High-Quality Speech Recovery Through Soundproof Protections via mmWave Sensing

Feng Lin, *Senior Member, IEEE*, Chao Wang, *Student Member, IEEE*, Tiantian Liu, *Student Member, IEEE*, Ziwei Liu, *Student Member, IEEE*, Yijie Shen, *Student Member, IEEE*, Zhongjie Ba, *Member, IEEE*, Li Lu, *Member, IEEE*, Wenyao Xu, *Senior Member, IEEE*, and Kui Ren, *Fellow, IEEE*

***Abstract*—Online voice communications are widely used nowadays. To protect speech from leakage, people tend to initiate the talk in sound-isolated environments. In this artcle, we reveal a novel attack that recovers high-quality speech from outside soundproof zones. The rationale of the attack is to leverage sound-sensitive characteristics of piezoelectric materials, i.e., a piezo film that can change the phase of reflected mmWaves when placed in a sound field. If the attacker transmits mmWaves and analyzes reflected signals from the piezo film, the speech information can be compromised. More importantly, the piezo film is paper-like and works without a power supply. We propose a new speech recovery methodology to transform sound waves into wireless signals and build an end-to-end eavesdropping system working as a through-wall "microphone" to recover high-quality speech stealthily. To combat signal attenuation and improve speech quality, we develop a speech-enhancement scheme based on generative adversarial networks and propose to use multi-antenna information for intelligible speech reconstruction. We conduct extensive experiments to evaluate the system. The results indicate that the system achieves over 98% accuracy for digit recognition and works well over 5 m away through the wall. We also test the system under complex scenarios and give countermeasures.**

***Index Terms*—Eavesdropping, mmWave sensing, speech recovery, through-obstacle perception.**

I. Introduction

ONLINE voice communication has been preferred by more and more people nowadays, especially with the increasing demand for free-of-contact communication due to the coronavirus [1], [2]. People can start personal conversation and virtual conferences at any time by using their electronic products, such as smartphones and personal computers. These devices play the machine-rendered speech during the communication. Considering that the speech can be closely related to personal privacy (e.g., passwords) or enterprise secrets (e.g., commercial decisions), the online conversation often takes place in a sound-isolated environment which can prevent the speech from leakage.

However, there are works revealing that attackers can leverage sound-induced vibrations on objects to recover speech and retrieve secret information. Typical methods include motion sensors [3], [4], [5], [6], RF signals [7], high-speed cameras [8], lidars [9], vibration motors [10], and hard drives [11]. Among these methods, the RF-based method is favored by researchers to exploit for long-range and through-wall eavesdropping because RF signals can penetrate obstructions (e.g., walls) and break through soundproof protections. On the other hand, researchers also find that the physical characteristics (e.g., the hardness and the material) of the vibrating objects have a vital impact on the vibration and thus influence the performance of speech recovery [12]. For example, some rigid objects (e.g., glasses) require a large sound pressure level of over 80-90 dB to induce measurable vibrations. Such a large sound volume is not often the case in daily life.

Based on the above consideration, we want to investigate whether there is a wireless-based method that can recover the speech by capturing the air-propagated sound waves directly rather than focusing on the induced vibration on objects. Moreover, how is the eavesdropping performance when penetrating the soundproof obstructions? To investigate this idea, we preliminarily find that the electromagnetic characteristics (e.g., reflection coefficients) of piezo films in the mmWave band can be changed by the incident sound waves due to the piezoelectric effect [13], [14]. Thus, if the mmWave signals are transmitted toward a piezo film when the piezo film is placed in a sound field, it is possible to recover the speech information from the reflected mmWave signals. For example, as shown in Fig. 1, when a victim is participating in an online conference and plays the speech via a loudspeaker, an adversary outside the soundproof room can use a mmWave sensor to transmit mmWave signals towards the room and analyze the reflected signals from the piezo film. If the speech information can be successfully decoded from the reflected signals, the secret speech can be compromised.

Manuscript received 28 September 2022; revised 15 September 2023; accepted 3 October 2023. Date of publication 5 October 2023; date of current version 11 July 2024. This work was supported in part by National Key Research and Development Program of China under Grant 2020AAA0107700, in part by National Natural Science Foundation of China under Grants 62032021, 62372406, 61972348, 62172359, and 62102354. *(Corresponding author: Feng Lin.)*

Feng Lin, Chao Wang, Tiantian Liu, Ziwei Liu, Yijie Shen, Zhongjie Ba, Li Lu, and Kui Ren are with the School of Cyber Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China, and also with the ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, Zhejiang 310027, China (e-mail: flin@zju.edu.cn; wangchao5001@zju.edu.cn; tiantian@zju.edu.cn; zivliu@zju.edu.cn; shenyijie@zju.edu.cn; zhongjieba@zju.edu.cn; li.lu@zju.edu.cn; kuiren@zju.edu.cn).

Wenyao Xu is with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14261 USA (e-mail: wenyaoxu@buffalo.edu).

Digital Object Identifier 10.1109/TDSC.2023.3322295

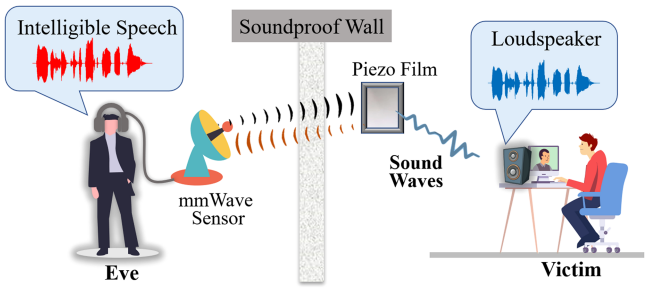


Fig. 1. mmPhone can recover intelligible speech emitted by loudspeakers in a soundproof room by sensing an in-room piezo film with mmWave signals.

However, it is nontrivial to realize such an attack. First, a sound-decoding scheme is required in order to recover intelligible speech from the reflected mmWave signals. Second, static and moving objects in the background can also induce reflected mmWave signals which pose challenges to decoding the speech. In particular, the moving objects can raise dynamic noise in the reflected mmWave signals and cause interference in the recovered speech. Third, the mmWave signals suffer from the increasing sensing distance and penetrating loss in the through-wall condition. The recovered speech can be distorted due to the low signal-to-noise ratio of the reflected mmWaves.

To solve these challenges, we develop an end-to-end eavesdropping system called mmPhone which can recover the speech protected by the soundproof room. Specifically, we first model the transformation between the sound waves and mmWave signals via the piezo film. Then we proposed a methodology to decode the speech from reflected mmWave signals. To eliminate the environmental interference, we develop a phase-calibration method to suppress the dynamic noise induced by moving objects in the background. To fight against the attenuation of mmWave signals during the through-wall sensing, we propose a speech enhancement scheme based on the generative adversarial network (GAN) to decrease the noise floor and enhance the speech quality. Finally, we adopt a harmonic-extension method based on the multiple receiving channels of the mmWave hardware and improve the intelligibility of recovered speech. Our contributions are summarized as follows:

- We establish a novel sound-mmWave transformation scheme based on the piezoelectric effect. We theoretically model the transformation and reveal a new speech eavesdropping that can break through soundproof protection and recover intelligible speech.
- We design an end-to-end eavesdropping system based on a commercial mmWave sensor. We develop a phase-calibration scheme to suppress the clutters and build a denoising neural network to mitigate the impact of through-wall attenuation and enhance the recovered speech. We also propose a harmonic-extension method to reconstruct high-intelligibility speech.
- We perform extensive experiments to evaluate the proposed attack system. We find that the attack can recover soundproof-protected speech with high quality from over 5 m away in the through-wall condition. The digit recognition accuracy can be over 98% which indicates that the attack can cause great threats to secret information such as passwords in the real world.

Compared with prior work [15], we highlight significant differences between this work and the prior work. 1) In this work, we propose a new denoising scheme to boost the performance. Previous work used a public dataset containing clean and noisy audio to train the denoising neural network, which cannot fully characterize the noise condition in the mmWave-based audio recovery, and thus has a limited performance to denoise the mmWave-recovered speech. This work solved this problem using a three-phase scheme (i.e., data synthesis, offline training, and online denoising) and achieves a better generality. 2) We proposed a solution for motion interference in this work. Previous work did not consider the impact of human movements on the attack but conducted experiments in a very limited condition with no moving subjects. In this work, we found that moving objects in the room can cause severe interference in the recovered speech which is distorted with poor speech quality. To improve the practicality of the attack in the real world, we investigated the impact of human movements and developed a clutter suppression method based on the correlation between the misaligned range bins. After the clutter suppression, the clutters induced by human movements can be eliminated, which makes the system resilient to human movements and able to calibrate the distorted speech. 3) We conducted more robustness studies in this work. Previous work only evaluated the system at a fixed sound pressure level (SPL) of around 67 dB, which cannot fully demonstrate the system's performance in real life. Besides, according to the proposed sound-mmWave transformation, the SPL can affect the phase of demodulated signals, and thus affect the performance. In this work, we further investigated the impact of the SPL on the system performance with quantitative experiments. The new result validates the effectiveness of the attack in causing threats to conversations with normal SPL. 4) We performed a detailed analysis of countermeasures in this work. Previous work mentioned countermeasures (e.g., blocking mmWave signals and wearing headsets) briefly without detailed discussion or analysis of their effectiveness. In this work, we added detailed discussions about the countermeasures against the proposed attack, i.e., shielding-based and jamming-based methods. In particular, we quantitatively analyzed the jamming-based method based on the parallel-interference and cross-interference in radar sensing, and discussed the effectiveness and practicality of the countermeasures in real life. 5) Previous work did not conclude the drawbacks of the proposed end-to-end attack system. To better understand the ability of the attack, in this work, we discussed the vulnerabilities of the attack based on our observations during the experiments, i.e., human-body blocking and ambient noise.

II. Background and Threat Model

A. *mmWave Sensing*

Nowadays frequency-modulated continuous wave (FMCW) sensors in the mmWave band are widely used in wireless sensing, such as industrial monitoring and automotive applications. As shown in Fig. 2, the FMCW sensor transmits FMCW (i.e.,

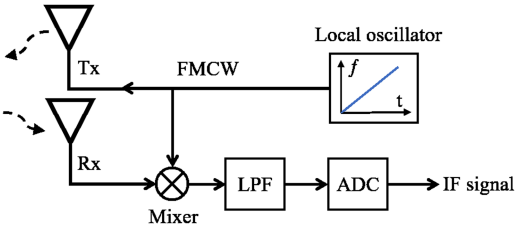


Fig. 2. Principle of the mmWave sensor. The sensor transmits frequency-modulated continuous-waves and demodulates the reflected signals. Then the signals are fed into a low-pass filter (LPF) to generate the intermediate-frequency (IF) signals and sampled by an analog-to-digital converter (ADC).

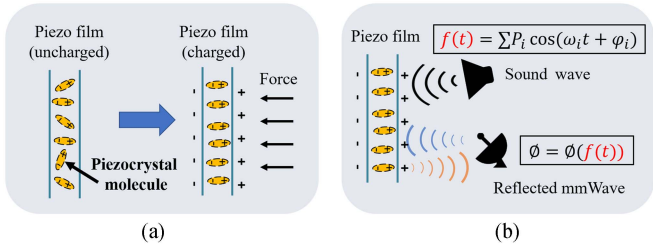


Fig. 3. (a) Applied force on the piezo film can cause the film to charge. (b) Sound waves can affect the phase of mmWaves reflected from the film.

chirp) and captures reflected mmWave signals from objects. The received signal is demodulated by a mixer and passed to a low-pass filter $LPF(\cdot)$. Then the demodulated signal is sampled by an analog-to-digital converter (ADC) to produce the intermediate frequency (IF) signal. Considering two sinusoidal inputs of the mixer $s_i = A_i \sin(\omega_i t + \phi_i)$ $(i = 1, 2)$ where s_1 and s_2 are the transmitted and received signals respectively, the output IF signal s_o:

$$s_o = LPF(s_1 \cdot s_2) = A_3 \cos((\omega_1 - \omega_2)t + \phi_1 - \phi_2), \quad (1)$$

where A_3 is the amplitude of s_o, $\omega_1 - \omega_2 = 2\pi f_c$, and f_c is the frequency of s_0. The reflected signal s_2 from an object can be taken as the replicate of the transmitted signal s_1. The time delay τ between s_1 and s_2 can be derived by $\tau = \frac{2 \cdot d}{c}$, where d is the distance to the object and c is the speed of light. Then we can derive the initial phase of s_o:

$$\phi_0 = 2\pi f_c \tau + \phi_1 - \phi_2 = \frac{4\pi d}{\lambda} + \phi_1 - \phi_2, \quad (2)$$

where f_c is the frequency of s_1 and λ is the wavelength of s_1. Considering a static object at distance d_0, the reflection coefficient of the object can be denoted as $\Gamma = |\Gamma| e^{j\phi_r}$. Given that s_2 is the echo of s_1 reflected by the object, i.e., $\phi_1 - \phi_2 = \phi_r$, we can derive

$$\phi_0 = \frac{4\pi d_0}{\lambda} + \phi_r, \quad (3)$$

where $\frac{4\pi d_0}{\lambda}$ is a constant. According to (3), if the Γ of the object changes with external stimulation (e.g., sound waves), then ϕ_0 can also change with the stimulation.

B. Piezoelectric Film

Polyvinylidene fluoride (PVDF) [16], [17] is a plastic material which is ubiquitous in daily life, such as folder covers and piping products. The PVDF film is flexible to cut into different shapes and attached to objects' surfaces. After the processing of polarization, the film can be piezoelectric which can transduce the sound pressure into electromagnetic signals [18]. Specifically, when a force is applied on the film surface, the film can charge as shown in Fig. 3(a). For a varying force, such as the force induced by the changing sound pressure in the air, the piezo film can charge and discharge alternatively. If the changes in the film are induced by the speech-related sound waves in the room, the piezo film can be a potential side channel that leaks the speech contents as shown in Fig. 3(b). It is worth mentioning that the piezo film is a passive component that can work in a power-free manner. It can be attached to objects' surfaces (e.g., book covers and cabinets). Thus, the passive film is more concealable and undetectable than electronic eavesdropping devices (e.g., microphones).

C. Threat Model

1) Attack Scenario: We consider a scenario where a victim participates in an online voice communication in a soundproof room, e.g., participating in an online conference at the company, and using loudspeakers to play the speech during the conference. To prevent the secret speech from leakage, we assume the victim is cautious about audio-recorded devices (such as microphones and smartphones) which can potentially be compromised for eavesdropping. These electric eavesdropping devices in the room can be easily found by commercial detectors which leverage emitted electromagnetic signals or wireless signals from malicious devices for detection. In such a scenario, the piezo film which is free of any electronic components like a power supply or analog-to-digital converters, can be in the disguise of a book cover or paper which cannot be detected by conventional detectors. Due to the aforementioned appearance, an executor can easily take the film into the room. For public-shared zones, such as enterprise conference rooms, the attacker can also be an insider adversary of the company and thus he/she can preset the film in the room. For private zones, the attacker can deploy the film on a book cover or a painting which is sent to the victim as a "gift". Note that the attacker can also a hired person to place the film.

2) Assumption: The victim broadcasts the conference participants' speech in the room via a loudspeaker. We assume that the piezo film is pre-placed in the room, which can be achieved by social engineering [19] as we mentioned above. For example, an adversary can pay for a hired person who can get into the room to place the film beforehand or the attacker can even be an adversarial college of the victim and easy to approach the room in advance. The adversary requires the film's location to transmit directional mmWave signals. He/She can get the location information from the one who places the film or by scanning the whole space until being able to decode the audio.

III. Modeling the Sound-mmWave Transformation

A. Sound Propagation Model

The sound field induced by the propagating sound waves through the air can be formulated as $P(x, t)$, where the x and

t are the position and time respectively [20]. Thus, the sound pressure at position x_0 can be formulated as

$$P(t) = \sum_i P_i \cos(\omega_i t + \phi_i), \tag{4}$$

where the P_i, ω_i, ϕ_i are the pressure amplitude, radian frequency and initial phase of i_{th} wave component at position x_0. [1] We can acquire the force induced by the sound waves on the film according to $F = P \cdot S$, where S is the film size and P is the sound pressure on the film. Due to the piezoelectric effect, the quantity of electric charges induced by F on the piezo film can be calculated as $Q = D_{33} \cdot F$, where D_{33} is the piezoelectric constant of the film. We get the relationship between sound waves and induced charges on the film

$$Q = \sum_i D_{33} S P_i \cos(\omega_i t + \phi_i). \tag{5}$$

B. mmWave Signal Reflected From a Piezo Film

When the mmWave signals are transmitted towards the room for through-wall sensing, the piezo film can partially reflect the mmWave signals. At the boundary between the air and the film surface, it is the electric properties of the two mediums that determine the phase change of reflected mmWave signals. According to the *Transmission Line* model [21], the mmWave reflection on the film can be characterized by the reflection coefficient

$$\Gamma = \frac{Z_2 - Z_1}{Z_2 + Z_1}. \tag{6}$$

where Z_1 and Z_2 are the intrinsic impedance of the air and the film, $Z_i = \frac{\omega \mu_i}{k_{z_i}}, k_i = \omega\sqrt{\mu_i \varepsilon_i} (i = 1, 2)$,$\omega$ is the radian frequency of the mmWave, μ_i, ε_i are the magnetic permittivity and medium permittivity. For a given incident angle θ of the mmWave signals, we can calculate the propagation coefficients k_{z_1}, k_{z_2} in the direction of propagation according to $k_{z_1} = k_1 \cos\theta, k_{z_2} = \sqrt{k_2^2 - k_x^2}$, where $k_x = k_1 \sin\theta$. Then the reflection coefficient can be calculated as

$$\Gamma = \frac{\cos\theta - \sqrt{\varepsilon_r - \sin^2\theta}}{\cos\theta + \sqrt{\varepsilon_r - \sin^2\theta}}, \tag{7}$$

Without loss of generality, we assume the transmitted mmWaves are vertically incident to the film, i.e., $\theta = 0$. Then the relative permittivity of the film can be calculated according to $\varepsilon_r = 1 - \frac{\omega_p^2}{\omega^2}$, where $\omega_p = \sqrt{\frac{Ne^2}{m\varepsilon_0}}, N = \frac{Q}{Sd}$ [21]. Now we can derive the approximate linearity between the phase of Γ and the quantity of sound-induced charges Q: $\Phi(\Gamma) = \Phi(\frac{1-\sqrt{a_1 Q - 1}j}{1+\sqrt{a_1 Q - 1}j}) = -\frac{1}{Q_0\sqrt{a_1 Q_0 - 1}} Q + C_0 + o(Q - Q_0)^2 = k_0 Q + C_0, |Q - Q_0| < \epsilon$, where k_0 and C_0 are determined by a_1 and Q_0. a_1 is the coefficient of induced charges and is determined by the relative permittivity ε_r of the film ($\varepsilon_r \approx 11$) and incident angle θ of soundwaves. Q_0 is the average quantity of induced charges by the incident soundwaves and is determined by the average sound pressure level of induced soundwaves. ϵ is a small value representing the quantity changes. According to (5), we derive the phase of Γ

$$\phi_r = k_0 D_{33} S \sum_i P_i \cos(\omega_i t + \phi_i) + C_0. \tag{8}$$

[1] When the space position of the sound wave is considered, (4) is a function of both position and time: $P(x, t) = \sum_i P_i \cos(\omega_i t + \phi_i - kx)$ where k is the wave number. For a specific position x_0 (e.g., the position of the piezoelectric film), we have $P(x_0, t) = \sum_j P_j \cos(\omega_j t + \phi_j - kx_0)$. For simplicity, it can be written as $P(t) = \sum_i P_i \cos(\omega_i t + \phi_i)$.

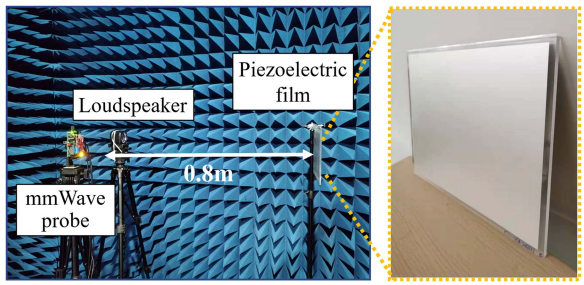


Fig. 4. Verification experiment in an anechoic chamber. The chirp audio is played by a loudspeaker. The A4-size piezoelectric film is stuck to an acrylic board with glue to avoid vibration. The acrylic board is fixed on a static pole.

C. Decoding Sound From Reflected mmWaves

According to (3) and (8), we derive the relationship between ϕ_0 and sound wave function:

$$\phi_0 = \Phi_1(\sum_i P_i \cos(\omega_i t + \phi_i)), \tag{9}$$

where $\Phi_1(\cdot)$ is a linear function. We assume there are n cycles of reflected mmWave signals $x_1, x_2, \ldots, x_n$. Here we denote the phase of the IF signals demodulated from the i_{th} one as ϕ_0^i. Then the phases of the IF signals can be written as $\phi_0^1, \phi_0^2, \ldots, \phi_0^n$. According to (9), the phase ϕ_0^i is determined by the sound waves at a specific time. In other words, we can take the demodulation of each cycle of the mmWave signal as a sampling of the sound waves where the sampling rate $f_s = \frac{1}{T_{chirp}}$. Up to now, we have modeled the relationship between the acoustic signal and the reflected mmWave signals. Next, we perform further experiments to validate the sound-mmWave transformation model.

D. Verification Experiment

1) Line-of-Sight Audio Recovery: We first performed experiments to validate the sound-mmWave transformation model in the line-of-sight (LoS) condition. We used a commercial mmWave sensor (AWR1843Boost) to interrogate the film in an anechoic chamber (Fig. 4). We stuck the film on an acrylic board with glue to avoid sound-induced vibration on the film. The loudspeaker played audio chirps (50∼2 kHz, 65 dB SPL) towards the piezoelectric film. Note that the displacement of a bag film under 80 dB SPL (within 0.5 m from the loudspeaker) is below 0.2 μm [8] which is far smaller than the resolution of the mmWave probe (tens of microns) [22]. We placed the mmWave sensor beside the loudspeaker which is beyond the field-of-view ($\pm 28^\circ$ horizontally) of the sensor, and there is no physical contact between them. We set the chirp rate of transmitted mmWave as 10.2 k chirp/sec (f_s=10.2 kHz) and used a microphone to record the played audio as the reference signals.

According to the aforementioned sound-mmWave transformation model, we applied N-point fast Fourier transform (FFT)

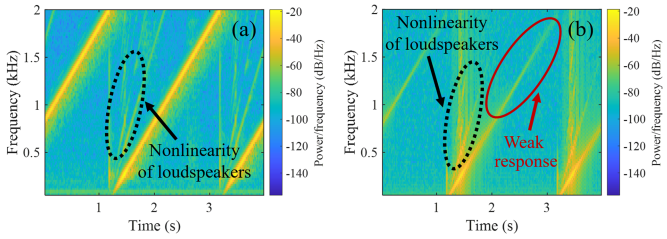


Fig. 5. (a) Audio recorded by a microphone and (b) Audio decoded by the proposed sound-mmWave transformation. The dotted circles indicate harmonics caused by acoustic nonlinearity of the loudspeaker. The harmonics above 250 Hz disappear and thus cause little interference to human speech.

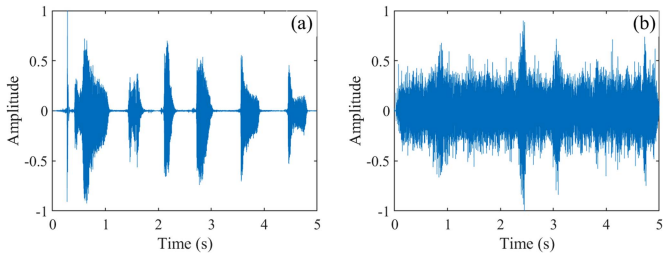


Fig. 6. (a) Recovered speech by mmPhone in a line-of-sight condition without blockage. (b) Recovered speech under the through-wall condition. The speech quality suffers from the downgraded signal-to-noise ratio induced by the wall. (Speech: *zero, one, two, three, four, five*).

to each period of demodulated mmWave signals (IF signals) and derived the phase spectrum $\Phi^i = \{\phi_1^i, \phi_2^i, \ldots, \phi_N^i\}(N = 512)$. Then we derive phase values across M cycles $\{\phi_j^1, \phi_j^2, \ldots, \phi_j^M\}$ for the j_{th} frequency point. According to (9), the speech audio can be decoded from the phase sequence $\{\phi_j^1, \phi_j^2, \ldots, \phi_j^M\}$. For all the derived N sequences, we calculated the correlation value [23] with the audio chirp (recorded by the mic) and chose the one that has the highest value as the decoded audio. Fig. 5(a) and (b) show the spectrograms of the microphone-recorded audio and mmWave-decoded audio, respectively. We can observe that the mmWave-decoded audio shows a high similarity with the microphone-recorded audio. This indicates that the played audio can be decoded successfully from the mmWave signals according to the proposed sound-mmWave transformation model. However, in such a LoS condition, we also find that *the mmWave-decoded audio has a weak response in the high-frequency band (1k-2.5 kHz) which affects the speech intelligibility [24].*

2) Through-Wall Audio Recovery: To investigate the impact of blockage quantitatively and qualitatively, we further performed a line-of-sight experiment without blockage and a through-wall experiment, respectively. The sensing distances are both set to $4\tilde{}m$. In the through-wall condition, we set the mmWave sensor outside a soundproof room and sensed the film through the soundproof glass wall. We chose samples of two speakers from AudioMNIST (100 traces in total) and used a loudspeaker to play the audio. We calculated the average Short-Time Objective Intelligibility (STOI) scores [25] of recovered speech to quantify the speech intelligibility. The result turns out that the average score achieves 0.76 for recovered speech w/o blockage. But the score in the through-wall condition only achieves 0.43. The reason for the degraded intelligibility is that the wall can decrease the SNR of received mmWave signals. Thus, the recovered speech is flooded by noise and has poor intelligibility. The recovered audio traces in the LoS and through-wall conditions are shown in Fig. 6(a) and (b), respectively. We can observe that although the speech audio can be recovered with a satisfying signal-to-noise ratio in the LoS condition, *the SNR of recovered speech degrades significantly in the through-wall condition due to the penetrating loss of the mmWave signals. The speech quality is required to be improved in order to cause practical threats to the speech protected by soundproof walls.*

IV. System Design

To solve the aforementioned challenges and compromise the speech protected by the soundproof room, we introduce an end-to-end eavesdropping system (called *mmPhone*) for through-wall speech recovery.

A. System Overview

The system overview of mmPhone is shown in Fig. 7. The mmWave sensor transmits FMCW signals through the soundproof wall to sense the piezo film. After the demodulation of reflected mmWave signals, the audio is decoded from multiple channels respectively according to the sound-mmWave transformation model (Section III). Then a GAN-based speech enhancement is applied to denoise decoded audio to improve the speech quality. To improve speech intelligibility, the enhanced audio is further fed into a speech reconstruction module to extend the harmonics and synthesize intelligible speech.

B. Clutter Suppression

The mmWave probe transmits mmWave signals periodically and demodulates the reflected mmWaves to generate IF signals. We apply 512-point FFT to each period of the IF signal. For M successive periods, we get an FFT matrix $S = [S_j^i], i = 1, \ldots, M, j = 1, \ldots, 512$ and the phase matrix $\Phi = \{\phi_j^i\}, i = 1, \ldots, M, j = 1, \ldots, 512$ from S. Considering there can be static and other participants in the soundproof room, the reflected mmWave signals from the room is the superposition of these clutters which should be eliminated. Specifically, the moving objects and human motions in the room can cause misalignment among the demodulated range profiles [26]. A consequence of the misalignment is that the range bin containing speech information does not match specific frequency points after the range-FFT. When we decode the speech from specific range bins, there can be loss of speech fragments after the decoding. For example, when there is someone moving around the path between the mmWave sensor and the piezo film, only part of the speech can be decoded from the reflected signals because of the misalignment of the range-bins. Thus, the clutter suppression is required to ensure the integrity of decoded speech. To solve this problem, we develop a correlation-based range-bin alignment solution. The rationale is that the shifted frequency points in the

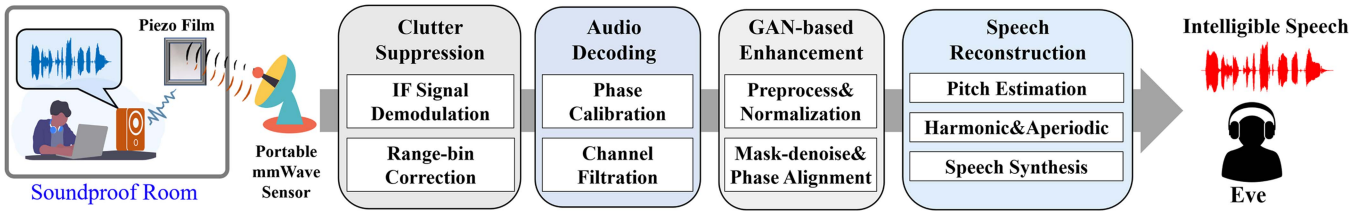


Fig. 7. Overview of the proposed through-wall eavesdropping system *mmPhone*.

misaligned range bins can always comprise a similar spectral envelope between successive bins. That is because the generation of the range bins are far faster than the object movement. Thus, we can think of the object as stationary during multiple chirps are transmitted and demodulated. This can result in a series of range bins that have similar spectral envelopes. So an intuitive solution to align the mismatched range-bins is to find the offset by calculating the cross-correlation between them, and then shift the latter range-bin with the calculated offset. Taking the demodulated range bins as $S = [S_j^i], i = 1, \ldots, M, j = 1, \ldots, N$ where M and N are the number of demodulated chirps and FFT points, respectively. Then the offset can be calculated according to $m_{offset} = \arg\max_m \sum_{n=0}^{N-1} S_n^i \cdot S_{n+m}^{i+1}$. After that, the misaligned range bin S_j^{i+1} can be rectified by $\widetilde{S}_j^{i+1} = S_{j-m}^{i+1}$ where S_{j-m}^{i+1} means shifting S_j^{i+1} by m FFT points in the frequency domain. We conduct this process until the last range bin is corrected. Finally, we derive a corrected S_j^i and the clutters are suppressed.

C. Decoding Audio From mmWave

1) Phase Calibration: After the clutter suppression, the phase sequence $\{\phi_j^1, \phi_j^2, \ldots, \phi_j^M\}$ is extracted from each of the N channels, where $j = 1, \ldots, N (N = 512)$. Here we denote each frequency point after the range-FFT as a channel. Considering that the derived signals are the phases within $[-\pi, \pi]$, the derived signals can suffer from the integer ambiguity problem. Phase unwrapping can be applied to all the derived phase sequences and solve this problem. However, the phase unwrapping is not a perfect solution and there can still be discontinuity points in the phase sequence. These discontinuity points can result in jitter noise in the phase sequence which affects the quality of recovered speech. Here we further perform an outlier detection to correct these discontinuity points and make the derived signals smoother. Specifically, we first apply a window with a size of 1024 to the derived signals and then calculate the triple median absolute deviation. The points that deviate from the calculated value are taken as the outliers. Finally, the outliers are replaced by the mean of the 1024 points within the window. The window slides with no overlaps until reaching the end of the sequence.

2) Channel Filtration: After the phase unwrapping and outlier detection, we acquire the phase sequences from all the channels. However, not all channels convey the speech information according to the sound-mmWave transformation model. To find the channels that contain speech information, decoding audio from all the channels and listening to the decoded audio one by one works but is a time-consuming method. Here we used a more efficient method called *channel filtration* to localize the desired channels automatically. The rationale is that the fundamental voice of human speech lies between 85 Hz and 255 Hz, which has a higher power density than the background noise. Thus, we can apply a high-pass filter on all the sequences, calculate their power density, and choose the Top-k ($k = 3$) ones as the desired traces that contain human speech.

D. GAN-Based Enhancement

1) Preprocessing&normalization: After decoding the audio from the mmWave, we subtract the mean from the decoded audio to correct the DC offset raised by the film and the probe (9). Then we design a high-pass filter with a cut-off frequency of $80Hz$ to eliminate residual noise. Then we apply an amplitude normalization to the decoded audio traces considering that the amplitude of the audio signal is within [-1,1]. The normalization can partly suppress the amplitude fluctuation due to different sensing distances and sound volumes. After the normalization, we acquire multiple noisy audio traces whose quality and intelligibility need to be improved.

2) Mask-Based Denoising: As mentioned before, the low SNR of reflected mmWave signals caused by the penetrating and propagating loss can cause poor-quality speech. To improve the speech quality, our key idea is to boost the SNR of the demodulated signals by suppressing the self-noise of the sensor. Traditional denoising methods, such as Wiener filter and spectral subtraction, can introduce residual noise and musical noise, damaging the quality of recovered speech. To solve this problem, we adopt a mask-based denoising method that can eliminate the broadband noise in the recovered speech. We used a deep neural network to estimate the mask that can characterize the noise pattern of the mmWave sensor. The estimated mask multiplies with the raw recovered speech in the joint time-spectral domain and suppresses the residual noise resulting from the hardware noise. This method benefits from the non-linear structures of deep neural networks that can characterize the complex mapping between noisy and clean speech, and thus can handle non-stationary noise. We adopt a generative adversarial network (GAN) as the neural networks to further improve the generality of the denoising model in the real world.

Specifically, the denoising network is trained with a dataset synthesized by a public dataset [27] and a mmWave-noise dataset. We note that the testing dataset in Section V-B is not included in the training phase in order to show the generalization ability of the denoising neural network. The workflow of the GAN-based speech enhancement is shown in Fig. 8, which

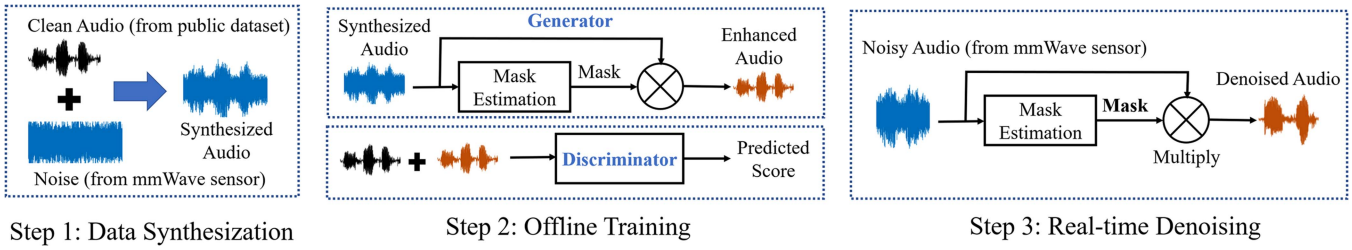


Fig. 8. Scheme of the GAN-based speech enhancement. The training set consists of noisy audio traces which are derived by the data synthesization. The denoising module is a generative adversarial network (GAN) and trained offline with the pre-build training set. After the training, the generator of the GAN can estimate the mask from inputted noisy speech and denoise the speech.

TABLE I
PARAMETERS OF THE GENERATOR

Block	Type	Size
Block[1]	BLSTM	x4, 200
Block[2]	Linear Layer	300
Block[3]	Leaky ReLu	300
Block[4]	Linear Layer	257
Block[5]	Learnable Sigmoid	257

TABLE II
PARAMETERS OF THE DISCRIMINATOR

Block	Type	Size
Block[1]	Conv2D Batch Normalization Leaky ReLU	[5,5], 15 15 15
Block[2]	Conv2D Batch Normalization Leaky ReLU	[11,11], 50 50 50
Block[3]	2D Average Pooling	50
Block[4]	Linear Layer Leaky ReLU Linear Layer Leaky ReLU	50 50 10 10
Block[5]	Linear Layer	10

consists of three steps. First, we set the mmWave sensor in four different scenarios (e.g., a conference room, an open square, a cafÃ©, and a classroom) with no moving objects in the sensor's field of view. We collected mmWave signals and extracted the phase sequence from random channels. After that, we acquired a mmWave noise dataset and superposed the noise dataset and the public audio dataset with a random signal-to-noise ratio (-9 dB∼6 dB). Then we acquired a training set consisting of synthesized noisy speech. In the second step, the synthesized audio traces are fed into the GAN for training. The parameters of the generator are shown in Table I. The discriminator is a Q-network [28] which is used to quantify the quality of estimated speech. The parameters of the discriminator are shown in Table II. The generator and discriminator are updated alternatively. The loss functions of the generator ($\mathcal{L}_G$) and the discriminator ($\mathcal{L}_D$) are shown in (10) and (11), respectively.

During the training process, the noisy speech is first segmented into fragments of the same size and transformed into spectrograms. Then the noisy fragments are fed into the generator to estimate a mask which can multiply with the noisy fragment to output the spectrogram of clean speech. After that, the spectrograms are transformed into time domain according to the Griffin-Lim algorithm. Then the estimated speech and original speech (clean speech) are fed into the discriminator to assess the speech quality. The loss of the generator and the discriminator are calculated according to (10) and (11), where x, y, s are the clean, enhanced speech, and desired score of quality, respectively. $Q(\cdot)$ is the function to calculate the true STOI score [25]. $G(\cdot)$ and $D(\cdot)$ represent the output of the generator and the discriminator. $MSE(\cdot)$ denotes the mean square error. When the training finishes, the generator can be deployed for real-time processing on the raw recovered speech (noisy speech) from the mmWave sensor. After the denoising, we further feed the audio traces decoded from multiple channels into a phase-alignment module to improve the speech quality.

$$\mathcal{L}_G = MSE(D(G(x), y), s) \tag{10}$$

$$\mathcal{L}_D = MSE(D(y, y), Q(y, y)) + MSE(D(x, y), Q(x, y)) + MSE(D(G(x), y), Q(G(x), y)) \tag{11}$$

3) Phase Alignment: After the denoising, we merge the multiple audio traces to further enhance the speech. An intuitive method is to align these traces in the time domain and then add up their amplitudes. However, this method can degrade the final SNR when these sequences are not phase-aligned. To solve this problem, we first choose the one that has the highest SNR as the baseline and apply the phase alignment on these traces. Then we merge them in the time-frequency domain. Specifically, we apply the FFT on the traces with a 512-size window and perform the phase alignment. Then we add up all the spectrums and acquire the corresponding time-domain signals by inverse FFT. Finally, we can acquire an enhanced speech with higher SNR.

E. Speech Reconstruction

Due to the weak response in the high-frequency band (Section III-D), the decoded audio loses formants, resulting in poor intelligibility. Speech synthesis can be a promising solution to recover the harmonics from the distorted audio. Training-based synthesis methods [29], [30] often have a great demand of training data to achieve a satisfying performance. So we turn to a training-free method to reconstruct the harmonics of the distorted speech and improve the intelligibility.

1) Pitch Estimation: Before the speech synthesis, the first step is to estimate the fundamental frequency (i.e., the *pitch*) of the human voice, which has the strongest power in the spectrum.

This basic step is vital for speech reconstruction because the harmonic extension relies heavily on the accuracy of the pitch estimation. To improve the estimating accuracy, our rationale is to take the four receiving antennas as a "microphone" array and estimate the pitch by a weighted result from all four antennas. Compared with existing methods using single audio for pitch estimation, we propose to use recovered traces from the four antennas to improve the estimation accuracy. Based on this idea, we apply the pitch (f_0^i) estimation on the audio trace of Antenna #i, where $i = 1, 2, 3, 4$, respectively, and calculate a calibrated f_0 based on the four estimated pitches. Specifically, we first segment all the audio traces with a size of 50 ms considering the short-time stability of the human voice, and then apply the f_0 estimation on the segments. To acquire the fundamental frequency of each segment coarsely, we first apply a high-pass filter with cut-off frequencies of 80 Hz and 260 Hz according to the pitch band of human voice (85-255 Hz). Then the estimated f_0 is chosen according to the highest magnitude after the FFT. Take SNR_i as the SNR of the segment from Antenna #i, then the calibrated f_0:

$$f_0 = \frac{\sum_{i=1}^{4} SNR_i \cdot f_0^i}{\sum_{i=1}^{4} SNR_i}. \tag{12}$$

2) Harmonic Extension: Formants are local maximums in the spectral envelope of human voice [31] and play a vital role in speech intelligibility. Speech with higher intelligibility can be acquired through a more accurate spectral envelope estimation. We recover the formants by performing the harmonic extension based on estimated spectral envelopes [32]. To improve the accuracy, we use the calibrated f_0 mentioned above for better spectral envelope estimation instead of feeding the estimated pitch of each audio trace.

3) Aperiodic Parameter Extraction: Aperiodicity has usually been used for natural speech synthesis. To improve the quality of synthesized speech, we adopt a group-delay-based method [33] to extract band aperiodicity parameters. The aperiodicity extraction consists of three steps, i.e., temporally static parameter calculation, parameter shaping calculation, and band-aperiodicity estimation. When extracted, the aperiodicity parameters along with the estimated spectral envelope are fed into the synthesis module.

4) Speech Synthesis: The aperiodic components of the speech are further combined with the calibrated pitch f_0 and extended harmonics to synthesize intelligible speech. The synthesis is achieved by the convolution of minimum phase response and excitation signals [32]. After the synthesis, the recovered speech has a bandwidth of up to 2.3 kHz, reaching a satisfying intelligibility for human hearing [34]. The speech synthesis is applied on all the antennas followed by the phase alignment in Section IV-D for further enhancement. Finally, we can acquire enhanced speech with high quality and intelligibility. Fig. 12(a), (b), and (c) show the spectrograms of the played audio, decoded (unprocessed) audio from a single channel, and processed audio by the scheme, respectively.

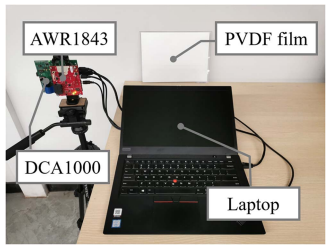


Fig. 9. System setup.

Fig. 10. Tested loudspeakers.

TABLE III
PARAMETER SETTING OF THE MMWAVE SENSOR

Start frequency	77 GHz	Freq. slope	43 MHz/μs
Ramp end time	90 μs	Bandwidth	4 GHz
ADC samples/sec	6000 k	Idle time	5 μs
RX Gain	30 dB	Chirps/frame	255
Frame period	25 ms	Samples/chirp	470

V. EVALUATION

A. System Setup

The system setup is shown in Fig. 9. The commercial mmWave sensor AWR1843Boost has a portable size of 6.5 cm $\times$ 8.5 cm $\times$ 2.0 cm with 12 dBm transmitting power. The frequency of transmitted FMCW signals ranges from 77 GHz to 81 GHz. The chirp rate is set to 10,200 chirps/sec. Detailed parameters are shown in Table III. The IF signal is collected by a DCA1000EVM and sent to a laptop (Thinkpad T490) for processing. The film has a size of 21 cm $\times$ 29.7 cm$\times 28\mu$ m and a piezo constant d_{33} of 3.3×10^{-11}C/N. It is stuck to an acrylic board with glue to avoid physical vibration. The denoising network is trained on a Linux server with a GeForce RTX 2060 GPU and deployed on a laptop for real-time processing. We adopt Adam as the optimizer (initial learning rate=0.001).

B. Dataset and Data Collection

To fully evaluate the proposed system for speech recovery, we used three public datasets widely adopted for speech testing, i.e., Harvard Speech Corpus (HSC) [35], AudioMNIST [36], and Open Speech Repository (OSR) [37]. HSC consists of 720 spoken sentences (*Harvard Sentences*) designed to feature phonemes at the same frequency they appear in spoken English. AudioMNIST contains 30,000 samples of spoken digits from 60 speakers. OSR includes *Harvard Sentences* from different speakers, from which we chose 100 samples. We chose samples collected from two speakers (AudioMNIST) to evaluate the

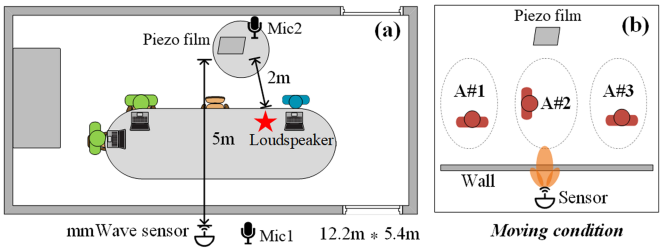


Fig. 11. (a) Experimental scenario (a conference room) with a soundproof glass wall. (b) Experiments with moving persons in the same room and eavesdropping through the soundproof wall.

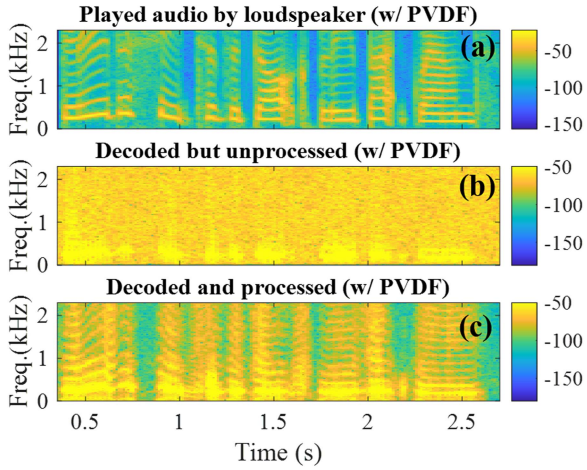


Fig. 12. (a), (b), and (c) show the played audio, raw decoded audio, and processed audio by mmPhone when we deployed the PVDF film (w/ PVDF).

robustness of mmPhone. The results are shown and analyzed in Section VI. All experiments were ensured to follow the institutional review board (IRB) protocol.

We played audio samples of the three datasets via a loudspeaker (Hp) in a soundproof room (Fig. 11). The SPL (measured by a sound level meter nearby the piezo film) was around 67 dB within the range of normal conversations [38]. We asked volunteers to type randomly on their laptops when the loudspeaker played audio. For comparison, we also deployed two microphones to record the played audio, one (Mic1) inside and the other (Mic2) outside the room. We set the mmWave sensor outside the room with a sensing distance of 5 m. The distance between the loudspeaker and the piezo film was 2 m. The soundproof wall consists of two layers of 1cm-thick board made of glass.

To ensure the recovered speech was from the piezo film, we performed a comparative experiment. We respectively placed (i.e., w/ PVDF) and removed (i.e., w/o PVDF) the film for speech recovery with exactly the same processing (e.g., the same decoding channel). The results under the two conditions are shown in Figs. 12 and 13, respectively. Comparing Figs. 12(b) with 13(b), we can find that mmPhone failed to recover the played audio when the film was removed, indicating that the recovered speech by mmPhone was from the piezo film. (Speech: *Glue the sheet to the dark blue background.*)

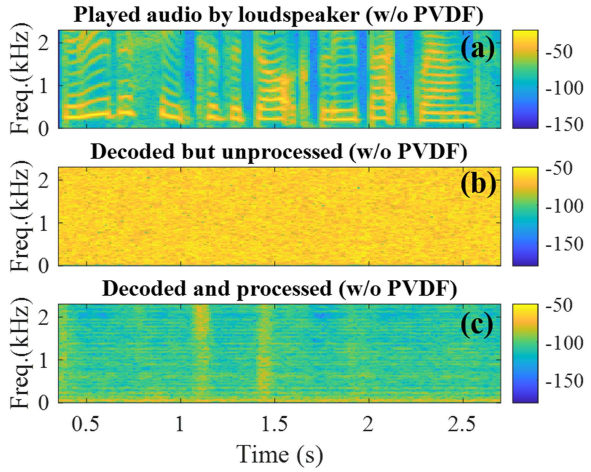


Fig. 13. (a), (b), and (c) show the played audio, raw decoded audio, and processed audio by mmPhone when we removed the PVDF film (w/o PVDF).

C. Metrics

1) Peak-Signal-to-Noise Ratio (PSNR): The PSNR is a commonly used metric to quantify speech quality [7], [8]. The empirical boundary of PSNR for human-audible speech is 0dB [7]. The PSNR is the ratio of peak signal power and mean square error (calculated from the original speech and degraded speech). To calculate the PSNR score, we take the original audio played by the sound source as the ground truth and the recovered audio as degraded audio. A higher PSNR score indicates better speech quality.

2) Short-Time Objective Intelligibility (STOI): The STOI score has a monotonic relationship with the subjective speech-intelligibility [25]. It varies within [0,1], of which a higher value indicates better speech intelligibility. According to Taal's work [25] on speech recognition, over 70% of words (Harvard Sentences) can be recognized correctly by human beings when the STOI score is larger than 0.6, which indicates a satisfying intelligibility. To calculate the STOI score, we take the original audio played by the sound source as the ground truth and the recovered audio as degraded audio.

D. Overall Performance

In this part, we evaluate the system and quantify the performance with PSNR and STOI scores of the recovered speech by mmPhone. Given that the digits (i.e., 0∼9) are often related to secret information, such as social passwords and security numbers, we also evaluate the system by applying both automatic speech recognition (ASR) and manual speech recognition (MSR) on the recovered digit speech. To better understand the performance of the proposed GAN-based scheme (i.e., GAN), we also applied the method (i.e., DNN) used in previous work [15] and calculated scores for comparison.

1) Sound Recovery: The PSNR scores of recovered audio signals by mmPhone and the microphones are shown in Fig. 14, and the STOI scores are shown in Fig. 15. From Fig. 14, we can observe that mmPhone outperforms the out-room microphone (Mic2) with a high PSNR above 20 dB on the three datasets. This

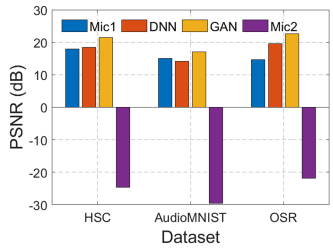


Fig. 14. Overall PSNR scores. DNN is the method proposed in previous work [15] while GAN is the one proposed in this work.

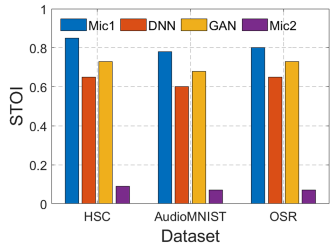


Fig. 15. Overall STOI scores. DNN is the method proposed in previous work [15] while GAN is the one proposed in this work.

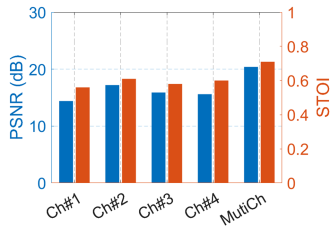


Fig. 16. Performance of single/ multiple channels.

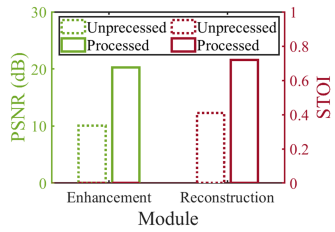


Fig. 17. Performance of each module.

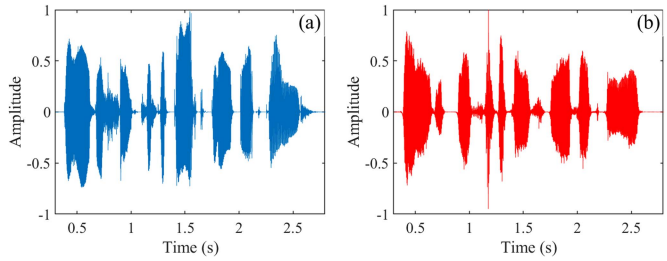


Fig. 18. (a) Original audio and (b) Recovered audio by mmPhone (Speech: *Glue the sheet to the dark blue background*).

is not surprising considering that the sound wave is constrained by the soundproof obstacles and thus, the out-room microphone hardly captured the speech but background noise.

We find that the recovered speech by mmPhone can achieve similar PSNR as the in-room microphone's (Mic1) on the datasets. The reasons are two folds. First, the mask-based denoising method can significantly improve the SNR of recovered speech. Second, mmPhone also enhances the recovered speech leveraging the four receiving antennas/channels, each of which acts as a separate microphone. The four receiving channels of mmPhone act like a "microphone array". To investigate the performance of the multi-channel technique, we performed a comparison experiment by re-running the speech recovery for every single channel without the multi-channel processing as introduced in Sections IV-D3 and IV-E1. The result is shown in Fig. 16. We can observe that performance varies across these channels (i,e., *Ch#1*, *Ch#2*, *Ch#3*, and *Ch#4*). The result of *MultiCh* indicates that by leveraging the multi-channel processing, we can acquire improved speech quality and intelligibility.

Fig. 15 indicates that the STOI of speech recovered by mmPhone is lower than the in-room acoustic microphone's (Mic1) but far larger than the out-room microphone's (Mic2). The lower STOI of mmPhone compared with the in-room microphone results from the limited sampling rate of mmPhone, which causes loss of high frequency (above 5.1 kHz according to Nyquist theorem) of human voice. As analyzed in Section III-C, the sampling rate of mmPhone is 10.2 kHz, smaller than the acoustic microphone's (48 kHz). So harmonics of speech above 5.1 kHz suffer from the severe undersampling and are unrecoverable for mmPhone but can be recovered by the acoustic microphone. Overall, the mean STOI score (0.73) of recovered speech by mmPhone is far larger than the out-room mic's, indicating that mmPhone can recover intelligible speech even though the original speech is protected by soundproof obstacles. Compared with the previous method (i.e., DNN) [15], the average PSNR and average STOI score of the proposed GAN-based scheme respectively increase by about 2.9 dB and 0.1, which indicates higher speech quality and intelligibility.

The original and recovered speech in the time domain are shown in Fig. 18(a) and (b), respectively. We can observe that the recovered speech shows a high similarity with the original speech and the noise is suppressed. The *Enhancement* (Section IV-D) and *Reconstruction* (Section IV-E) modules aim to improve speech quality and intelligibility. To quantify the improvement, we calculated the metrics of the input (i.e., Unprocessed) and output (i.e., Processed) audio to investigate the module gain. Fig. 17 shows the performance gain of the system modules. We find that the PSNR score increases by 12.7 dB after the *Enhancement*. The STOI score increases by 57.4% after the *Reconstruction*. The results validate the effectiveness of the proposed scheme in Section IV. To put it intuitively, from Fig. 12(b), we can observe that the raw recovered speech loses formants in the high-frequency band (1 k∼2.5 kHz) which causes poor intelligibility for human hearing. After the speech enhancement and reconstruction of mmPhone, we find that the formants above 1 kHz are recovered in the processed speech as shown in Fig. 12(c). Overall, the proposed system is effective in compensating the distorted speech and improving the quality and intelligibility significantly which can cause threats to speech protected by soundproof walls.

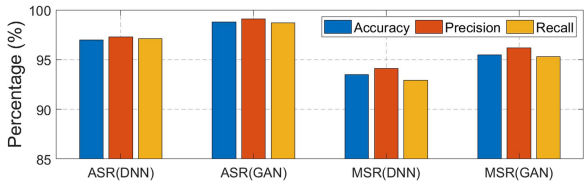


Fig. 19. Digit-recognition results. ASR(DNN) and MSR(DNN) indicate the performance of the method in previous work [15]. ASR(GAN) and MSR(GAN) indicate the performance of the method in this work.

2) Digit Recognition: We define a true positive as a correctly predicted digit and a false negative as a digit wrongly classified into other classes. *For the ASR*, we trained a digit recognition model based on ResNet-50. Specifically, we got the spectrograms of the recovered speech from the AudioMNIST (30,000 traces in total) corresponding to each digit by applying the STFT. We randomly separated the spectrograms ($3 \times 224 \times 224$) into 80 % training data and 20% testing data for model training and testing, respectively. *For the MSR*, we randomly chose 10 recovered audio traces from each digit class and invited 15 volunteers to listen to the 100 audio traces. We asked the volunteers to pick up the most likely one from the ten digits (0∼9). The recognition results are shown in Fig. 19. For the previous method [15] (i.e., DNN), the accuracy, precision, and recall scores of ASR are 97.0%, 97.3%, 97.1%, and MSR achieves 93.5%, 94.1%, and 92.9%, respectively. For the proposed GAN-based scheme, the scores of ASR are 98.8%, 99.1%, and 98.7%, and MSR achieves 95.5%, 96.2%, and 95.3%, respectively. We find that the ASR of GAN can achieve a higher accuracy for digit recognition which indicates the attack can cause greater threats to uttered digits by the victim. For both DNN-based and GAN-based methods, the scores of ASR are slightly lower than the MSR's. The possible reason is that some words have a similar pattern in the low bands, which are difficult for human beings to differentiate. Besides, the accents can also pose challenges to the MSR. But overall, the GAN-based scheme can achieve over 95% accuracy for both automatic speech recognition and manual speech recognition.

VI. Complex Scenario

In this part, we quantitatively evaluate the system performance under complex environments. The evaluated speech are chosen from AudioMNIST (two speakers, i.e., a male and a female). To show the effectiveness of the proposed processing scheme in Section IV, we compare the results of the raw retrieved speech and enhanced speech by mmPhone.

A. Sensing Distance

The mmWave signal can decay with increasing distance, causing a downgrading SNR. Thus, we performed experiments to evaluate mmPhone under different sensing (probe-film) distances from 2 m to 7 m in the through-wall scenario (Fig. 11). Other settings are the same as in Section V-B. We compare the results without and with our proposed processing scheme (Section IV) to show the performance gain. As shown in Fig. 20, the PSNR of raw decoded speech (Unprocessed, green dashed line) is steady above 10 dB when the distance is within 5 m

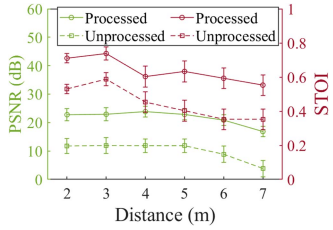


Fig. 20. Impact of sensing distance.

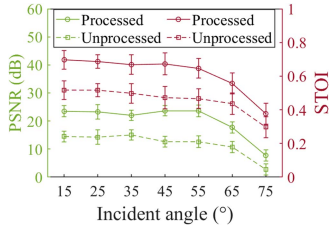


Fig. 21. Soundwave incident angle.

but reduces to 3.9 dB when the distance is 7 m. This results from the declining power density of mmWave when the distance increases. To cope with this problem, we propose the scheme in Section IV to suppress the noise and merge decoded audio traces from multiple channels and antennas for speech enhancement. With our proposed enhancement scheme, the PSNR has a gain of 12 dB∼16 dB. We also observe that the STOI reduces to 0.41 when the distance is 5 m. The reason is that mmWaves suffer larger attenuation as the distance increases, resulting in a low SNR at the receiver. Thus, the high-frequency components of decoded audio can be partially flooded by noise, which reduces speech intelligibility. However, With the GAN-based enhancement scheme, the STOI of reconstructed audio increases to over 0.6 when the sensing distance is 5 m.

B. Incident Angle of Propagating Sound Waves

The pressure amplitude (P_i in (4)) applied to the film can be different with respect to different incident angles of sound waves. This can further influence the reflection coefficient Γ of the film according to (8). Here we define the incident angle of sound waves as the angle between the soundwave-propagating direction and the normal of the film surface. We deployed the mmWave sensor outside the soundproof room for through-wall eavesdropping and changed the incident angle with other settings the same as in V-B. The results are shown in Fig. 21. We can observe that mmPhone is resilient to the incident angles of sound waves within 55°, but the performance declines significantly when the incident angle is above 65°. The recovered speech is still audible (7.1 dB) but with poor intelligibility. We also find that the gain of intelligibility (i.e., the difference between red lines) goes down as the angle increases. Considering that the spectral envelope (Section IV-E) plays a vital role in speech intelligibility, the possible reason is that the low PSNR score at the large incident angle makes high-frequency components of decoded speech ambiguous in the spectrum. Thus, the estimated spectral envelope can be partly distorted, resulting in the limited performance gain under a large incident angle.

TABLE IV
IMPACT OF LOUDSPEAKER TYPE

Loudspeaker Type	PSNR Score	STOI Score
Hp	25.5 dB	0.76
Xiaomi	25.2 dB	0.68
Philips	27.5 dB	0.74
Sanag	25.1 dB	0.70
Hivi	28.4 dB	0.77

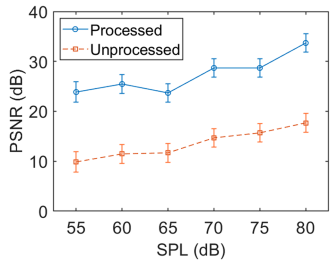


Fig. 22. PSNR under different SPLs.

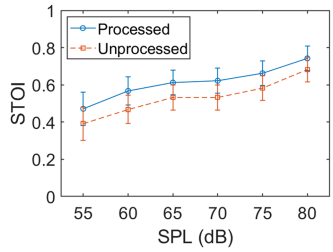


Fig. 23. STOI under different SPLs.

TABLE V
IMPACT OF MOTION INTERFERENCE

Area	PSNR / STOI	
	w/o clutter suppression	w/ clutter suppression
A#1	8.3 dB / 0.43	21.8 dB / 0.61
A#2	5.6 dB / 0.42	10.2 dB / 0.52
A#3	7.1 dB / 0.46	24.4 dB / 0.65

C. Loudspeaker Type

The structure has a vital impact on the frequency response of the loudspeaker [39], which means loudspeakers with different structures can have different frequency response curves (especially the high-frequency band of 500 Hz∼2 kHz closely related to speech intelligibility). Different physical structures of loudspeakers may affect the generated sound fields formulated by (4). To investigate the impact of loudspeaker structures, we chose five different types of commodity loudspeakers shown in Table IV. We played speech samples in the AudioMNIST dataset with the same experimental setting as in Section V-B. The PSNR and STOI of recovered speech are shown in Table IV. We find that the PSNR score of recovered speech (blue bars) varies from 25.1 dB to 28.4 dB with a fluctuation of 3.3 dB, and the STOI score (red bars) varies from 0.68 to 0.77. We observe that Hivi M200 has higher STOI than other loudspeakers because the speaker (costs more than $430) has a more flat frequency-response curve than others [40]. The results indicate that mmPhone can recover intelligible speech with a slight intelligibility fluctuation among different loudspeakers. Overall, mmPhone is robust to recover intelligible speech played by different loudspeakers.

D. Sound Pressure Level

To demonstrate the ability of mmPhone to recover human speech, we quantitatively evaluated the speech retrieval performance in a controlled environment. We adjusted the sound volume and placed a sound level meter beside the piezo film to measure the sound pressure level (SPL). The sensing distance is set to 5 m and the incident angle of sound waves is about 15°. As shown in Figs. 22 and 23, we can observe that the PSNR and STOI scores increase with the sound pressure level. According to (9), a higher SPL of propagated sound waves can cause a larger phase change of IF signals from which we decode the audio signal. This results in the decoded audio signal with a higher SNR, which benefits the quality and intelligibility of the recovered speech. Considering that the sound pressure level of normal human speech varies from 60 dB∼ 70 dB [38], [41], mmPhone works well to recover high-quality and intelligible human speech in a normal conversation.

E. Motion Interference

Considering that there can be moving people around the room, the recovered speech may be interfered by the induced clutters. Specifically, we find that the movement of humans near the direct path of transmitted mmWave signals can cause interference. Accordingly, we proposed the clutter suppression in Section IV-B to solve this problem. To validate the effectiveness of the clutter suppression, we rearranged the layouts and asked three volunteers to move randomly in three areas (i.e., *A#1, A#2, A#3*) of the room, respectively, as shown in Fig. 11(b). We set the sensing distance to 5 m and kept other settings the same as in Section V-B. The results are shown in Table V. We find that the recovered speech (without clutter suppression) suffer from the dynamic noise when volunteers were moving in the three areas. The PSNR and STOI scores are below 10 dB and 0.5, which indicates poor quality and intelligibility of recovered speech. After the clutter suppression, we can observe that the PSNR and STOI scores are improved to acceptable ranges for *A#1* and *A#3*. However, the improvement in *A#2* is smaller than in the other two conditions. The reason is that the volunteer's body sometimes blocked the transmitted mmWave signals during the movement and thus, the SNR of demodulated signals deteriorated significantly. Overall, except for the blocking situation, the clutter suppression design is effective in suppressing the impact of human movements.

F. Soundproof Wall

To investigate the impact of different walls, we used different materials to block the mmWave sensor, i.e., double-layer glass (2 × 1.0 cm thick), single-layer glass (1.0 cm thick), sound-absorbing sponges (3.1 cm thick), and wood (1.0 cm thick). The speech samples were chosen from AudioMNIST (100 male

TABLE VI
IMPACT OF WALL THICKNESS AND MATERIAL

Material	Thickness	PSNR Score	STOI Score
Double-layer Glass	2.0 cm	22.1 dB	0.62
single-layer Glass	1.0 cm	25.9 dB	0.68
Sponge	3.1 cm	26.3 dB	0.71
Wood	1.0 cm	22.5 dB	0.64

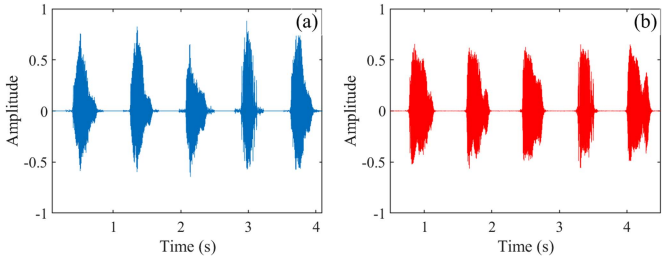


Fig. 24. (a) Recorded audio by a microphone and (b) Recovered speech by mmPhone (Speech: *zero* repeated five times).

utterances and 100 female utterances). We set the sensing distance to 5 m. The PSNR and STOI scores of recovered speech are shown in Table VI. We find that the performance varies across different materials and thicknesses. The performance degrades when penetrating the double-layer glass wall due to the larger attenuation of mmWave signals. But overall, the recovered speech achieves the PSNR score of over 22 dB and the STOI score of over 0.61, which means the attack is resilient to common soundproof materials.

G. Human Speech

To investigate the feasibility of human speech recovery, we asked a male volunteer to sit on a chair in the same soundproof room as shown in Fig. 11(a). The volunteer was asked to speak each digit (from *zero* to *nine*) ten times towards the piezo film from a distance of around 1.8 m. We used a microphone beside him to record the ground-truth audio and also used the mmWave sensor to perform the through-wall attack with a sensing distance of 5 m. We used the proposed scheme to recover the speech from received mmWave signals. We took the microphone-recorded audio as the ground truth and calculated the PSNR and STOI scores of recovered speech. The average scores are 23.7 dB and 0.69, respectively. Samples of the ground-truth data (microphone) and recovered audio by mmPhone are shown in Fig. 24.

VII. COUNTERMEASURES

A. Shielding

Considering that the key rationale of mmPhone is based on the sound-mmWave transformation, the shielding-based defense can be two folds. First, a potential countermeasure is to wear a headset or earphones to prevent the soundwaves from propagating through the air rather than playing the audio on loudspeakers. But it is common that people tend to use loudspeakers to play the audio in a multi-participant conference in which case the attack can still cause practical threats to the speech. Besides, recent work [42] reveals that the earphones are not secure due to the speech leakage via the magnetic side channel. Second, blocking the transmitted mmWave signals can defend against mmPhone, such as deploying electromagnetic shielding materials around the room. However, it costs a lot to build a soundproofing conference room also with electromagnetic shielding protections for a large zone in real-world scenarios.

B. Jamming

Considering the increasing deployment of mmWave sensors, the user may use a mmWave jammer to defend against the attack. An effective jamming can confuse and even fail the adversarial device. But we find that a practical and cost-effective jamming is not easy to achieve. To jam the adversarial device, the user needs to know the parameters (e.g., the operating frequency, period, and chirp rate) of the malicious device in order to interfere with the demodulated mmWave signals by it. Here we call the jamming works when transmitted jamming signals are successfully demodulated by the malicious sensor. The success rate of the jamming $P_{parallel-interf}$ can be calculated according to the interference analysis of FMCW radars [43]:

$$P_{parallel-interf} = 1 - \left(1 - \frac{t_d}{t_r}\right)^{N_p - 1}, \tag{13}$$

where t_d is the chirp delay, t_r is the chirp period, and N_P is the number of mmWave sensors ($N_p = 2$) in the same scene. In this work, we have $t_d = 5\mu$s, $t_r = 95\mu$s. Considering one jamming device (i.e., $N_p = 2$), the success rate of jamming is 5.26% which is far from practical in real life.

Based on the above analysis, a wise strategy is to change the frequency slope and idle time of transmitted mmWave signals, which can cause glitches in the malicious mmWave sensor (also known as *crossing interference* in mmWave sensing). But this strategy cannot steadily jam the malicious device because the glitch duration is typically small (often at the μs level). Thus, it is potentially more effective to change the slope of transmitted chirps randomly with short idle time. Specifically, the glitch duration τ_{glitch} depends on the difference between the slope of the jamming signals and the slope of malicious signals:

$$\tau_{glitch} = \frac{BW}{|S_{malicious} - S_{benign}|}, \tag{14}$$

where BW is the bandwidth of the IF signals, $S_{malicious}$ and S_{benign} are the slopes of the malicious and benign mmWave signals, respectively. The used mmWave sensor in this article has an IF bandwidth of 10 MHz and we assume the slope difference is 1 MHz/μs (close to the slope setting of the malicious device, a beneficial condition to the user), then the duration of the interfered signal is only 10 μs which has little impact on the intelligibility of recovered speech considering the short-term invariance of human speech (30 ms$\sim$50 ms).

VIII. Related Work

A. Vibrometry-Based Speech Eavesdropping

Non-acoustic sensors, such as motion sensors [3], [4], [5], [6], [44], [45], wireless signals [7], [12], [46], lidars [9], [47], high-speed cameras [8], vibration motors [10], wireless sensors [48], [49] and hard drives [11] can capture objects' vibration for sound recover. When targeting surrounding objects, the sound-induced vibration on objects in normal conversations (60 dB ∼70 dB) can be extremely delicate, requiring μm-level resolution for accurate vibration measurement, such as a laser vibrometer. However, existing work [9] reveals that the rigidness and transparency of vibrating objects can significantly affect the performance of laser-based methods. Vibration-damping and sound-absorbing materials, such as sponges and glossy plywood, can dampen the sound-induced vibration on objects and thus degrade the performance of laser-based methods and make them prone to failure. Due to mmWave's penetrating ability, mmPhone can penetrate common soundproof and sound-absorbing materials and recover the speech. mmPhone eavesdrops on the propagating sound waves directly via the sound-mmWave transformation and thus can be less affected by the diaphragm material. Wei et al. [7] developed a radio-based vibrometry using the multipath effect in the WiFi band to measure the loudspeaker's vibration. Our used attack device transmits fast chirps and operates in the band of 77–81 GHz which has a higher resolution and can be less influenced by the uncertain wireless traffic in real scenarios. Davis et al. used a high-speed video camera to capture the delicate changes in images and recovered speech. Their method achieved a mean STOI score of 0.68. Nassi et al. [50] used a light sensor to capture the sound-induced vibration on a lamp in a line-of-sight condition. The mean STOI score of recovered speech using their method was 0.69. Compared with prior work, the overall performance of our work achieves a mean STOI score of 0.73 as analyzed in Section V-D, which indicates better intelligibility.

B. mmWave-Based Speech Recovery

mmWave draws more and more attention in both security areas and noise-resistant speech applications [26], [51], [52], [53], [54], [55]. Xu et al. [54] developed a noise-resilient speech recovery scheme based on mmWave sensing. The rationale of their work is to use the captured vocal vibration signals to recover human speech. Liu et al. [55] proposed a multi-modal speech recognition system by fusing mmWave and audio signals, which achieved a high accuracy and robustness in the real world. Li et al. [26] developed a noise-resilient user authentication system by interrogating users' vocal vibration with mmWave signals. These works leverage the vocal vibration to extract information related to speech contents and the speaker's identity. Targeting human vocal vibration for eavesdropping needs to focus the mmWave beam on specific throat area of the speaker and thus face practical challenges, such as the target's unknown orientation and random movement. mmPhone recover the propagating soundwaves without the requirement of focusing on the small throat area. Recently, Basak et al. [56] and Wang et al. [57] found that mmWave sensors can be used to eavesdrop on phone calls by capturing the delicate vibrations on the smartphone's body. The rationale of their work is based on the delicate vibration on the smartphone surface resulting from the smartphone earpiece. mmPhone focuses on a new acoustic side channel, leveraging the mmWave-characterized piezoelectric effect for eavesdropping. What's more, another difference between mmPhone and these works is that the above works focus on line-of-sight sensing without blockage in between while we face the challenge of the penetrating loss in the non-line-of-sight (through the wall) condition.

C. mmWave-Based Attack

Except for speech recovery, Li et al. [52] revealed a new side-channel attack to compromise the screen contents protected in an isolated zone. They leveraged a customized mmWave probe to infer the screen contents of victims behind a wall. Attacks targeting mmWave sensors on vehicles have also been explored by researchers, such as jamming and spoofing attacks. Sun et al. [53] designed and implemented practical physical layer attacks and defense strategies based on a mmWave testbed. They revealed that adversaries can use mmWave sensors to jam and spoof the equipped mmWave sensors on automotive vehicles. Our work mainly focuses on the acoustic side channel, which leverages the mmWave to characterize the piezoelectric effect of piezoelectric materials for speech eavesdropping.

IX. Discussion

A. Blocking Condition

Penetrating loss affects the SNR of propagating mmWave signals and thus downgrades the performance of mmPhone. Although mmPhone can penetrate sound-isolation obstacles (e.g., soundproofing glasses, sound-absorbing sponges, and wood), objects with larger penetrating loss of mmWave signals may totally block the mmWave signals and disable mmPhone. For example, the human body between the mmWave sensor and the piezo film can block the transmitted mmWave signals towards the piezo film. In such a blocking condition, mmPhone behaves badly due to the poor SNR of demodulated signals and hardly recovers intelligible speech.

B. Ambient Noise Cancellation

In Section IV, we develop the GAN-based speech enhancement to suppress the sensor's noise and improve the quality of recovered speech. For more complex scenarios in the real world, there can be acoustic noise in the background. Thus, the quality of recovered speech may be affected. Considering that there have been acoustic denoising methods, such as spectral subtraction and Wiener filtering, the system design of mmPhone focuses on eliminating the electromagnetic noise along with the mmWave signals. The performance can be further improved when combining the current solution in this article with traditional methods of acoustic denoising.

X. Conclusion

In this article, we present a novel attack to compromise the speech protected by the soundproof environment. We established a sound-mmWave transformation scheme that can decode the speech from mmWave signals reflected by a piezo film. To reveal the threats of the attack, we proposed an end-to-end attack system that can improve the speech quality and intelligibility significantly. We conducted extensive experiments to evaluate the proposed attack system. The results on public datasets indicate that mmPhone can infer digit contents with over 98% recognition accuracy.

References

[1] Parks Associates, "Voice and video calls more than trapled during COVID-19 pandemic," 2020. Accessed: Jul. 17, 2021. [Online]. Available: https://www.parksassociates.com/blog/article/pr-08262020

[2] Businesswire, "Video calls fast becoming as popular as voice calls, reaching almost universal adoption for social use, according to vonage study," 2029. Accessed: Jul. 17, 2021. [Online]. Available: https://www.businesswire.com/news/home/

[3] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *Proc. 23rd {USENIX} Secur. Symp.*, 2014, pp. 1053–1067.

[4] Z. Ba et al., "Learning-based practical smartphone eavesdropping with built-in accelerometer," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2020, pp. 23–26.

[5] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in *Proc. IEEE Symp. Secur. Privacy*, 2018, pp. 1000–1017.

[6] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, "Spearphone: A lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers," in *Proc. 14th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, 2021, pp. 288–299.

[7] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 130–141.

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

[9] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: Eavesdropping via lidar sensors," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, 2020, pp. 354–367.

[10] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *Proc. 14th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2016, pp. 57–69.

[11] A. Kwong, W. Xu, and K. Fu, "Hard drive of hearing: Disks that eavesdrop with a synthesized microphone," in *Proc. IEEE Symp. Secur. Privacy*, 2019, pp. 905–919.

[12] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M.B. Srivastava, "UWHear: Through-wall extraction and separation of audio vibrations using wireless signals," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, 2020, pp. 1–14.

[13] C. Cochard, T. Spielmann, and T. Granzow, "Dielectric tunability of ferroelectric barium titanate at millimeter-wave frequencies," *Phys. Rev. B*, vol. 100, no. 18, 2019, Art. no. 184104.

[14] G. Srinivasan, A. Tatarenko, V. Mathe, and M. Bichurin, "Microwave and mmWave magnetoelectric interactions in ferrite-ferroelectric bilayers," *Eur. Phys. J. B*, vol. 71, no. 3, pp. 371–375, 2009.

[15] C. Wang et al., "mmPhone: Acoustic eavesdropping on loudspeakers via mmWave-characterized piezoelectric effect," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 820–829.

[16] Q. Zhang, V. Bharti, and G. Kavarnos, "Poly (vinylidene fluoride)(PVDF) and its copolymers," *Encyclopedia Smart Mater.*, p. 234, 2002, ch. 44.

[17] A. V. Shirinov and W. K. Schomburg, "Pressure sensor from a PVDF film," *Sensors Actuators A: Phys.*, vol. 142, no. 1, pp. 48–55, 2008.

[18] V. S. Bystrov, E. V. Paramonova, I. K. Bdikin, A. V. Bystrova, R. C. Pullar, and A. L. Kholkin, "Molecular modeling of the piezoelectric effect in the ferroelectric polymer poly (vinylidene fluoride)(PVDF)," *J. Mol. Model.*, vol. 19, no. 9, pp. 3591–3602, 2013.

[19] C. Hadnagy, *Social Engineering: The Art of Human Hacking*. New York, NY, USA: Wiley, 2010.

[20] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Berlin, Germany: Springer, 2012.

[21] D. K. Cheng, *Field and Wave Electromagnetics*. Noida, Uttar Pradesh, India: Pearson Education India, 1989.

[22] S. Rao, "Introduction to mmWave sensing: FMCW radars," *Texas Instrum. (TI) mmWave Training Ser.*, 2017. [Online]. Available: https://www.ti.com.cn/content/dam/videos/external-videos/2/3816841626001/5415203482001.mp4/subassets/mmwaveSensing-FMCW-offlineviewing0.pdf

[23] I. Cohen et al., "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009, pp. 1–4.

[24] O. Lapteva, *Speaker Perception and Recognition. An Integrative Framework for Computational Speech Processing*. Kassel, Germany: Kassel Univ. Press, 2011.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[26] H. Li et al., "VocalPrint: Exploring a resilient and secure voice authentication via mmWave biometric interrogation," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, 2020, pp. 312–325.

[27] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Held Jointly Conf. Asian Spoken Lang. Res. Eval.*, 2013, pp. 1–4.

[28] S.-W. Fu et al., "MetricGAN+: An improved version of metricgan for speech enhancement," 2021, *arXiv:2104.03538*.

[29] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5429–5433.

[30] P. Bachhav, M. Todisco, and N. Nicholas, "Exploiting explicit memory inclusion for artificial bandwidth extension," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5459–5463.

[31] "Formant," 2021. Accessed: Jul. 14, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Formant&oldid=1031036735

[32] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[33] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.

[34] P. Pearson, "Sound sampling," 1993. [Online]. Available: https://www.hitl.washington.edu/projects/knowledgebase/virtual-worlds/EVE/I.B.3.a.SoundSampling.html

[35] P. Demonte, "Harvard speech corpus–audio recording 2019," University of Salford Collection, 2019.

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

[37] V. Troubleshooter, "The open speech repository," 2010. [Online]. Available: http://www.voiptroubleshooter.com/open_speech/american.html

[38] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *Proc. 13th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2015, pp. 301–315.

[39] W. M. Leach Jr., "Loudspeaker voice-coil inductance losses: Circuit models, parameter estimation, and effect on frequency response," *J. Audio Eng. Soc.*, vol. 50, no. 6, pp. 442–450, Jun. 2002.

[40] Swans, "HiVi M200MKIII," Accessed: Jul. 10, 2021. [Online]. Available: https://swanspeakers.com/product/m200mkiii/

[41] Engineering ToolBox, "Sound pressure," 2004, Accessed: Jul. 05, 2021. [Online]. Available: https://www.engineeringtoolbox.com/sound-pressure-d_711.html

[42] Q. Liao, Y. Huang, Y. Huang, Y. Zhong, H. Jin, and K. Wu, "MagEar: Eavesdropping via audio recovery using magnetic side channel," in *Proc. 20th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2022, pp. 371–383.

[43] S. Rao and A. V. Mani, "Interference characterization in FMCW radars," in *Proc. IEEE Radar Conf.*, 2020, pp. 1–6.

[44] J. Han, A. J. Chung, and P. Tague, "Pitchln: Eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in *Proc. 16th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2017, pp. 181–192.

[45] P. Hu et al., "AccEar: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 1757–1773.

[46] W. McGrath, "Technique and device for through-the-wall audio surveillance," US Patent App. 11/095,122, 2005.

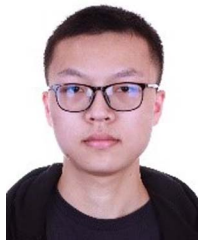
[47] R. P. Muscatell, "Laser microphone," *J. Acoustical Soc. Amer.*, vol. 76, no. 4, pp. 1284–1284, 1984.

[48] P. Hu, W. Li, R. Spolaor, and X. Cheng, "mmEcho: A mmWave-based acoustic eavesdropping method," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 836–852.
[49] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 11–20.
[50] B. Nassi, Y. Pirutin, R. Swisa, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Passive sound recovery from a desk lamp's light bulb vibrations," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 4401–4417.
[51] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 315–328.
[52] Z. Li et al., "WaveSpy: Remote and through-wall screen attack via mmWave sensing," in *Proc. IEEE Symp. Secur. Privacy*, 2020, pp. 217–232.
[53] Z. Sun, S. Balakrishnan, L. Su, A. Bhuyan, P. Wang, and C. Qiao, "Who is in control? practical physical layer attack and defense for mmWave-based sensing in autonomous vehicles," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 3199–3214, 2021.
[54] C. Xu et al., "WaveEar: Exploring a mmWave-based noise-resistant speech sensing for voice-user interface," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2019, pp. 14–26.
[55] T. Liu et al., "Wavoice: A noise-resistant multi-modal speech recognition system fusing mmWave and audio signals," in *Proc. 19th ACM Conf. Embedded Netw. Sensor Syst.*, 2021, pp. 97–110.
[56] S. Basak and M. Gowda, "mmSpy: Spying phone calls using mmWave radars," in *Proc. IEEE Symp. Secur. Privacy*. Los Alamitos, CA, USA, 2022, pp. 995–1012.
[57] C. Wang et al., "mmEve: Eavesdropping on smartphone's earpiece via cots mmwave device," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 338–351.

Feng Lin (Senior Member, IEEE) received the PhD degree from the Department of Electrical and Computer Engineering, Tennessee Technological University, USA, in 2015. He is currently a professor with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. He was an assistant professor with the University of Colorado Denver, USA, a research scientist with the State University of New York (SUNY), Buffalo, USA, and an engineer with Alcatel-Lucent (currently, Nokia). His current research interests include mobile sensing, wireless sensing, Internet of Things security, biometrics, and AI security. He was a recipient of the ACM SIGSAC China Rising Star Award, the Best Paper Awards from ACM MobiSys'20, IEEE Globecom'19, IEEE BHI'17, the Best Demo Award from ACM HotMobile'18, and the Best Paper Award Nomination from SenSys'21 and INFOCOM'21. He serves as an editor for IEEE Network.

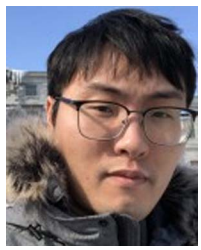
Chao Wang (Student Member, IEEE) received the bachelor degree of information engineering from Zhejiang University, in 2019. He is currently working toward the PhD degree with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. His research interests are wireless sensing and IoT security.

Tiantian Liu (Student Member, IEEE) received the BS degree in information engineering from Zhejiang University, in 2020. She is currently working toward the PhD degree with the school of Cyber Science and Technology, Zhejiang University. Her research interests include wireless sensing and cyber-physical security.



Ziwei Liu (Student Member, IEEE) received the BS degree in information security from Zhejiang University, in 2020. He is currently working toward the PhD degree with the school of Cyber Science and Technology, Zhejiang University. His research interests lie in IoT security and cyber-physical security.

Yijie Shen (Student Member, IEEE) received the bachelor degree in computer science and technology from HeiFei University of Technology, in 2015. He is currently working toward the PhD degree with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. His research interests are biometrics, mobile security, Internet of Things, machine learning and deep learning.



Zhongjie Ba (Member, IEEE) received the PhD degree in computer science and engineering from the State University of New York at Buffalo, USA, in 2019. He is currently a professor with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. He was a postdoctoral researcher in the School of Computer Science with McGill University, Canada. His current research interests include the security and privacy aspects of Internet of Things, forensic analysis of multimedia content, and privacy-enhancing technologies in the context of collaborative deep learning. Results have been published in peer reviewed top conferences and journals, including CCS, NDSS, INFOCOM, ICDCS, and *IEEE Trans. Inf. Forensics Security*. Currently, He serves as an associate editor of IEEE Internet of Things Journal and the technical program committee of several conferences in the field of Internet of Things and wireless communication.

Li Lu (Member, IEEE) received the BE and PhD degrees in computer science and technology from Shanghai Jiao Tong University and Xi'an Jiaotong University, respectively. He is a tenure-track research professor in the School of Cyber Science and Technology and College of Computer Science and Technology at Zhejiang University. He was also a visiting research student in Wireless Information Network Laboratory (WINLAB) and Department of Electrical and Computer Engineering with Rutgers University. His research interests include IoT security, voice security, mobile sensing, and ubiquitous computing. He is the recipient of ACM China SIGAPP Chapter Doctoral Dissertation Award, and First Runner-up Poster Award from ACM MobiCom 2019.

Wenyao Xu (Senior Member, IEEE) received the PhD degree from the University of California with Los Angeles, Los Angeles, USA, and both the master and bachelor degree from Zhejiang University, China. He is an associate professor with tenure of the Computer Science and Engineering Department, University, Buffalo (SUNY). His research has focused on exploring novel sensing and computing technologies to build up innovative Internet-of-Things (IoT) systems for high-impact human-technology applications in the fields of Smart Health and Cyber-Security. Results have been published in peer reviewed top research venues across multiple disciplines, including Computer Science conferences (e.g., ACM MobiCom, SenSys, MobiSys, UbiComp, ASPLOS, ISCA, HPCA, Oakland, NDSS and CCS), Biomedical Engineering journals (e.g., IEEE TBME, TBioCAS, and JBHI), and Medicine journals (e.g., LANCET). To date, his group has published over peer-reviewed 180 papers, won nine best paper awards, two best paper nominations and three international best design awards. His inventions have been filed within U.S. and internationally as patents, and have been licensed to industrial players. His research has been reported in high-impact media outlets, including the Discovery Channel, CNN, NPR and the Wall Street Journal. Currently, Dr. Xu serves as an associate editor of IEEE Transactions on Biomedical Circuits and Systems (TBCAS), the technical program committee of numerous conferences in the field of Smart Health and Internet of Things, and has been a TPC co-chair of IEEE Body Sensor Networks in 2018.

Kui Ren (Fellow, IEEE) received the PhD degree in electrical and computer engineering from Worcester Polytechnic Institute. He is a professor and associate dean of college of Computer Science and Technology with the Zhejiang University, where he also directs the Institute of Cyber Science and Technology. Before that, he was with State University of New York, Buffalo. Kui's current research interests include Data Security, IoT Security, AI Security, and Privacy. He received Guohua Distinguished Scholar Award from ZJU, in 2020, IEEE CISTC Technical Recognition Award, in 2017, SUNY Chancellor's Research Excellence Award, in 2017, Sigma Xi Research Excellence Award, in 2012 and NSF CAREER Award, in 2011. He has published extensively in peer-reviewed journals and conferences and received the Test-oftime Paper Award from IEEE INFOCOM and many Best Paper Awards from IEEE and ACM including MobiSys'20, ICDCS'20, Globecom'19, ASIACCS'18, ICDCS'17, etc. His h-index is 74, and his total publication citation exceeds 32,000 according to Google Scholar. He is a Fellow of ACM and a Clarivate Highly-Cited Researcher. He is a frequent reviewer for funding agencies internationally and serves on the editorial boards of many IEEE and ACM journals. He currently serves as Chair of SIGSAC of ACM China.