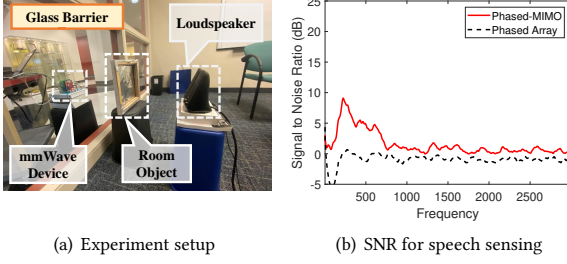
(b) SNR for speech sensing

Figure 7: Comparing the vibration sensing sensitivity (SNR) of the proposed phased-MIMO radar and single-channel phased array radar under the thru-the-wall attack scenario.

$P \times N_r$ elements with steering vector:

$$\mathbf{a}_v(\theta) = \mathbf{a}_r(\theta) \otimes [1, \ldots, e^{j2\pi(P-1)\alpha(\theta)}]^T \in \mathbb{C}^{PN_r \times 1} \quad (8)$$

where $\mathbf{a}$ is the receive steering vector and $\otimes$ denotes the Kronecker product. It provides a larger aperture than that of the physical receiver array. The output of the virtual array is formulated as: $\mathbf{y}_v(t) = N_t \mathbf{a}_v(\theta_0)x(t-\tau)x^H(t)$,

where θ_0 is the direction of the object. On the outputs of the virtual array, receiver digital beamforming is further leveraged to focus on the energy of echoed signals coming from the desired direction. By applying the digital beamformer $\mathbf{w}_v = \mathbf{a}_v(\theta_0)$, the final output signal can be represented as:

$$z(t) = \mathbf{w}_v^H \mathbf{y}_v(t)$$
$$= N_t PN_r x(t-\tau)x^H(t). \quad (9)$$

By combining transmit beamforming, virtual array, and receive beamforming, the SNR for sensing speech-induced vibrations can be significantly increased. Our attack system then applies FMCW techniques to the mmWave signals $z(t)$ to detect the distance to the vibration object and then extract phases from it, as we introduced in Section 2.2.

4.3 TDM Phased-MIMO Radar Implementation

Our TDM phased-MIMO radar is a software-defined scheme to compute/optimize the phase applied to each transmitting/receiving antenna for beamforming. It can be deployed on commercial mmWave devices (e.g., mmWave WiFi routers/sensors) with an antenna array. We showcase our implementation on a single off-the-shelf automotive mmWave device (i.e., TI mmWave module AWR2243 [49]) with a frequency range of $76 \sim 81GHz$. The layout of our implementation is illustrated in Figure 5. The mmWave device has 3 transmitter antennas (TXs) spaced apart by a wavelength (i.e., 3.89mm), and 4 receiver antennas (RXs) spaced apart by half of the wavelength 1.95mm. We implement the analog transmit beamforming on the 3 TXs ($N_t = 3$) by utilizing the built-in phase shifters. The phases obtained with Equation (5) are applied to the phase shifters, rendering the mmWave beam focusing on the direction of interest. We synthesize two virtual antenna arrays (i.e., $P = 2$ and $N_r = 4$) by using time slot 0 and time slot 1, respectively. We thus have 8 receiver antennas in total. Then, we apply the receive beamforming as we have shown in Equation (9).

Table 2: Chirp configurations used for mmWave sensing

Radar Parameter	Value
Frequency Slope, S	$180.470MHz/\mu s$
Idle Time	$5\mu s$
ADC Samples	256
ADC Sample Rate	$23MHz$
Ramp End Time	16.8μ
Number of chirp Per frame	11000
Chirp period, T_s	$21.8\mu s$
Slow-time Sampling Frequency, $f_s = 1/T_s$	$11467.89Hz$
Chirp sweeping bandwidth	$2.008GHz$

We apply FMCW upon the implemented phased-MIMO radar to measure speech-induced vibrations. The chirp parameters setting are summarized in Appendix Table 2. We conduct experiments to compare the performance of phased-MIMO and traditional single-channel phased-array. We use the same experimental setting as our preliminary study in Section 2.3. We show the frequency-specific SNR [54] of the phased-MIMO and traditional phased array that only employs transmit beamforming. The higher the SNR the better the sensing capability will achieve. We can observe that phase-MIMO has around $10dB$ higher SNR compared to phased array, meaning that it is more sensitive to minute speech-induced vibrations. We further examine the capability of our phased-MIMO radar on picking up speech-induced vibrations through a glass wall, as we showed in Figure 7(a). Under this setting, we can find that the single-channel phased array has an SNR close to 0. In comparison, our phased-MIMO radar has over 2dB in a frequency range of $94 \sim 794Hz$, where richer speech information is captured.

5 VIBRATION FEATURE EXTRACTION

5.1 Speech Vibration Extraction via Object Identification

Angle Derivation. In practical attack scenarios, some room objects will capture stronger speech-induced vibrations, which can be leveraged as target objects. To realize the angle detection of such objects, we utilize Sparse Asymptotic Minimum Variance (SAMV) [1], a super-resolution angle estimation algorithm to detect the angle of strong reflectors. Compared to the traditional multiple signal classification (MUSIC) algorithm, the SAMV algorithm is more feasible on devices with a small number of receiver antennas, which suites commercial mmWave devices normally equipped with a smaller size receiver array compared to dedicated radars.

Distance Derivation. Given the angle of interest, our attack further determines the distance an object of interest with prominent vibration responses and then only extracts phases from the distance (range) of the object. It allows the extracted phases to contain strong speech vibrations while removing the impacts of all other objects. Our attack system determines the distance of an object with the strongest vibrations through two steps. We first determine a set of candidate distances potentially having objects with strong reflections by using the range-FFT. We show an example range profile in Figure 8, where some objects that have strong signal reflections (e.g., tinfoil, steel cabinet, and refrigerator) exhibit
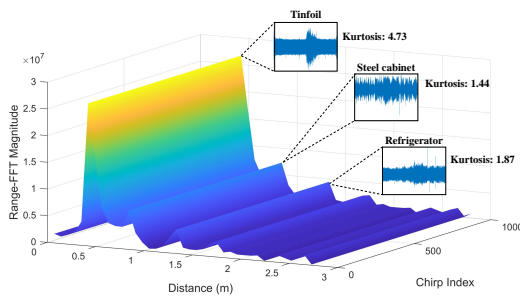
Figure 8: Illustration of identifying object distance (i.e., range bin) with the strongest vibration responses through examining the Kurtosis of the phase.

high Range-FFT magnitude. This observation motivates us to use a threshold-based method upon range-FFT magnitude to detect these objects. As we discuss in Section 2.1, solely relying on signal strength does not guarantee strong speech-induced vibrations, as some objects with large Young's modulus (e.g., steel cabinet) have strong reflections but are not sensitive to speech. To deal with this problem, we design a second step to examine the vibration responses on all candidate objects and select one with the strongest vibration responses for speech-induced vibration extraction. We find that phase segments are normally heavy-tailed when the object is vibrating (e.g., a human subject is speaking). We use Kurtosis [57] to capture and quantify the degree of tailedness of vibrations as Kurtosis is a common statistical metric to quantify the tailedness of data. In our case, the larger the Kurtosis, the more likely the object has strong vibration responses to speech. Note that an adversary could monitor the mmWave phase within a long period to ensure that the analysis involves phase segments of human speeches. We show the Kurtosis values of phases from the three objects in Figure 8, and we observe that the tinfoil with the highest Kurtosis value shows stronger and clearer speech vibrations compared to the other two objects, which confirms its effectiveness.

5.2 mmWave Phase Calibration

The extracted mmWave phases containing speech-induced vibrations could experience phase drifts caused by temperature and humidity variations, which normally exceed the normal range of speech vibrations. We calibrate these phase drifts based on the phase differences across time points: $\delta\phi(t) = \phi(t) - \phi(t-1)$, where $\phi(t)$ denotes the phase value at time point t. If the absolute value of $\delta\phi(t)$ exceeds a pre-defined threshold, which we empirically choose 0.022, $\phi(t)$ will be replaced by a new value computed by the Lagrange interpolation [9] using previous three phases $\phi(t-3), \phi(t-2), \phi(t-1)$. We further calibrate the phase values by applying a bandpass filter with a cut-off frequency of $85Hz \sim 3000Hz$. The fundamental frequency of the human voice is over $85Hz$. The threshold 0.022 is associated with a high vibration magnitude of $1.42mm$. Human speech can hardly exceed this value, and it can thus be used as an upper bound to exclude abnormal phase values. In addition, we find that speech-induced vibrations normally have vibration responses below $3000Hz$, though human voice captured by microphones may reach higher frequencies. The reason is that the diaphragm used in microphones has a much smaller Young's modulus compared

to common room objects. It is more sensitive to high-frequency parts of speech which could be significantly attenuated during over-the-air propagation. We thus use $3000Hz$ as the upper bound.

5.3 Time-frequency Feature Extraction

We extract spectrograms from the calibrated phases as the time-frequency features, which has been shown effective in various acoustic sensing tasks. Our system first detects phase segments involving human speech by examining the moving variance of the phases. The regions of human speech have relatively higher variances compared, and thus they can be segmented with a threshold. Then, our system computes the short-time Fourier transform representations of the phase segment using a sliding time window. We use a sliding window with a width of $25ms$, shifting $10ms$ each step. For each time window, we apply $512 - point$ FFT to derive energy distribution across frequencies. The magnitude of the extracted spectrogram is used as the feature for speech content inference.

6 DEEP-LEARNING-BASED PRIVACY INFORMATION INFERENCE

6.1 Model Overview

To remove the requirement on training labels, our idea is to leverage domain adaptation techniques to transfer the speech knowledge learned from other people's labeled mmWave data to the victim's feature space in an unsupervised fashion. Regarding labeled mmWave data, the adversary may obtain public audio datasets and replay the audio samples to generate vibrations. The vibration data can then be collected using the adversary's mmWave device and labeled by the adversary. We develop a deep-learning-based framework to realize the proposed attack as illustrated in Figure 9. The framework takes as input unlabeled mmWave spectrograms $X_t = \{x_{t,1}, ..., x_{t,n_t}\}$ from the victim and labeled mmWave spectrograms $X_s = \{x_{s,1}, ..., x_{s,n_s}\}$, $Y_s = \{y_{s,1}, ..., y_{s,n_s}\}$ from other people. The ground-truth labels of the victim's mmWave spectrograms, denoted as $Y_t = \{y_{t,1}, ..., y_{t,n_t}\}$, may not be available. Both of the unlabeled and labeled spectrograms are first encoded into a set of low-rank speech representations using a representation extractor $F(\cdot)$. To align the speech representations of the victim and other people, the framework trained an unsupervised domain discriminator $G(\cdot)$ to remove the differences between $F(x_s)$ and $F(x_t)$. A speech classifier trained on X_s and Y_s can then be applied to the unlabeled spectrograms X_t for speech content recognition.

6.2 Representation Extractor

We consider using MobileNet [43] as the representation extractor. The key advantage of MobileNet over traditional architectures (e.g., ResNet and DenseNet) is the use of depth-wise separable convolution and inverted residual blocks, which significantly reduce the number of weights and thus improve the speed of learning, especially when the network is large and deep. To improve the learning efficiency, we resize spectrograms into 2D inputs with a size of 96×96, which have the same height and width. To feed the resized spectrograms into MobileNet, which requires three input channels, our representation extractor uses a convolution layer with 3 filters
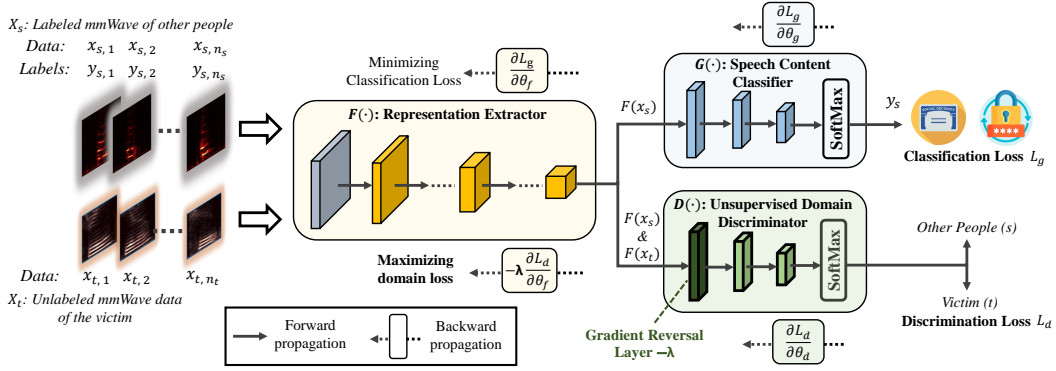
Figure 9: Architecture of our deep-learning framework based on domain adaptation.

to expand the spectrograms into representations with three channels. Then, a series of convolutional layers and MBConv blocks are constructed to project the representations into a hidden space.

6.3 Speech Content Classifier

To infer privacy information in speech, the framework passes the representations to a classifier $G(\cdot)$ consisting of a global average pooling layer, two fully-connected layers, and a SoftMax layer. We use global average pooling [24] as the first layer to average 2D representations of each channel, which helps to avoid overfitting. Two fully-connected layers with 512 and 256 neurons are further leveraged to process the averaged representations. We use the representations of labeled spectrograms X_s and their labels Y_s to optimize the classifier. The classification loss is defined as:

$$L_g = -\frac{1}{n_s} \sum_{i=0}^{n_s} y_{s,i} log(G(F(x_{s,i}))), \tag{10}$$

where $F(x_{s,i})$ denotes the extracted representations from the spectrogram $x_{s,i}$. Note that our attack will train on labeled mmWave spectrograms if the victim's speech labels are available.

6.4 Unsupervised Domain Discriminator

To enable the classifier $G(\cdot)$ applicable to a victim's unlabeled data, we design an unsupervised domain discriminator $D(\cdot)$ as shown in Figure 9. The discriminator is trained to classify the representations as belonging to the victim or other people. The key difference between the discriminator and traditional classifier (e.g., $G(\cdot)$) is the replacement of the first fully-connected layer with a gradient reversal layer (GRL) [13]. During the forward propagation in training, the GRL performs the same mapping as a fully-connected layer. But during the back-propagation, it multiplies the gradient with a negative factor of $-\lambda$ before passing it to the preceding representation extractor $F(\cdot)$. By optimizing the representation extractor with the "reversed" gradients, the extracted representations corresponding to the victim and other people become similar, thereby aligning their distributions. Besides the gradient reversal layer, we use the same architecture of the speech classifier in the domain discriminator. We optimize the domain discriminator based on a mixed dataset of the labeled and unlabeled spectrograms, which we referred to as domain dataset X_d, Y_d. The domain labels Y_d are pseudo labels generated based on the domain (i.e., the victim or other people) of
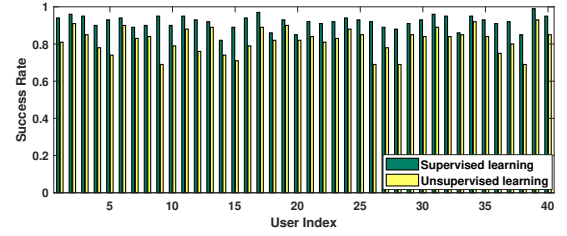

Figure 10: Performance of speech eavesdropping attack using tinfoil as the object to capture speech-induced vibrations.

each spectrogram. The domain loss is defined as:

$$L_d = -\frac{1}{n_s + n_t} \sum_{i=0}^{n_s+n_t} y_{d,i} log(G(F(x_{d,i}))), \tag{11}$$

where $y_{d,i}$ is the domain label of $x_{d,i}$. n_t and n_s represent the number of spectrograms of the victim and other people, respectively.

6.5 Training Process

Our framework only trains the representation extractor and the speech content classifier through backpropagation if the victim's speech labels are available. The network weights θ_f and θ_g are optimized to reduce the classification loss L_d defined in Equation 10. In cases where the victim's speech labels are not accessible, our framework enables the domain discriminator (with weights θ_d) and performs domain adaptation based on the two loss functions defined in Equation 10 and 11. Particularly, the discriminator is trained with both labeled and unlabeled spectrograms with pseudo-domain labels, while the classifier is still trained with labeled spectrograms. With the gradient reversal layer, as we introduced in Section 6.4, the representation extractor learns to maximize the domain loss (i.e., confusing the domain discriminator), rendering similar representations across the victim and other people. The overall loss for optimizing the representation extractor can be formulated as:

$$L_{adv} = L_g - \lambda L_d. \tag{12}$$

By minimizing the L_g while maximizing L_d, the learned representations are indistinguishable between the two sets of spectrograms while being distinct across speech content. Our framework trains the three networks for 200 epochs using a batch size of 16. During each epoch, we initially freeze the discriminator, θ_d, and update θ_f and θ_g. This process accelerates the convergence of the classifier.

(a) Cardboard box bin (setup) (b) Paper bag (setup) (c) Plastic storage bin (setup) (d) Printer (Setup)

(e) Cardboard box bin (results) (f) Paper bag (results) (g) Plastic storage bin (results) (h) Printer (results)

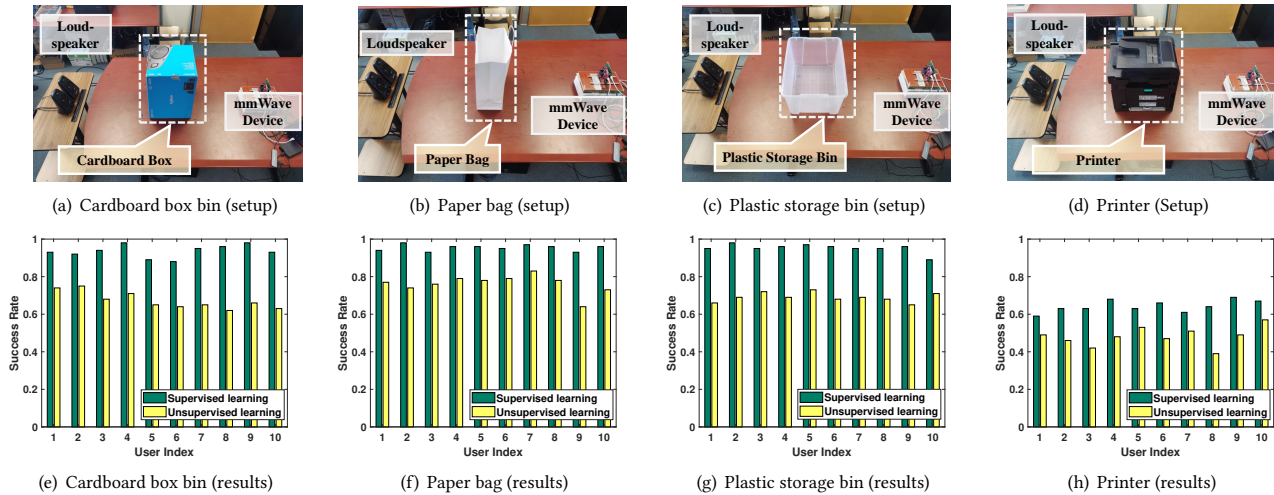**Figure 11: Performance of our attack using different types of common room objects to capture speech-induced vibrations.**

Subsequently, we fix θ_g and update θ_f and θ_d. To update θ_f, θ_g, and θ_d, we employ three separate Adam optimizers [20] with an initial learning rate of 0.01.

7 EVALUATION I: ATTACKS WITH IN-ROOM MMWAVE DEVICES

Speech datasets. We validate the proposed attack by replaying audio samples of 40 speakers (i.e., 34 males and 6 females) from the AudioMNIST dataset [8] to generate speech-induced vibrations by replaying the audio samples using a loudspeaker, which contains 20, 000 samples of spoken digits (0∼9). The sound pressure level measured at a distance of 0.3m to the loudspeaker is 70dB unless specified. We show our studies on more sound volumes in Section 7.5. The speakers have different ages (ranging from 22 to 41 years), gender (6 females and 34 males), and accent (e.g., German, English, Chinese, etc). We focus on digits as they appear in extensive sensitive information, including birthdays, SSN, etc.

Evaluation methodology. We evaluated our attack under the two training requirements introduced in Section 3.3. *1) supervised learning and testing:* we used the collected data of a victim (mmWave data with speech labels) for training and testing. The victim's mmWave data is split into a training set and a testing set with a ratio of 6:4. As the deep learning model needs to use a sufficiently large dataset for effective training, we mix the victim's and other people's labeled data to train the representation extractors and privacy information classifier as we introduced in Section 6. The victim's testing set is then used for evaluation. *1) unsupervised learning and testing:* we used a victim's unlabeled data (mmWave data) and other people's labeled data (mmWave/audio data with speech labels) for domain adaptation as we described in Section 6. We do not involve the victim's speech labels in this setting.

Evaluation metrics. *1) Attack success rate:* For digit recognition, we report the attack success rate defined as the ratio between the number of correctly classified mmWave samples of speech-induced vibration against the total number of mmWave samples. We consider evaluating our attack under both the aforementioned supervised and unsupervised learning settings. *2) Peak signal-to-noise ratio (PSNR):* In addition, we use PSNR [14, 56] to quantify

the quality of speech captured in terms of mmWave data. A higher PSNR means better speech quality. *PSNR* = 1 is used as the benchmark, which means the speech signals are clearly audible for human perception [56] (e.g., after packing the signals into an audio file).

7.1 General Attack Performance

Setup. We use tinfoil as the room object to capture speech-induced vibrations, which are generated by a loudspeaker under the experimental setup shown in Figure 6(a). We collect mmWave data from 40 speakers in the AudioMNIST dataset. For each speaker, we collect mmWave data of 50 repeats per digit, and we collected 20, 000 mmWave samples in total. We take turns considering each speaker as the victim, and the remaining 39 speakers as the other people whose mmWave data are accessible by the adversary.

Results of supervised learning. The speech recognition accuracies for the 40 speakers are shown in Figure 10. We find that our attack has an average success rate of 93.2%. It indicates that our attack is effective when labeled data are available for training.

Results of unsupervised learning. Next, we examine the attack under unsupervised learning settings, where the victim's speech labels are not available. We show the success rates of the 40 speakers in Figure 10. We find that our attack can achieve average success rates of 77.3% under this setting. The results demonstrate that an adversary can still reveal the speech content without using the victim's speech labels for training.

Case Study: unsupervised learning with public audio data. We further conduct a case study to evaluate our attack under an even more practical scenario: the adversary directly uses the audio samples of other people using the public audio dataset (AudioMNIST) with the proposed domain adaptation techniques, without the need of collecting labeled mmWave data. Our attack has 72.1% average attack success rate, which shows that it is also feasible to use public audio data for domain adaptation.

7.2 Attack Using Different Room Objects

Setup. We examine the attack performance when using various types of common room objects, including a cardboard box 0.3×0.3×0.4 m^3, a paper bag 0.4×0.15×0.4 m^3, a plastic storage bin 0.54×0.39×0.24
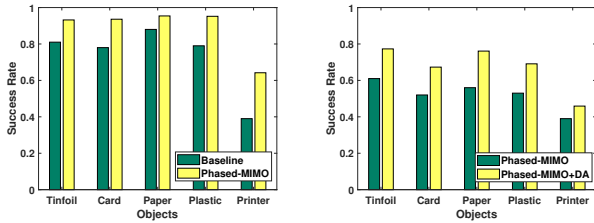
(a) Supervised learning with and without phased-MIMO

(b) Unsupervised learning with and without domain adaptation (DA)

Figure 12: Ablation study: (a) attack performance with (phased-MIMO) and without phased-MIMO (baseline); (b) attack performance with (phased-MIMO+DA) and without domain adaptation (phased-MIMO).

m^3, and a printer 0.4×0.35×0.4 m^3 as shown in Figure 11(a), 11(b), 11(c) and 11(d). The distances between the mmWave device and the object and between the object and the loudspeaker are all 0.5m. For each object, we collect mmWave data from 10 speakers (i.e., speakers 1 ∼ 10 in the AudioMNIST), with 50 repeats per digit. In total, 5,000 mmWave samples are collected for each object.

Results of the cardboard box. We show the success rates of using a cardboard box for our proposed attack in Figure 11(e). Our attack can achieve average success rates of 93.6% for supervised learning scenarios. In addition, the success rate is 67.3% for unsupervised learning. The results show the feasibility of eavesdropping via cardboard boxes, which are common in indoor environments.

Results of the paper bag. We show the results of a paper bag in Figure 11(f), another type of representative room object. When the victims' labeled data is used for training, the average success rate is over 95.4%. For unsupervised learning without speech labels, the attack success rate reaches 76.1%. The results show that our speech eavesdropping can also be applied to paper bags.

Results of the plastic bin. The results of the plastic storage bin are shown in Figure 11(g). The average success rate can reach up to 95.2% using labeled data from victims. The results of unsupervised learning also achieve a high average success rate of 69.1%, showing the attack effectiveness on plastic objects.

Results of the printer. We also make an evaluation of the printer under both supervised and unsupervised learning cases. In Figure 11(h), the attack with supervised learning has an average accuracy of 64.5%. The accuracy of unsupervised learning case can also reach an average of around 45.9%. Although the accuracy of the printer is not high as other objects, it is still much higher than the random guess with 10%. The lower success rate could be attributed to the printer's complicated inner structure.

7.3 Ablation Study

Setup. We examine the importance of phased-MIMO radar and domain adaptation scheme, which are two key components in our attack, through comparisons with a baseline without applying both techniques. We use the data of 10 speakers collected from the tinfoil, cardboard (card), paper bag (paper), plastic storage bin (plastic), and printer with the same setup in Section 7.2.

Role of Phased-MIMO Radar. We demonstrate the attack performance with and without the implementation of the proposed
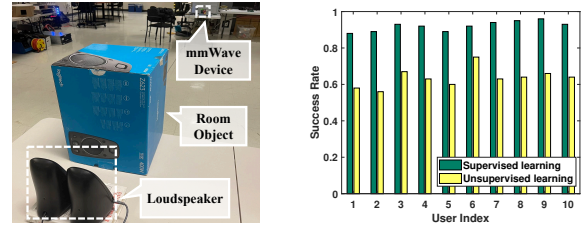
(a) Setup for long-distance attack

(b) Cardboard (result)
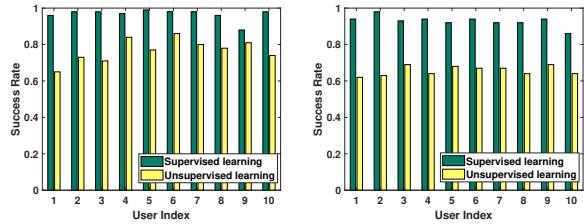
(c) Paper bag (result)

(d) Plastic bin (result)

Figure 13: Performance of our attack under a long sensor-to-object distance of 5m.

phased-MIMO radar. In Figure 12(a), it can be observed that the average attack success rate is enhanced by 15.3% in the supervised learning scenario. This enhancement is more pronounced for the cardboard box (increased by 15.6%), plastic storage bin (increased by 16.2%), and printer (increased by 25.2%), which possess greater thickness and smaller vibration magnitudes. These findings indicate that the proposed phased-MIMO radar can considerably improve the attack success rates, especially for thicker objects.

Role of Domain Adaptation. The attack success rates with and without domain adaptation in Figure 12(b). We observe that for all five objects, the success rates are markedly enhanced after employing the domain adaptation technique, yielding an average improvement of 14.9%. These advancements indicate that the domain adaptation technique more effectively aligns the representations of the victim and other individuals.

7.4 Long-distance Attack

Setup. We further examine our attack under scenarios where the mmWave device is at a long distance to the object. The experiment setup is shown in Figure 13(a), and the radar-to-object distance is 5m. mmWave data of the aforementioned cardboard box, paper bag, and plastic storage bin are collected. For the printer, we do not observe speech patterns at 5m, potentially due to the significant attenuation caused by the embedded electronic components.

Deriving speech content. We show the performance of deriving speech content under such a long-distance attack scenario in Figure 13(b), 13(c), and 13(d). The attack success rates are 92.1%, 93.2%, 94.1% for the cardboard box, paper bag, and plastic storage bin. For unsupervised learning, our attack can still retain high success rates of 63.6%, 74.2%, and 67.3%. We find that the success rates of both supervised and unsupervised learning approximate the evaluation results under a radar-to-object distance of 0.5m in Section 7.2. It confirms the effectiveness of the proposed phased-MIMO scheme in retaining attack effectiveness under long distances.

Table 3: PSNR comparison of four different objects under object-to-sensor distances from 7m to 11m.

Object Distance	Tinfoil	Cardboard	Paper bag	Plastic bin
7.0m	9.02dB	8.03dB	12.61dB	3.72dB
9.0m	5.27dB	4.53dB	6.85dB	0.83dB
11.0m	2.67dB	1.99dB	2.46dB	0.30dB

Speech quality assessment. We further study the maximum distance of our attack by examining the PSNR under $7m$, $9m$, and $11m$. The PSNRs of the tinfoil and the three objects are shown in Table 3. We observe that at a distance of $11m$, our attack still has over $1dB$ PSNR for tinfoil, cardboard box, and paper bag, meaning that speech patterns can still be captured. For the plastic storage bin, the maximum attack distance is around $9m$. Such long attack distances allow effective attacks in common home and office rooms.

7.5 Attack under Different Practical Factors

We further study the attack performance under different realistic attack scenarios, where the sound source (loudspeaker) and the object are not aligned with each other. We examine the PSNR of extracted speech signals (phases of mmWave data) under different distances and angles. We consider three practical volumes for the sound source, including $65dB$, $70dB$, and $75dB$.

Speaker-to-object Distance. We show the PSNRs of the tinfoil and paper bag under speaker-to-object distances of $0.1m \sim 1m$ in Figure 14 (a) and (b), respectively. We observed that for tinfoil, even with a low sound volume of $65dB$, our attack still obtain PSNR over 1 within a distance of $0.6m$. The PSNRs of the paper bag are much higher for all distances and sound volumes. This is because the lightweight surface of the paper bag is easier to vibrate. The results show that our attack can elicit speech when the object is in proximity (e.g., less than $1m$) to a sound source.

Speaker-to-object Orientation. Similarly, we examine the PSNRs with the loudspeaker placed at different angles to the object (i.e., $0° \sim 50°$). As shown in Figure 15 (a), the attack achieves PSNRs of around $2dB$, $4dB$, and $6dB$ for the three sound volumes across different angles. We have a similar observation for the paper bag in Figure 15 (b), where the PSNRs are consistently over $6dB$, $10dB$, and $11dB$ for the three sound volumes at different angles. The consistency of the PSNRs for different angles indicates the orientation variations have a minor impact on the attack.

7.6 Attack on Live Human Speech

Setup. We further evaluate the effectiveness of our attack on the live human speech by performing a case study. We use tinfoil as the room object to pick up the speech of 10 human subjects (8 male and 2 female), whose sound volumes are around $70dB$, a common sound volume used in voice communication. The distance between the human speaker and the tinfoil is $0.2m$, while the distance between the tinfoil and the mmWave device is $0.5m$. The data collection procedures were approved by our university's IRB.

Results to attack live speech. Figure 16(b) shows the success rates on live human speech. The average success rate is 87.2%. The success rate is around 10% lower compared to those of using loudspeakers as the sound source, which is partially caused by the lower
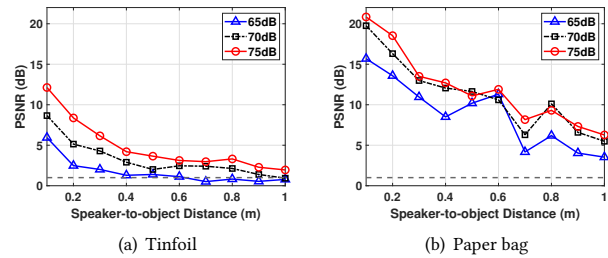
Figure 14: Object-to-speaker distance study: different distances between the surface and the object (at a distance of $0.5m$ and a degree of $0°$ to the mmWave device).
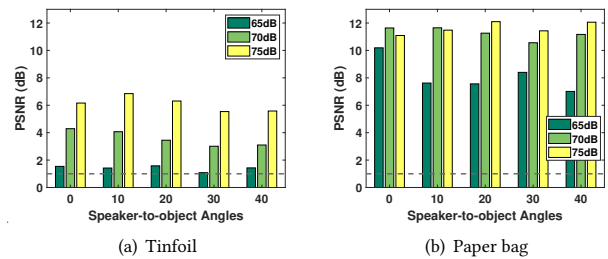
Figure 15: Object-to-speaker angle study: different angles between the loudspeaker and the object (at a distance of $0.5m$ and a degree of $0°$ to the mmWave device).

sound volumes and training set with smaller sizes. Under an unsupervised learning scenario, our attack achieves an average success rate of 49.8% (the random guess is 10%). The results show the potential of eliciting private information from live speech leveraging our proposed eavesdropping attack scheme.

8 EVALUATION II: THRU-THE-WALL ATTACKS

8.1 Attack Through Thick Wooden Wall

Setup. In this scenario, the same tinfoil is utilized as the target object for capturing passive speech vibrations. As shown in Appendix Figure 17, the tinfoil and the radar are separately placed on two sides of a wall, which is made of composite wood boards with a thickness of 0.33m. The distance between the radar/object and wall is set to be 0.1m/0.2m and the loudspeaker is placed 0.5m away from the tinfoil. During the experiment, we collect mmWave data from 10 users in the AudioMNIST dataset and each digit is repeated for 50 times. We take turns treating each speaker as the victim and all 9 remaining speakers as the other people.

Result. The results of supervised and unsupervised speech recognition are shown in Figure 18(a), respectively. The success rate of supervised speech recognition can reach more than 92.9%. For more challenging unsupervised scenarios, the success rate still remains at a high level of more than 69.8%. Although the wooden wall induces attenuation on both audio signals and mmWave signals, high success rates in both supervised and unsupervised scenarios demonstrate that thru-the-wall speech eavesdropping attacks can be achieved through our attack scheme.
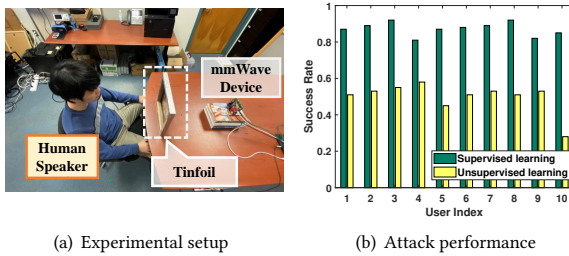
(a) Experimental setup   (b) Attack performance

Figure 16: Experimental setup and performance of attack on live human speech-induced vibrations upon tinfoil.
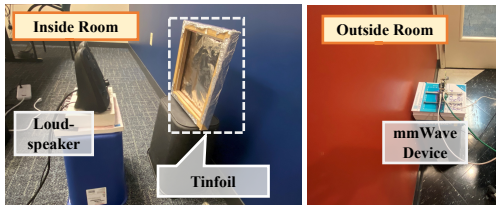

Figure 17: Experimental setup and attack performance of eavesdropping through a thick (33cm) wooden wall.

8.2 Attack Through a Glass Wall

Setup. We also apply a similar setup with a glass wall as the barrier to deploy our speech privacy eavesdropping attack, which is shown in Figure 7(a). For the data collection, we follow the setup in 8.1.

Result. As shown in Figure 18(b), supervised speech recognition delivers a high success rate of more than 80.1%. For the results of the unsupervised scenario, the speech recognition accuracy is about 53.3%. The possible reason for the performance degradation on glass walls compared to the wooden wall is that the glass barrier holds a relatively smooth and flat surface, which helps to reflect the majority of the mmWave signal before it can reach the room objects and bring back passive vibration.

9 RELATED WORK

Eavesdropping using microphones. An adversary may deploy or compromise audio recording devices in a target environment (e.g., hidden microphones, mobile phones) to eavesdrop on voice communications [21, 29]. Different from these traditional attacks, which directly capture sounds, our attack uses mmWave sensing to remotely turn room objects into sound sensors. It exhibits two key advantages over microphone-based eavesdropping: (1) by using mmWave/electromagnetic signals, our attack can bypass many defenses preventing audio eavesdropping. For example, our attack can sense through vacuum-insulation glasses/walls, which is impossible for eavesdropping via microphones. (2) with phased-MIMO, our attack works through a much longer distance compared to sounds, and it does not suffer from sound loss/interference.

Eavesdropping using vision sensors. Besides audio recording devices, research studies demonstrate the possibility of capturing speech-induced vibrations using vision sensors, such as lasers [48], high-speed cameras [11], and Lidars [41]. For example, Davis et al. [11] utilizes a high-speed camera to capture video streams to recover vibrations from some room objects (e.g., a bag of chips). These
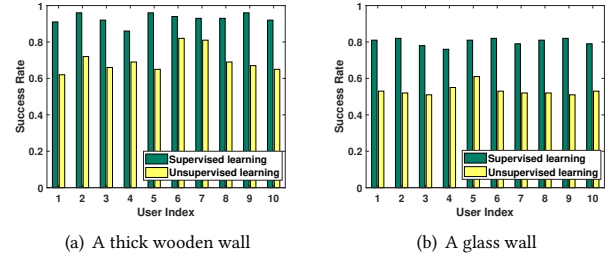

(a) A thick wooden wall   (b) A glass wall

Figure 18: Attack performance of eavesdropping through a glass wall and a thick (33cm) wooden wall.

vision-based attacks exploit surface vibrations induced by speeches, which share some similarities with our attack. However, these attacks all rely on a visual line-of-sight between the laser/camera and the surface, which precludes scenarios with visual occlusion. Differently, our attack does not have such a limitation. We confirmed this via through-wooden(opaque)-wall attacks in Section 8.1.

Speech sensing using radio frequency signals. Researchers have exploited the wireless nature of radio frequency (RF) signals for speech eavesdroppings, such as those using WiFi [53], Ultra-Wideband Inpulse Radio [54], and mmWave [6, 22, 35]. For example, Wei *et al.* examine the variations of channel state information (CSI) in WiFi signals to derive patterns of speeches played by loudspeakers. UWHear [54] utilizes an impulse Radio Ultra-Wideband-based radar to sense vibrations upon the surface of loudspeakers to infer speech content. More recent studies show the potential of mmWave sensing to capture vibrations on human throat [22] or replayed by the smartphone earpiece [6]. These prior attacks show promising results, but they are limited to sensing conductive vibrations propagating through a solid medium. Differently, we target vibrations induced by airborne sounds, which is far more challenging yet practical. Ozturk et al. [35] show the potential of using mmWave to sense surface vibrations generated by audio chirps. It does not reveal privacy speech leakage. In addition, the high propagation loss rooted in mmWave limits the effective sensing distance and precludes attacks through walls. Differently, in this work, we design a software-defined phased-MIMO radar to significantly extend the distance and penetration capability of mmWave sensing.

10 DEFENSE AGAINST MMWAVE-BASED EAVESDROPPING

Defense against in-room attack. An adversary might compromise mmWave/5G-enabled IoT devices to initiate an attack. One intuitive solution is to isolate the devices from the Internet or other communication channels during critical voice communication. However, this approach may also impact the functionality of applications, such as remote/virtual meetings, since high-throughput wireless communication is a key feature of mmWave. Another potential method involves programmatically injecting random amplitude and phase noise into the mmWave signals at the receiver. This approach can effectively scramble mmWave signals associated with high-frequency vibrations while preserving normal communication and sensing functionalities (e.g., localization and gesture recognition). Furthermore, constraining the sampling rate of mmWave sensing devices could limit their ability to pick up human speech.

Human speech generally has frequency responses above $85Hz$. According to the Nyquist theorem, limiting the device's sampling rate to below $170Hz$ would render it incapable of capturing speech signals, although the mmWave sensing capability may be degraded.

Defense against thru-the-wall attack. RF shielding using electromagnetic absorbers [39] could provide an effective defense against the thru-the-wall attacks proposed in this paper. For instance, pyramidal absorbers [37] typically absorb electromagnetic signals within a broad frequency range [45, 46] (e.g., $0.8 \sim 110$ GHz), thereby blocking the mmWave signals and preventing the attack. Various other more common materials, such as carbon, metallic, and metal oxide, can be also utilized to isolate the room from mmWave signals [16]. Additionally, constructing walls using multiple layers of varying materials and thicknesses can effectively attenuate mmWave signals and change the direction of the transmitted wave, making mmWave-based eavesdropping challenging. The selection of materials and thicknesses can be based on the permittivity and attenuation characteristics of specific building materials [16]. Furthermore, topological periodic patterns can be etched onto the metal layers of the wall to create a frequency-selective surface [19, 32, 58], which allows waves of particular frequencies (e.g., WiFi signals, Bluetooth) to pass through to maintain application functionalities while blocking waves of other frequencies to defend against mmWave eavesdropping attacks launched from the outside.

11 DISCUSSION

Attack Complexity vs. Payload. In this paper, we demonstrate both in-room and through-the-wall attacks with different levels of complexity. Our evaluations presented in Sections 7 and 8 reveal that in-room scenarios yield higher attack success rates (e.g., 93.2% for tinfoil), particularly under the unsupervised learning setting (e.g., 77.2%). In comparison to the thru-the-wall attacks, in-room attacks necessitate more complicated procedures, such as compromising and reprogramming mmWave-enabled IoT devices. The increased complexity of in-room attacks produces higher attack success rates, which can be considered as the attack's payload. In contrast, through-the-wall attacks are less complex and are more favorable for adversaries. These attacks require only a commercial mmWave device, without accessing or modifying any devices within the room. The wall's occlusion also prevents users from noticing the attack, resulting in higher stealthiness. But this type of attack must contend with the significant signal propagation loss caused by the wall. Our results in Section 8 show that the attack can maintain an accuracy of 85.1% for supervised learning using tinfoil, though the accuracy declines to 53.3% for unsupervised learning. With such a capability, the adversary still has a substantial likelihood of inferring sensitive information. In general, a higher attack complexity yields a larger payload for the proposed attack.

Potential Attack Improvement. As the first work in this line of research, we conduct a thorough investigation of speech privacy attacks using the proposed phased-MIMO radar sensing techniques in various practical scenarios. Our study demonstrates the feasibility of recognizing isolated words, specifically simple digits, which are often used to reveal and quantify potential speech leakage [6, 31, 60]. We believe that extending the dataset (e.g., by incorporating a larger vocabulary) will further enhance our attack

performance. However, collecting a large labeled mmWave dataset might necessitate substantial manpower, making it challenging in practice. Thus, we intend to develop a speech reconstruction algorithm capable of converting the vibrations into audio signals resembling microphone data. This conversion can be achieved using deep encoder-decoder networks, such as autoencoders [17] and U-nets [40]. By utilizing the reconstructed signals, we can employ pre-trained speech recognition models (e.g., Google Speech-to-Text, Microsoft Azure) built with extensive speech datasets.

Thru-the-wall Attack for Classified Environments. As an initial demonstration, we show that our phased-MIMO-based attack can extract speech through glass and wooden walls with 92.9% and 80.1% success rates (supervised learning), respectively. Since our attack is deployed on a commercial mmWave device with a relatively smaller antenna array (i.e., 3 transmitter antennas and 4 receiver antennas), it may not be effective for classified environments equipped with thick acoustic insulation layers (e.g., mineral wool, fiberglass, or acoustic foam) and multi-layered walls (e.g., Mass-loaded vinyl). The substantial wall thickness and complex materials can significantly attenuate mmWave signals. To enhance the attack's effectiveness, the adversary could improve the phased-MIMO radar sensing approach by utilizing a larger transmitter array or employing the TDM phased-MIMO used in our paper, but with additional time slots. This enables the formation of a larger virtual array, facilitating the generation of a more focused beam to counter the severe attenuation caused by through-the-wall propagation.

12 CONCLUSION

In this paper, we proposed a remote, long-distance, and thru-the-wall eavesdropping attack that elicits privacy information from minute vibrations upon objects' surfaces through mmWave sensing. To overcome the challenge of high propagation loss of mmWave, we designed a software-defined high-resolution phased-MIMO radar that allows mmWave signals to pick up minute speech-induced vibrations. We further developed an attack system that integrates object identification/localization, time-frequency feature extraction, and deep learning techniques to infer sensitive speech content. As mmWave/5G devices are increasingly deployed, we believe the attack remotely turning room objects into passive "microphones" via mmWave signals will be a critical security concern.

13 ACKNOWLEDGMENT

REFERENCES

[1] Habti Abeida, Qilin Zhang, Jian Li, and Nadjim Merabtine. 2013. Iterative Sparse Asymptotic Minimum Variance Based Approaches for Array Processing. *IEEE Transactions on Signal Processing* 61, 4 (2013), 933–944.

[2] Samir Allaoui, Z Aboura, and ML Benzeggagh. 2011. Contribution to the modelling of the corrugated cardboard behaviour. *arXiv:1110.5417* (2011).

[3] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2019. Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. *arXiv preprint arXiv:1907.05972* (2019).

[4] Saniwise Automatic. 2014. SANIWISE Automatic Sensor Trash Can. (2014). https://www.saniwise.com/products/saniwise-automatic-sensor-trash-can-

with-lid-50-liter-13-gallon-stainless-steel-garbage-bin-powered-by-batteries-not-included-for-kitchen-office-home-silent-and-gentle-open-and-close-1/.

[5] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *Proceedings of NDSS*.

[6] Suryoday Basak and Mahanth Gowda. 2022. mmSpy: Spying Phone Calls using mmWave Radars. In *IEEE Symposium on Security and Privacy (SP)*. 1211–1228.

[7] E Baumgart. 2000. Stiffness—an unknown world of mechanical science. *Injury* 31, Suppl 2 (2000), B14–23.

[8] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2018. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418* (2018).

[9] Jean-Paul Berrut and Lloyd N Trefethen. 2004. Barycentric lagrange interpolation. *SIAM review* 46, 3 (2004), 501–517.

[10] Leonid M Brekhovskikh and Oleg A Godin. 2012. *Acoustics of layered media I: Plane and quasi-plane waves*. Vol. 5. Springer Science & Business Media.

[11] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. 2014. The Visual Microphone: Passive Recovery of Sound from Video. *ACM Transactions on Graphics* 33, 4 (2014), 79:1–79:10.

[12] Takahiro Fukumori, Chengkai Cai, Yutao Zhang, Lotfi El Hafi, Yoshinobu Hagiwara, Takanobu Nishiura, and Tadahiro Taniguchi. 2022. Optical laser microphone for human-robot interaction: speech recognition in extremely noisy service environments. *Advanced Robotics* 36, 5-6 (2022), 304–317.

[13] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. 1180–1189.

[14] John HL Hansen and Bryan L Pellom. 1998. An effective quality evaluation protocol for speech enhancement algorithms. In *Fifth international conference on spoken language processing*.

[15] Honeywell Home. 2014. 5853 WIRELESS GLASSBREAK DETECTOR. (2014). https://www.honeywellhome.com/us/en/products/security/security-accessories/5853-wireless-glassbreak-detector-5853/.

[16] Nozhan Hosseini, Mahfuza Khatun, Changyu Guo, Kairui Du, Ozgur Ozdemir, David W Matolak, Ismail Guvenc, and Hani Mehrpouyan. 2021. Attenuation of several common building materials: millimeter-wave frequency bands 28, 73, and 91 GHz. *IEEE Antennas and Propagation Magazine* 63, 6 (2021), 40–50.

[17] Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi, and Shingo Kuroiwa. 2013. Reverberant speech recognition based on denoising autoencoder.. In *Interspeech*. 3512–3516.

[18] Chang-Hoi Kim, Seong-In Moon, Jae-Boons Choi, Young-Jin Kim, Jeoung-Gwen Lee, and Ja-Choon Koo. 2005. Elastic Modulus Measurement of a Large Size Digital TV Display Unit. *Journal of the Korean Society for Precision Engineering* 22, 3 (2005), 115–122.

[19] Min-Sung Kim and Sung-Soo Kim. 2019. Design and Fabrication of 77-GHz Radar Absorbing Materials Using Frequency-Selective Surfaces for Autonomous Vehicles Application. *IEEE Microwave and Wireless Components Letters* 29, 12 (2019), 779–782.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Akvile Kiskis. 2020. Android Microphone Eavesdropping. In *17th International Conference on Information Technology–New Generations*. Springer, 39–43.

[22] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, and Wenyao Xu. 2020. VocalPrint: Exploring a Resilient and Secure Voice Authentication via MmWave Biometric Interrogation *(SenSys '20)*. 312–325.

[23] Rui Li, Tao Wang, Zhigang Zhu, and Wen Xiao. 2010. Vibration characteristics of various surfaces using an LDV for long-range voice acquisition. *IEEE Sensors Journal* 11, 6 (2010), 1415–1422.

[24] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).

[25] AZO Materials. 2001. Float Glass - Properties and Applications. (2001). https://www.azom.com/properties.aspx?ArticleID=89.

[26] AZO Materials. 2012. ASTM A36 Mild/Low Carbon Steel. (2012). https://www.azom.com/article.aspx?ArticleID=6117.

[27] MatWeb. 2021. Aluminum, Al. (2021). https://www.matweb.com/search/DataSheet.aspx?MatGUID=0cd1edf33ac1456e93a0aa6fc666c0e0.

[28] MatWeb. 2021. Overview of materials for Polypropylene, Molded. (2021). https://www.matweb.com/search/DataSheet.aspx?MatGUID=08fb0f47ef7e454fbf7092517b2264b2.

[29] Declan McCullagh and Anne Broache. 2006. FBI taps cell phone mic as eavesdropping tool. *Cnet News* 1 (2006), 2100–1029.

[30] Francesca Meneghello, Matteo Calore, Daniel Zucchetto, Michele Polese, and Andrea Zanella. 2019. IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices. *IEEE IoT Journal* 6, 5 (2019), 8182–8201.

[31] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *Proceedings of USENIX Security Symposium*. 1053–1067.

[32] Raj Mittra, Chi H Chan, and Tom Cwik. 1988. Techniques for analyzing frequency selective surfaces-a review. *Proc. IEEE* 76, 12 (1988), 1593–1615.

[33] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. 2013. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine* 1, 1 (2013), 6–43.

[34] Philip McCord Morse and K Uno Ingard. 1986. *Theoretical acoustics*. Princeton university press.

[35] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2021. Sound recovery from radio signals. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8022–8026.

[36] Luiz Claudio Pardini and Luis Guilherme Borzani Manhani. 2002. Influence of the testing gage length on the strength, Young's modulus and Weibull modulus of carbon fibres and glass fibres. *Materials research* 5 (2002), 411–420.

[37] M. Park, Y. Ryu, T. Liu, and S. Kim. 2015. Design of wide bandwidth pyramidal microwave absorbers with ferrite composites of broad magnetic loss spectrum. In *2015 IEEE International Magnetics Conference (INTERMAG)*. 1–1.

[38] Sandeep Rao. 2017. Introduction to mmWave sensing: FMCW radars. *Texas Instruments (TI) mmWave Training Series* (2017), 1–11.

[39] Vince Rodriguez, Brett Walkenhorst, and Jorgen Bruun. 2019. A Method for the Measurement of RF Absorber using Spectral Domain Transformations. In *2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting*. 347–348.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[41] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with Your Robot Vacuum Cleaner: Eavesdropping via Lidar Sensors *(SenSys '20)*. 354–367.

[42] Samsung. 2014. Jet Bot+ Robot Vacuum with Clean Station. (2014). https://www.samsung.com/levant/vacuum-cleaners/robot/vr8500t-white-vr30t85513w-eu/.

[43] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE CVRP*. 4510–4520.

[44] Sabrina Sicari, Alessandra Rizzardi, and Alberto Coen-Porisini. 2020. 5G in the internet of things era: an overview on security and privacy challenges. *Computer Networks* 179 (2020), 107345.

[45] EC RF solutions. 2015. For microwave anechoic chambers - E&C Engineering. (2015). https://ece.co.jp/assets/pdf/products/EC-SORB%20ECP-3_en.pdf.

[46] TDK RF solutions. 2015. For microwave anechoic chambers - TDK. (2015). https://www.tdkrfsolutions.tdk.com/images/uploads/data-sheets/TDK-IS-Absorber-Series.pdf.

[47] Andrew G Stove. 1992. Linear FMCW radar techniques. In *IEE Proceedings F (Radar and Signal Processing)*, Vol. 139. IET, 343–350.

[48] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems. In *Proceedings of USENIX Security Symposium*. 2631–2648.

[49] Texas Instrument. 2020. AWR2243 Evaluation Module (AWR2243BOOST) mmWave Sensing Solution. (2020).

[50] Chandrakumar Thangavel and Parthasarathy Sudhaman. 2017. Security challenges in the IoT paradigm for enterprise information systems. In *Connected environments for the internet of things*. Springer, 3–17.

[51] Payton Walker and Nitesh Saxena. 2022. Laser Meager Listener: A Scientific Exploration of Laser-based Speech Eavesdropping in Commercial User Space. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. 537–554.

[52] Chuyu Wang, Lei Xie, Yuancan Lin, Wei Wang, Yingying Chen, Yanling Bu, Kai Zhang, and Sanglu Lu. 2021. Thru-the-wall Eavesdropping on Loudspeakers via RFID by Capturing Sub-mm Level Vibration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–25.

[53] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. We can hear you with wifi!.(2014). In *Proceedings of MobiCom*. 593–604.

[54] Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani B Srivastava. 2020. UWHear: through-wall extraction and separation of audio vibrations using wireless signals. In *Proceedings of SenSys*. 1–14.

[55] Dongdi Wei and Xiaofeng Qiu. 2018. Status-based detection of malicious code in Internet of Things (IoT) devices. In *2018 IEEE CNS*. 1–7.

[56] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of ACM MobiCom*.

[57] Peter H Westfall. 2014. Kurtosis as peakedness, 1905–2014. RIP. *The American Statistician* 68, 3 (2014), 191–195.

[58] Te-Kao Wu. 1995. Frequency selective surfaces. *Encyclopedia RF Microwave Engineering* (1995).

[59] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of ACM MobiSys*.

[60] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of ACM MobiSys*. 301–315.