IEEE INFOCOM 2022 - IEEE Conference on Computer Communications | 978-1-6654-5822-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INFOCOM48880.2022.9796806

mmPhone: Acoustic Eavesdropping on Loudspeakers via mmWave-characterized Piezoelectric Effect

Chao Wang[1], Feng Lin[1]*, Tiantian Liu[1], Ziwei Liu[1], Yijie Shen[1], Zhongjie Ba[1], Li Lu[1], Wenyao Xu[2], Kui Ren[1]
[1]Zhejiang University, Hangzhou, China
[2]University at Buffalo, SUNY, USA
{wangchao5001, flin, tiantian, zivliu, shenyijie, zhongjieba, li.lu, kuiren}@zju.edu.cn, wenyaoxu@buffalo.edu

***Abstract*—More and more people turn to online voice communication with loudspeaker-equipped devices due to its convenience. To prevent speech leakage, soundproof rooms are often adopted. This paper presents *mmPhone*, a novel acoustic eavesdropping system that recovers loudspeaker speech protected by soundproof environments. The key idea is that properties of piezoelectric films in mmWave band can change with sound pressure due to the piezoelectric effect. If the property changes are acquired by an adversary (i.e., characterizing the piezoelectric effect with mmWaves), speech leakage can happen. More importantly, the piezoelectric film can work without a power supply. Base on this, we proposed a methodology using mmWaves to sense the film and decoding the speech from mmWaves, which turns the film into a passive "microphone". To recover intelligible speech, we further develop an enhancement scheme based on a denoising neural network, multi-channel augmentation, and speech synthesis, to compensate for the propagation and penetration loss of mmWaves. We perform extensive experiments to evaluate mmPhone and conduct digit recognition with over 93% accuracy. The results indicate mmPhone can recover high-quality and intelligible speech from a distance over 5m and is resilient to incident angles of sound waves (within 55 degrees) and different types of loudspeakers.**

***Index Terms*—Eavesdropping, mmWave sensing, loudspeakers**

I. Introduction

Voice telecommunication plays an important role in our daily life. Voice and video calls are becoming more and more popular for social and business communications [1], such as web conferences. Especially due to the COVID-19 pandemic, online voice communication has shown increasing uptake on a global scale [2]. More and more people choose to have online conversations leveraging their electronic communication equipment, such as smartphones, personal computers, and intercom screens in meeting rooms. These devices play the machine-rendered speech on loudspeakers during conversations. Because the speech can be related to personal privacy (e.g., social passwords) and business secrets (e.g., enterprise conferences), soundproof obstacles are often deployed around a room to prevent speech leakage.

*Feng Lin is the corresponding author.

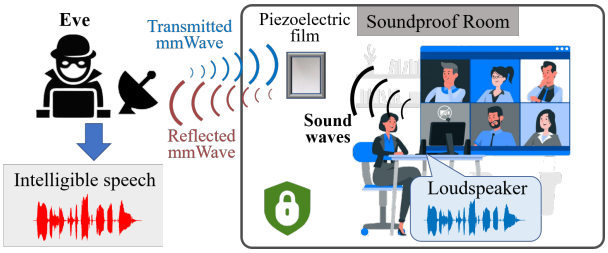
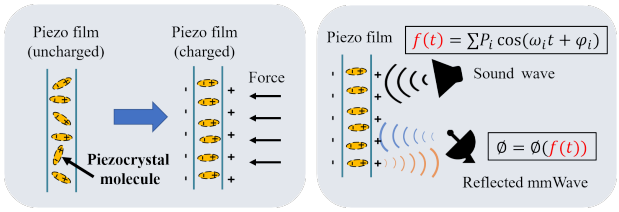


Fig. 1. mmPhone can recover speech emitted by loudspeakers in a soundproof room via interrogating an in-room piezoelectric film with mmWaves.

Nonetheless, researchers reveal that such machine-rendered speech can be compromised by leveraging sound-related vibration. When a loudspeaker generates sound waves, speech leakage can happen via the physical vibration of objects, including the loudspeaker itself or nearby objects. For the former, speech information can be retrieved by the motion sensors [3]–[6] and RF signals [7]. For the latter, adversaries can use non-acoustic sensors to sense the vibration of objects induced by propagating sound waves for sound recovery, such as lidars [8], high-speed cameras [9], vibration motor [10], and hard drives [11]. However, the physical properties of objects (e.g., materials, stiffness, and structures) can affect the vibration, which further influences the eavesdropping performance.

Considering the pervasiveness of sound waves in the air, we wonder if there exists a methodology to eavesdrop on the propagating sound waves directly instead of targeting particular vibrating objects. Can we use the methodology to recover high-quality and intelligible speech even though the speech is protected by a soundproof environment? Based on our preliminary study, we find that the electromagnetic property (e.g., reflection coefficients) of piezoelectric materials in the millimeter-wave band can change with sound pressure due to the piezoelectric effect [12], [13]. If an adversary leverages a mmWave sensor to sense such changes remotely, the adversary can acquire sound-related information, which may cause threats to the speech contents. Specifically, as shown in Figure 1, the propagating sound waves hit the



(a) (b)

Fig. 2. Piezoelectric effect of the piezo (piezoelectric) film. (a) The piezo film charges when a force is applied to its surface. (b) The changing sound pressure of sound waves can cause the film charging/discharging, influencing the phase of reflected mmWave signals.

surface of a piezoelectric film and change the film's reflection coefficient by the applied sound pressure. An attacker uses a mmWave probe to interrogate the reflection coefficient changes of the film penetrating the soundproof wall and analyzes the reflected mmWaves to recover the speech remotely.

To realize the speech eavesdropping, there are some challenges to be addressed. First, the sound pressure level (SPL) of propagating sound waves in a normal conversation is limited ($60 \sim 70dB$). Building a remote sensing methodology based on the sound-sensitive material to recover the low-energy soundwaves is vital to the eavesdropping. Second, the mmWave signal can decay with the sensing distance, causing a decreasing signal-to-noise ratio (SNR), especially when penetrating obstacles (e.g., soundproof walls). This poses a great challenge for a remote attacker to recover high-quality speech. Third, considering the spectral complexity of human speech, higher frequency components of sound waves suffer more attenuation during the propagation [14] and are easier to be flooded by noise, which results in poor speech intelligibility. An intelligible voice recovery is required to cause practical threats to the speech contents.

In this paper, we propose **mmPhone**, an eavesdropping system that can recover the speech protected by a soundproof room. Our work focuses on speech threats exposed by loudspeakers. We first theoretically model the piezoelectric film in the sound fields, transforming the speech from sound fields to electromagnetic fields. Then we build a mmWave-based methodology to interrogate the film and decode the speech from reflected mmWave signals. To fight against attenuation and noise interference, we design a speech enhancement scheme containing a denoising neural network and a multi-channel augmentation to improve recovered speech quality (i.e., how comfortable the speech sounds, such as clean or noisy). To increase the intelligibility (i.e., how comprehensible the speech is) of recovered speech, we further develop a training-free method based on pitch estimation and harmonic extension to recover harmonics of the human voice and reconstruct intelligible speech. Overall, we make the following contributions in this work:

- We build a methodology to sense acoustic waves directly leveraging the piezoelectric effect and mmWave signals. We theoretically model the sound-mmWave transformation and reveal a novel attack that can eavesdrop on the speech protected by soundproof obstacles.
- We propose an eavesdropping system *mmPhone* based on a COTS mmWave probe. To increase speech quality, we design an enhancement scheme containing a denoising neural network and a phase-alignment-based multi-channel augmentation. To improve speech intelligibility, we develop a training-free method based on speech synthesis which leverages recovered audio from the four receiving antennas for accurate speech reconstruction.
- We evaluate mmPhone with extensive experiments. The results indicate that mmPhone can recover high-quality and intelligible speech remotely (5m) with digit recognition accuracy of over 93%. We quantitatively evaluate robustness of mmPhone which shows resilience to incident angles of soundwaves ($< 55°$) and different loudspeakers.

II. Background and Threat Model

A. Piezoelectric Film

Polyvinylidene fluoride (PVDF) [15] is a specialty plastic ubiquitous in daily life, such as folder covers and piping products. After polarization, a PVDF film can become piezoelectric without appearance changes. We use a PVDF piezoelectric film in this paper. Due to the piezoelectric effect, the film can work as a passive transducer to transform sound pressure into electromagnetic signals [16]. As shown in Figure 2(a), the film charges when a force is applied to the film surface. A changing force, such as the force induced by the changing sound pressure, can cause the film to charge and discharge alternatively. We theoretically model the piezoelectric effect when sound waves hit the film surface in Section III.

B. mmWave Sensing

The frequency-modulated continuous-wave (FMCW) radar system transmits FMCW (also called chirp) and captures reflected signals from objects. The received signal is demodulated by a mixer, passed to a low-pass filter $LPF(\cdot)$, and then sampled by the analog-to-digit converter (ADC) to produce an intermediate frequency (IF) signal. For two sinusoidal inputs x_1 (transmitted signal) and x_2 (received signal) of the mixer $x_i = A_i \sin(\omega_i t + \phi_i)$ $(i = 1, 2)$, the output IF signal x_o can be calculated as

$$x_o = LPF(x_1 \cdot x_2) = A_3 \cos((\omega_1 - \omega_2)t + \phi_1 - \phi_2), \quad (1)$$

where A_3 is the amplitude of x_o, $\omega_1 - \omega_2 = 2\pi f_c$, and f_c is the frequency of x_0. The reflected signal x_2 from an object can be taken as the replicate of the transmitted signal x_1. The time delay τ of the two signals can be calculated by $\tau = \frac{2d}{c}$, where d is the distance to the object and c is the speed of light. Then the initial phase of x_o can be written as

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

where f_c is the frequency of x_1 and λ is the wavelength of x_1. For a static object at distance d_0, we denote the reflection

coefficient of the object as $\Gamma = |\Gamma|e^{j\phi_r}$. Given that x_2 is the echo of x_1 reflected by the object, i.e., $\phi_1 - \phi_2 = \phi_r$, we have

$$\phi_0 = \frac{4\pi d_0}{\lambda} + \phi_r, \tag{3}$$

where $\frac{4\pi d_0}{\lambda}$ is a constant. According to Eq. 3, the phase of the IF signal x_0 and the phase of the reflection coefficient Γ has the following relationship: 1) If the Γ of the object is constant, i.e., $\phi_r = \phi_a$, then ϕ_0 is also a constant. 2) If the Γ of the object varies with external stimulation like soundwaves, then ϕ_0 can also change with the stimulation.

C. Threat Model

We consider a scenario where a victim uses loudspeakers in a soundproof room, e.g., attending an online conference in a conference room at the company. To ensure speech confidentiality, the victim is vigilant to the in-room eavesdropping devices and thus forbids any **active** electronic devices into the room, such as microphones or smartphones. This can be achieved by electronic device detectors which rely on the detection of emitted electromagnetic signals or wireless signals from malicious devices. In such a scenario, the piezoelectric film, a **passive** element free of any electronic components like butteries or ADCs, can be disguised as the cover of a book or papers and token by an executor into the room. The executor can be a hired person by the adversary. The attacker can also be an inner adversary of the company and thus he/she can pre-install the film in the room. We have the following assumptions: 1) The victim uses a loudspeaker to emit the participants' speech in the online communication. 2) The piezoelectric film is pre-placed in the room, which can be achieved by social engineering [17]. 3) The adversary knows the film's location to transmit directional mmWaves.

III. Modeling sound-mmWave transformation

A. Sound Propagation Model

Sound waves propagating through the air can be formulated by a function of position and time $P(x,t)$ which is also known as sound fields [18]. The sound pressure at position x_0 can be taken as the superposition of a series of single-frequency waves varying with time t:

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

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. When soundwaves hit a piezoelectric film, the induced force on the film can be calculated by $F = P \cdot S$, where S is the size of the film and P is the sound pressure. Due to the piezoelectric effect of the film, the quantity of electric charges induced by the applied force F can be calculated as $Q = D_{33} \cdot F$, where D_{33} is the piezoelectric constant of the piezoelectric film. Then we get the relationship between sound waves and induced charges on the film

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

B. mmWave Reflected from A Piezoelectric Film

When propagating mmWaves in the air impinges upon a piezoelectric film, the mmWave can be partially reflected. The initial phase of reflected waves (i.e., the phase of reflected mmWaves on the boundary between the air and the film) is determined by electric properties of the two mediums. According to the *Transmission Line* model [19], the reflection of mmWaves at the boundary can be formulated by the reflection coefficient

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

Z_1 and Z_2 are the intrinsic impedance of the air and the film, where $Z_i = \frac{\omega\mu_i}{k_{z_i}}, k_i = \omega\sqrt{\mu_i\varepsilon_i} (i = 1, 2)$, ω 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, the propagation coefficients k_{z_1}, k_{z_2} in the direction of propagation can be calculated by $k_{z_1} = k_1 \cos\theta, k_{z_2} = \sqrt{k_2^2 - k_x^2}$, where $k_x = k_1 \sin\theta$. Then we can derive the reflection coefficient

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

We consider vertically incident mmWaves towards the film without loss of generality, i.e., $\theta = 0$. The relative permittivity of the film $\varepsilon_r = 1 - \frac{\omega_p^2}{\omega^2}$, where $\omega_p = \sqrt{\frac{Ne^2}{m\varepsilon_0}}, N = \frac{Q}{Sd}$ [19]. Now we can derive the relationship between the phase of Γ and the quantity of sound-induced charges Q:

$$\Phi(\Gamma) = \Phi\left(\frac{1 - \sqrt{a_1 Q - 1}j}{1 + \sqrt{a_1 Q - 1}j}\right) = -2\arctan\sqrt{a_1 Q - 1}, \tag{8}$$

where $a_1 = \frac{e^2}{\omega^2 m\varepsilon_0 Sd}$. We further apply *Taylor expansion* on Eq.8 at $Q = Q_0$:

$$\begin{aligned}\Phi(\Gamma) &= -\frac{1}{Q_0\sqrt{a_1 Q_0 - 1}}Q + C_0 + o(Q - Q_0)^2 \\ &= k_0 Q + C_0, |Q - Q_0| < \epsilon,\end{aligned} \tag{9}$$

where k_0 and C_0 are determined by a_1 and Q_0, Q_0 is the average quantity of induced charges by the incident sound waves. ϵ is the quantity changes determined by the changes of sound waves, which is a small value. According to Eq. 5 and Eq. 9, the phase of Γ

$$\phi_r = k_0 D_{33} S \textstyle\sum_i P_i \cos(\omega_i t + \phi_i) + C_0. \tag{10}$$

C. Decoding Sound Waves from Reflected mmWaves

As introduced in Section II-B, an object with varying reflection coefficient $\Gamma = |\Gamma|e^{j\phi_r}$ can change the ϕ_0 of the IF signal. According to Eq. 3 and Eq. 10, we can easily derive

$$\phi_0 = \Phi_1(\textstyle\sum_i P_i \cos(\omega_i t + \phi_i)), \tag{11}$$

where $\Phi_1(\cdot)$ is a linear function. For successively reflected chirps $x_1, x_2, ..., x_n$, we denote the phase of the IF signal demodulated from i_{th} chirp as ϕ_0^i. Then the phases of successively demodulated chirps can be written as $\phi_0^1, \phi_0^2, ..., \phi_0^n$. According to Eq. 11, the phase ϕ_0^i is determined by the soundwave state at a specific time. In other words, we can

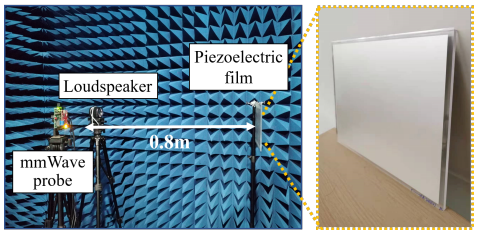


Fig. 3. Verification experiment setting in an anechoic chamber. The A4-size piezoelectric film is stuck to an acrylic board with glue to avoid vibration.

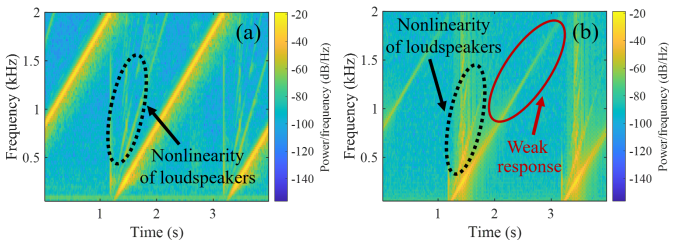


Fig. 4. We compare (a) the audio recorded by a microphone and (b) the audio decoded by the proposed sound-mmWave transformation model.

take each demodulation of a mmWave chirp as a sampling of the sound waves where the sampling rate $f_s = \frac{1}{T_{chirp}}$.

D. Verification Experiment

Experimental Setting: As shown in Figure 3, we conducted experiments in an anechoic chamber. We placed a loudspeaker playing audio chirps ($50 - 2k$ Hz, 65 dBSPL) with a period of 2s towards the piezoelectric film and used a mmWave probe (AWR1843) to interrogate the film from a distance of $0.8m$. The whole film was stuck on an acrylic board with glue to avoid physical vibration caused by sound waves. To avoid vibration interference from the loudspeaker, we placed the loudspeaker side by side with the probe without any physical contact between the two. Note that the loudspeaker was beyond the field-of-view ($\pm 28^\circ$ horizontally) of the probe. The chirp rate of transmitted mmWave is 10.2k per second ($f_s = 10.2kHz$). We also used a microphone to record the played audio as the reference.

Decoding Audio from mmWave: For each period of demodulated mmWave signals (IF signals), we derived the phase spectrum $\Phi^i = \{\phi_1^i, \phi_2^i, ..., \phi_{512}^i\}$ by a 512-point fast Fourier transform (FFT). For M successively demodulated mmWave signals, we got a set of phase spectrums $\{\Phi^1, \Phi^2, ..., \Phi^M\}$. Then we derived the phase value across different periods $\{\phi_j^1, \phi_j^2, ..., \phi_j^M\}$ for the j_{th} frequency point of the phase spectrum, where $j = 1, 2, ..., 512$. According to Eq.11, the sequence $\{\phi_j^1, \phi_j^2, ..., \phi_j^M\}$ has a linear relationship with the sound waves. So we searched for all the 512 derived sequences and chose the one that has the largest correlation value [20] with the audio chirp (recorded by the mic) as the decoded result. The spectrograms of the mic-recorded audio and mmWave-decoded audio are shown in Figure 4(a) and (b), respectively. We observe that the audio can be successfully decoded from the phase sequences which is consistent with our theoretical model. However, we also find that:

- The power density of recovered audio in Figure 4(b) is weaker than the mic's in the unobstructed case. For a remote through-wall eavesdropping, the attenuation of mmWave would be larger and worsen recovered speech quality.
- The weak response in the high-frequency band of mmWave-decoded audio (indicated by the red circle in Figure 4(b)) can cause loss of speech formants which play a vital role in the intelligibility of speech [21].

E. Observations and Potential Solution

Audio Decoded from Multiple Channels: Hereafter, we denote each *frequency point* in the FFT result of IF signals as a *channel*, which can modulate the audio information. During the experiments, we found that there was more than one channel from which we could successfully decode the played audio. A premier observation is that these frequency points are successive. This is probably because the A4-size ($21.0cm \times 29.7cm$) film can induce multiple frequency points in the FFT results considering that the range resolution of the mmWave probe is only $3.75cm$ for a 4GHz bandwidth. Figure 5 shows the decoded audio from another two channels adjacent to the one of Figure 4(b). We can observe that the audio can be decoded from multiple channels but with different SNR. Based on this observation, we proposed a phase-alignment-based method to merge audio traces decoded from multiple channels to improve speech quality (Section IV-D).

Audio Recovered from Multiple Receiving Antennas: The COTS mmWave probe we used has an antenna array with three transmitting antennas (Tx) and four receiving antennas (Rx). The four receiving antennas can act as four separate microphones with different SNR considering that they have respective mixers and analog-to-digital converters (ADCs). Thus, we can leverage the "microphone array" for further speech enhancement as introduced in Section IV-E.

Acoustic Nonlinearity of Loudspeakers: Due to the acoustic nonlinearity of loudspeakers [22], there are chirp harmonics in both mic-recorded and mmWave-decoded audio, as dotted circles in Figure 4(a)(b) show. However, the harmonics in Figure 4(b) disappear when the chirp frequency is above 250 Hz, which cause little interference to the intelligibility of

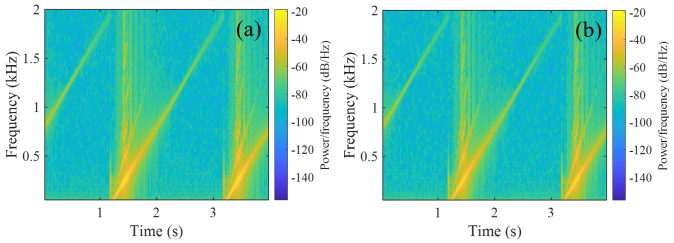


Fig. 5. (a) and (b) show decoded results of two channels adjacent to the one in Figure 4(b), which can be leveraged for enhancement (detailed in IV-D).

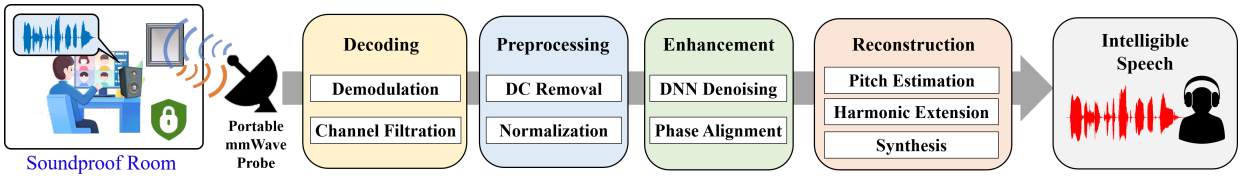


Fig. 6. System overview of the proposed system *mmPhone*.

recovered speech (only two vowels, i.e., i and y, have formants below 250 Hz that may be interfered with [23]). So we do not take further exploration about this in the following pages.

IV. System Design

A. System Overview

The framework of mmPhone is shown in Figure 6. The mmWave probe transmits FMCW to interrogate the piezoelectric film through the soundproof wall and demodulate reflected mmWave signals. The audio is decoded from multiple channels respectively according to the sound-mmWave transformation model (Section III). Then a preprocessing is applied to the decoded audio traces to eliminate clutters preliminarily. To recover high-quality speech, we merge the multiple audio traces by a speech enhancement scheme which consists of a denoising neural network and a multi-channel augmentation. Finally, the enhanced audio is fed into a training-free speech reconstruction module to improve the speech intelligibility.

B. Decoding Audio from mmWave

Demodulation: 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 $S = [S_j^i], i = 1, ..., M, j = 1, ..., 512$ and the phase matrix $\Phi = \{\phi_j^i\}, i = 1, ..., M, j = 1, ..., 512$ from S. We further extract the phase sequence $\{\phi_j^1, \phi_j^2, ..., \phi_j^M\}$ for each of the 512 channels, where $j = 1, ..., 512$. Then we perform phase unwrapping to all the derived phase sequences to eliminate the impact of *integer ambiguity*.

Channel Filtration: After the phase unwrapping, we get the decoded audio of each channel, i.e., the unwrapped phase sequence. However, not all channels convey the speech information as analyzed in Section III-E. To localize the channels that contain speech information, an intuitive but time-consuming way is listening to the decoded audio one by one to judge if it contains speech. Here we propose a more efficient method called *channel filtration* to localize the desired channels automatically. The core idea is that the human voice has a higher power density than background noise in the frequency band of $85 \sim 255$ Hz. Thus, we can use this feature of the human voice to identify if a decoded trace contains speech. Specifically, we first apply a band-pass filter with cut-off frequencies of 80 Hz and 260 Hz to all the decoded traces and then calculate power spectral density for each trace. The top k traces with the highest density values are chosen (k=3).

C. Speech Preprocessing

DC Removal and Filtering: After decoding the audio from the mmWave, we subtract the mean from the decoded audio trace to correct DC offset raised by the film (Eq. 11) and the probe. Then we design a fourth-order highpass Butterworth filter with a cut-off frequency of 80 Hz to eliminate clutters, such as human wiggles and wind blowing.

Normalization: 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 played volumes. After the normalization, we acquire multiple noisy audio traces whose quality and intelligibility need to be improved.

D. Speech Enhancement

Denoising Neural Network: The propagation and penetration loss of mmWave can result in a low SNR of reflected signals and thus cause a poor-quality speech that requires denoising. Traditional denoising methods, such as Wiener filter and spectral subtraction, will introduce broadband residual noise and musical noise, damaging speech quality. In this part, we use a deep neural network (DNN) based denoising method. One of the advantages of such methods is that non-linear structures of DNN contribute to learning the more complex mapping between noisy and clean speech. This allows the network to handle non-stationary noise in real-world scenarios. Specifically, the denoising approach leverages spectral masking [24] to map noisy speech into clean ones. The noisy speech segment is first transformed into a spectrogram and multiplied with the spectral mask to generate a new spectrogram. Then the generated spectrogram is resynthesized into a clean speech with the phase information derived from the noisy speech. The spectral mask is estimated by a DNN, as shown in Figure 7. The input is the magnitude spectrum of a speech segment after a short-time Fourier transform (STFT) with a size of 240×513. The first two 3×3 convolution layers consist of 1024 and 512 filters with a stride of 1. The eight encoder layers adopt a multi-head self-attention mechanism [25] which allows the model to jointly attend to information from different subspaces at different positions. The output layer applies a linear transformation to the incoming vector, and a following ReLU layer is used as the activation function. The denoising neural network takes the MimicLoss [24] as the loss function which achieves better enhancement using the information derived from a speech recognizer. We trained the

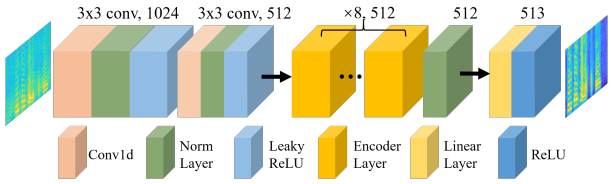


Fig. 7. The structure of the denoising neural network.

network with public datasets [26] and deployed the model on a laptop for signal denoising. After denoising, the audio traces decoded from multiple channels are fed into a phase-alignment module to further improve the speech quality.

Phase-alignment-based Enhancement: As demonstrated in Section III-E, we can decode the played audio from multiple channels respectively. The decoded audio traces contain the same speech contents but have different SNR. A simple method to merge the multiple traces for enhancement is aligning the signals in the time domain and adding up their amplitudes directly. However, this method has poor performance when the phases of signals are not aligned. Thus, we propose a phase-alignment-based enhancement as shown in Algorithm 1. We first choose the one with the highest SNR from multiple decoded traces of each Rx (note that the probe has four Rx) as the baseline. Then we apply FFT to all the traces with a window size of 512 to get the spectrums. We set the phases of the other traces to the baseline's and add up all the spectrums (complex values). Then we perform the inverse FFT (iFFT) to get time-domain signals. The time-domain segments are overlap-added when the window slides with a step of 128 until all the segments are phase-aligned.

E. Speech Reconstruction

As aforementioned in Section III-D, the decoded audio traces suffer weak response in the high-frequency band due to the attenuation of mmWave, which damages speech intelligibility. This can be mitigated by speech synthesis to recover the harmonic bandwidth from the distorted audio traces. Training-based methods [27], [28] require a large amount of training data to achieve a satisfying performance which costs a lot. Based on this consideration, we turn to a training-free speech reconstruction scheme to improve the speech intelligibility.

Pitch Estimation: This step aims to estimate the fundamental frequency (i.e., the *pitch*, which has the strongest power in the spectrum) of the human voice. This is the basic step for speech reconstruction. As analyzed in Section III-E, the four receiving antennas can act as a "microphone" array. However, different SNR of the four antennas can cause varying estimation errors. To improve the estimation accuracy, we first estimate the pitch f_0^i for the audio trace of Antenna #i, where $i = 1, 2, 3, 4$. Then a calibrated f_0 is calculated based on the four estimated pitches. Specifically, we first segment the audio of each antenna into 50-ms segments considering the short-time stability of the human voice and then apply the f_0 estimation on the segments. We use a band-pass filter to get the fundamental frequency of each segment coarsely. Considering

Algorithm 1: Speech Enhancement for A Single Rx

Input: Preprocessed audio $\{s_i\}$, $i = 1, ..., I$, where I is the number of channels.
Output: Enhanced audio s_{en}

Initialize $N = 512, hop = \frac{N}{4}, win = hanning(N)$;
$\{s_i\} \leftarrow DNN(\{s_i\}), slen = length(s_i)$
for $i \leftarrow 1 : I$ **do**
  $s_{en} \leftarrow zeros(slen), s_{ref} \leftarrow \arg\max_s SNR(s)$
  for $k \leftarrow 0 : floor(\frac{slen-N}{hop})$ **do**
    $S_i \leftarrow FFT_N(s_i(1+k*hop : N+k*hop)*win)$
    $S \leftarrow \sum_i \frac{|S_i|}{|S_{ref}|} * S_{ref}$
    $s_{en}(1 + k * hop : N + k * hop) += iFFT(S)$
return s_{en}

that the fundamental frequency of the human voice is within 85-255 Hz, we set the cut-off frequency of the band-pass filter to 80 Hz and 260 Hz. After the filtering, we apply FFT to the segments and take the frequency with the highest magnitude as the estimated f_0. We denote SNR_i as the SNR of the segment from Antenna #i, then the calibrated f_0 is calculated by $f_0 = \sum_{i=1}^{4}(snr_i * f_0^i) / \sum_{i=1}^{4} snr_i$.

Harmonic Extension: The formants, which refer to local maximums in the spectral envelope of human voice [23], play a vital role in speech intelligibility. We adopt a spectral envelope estimation algorithm [29] which applies harmonic extension to recover the distorted spectral envelope. Instead of feeding the estimated pitch of each audio trace respectively, we use the calibrated f_0 for better spectral envelope estimation. After the harmonic extension, the formants in the spectral envelope are recovered, which helps improve speech intelligibility.

Synthesis: After harmonic extension, we use D4C algorithm [30] to recover aperiodic components of the speech and combine it with the calibrated pitch f_0 and extended harmonics for intelligible speech synthesis. The synthesis is achieved by the convolution of minimum phase response and excitation signals. The synthesized speech has a bandwidth of up to 2.3 kHz, intelligible for human hearing [31]. After speech synthesis for all the antennas, we apply the phase-alignment method in Section IV-D to the four reconstructed audio for further enhancement. Finally, we get an enhanced speech with high quality and intelligibility. The spectrograms of the played audio, decoded (unprocessed) audio from a single channel, and processed audio by the scheme are shown in Figure 8(a)(b)(c).

V. Evaluation

A. System Setup

The system setup is shown in Figure 9. The COTS mmWave probe AWR1843Boost has a transmitting power of 12dBm with a portable size of $6.5cm \times 8.5cm \times 2.0cm$ to interrogate the piezoelectric film. The transmitted chirp has a frequency range of 77-81 GHz. We set the chirp rate to 10,200 chirps/sec. The IF signal is collected by a DCA1000EVM and sent to a laptop (Thinkpad T490) for processing. The film has a size

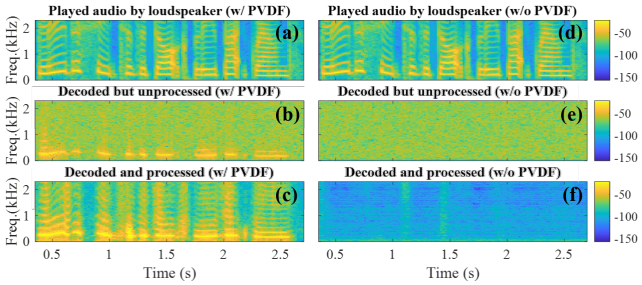


Fig. 8. (a), (b), and (c) show the played audio, raw decoded audio, and processed audio by mmPhone when we deployed the PVDF film. (d), (e), and (f) show the results when we removed the film.

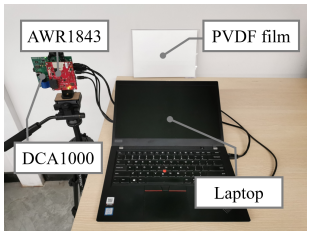


Fig. 9. System setup (mmPhone).

Fig. 10. Tested loudspeakers.

of $21cm \times 29.7cm \times 28\mu m$ 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. We adopted Adam as the optimizer (Lr=0.001,epoch=500).

B. Datasets and Data Collection

We used three public datasets that are widely adopted for speech testing to evaluate our system, i.e., Harvard Speech Corpus (HSC) [32], AudioMNIST [33], and Open Speech Repository (OSR) [34]. The HSC consists of 720 sentences (known as *Harvard Sentences*) designed to feature phonemes at the same frequency they appear in spoken English. The AudioMNIST contains 30,000 samples of spoken digits from 60 speakers. The OSR includes *Harvard Sentences* from different speakers, from which we chose 100 samples. We evaluated mmPhone robustness in V-E, V-F, and V-H with two speakers' samples in AudioMNIST. All experiments were ensured to follow the institutional review board (IRB) protocol.

We asked a volunteer to play audio samples of the three datasets via a loudspeaker (Hp) in a soundproof room (Figure 11). The SPL (measured by a sound level meter placed nearby the piezoelectric film) was around 67dB within the range of normal conversations [35]. When the loudspeaker played audio, the volunteer and two other volunteers typed randomly on their laptops. We deployed the mmWave probe outside the room with a sensing distance of 5m. The distance between the loudspeaker and piezoelectric film is 2m. For comparison, we also deployed two microphones to record the played audio, one (Mic1) inside and the other (Mic2) outside the room.

Interrogated Source Validation: We performed a controlled experiment to ensure the recovered speech was from the PVDF film rather than the vibrating loudspeaker. We respectively placed (i.e., w/ PVDF) and removed (i.e., w/o PVDF) the film for speech recovery with exactly the same processing steps (e.g., the same decoding channel). The results are shown in Figure 8. Comparing (c) and (f), we observe that mmPhone failed to recover the played audio when we removed the film, indicating the recovered speech by mmPhone was from the PVDF film rather than the vibrating loudspeaker.

C. Metrics

Peak-signal-to-noise ratio (PSNR): The PSNR is a commonly used metric to quantify speech quality [7], [9]. The empirical boundary of PSNR for human-audible speech is $0dB$ [7]. A higher PSNR indicates a better speech quality.

Short-Time Objective Intelligibility (STOI): The STOI has a monotonic relationship with the subjective speech-intelligibility [36]. The STOI varies within [0,1], of which a higher value indicates better speech intelligibility.

D. Overall Performance

In this part, we evaluate the system performance by calculating the PSNR and STOI of the speech recovered by mmPhone. Given that the digits (i.e., $0 \sim 9$) are often related to secret information, such as passwords and security numbers, we also apply automatic speech recognition (ASR) and manual speech recognition (MSR) to the recovered digit speech.

*1) **Sound Recovery**:* The results of PSNR and STOI are shown in Figure 12 and Figure 13, respectively. From Figure 12, we can observe that mmPhone outperforms the out-room microphone (Mic2, yellow bars) with high PSNR above $14dB$ on the three datasets. We find that the PSNR of recovered speech by mmPhone can even be higher than the in-room microphone's (blue bars) on the OSR datasets. The reasons are two folds. First, the DNN-based denoising and phase-alignment-based enhancement can significantly improve the SNR of recovered speech. Second, compared with the single in-room microphone, mmPhone also enhances the recovered speech leveraging the four receiving antennas, each of which acts as a separate microphone. Thus, the "microphone array" of mmPhone can outperform the single acoustic microphone by leveraging the proposed scheme in Section IV.

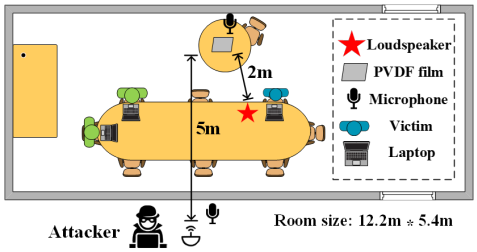


Fig. 11. Experimental scenario of a conference room with a soundproof glass wall (two layers of glass and 1cm-thick for each).

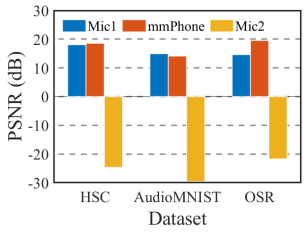


Fig. 12. Overall performance(PSNR).

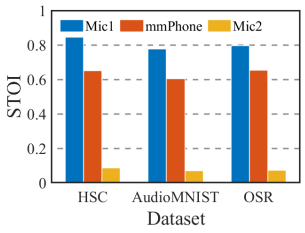


Fig. 13. Overall performance(STOI).

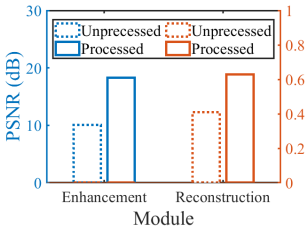


Fig. 14. Module performance.

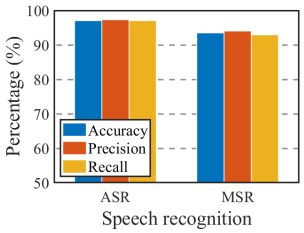


Fig. 15. Digit recognition.

Figure 13 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 cause loss of high frequency (above 5.1kHz according to Nyquist theorem) of human voice. Figure 14 shows the performance gain of system modules. The *Enhancement* (Section IV-D) and *Reconstruction* (Section IV-E) modules aim to improve speech quality and intelligibility, respectively. For each module, we calculated the metrics of the input (i.e., Unprocessed) and output (i.e., Processed) audio for comparison. We find that the PSNR increases by $8.2dB$ after the *Enhancement*. The STOI increases by 53.6% after the *Reconstruction*. The results validate the effectiveness of the proposed scheme in Section IV.

*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 (Chinese) to listen to the 100 audio traces. We asked the volunteers to pick up the most likely one from the ten digits ($0 \sim 9$). The recognition results are shown in Figure 15. The accuracy, precision, and recall of ASR are 97.0%, 97.3%, and 97.1%, and MSR achieves 93.5%, 94.1%, and 92.9%, respectively. We find that the digits "four" and "five" are more likely to be misclassified by the volunteers due to distorted phonemes. Besides, the diverse accents and tones can also pose challenges for the MSR. Overall, mmPhone can achieve over 93% accuracy for both ASR and MSR.

E. Sensing Distance

Considering that mmWave signals can decay with distance, we evaluated mmPhone with different sensing (probe-film) distances from 2m to 7m in the through-wall scenario (Figure 11). Other settings are the same as in V-B. As shown in Figure 16, the PSNR of raw decoded speech (Unprocessed, blue dashed line) is steady above $10dB$ when the distance is within $5m$ but reduces to $3.9dB$ when the distance is $7m$. This results from the declining power density of mmWave when the distance increases. With our proposed enhancement scheme, the PSNR has a gain of $9 \sim 11dB$. We also observe that the STOI reduces to 0.53 when the distance is 6m. The possible reason is that mmWaves suffer larger attenuation as the distance increases, resulting in a lower SNR at the receiver. With the enhancement scheme, the STOI of reconstructed audio increases to 0.61 when the distance is less than $6m$.

F. Incident Angle of Propagating Sound Waves

The pressure amplitude P_i (in Eq. 4) applied to the film can be different with respect to different incident angles of sound saves, which can influence the reflection coefficient Γ of the film according to Eq. 10. *Here we define the incident angle of sound waves as the angle between the wave-propagating direction and the normal of the film surface.* We changed the incident angle of sound waves and kept other settings the same as in V-B. We deployed the probe outside the soundproof room for through-wall eavesdropping. The results are shown in Figure 17. 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 ($5.8dB$) but with poor intelligibility. We also find that the gain of intelligibility (i.e., the difference between orange lines) goes down as the angle increases. Considering that the spectral envelope (Section IV-E) plays an important role in the speech intelligibility, the possible reason is that the low PSNR at large incident angles 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.

G. Incident Angle of the mmWave

We quantitatively study the impact of the incident angle of mmWave in the soundproof scene (Figure 11). We define the incident angle of mmWave as the angle between the mmWave-propagating direction (main lobe) and the normal of the film surface. Other settings are the same as in V-B. The results are shown in Figure 18. We observe that the PSNR varies from $21.5dB$ to $23.1dB$ and the STOI varies from 0.61 to 0.63. Both are steady with little fluctuation under different incident angles of mmWave ($15°/30°/45°$). The results indicate that mmPhone has a certain tolerance for different incident angles of mmWave to recover high-quality and intelligible speech.

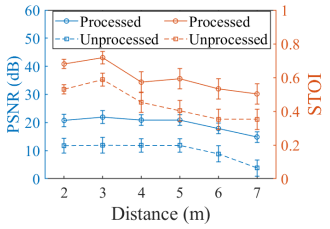


Fig. 16. Impact of distance.

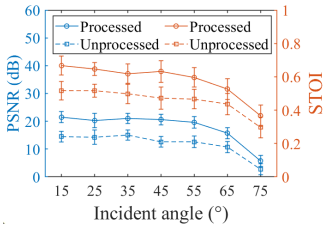


Fig. 17. Soundwave incident angle.

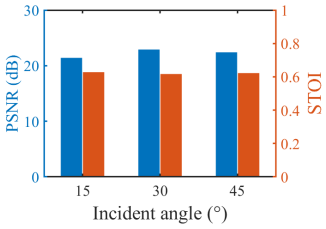


Fig. 18. Incident angle of mmWave.

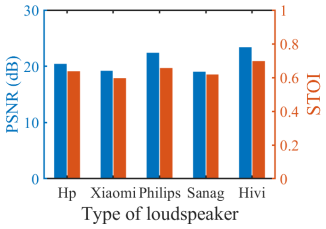


Fig. 19. Impact of loudspeaker type.

H. Different Loudspeakers

To investigate the impact of loudspeaker structures, we chose five different types of commodity loudspeakers shown in Figure 10. 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 Figure 19. We find that the PSNR of recovered speech (blue bars) varies from $19.1dB$ to $23.5dB$ with a fluctuation of $4.4dB$, and the STOI (red bars) varies from 0.59 to 0.70. 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 [37]. The results indicate that mmPhone can recover intelligible speech with a slight intelligibility fluctuation among different loudspeakers.

VI. Related Work

A. Vibrometry-based Speech Eavesdropping

Non-acoustic sensors, such as motion sensors [3]–[6], [38], wireless signals [7], [39], [40], lidars [8], [41], high-speed cameras [9], vibration motors [10], and hard drives [11] can measure the physical vibration to recover sound information. When targeting surrounding objects, the sound-induced vibration on objects in normal conversations (60-70dB) can be extremely delicate, requiring μm-level resolution sensors for vibration measurement, such as a laser. However, the rigidness and transparency of vibrating objects can significantly impact the performance of laser-based methods [8]. Soundproof materials, such as wool board and glossy plywood, can degrade the performance but cannot prevent mmPhone due to mmWave's penetrating property. Wei *et al.* [7] proposed a radio-based vibrometry leveraging the multipath effect in the WiFi band to measure the loudspeaker vibration for sound retrieval. Our work relies on different principles and eavesdrops on the propagating sound waves rather than the vibration of the loudspeaker. Wei's work leveraged wireless traffic in 2.4GHz band, which is prone to be detected and raise the awareness of the user. Our used attack device transmits fast chirps and operate in the band of 77-81GHz. To detect malicious signals, the user requires a strict synchronization with the adversarial device and pre-knowledge about the operating frequency of the malicious device, which is unlikely to happen in real world.

B. mmWave-based Speech Recovery

mmWave is raising more and more attention in both security areas and noise-resistent speech applications [42]–[48]. Xu *et al.* [45] proposed a noise-resistent speech recovery scheme based on mmWave sensing. Liu *et al.* [46] proposed a multi-modal speech recognition system with a high recognition accuracy and robustness in real world by fusing mmWave and audio signals. Li *et al.* [47] developed a noise-resilient user authentication system based on interrogating users' vocal vibration. Hu *et al.* [48] proposed a mmWave-based speech eavesdropping scheme which could reconstruct the original audio. Our work focus on a new acoustic side-channel, leveraging the mmWave-characterized piezoelectric effect for eavesdropping.

VII. Countermeasures

The defense methods can be two folds. First, blocking or interfering with the mmWave can defend against mmPhone, such as deploying electromagnetic shielding materials around the room and jamming with mmWave signals. However, the jamming method requires the user to know the operating frequency of malicious devices. Second, considering that mmPhone eavesdrops on the propagating sound waves for speech recovery, it is an efficient countermeasure to prevent the sound waves from propagating through the air, such as wearing a headset or earphone when people are involved in an online conversation rather than playing the audio on loudspeakers.

VIII. Conclusion

This paper presents *mmPhone*, a novel eavesdropping system recovering speech emitted by loudspeakers in a soundproof room. We built a methodology to decode the speech via reflected mmWaves from a piezoelectric film and proposed an enhancement scheme to improve the speech quality and intelligibility. The results of extensive experiments indicate that mmPhone can recover high-quality and intelligible speech with digit recognition above 93%.

Acknowledgment

This work is supported by National Key Research and Development Program of China under grant 2020AAA0107700, National Natural Science Foundation of China under grant 62032021, 61772236, 61972348, 62172359, 62102354, Zhejiang Key R&D Plan under grant 2019C03133, Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang under grant 2018R01005, Fundamental Research Funds for the Central Universities under grant 2021FZZX001-27, Research Institute of Cyberspace Governance in Zhejiang University.

REFERENCES

[1] Businesswire, "Video calls fast becoming as popular as voice calls, reaching almost universal adoption for social use, according to vonage study," https://www.businesswire.com/news/home/20190117005173/en, 2019, [Online; accessed 17-July-2021].

[2] Parks Associates, "Voice and video calls more than trapled during covid-19 pandemic," https://www.parksassociates.com/blog/article/pr-08262020, 2020, [Online; accessed 17-July-2021].

[3] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 2014, pp. 1053–1067.

[4] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer," in *Proceedings of the Network and Distributed Systems Security (NDSS) Symposium*, 2020, pp. 23–26.

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

[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 *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, pp. 288–299.

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

[8] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: eavesdropping via lidar sensors," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 354–367.

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

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

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

[12] C. Cochard, T. Spielmann, and T. Granzow, "Dielectric tunability of ferroelectric barium titanate at millimeter-wave frequencies," *Physical Review B*, vol. 100, no. 18, p. 184104, 2019.

[13] G. Srinivasan, A. Tatarenko, and etc., "Microwave and mm-wave magnetoelectric interactions in ferrite-ferroelectric bilayers," *The European Physical Journal B*, vol. 71, no. 3, pp. 371–375, 2009.

[14] P. M. Morse and K. U. Ingard, *Theoretical acoustics*. Princeton university press, 1986.

[15] Q. Zhang, V. Bharti, and G. Kavarnos, "Poly (vinylidene fluoride)(pvdf) and its copolymers," *Encyclopedia of Smart Materials*, 2002.

[16] V. S. Bystrov and etc., "Molecular modeling of the piezoelectric effect in the ferroelectric polymer poly (vinylidene fluoride)(pvdf)," *Journal of molecular modeling*, vol. 19, no. 9, pp. 3591–3602, 2013.

[17] C. Hadnagy, *Social engineering: The art of human hacking*. John Wiley & Sons, 2010.

[18] J. Ahrens, *Analytic methods of sound field synthesis*. Springer Science & Business Media, 2012.

[19] D. K. Cheng *et al.*, *Field and wave electromagnetics*. Pearson Education India, 1989.

[20] J. Benesty, J. Chen, and etc., "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[21] O. Lapteva, *Speaker Perception and Recognition. An Integrative Framework for Computational Speech Processing*, 2011.

[22] W. Klippel, "Loudspeaker nonlinearities – causes, parameters, symptoms," *Journal of The Audio Engineering Society*, october 2005.

[23] "Formant," https://en.wikipedia.org/w/index.php?title=Formant&oldid=1031036735, 2021, [Online; accessed 14-July-2021].

[24] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5609–5613.

[25] A. Vaswani, N. Shazeer, and etc., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[26] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," 2013, pp. 1–4.

[27] P. Bachhav, M. Todisco, and N. Evans, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5429–5433.

[28] P. Bachhav, M. Todisco, and N. Evans, "Exploiting explicit memory inclusion for artificial bandwidth extension," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5459–5463.

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

[30] M. Morise, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[31] P. Pearson, "Sound sampling," 1993.

[32] P. Demonte, "Harvard speech corpus—audio recording 2019," *University of Salford Collection*, 2019.

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

[34] V. Troubleshooter, "The open speech repository," 2010.

[35] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 301–315.

[36] C. Taal, R. Hendriks, and etc., "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[37] Swans, "Hivi m200mkiii," [Accessed 10-July-2021]. [Online]. Available: https://swanspeakers.com/product/m200mkiii/

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

[39] 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 *SenSys'20*, 2020, pp. 1–14.

[40] W. McGrath, "Technique and device for through-the-wall audio surveillance," Oct. 6 2005, US Patent App. 11/095,122.

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

[42] 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 *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 315–328.

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

[44] 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 Transactions on Information Forensics and Security*, vol. 16, pp. 3199–3214, 2021.

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

[46] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren, "Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 97–110.

[47] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren *et al.*, "Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 312–325.

[48] P. Hu, Y. Ma, S. S. Panneer, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 5 2022.