mmEve: Eavesdropping on Smartphone's Earpiece via COTS mmWave Device

Chao Wang[1,2], Feng Lin[1,2*], Tiantian Liu[1,2], Kaidi Zheng[1,2], Zhibo Wang[1,2], Zhengxiong Li[3],
Ming-Chun Huang[4], Wenyao Xu[5], Kui Ren[1,2]
[1]Zhejiang University, Hangzhou, China
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China
[3]University of Colorado Denver, Denver, Colorado, USA
[4]Duke Kunshan University, Kunshan, China
[5]SUNY Buffalo, Buffalo, New York, USA
{wangchao5001,flin,tiantian,zhkheix,zhibowang,kuiren}@zju.edu.cn
zhengxiong.li@ucdenver.edu,mh596@duke.edu,wenyaoxu@buffalo.edu

ABSTRACT

Earpiece mode of smartphones is often used for confidential communication. In this paper, we proposed a remote(>2m) and motion-resilient attack on smartphone earpiece. We developed an end-to-end eavesdropping system *mmEve* based on a commercial mmWave sensor to recover speech emitted from smartphone earpiece. The rationale of the attack is based on our observation that, *soundwaves* emitted from the smartphone's earpiece have a strong correlation with reflected *mmWaves* from the smartphone's rear. However, we find the recovered speech suffers from the sensor's self-noise and smartphone user's motion which limit attack distance to less than 2m, causing limited threats in real world. We modeled the motion interference under mmWave sensing and proposed a motion-resilient solution by optimizing the fitting function on I/Q plane. To achieve a practical attack with reasonable attack distance, we developed a GAN-based denoising scheme to eliminate the noise pattern of the sensor, which boosted the attack range to 6-8m. We evaluated mmEve with extensive experiments and find 23 different models of smartphones manufactured by Samsung, Huawei, etc. can be compromised by the proposed attack.

CCS CONCEPTS

• **Security and privacy** → **Mobile and wireless security**; • **Human-centered computing** → **Smartphones**.

KEYWORDS

Eavesdropping, smartphone, earpiece, mmWave sensing

ACM Reference Format:
Chao Wang[1,2], Feng Lin[1,2*], Tiantian Liu[1,2], Kaidi Zheng[1,2], Zhibo Wang[1,2], Zhengxiong Li[3], Ming-Chun Huang[4], Wenyao Xu[5], Kui Ren[1,2]. 2022. mmEve: Eavesdropping on Smartphone's Earpiece via COTS mmWave Device. In *The 28th Annual International Conference On Mobile Computing And Networking (ACM MobiCom '22), October 17–21, 2022, Sydney, NSW, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3495243.3560543

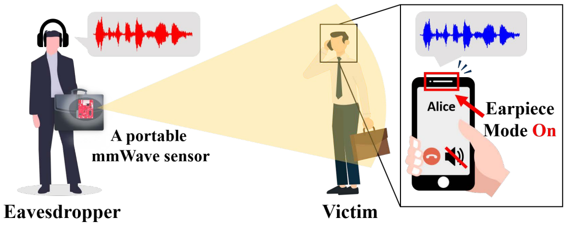

Figure 1: An eavesdropper can use a portable mmWave sensor to recover speech of a victim's phone call remotely when the victim uses the *Earpiece Mode* of his/her smartphone.

1 INTRODUCTION

According to Statista's report, the number of smartphone users has surpassed 7 billion up to 2021 and is forecast to further grow by several hundred million in the next few years [18]. People use smartphones for their daily voice communications, so the speech security of smartphones is raising more and more attention.

Previous studies have revealed side-channel attacks on loudspeakers leveraging non-acoustic sensors, such as lasers [32, 44], high-speed cameras [12], vibration motors [43], hard drives [24], optical sensors [33, 34], RF signals [50, 52, 53], electromagnetic radiation [11], and motion sensors [4, 29]. In recent years, researchers found motion sensors on smartphones can recover speech emitted by the inbuilt loudspeaker of the same smartphone [5, 6]. These attacks reveal speech risks on the **loudspeaker mode** of smartphones. However, another more often-used speech mode of smartphones, i.e., the **Earpiece Mode**, is rarely studied.

The earpiece of a smartphone is often mounted on the top area of the smartphone's motherboard [40, 41]. Figure 1 shows a common case to guarantee speech confidentiality of a phone call or voice message, i.e., the user disables the loudspeaker mode, holds the smartphone to his/her ear, and listens to speech emitted from the earpiece. Compared with the loudspeaker of smartphones, a key feature of the earpiece is that the sound pressure level (SPL) of its emitted speech is far lower to avoid soundwave-propagation

[*] Feng Lin is the corresponding author.

ACM MobiCom '22, October 17–21, 2022, Sydney, NSW, Australia
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9181-8/22/10…$15.00
https://doi.org/10.1145/3495243.3560543

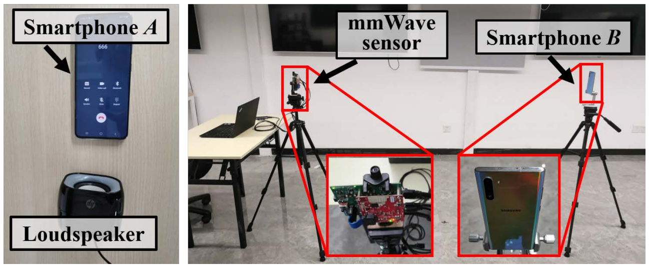

Figure 2: Preliminary study. We used Smartphone A to call Smartphone B. A loudspeaker played human utterances and chirp audio towards Smartphone A.

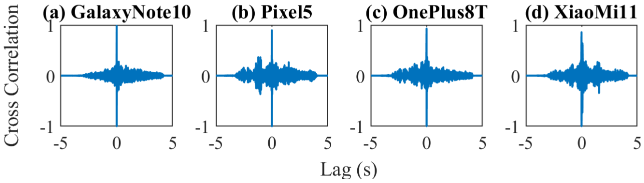

Figure 3: The sharp peaks of the cross correlation when *lag* = 0 indicate the strong correlation between the mic-recorded and mmWave-recovered speech.

through the air for confidentiality consideration, which is hardly perceived by acoustic sensors or humans from several meters away.

In this paper, we aim to investigate the speech security of smartphones' earpiece mode and answer the following questions: ***Is the earpiece mode of smartphones secure enough to ensure the speech security? If it is not, is it possible to recover intelligible speech via a portable and easy-to-acquire attack setup? Can the attacker achieve a practical attack considering a reasonable attack distance(>2m) and the handheld target in real life?*** Based on our observation, we find that the reflected mmWave signals from smartphone's rear has a strong correlation with the emitted speech by the earpiece, which can be a side channel of confidential speech. However, we also investigated that the sensing distance was limited to less than 2m so as to acquire enough signal-to-noise ratio for audible speech recovery. What's worse, the quality of recovered speech can be severely interfered with by users' movement when they hold the smartphone. So we aim to solve the challenges of attack distance and human motion and develop an end-to-end attack system to achieve a practical side-channel attack.

Our work is mainly carried out in the following aspects. We first identified the side channel by the correlation between the speech emitted from the smartphone's earpiece and the reflected mmWave from the smartphone's rear via a commercial mmWave sensor. Then we investigated the characteristics of the side channel in a controlled environment and demonstrated the limitation of the attack, i.e., limited attack distance and motion interference. Furthermore, we detailed our software-based solution (i.e., **mmEve**) to enlarge the attack distance (6-8m) and eliminate the interference of smartphone users' motion. Specifically, 1) we modeled the static and body-motion components of the demodulated mmWave signals on the I/Q plane and proposed a segmental optimization and outlier detection mechanism to eliminate these phase components. 2) We proposed a GAN-based denoising network to characterize the self-noise pattern of hardware components and improve the intelligibility of recovered speech, which boosted the sensing distance of the used commercial off-the-shelf (COTS) mmWave sensor to 6-8m. We performed experiments on 23 different models of smartphones to validate the performance of our proposed attack system. The results indicated that mmEve can recover intelligible speech with an enlarged distance of 6-8m and resilient to the smartphone user's motion. The recovered speech can be distorted and unrecognizable by human but by the machine. To further investigate the threats, we performed speech recognition to recognize the distorted speech. Overall, our contributions are summarized as follows:

- Our work reveals a remote(>2m) and motion-resilient eavesdropping on smartphone's earpiece, i.e., a remote adversary can recover intelligible speech via a COTS mmWave device. The attack does not require any installed malware on targeted smartphones.
- We solved several technical challenges and proposed an end-to-end system to achieve the practical attack with speech recovery, including a dynamic clutter suppression method to eliminate human motion interference and a GAN-based denoising scheme to boost the attack range to 6-8m.
- We performed extensive experiments to evaluate the proposed attack on different smartphone models manufactured by Samsung, Huawei, Apple, etc. We find that 23 different models of smartphones can be compromised by the proposed attack for speech recovery.

2 RELATED WORK

2.1 Attack on Smartphone Speaker

Recent studies have revealed that motion sensors can capture sound-induced vibrations and leak speech information [4–6, 29, 57]. State-of-the-art works [5, 6] reveal that motion sensors built in smartphones can compromise the speech emitted from the same smartphone's loudspeaker. Although these attacks pose great threats to the smartphones' speech, they require preinstalled malware to obtain data from the targeted smartphone and rely on pre-collected data from the targeted smartphone to train a user-specific model for speech recognition or reconstruction. In this paper, we seek a general attack without assumptions of preinstalled malware or a user-specific model. Another main difference from the aforementioned motion sensor-based works is that we target another speech mode of the smartphone, i.e., the earpiece mode which is often used in people's daily life. Recently, Basak *et al.* [7] used mmWave

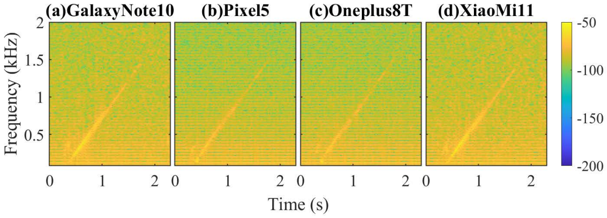

Figure 4: The results of the chirp-audio recovery.

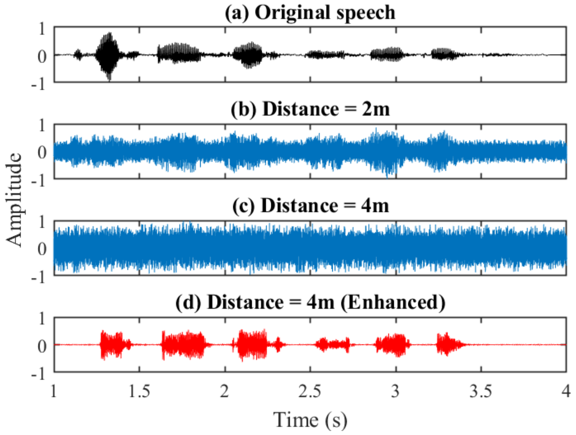

Figure 5: (c) Recovered speech from a targeted smartphone (4m away) is overwhelmed by the noise. (d) The speech quality can be improved significantly with our proposed solution in Section 6.4.

signals to sense the static smartphone when it played sound via the earpiece. Our work has four main differences. First, we focus on recovering continuous sentences while Basak's focuses on reconstructing and recovering isolated words on a predefined vocabulary. Second, we proposed a dedicated solution to boost the limited attack distance from 2m to 6-8m and solve the challenge of human motion but Basak's did not solve these problems (limited distance of less than 2m and targetting on a static smartphone). Third, we performed extensive experiments on 23 different models of smartphones while Basak's only tested on 2 smartphone models. Fourth, we used two different public datasets to train and test our model, respectively. Basak's used audio samples from all tested speakers to train the model. Their training and testing words are from the same person which is a stronger assumption than ours.

2.2 Attack on External Loudspeakers

Vibration-based: Many studies have investigated the side effect of the loudspeaker or sound-induced vibrations for eavesdropping. mmVib [22] presented a mmWave-based method for micro-level vibration monitoring on industry machines. Vibrating objects without movements/location changes (e.g., a hanging bulb or a static loudspeaker placed on the desk) [12, 27, 32, 34, 44, 51] can leak sound information. Adversaries can leverage preinstalled malware or pre-collected data from the victim's device to recover sound by non-acoustic sensors [4, 29, 44], wireless signals [52, 53], reprogrammed audio port [17], hard drive [24], and vibration motors [43]. Compared with these vibrometry-based work that poses threats to external loudspeakers, **our work mainly has three differences: (1)** Our proposed attack requires no prior knowledge (e.g., any malware or pre-collected data) about the target. **(2)** The vibrometry-based attacks often require a large sound pressure (70-110dB) of loudspeakers or a close distance (centimeters) between the vibrating object and the loudspeaker to induce the vibration. But our target, i.e., the small earpiece of smartphones, has an extremely lower sound volume (48-50dBSPL) than large external loudspeakers (60-110dBSPL). **(3)** Our target is a moving sound source due to the movement of the smartphone holder (human). The movements cause great challenges to capture delicate vibrations induced by the earpiece for speech recovery.

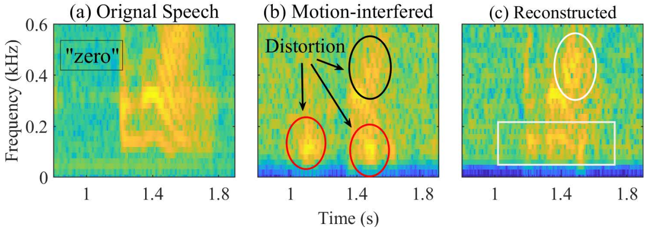

Figure 6: Compared with (a) original spectrogram, the low-band (60-180Hz) and high-band of (b) recovered speech is overwhelmed by the victim's motion as the red ellipses indicate. (c) The motion interference is eliminated by the proposed method (Section 6.3).

TEMPEST-based: Recently, Nassi *et al.* [33] revealed an optical TEMPEST attack that power indicators on the static loudspeaker or connected hubs can leak speech information due to the variation of power consumption. Via a telescope, sound can be recovered from a distance of 15-35m. Nassi's work targeted *static* objects that have power indicators. We contribute to eavesdropping on a moving target (i.e., a handheld smartphone by the user) free of power indicators. Compared with the visible light of the optical side channel, mmWave suffers more attenuation during the air propagation. Due to the limited transmitting power (18mW) and resolution of our used COTS mmWave sensor, we found that intelligible speech can only be recovered within an attack distance of 2m. So one of our contributions lies in boosting the performance of COTS mmWave sensors and recovering intelligible speech with an attack distance up to 6-8m which is larger than a normal distance between strangers in life. Besides, an advantage of mmWave compared with visible lights is that mmWave can easily penetrate opaque objects (e.g., a thin paperboard) which hides the device from victims' view for stealthy eavesdropping. Choi *et al.* [11] revealed an EMR (electromagnetic radiation) TEMPEST attack against devices that have a mixed-signal system-on-chip, such as earbuds. Our work eavesdrops on the smartphone itself and focuses on a different threat model that earbuds are not used.

3 BACKGROUND

Frequency-modulated continuous-wave (FMCW) mmWave radars have been widely applied in automotive and industrial applications. By transmitting a series of FMCW signals (called chirps), the radar receives and demodulates the reflected wave signals to produce the intermediate-frequency (IF) signals composed of the In-phase and Quadrature-phase (I/Q) data. Then the range and displacement of the objects can be calculated by applying range-FFT on IF signals.

Range Estimation: Take IF signals as $Asin(2\pi f_0 t + \phi_0)$, then the estimated range of the object d has following relationship with the frequency f_0 of IF signals: $d = \frac{cf_0}{2S}$, where c and S are the light speed in a vacuum and the slope of the transmitted FMCW. A mmWave device with 4 GHz bandwidth can achieve a range resolution of 3.75cm. Leveraging the range estimation, the attacker

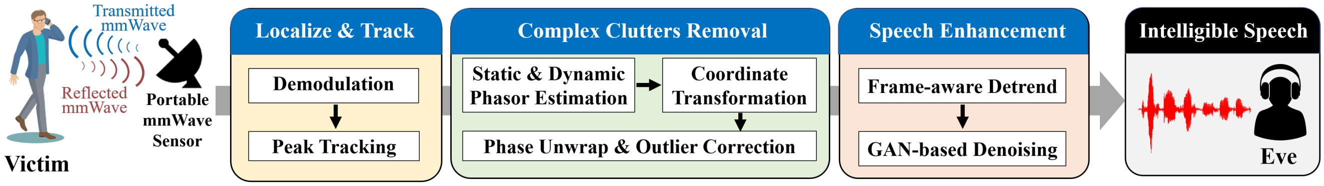

Figure 7: System framework of mmEve.

can locate the targeted smartphone and steer the mmWave beam to the specific direction to transmit and receive mmWaves remotely. With a horizontal angle of ±50 degrees and an elevation angle of ±20 degrees [20], the large Field-of-View (FoV) of the used mmWave sensor contributes to compromising smartphones in a large attack area with little effort to align the beam to the specific area of a **moving** target.

Audio Extraction from mmWave Signals: Take the small displacement of the smartphone's rear surface resulting from the earpiece as Δd, then Δd can be calculated by $\Delta d = \frac{\lambda \Delta \phi}{4\pi}$, where $\Delta \phi$ is the phase change of the IF signal corresponding to the object, $\lambda \approx 4mm$ is the wavelength of the transmitted FMCW. Every derived phase change of demodulated chirps can measure a corresponding displacement Δd of the object. The audio can further be extracted from the phase ϕ where ϕ is the phase of the Range-FFT point corresponding to the smartphone. The sampling rate of the mmWave sensor can be calculated by $f_s = f_{chirp}$, where f_{chirp} is the chirp rate of the mmWave device (10,200 chirps per second in this paper). Considering that intelligible speech can be distinguished from a bandwidth of 2.5 kHz [38], the f_s of the mmWave device satisfies the *Nyquist theorem* for intelligible speech recovery theoretically ($f_s = 10.2kHz > 2 * 2.5kHz$). However, the speech quality and intelligibility can also be affected by factors like SNR of received signals and clutter interference in remote eavesdropping.

4 THREAT MODEL

Attack Scenario: We consider a scenario when a victim makes a smartphone call or listens to voice messages. The victim puts his/her ear close to the smartphone **Earpiece** to ensure the confidentiality of the speech contents which can be secrets related to the victim's privacy. An adversary, interested in the speech contents, leverages a portable mmWave device to launch remote eavesdropping.

Attack Goal: Considering that audible speech is often taken as the first-hand data for speech analysis, the attacker's goal is to recover audible and intelligible speech. To ensure stealthiness, the attack aims to eavesdrop from several meters (>2m) away from the victim to avoid the victim's awareness.

Assumption: We assume the attacker and the victim's smartphone are at the same scene without blockages in between, such as an open square or an office, so the attacker can transmit mmWave towards the victim's smartphone. Considering that the victim's head can block the mmWave towards the smartphone when he/she holds the smartphone close to the ear, we focus on the rear surface of the victim's smartphone to achieve the eavesdropping in this paper. **Note that we do not assume the attacker has any installed malware or prior knowledge about the targeted smartphone.**

5 PRELIMINARY STUDY

5.1 Characterizing the Side Channel on Smartphones with A mmWave Sensor

Correlation analysis: We used a COTS mmWave sensor (AWR1843-Boost) to sense the smartphone's rear when the smartphone's earpiece plays audio signals. For each tested smartphone, we used two tripods to hold the mmWave device and the smartphone respectively as shown in Figure 2. We used *Smartphone A* to call the fixed *Smartphone B* (i.e., the target). And then the loudspeaker near *Smartphone A* played audio so that *Smartphone B* could also replay the same audio via its earpiece mode. To avoid unwanted vibration sources, we placed the loudspeaker and *Smartphone A* in a conference room about 200m away from *Smartphone B* and the mmWave sensor. We controlled the loudspeaker-connected laptop to play the audio via a remote desktop software [19]. The loudspeaker near *Smartphone A* played the first sentence of Harvard sentences [13]: *The birch canoe slid on the smooth planks.* In the meantime, we used the mmWave sensor to transmit mmWave towards the *Smartphone B*'s rear and extracted speech audio from the phase of reflected mmWave signals as introduced in Section 3. Unfortunately, we found that the extracted audio from mmWave signals had poor intelligibility and is corrupted by noises. To quantify the similarity between the mmWave-recovered and original audio, we calculated cross correlation between the two audio[24]:

$$(f * g)[n] \triangleq \sum_{m=-\infty}^{\infty} \overline{f[m]} \cdot g[m+n], \tag{1}$$

where $\overline{f[m]}$ is the complex conjugate of $f[m]$ and n is the time displacement between the two series. A larger value of Eq. 1 indicates a stronger correlation between the two audio series. Considering that we have aligned the two series in the time domain for better display, there should be a peak when $n = 0$ (lag=0) if the audio recovered by mmWave has a correlation with the audio played by the earpiece of the smartphone. The results are shown in Figure 3. We observed that **the recovered speech via mmWave has a significant correlation with the original speech audio in the controlled experiment (with an attack distance of 2m)**. Although the speech has poor intelligibility, the results confirm the existence of the side channel in the studied smartphones' earpiece.

Frequency Response of the Side Channel. To further investigate the characteristics of the side channel, we placed the smartphone 1m away from the mmWave sensor and study the frequency response via chirp audio. We kept other settings unchanged and made a phone call by *Smartphone A* to the targeted *Smartphone B*. We played the audio chirp (80Hz-2kHz) towards *Smartphone A* via

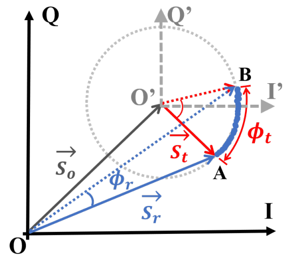

Figure 8: An illustration of basics of the clutter suppression.

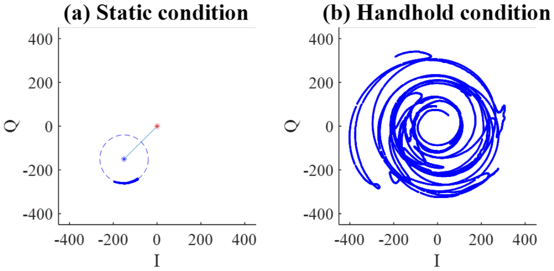

Figure 9: Trajectory of $\vec{S_r}$ on I/Q plane is (a) a regular arc for a static smartphone fixed on a tripod but (b) helical curves for a handheld smartphone.

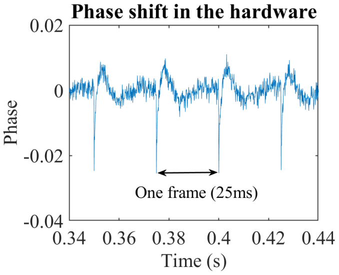

Figure 10: There is a step at the junction of every two successive frames.

the loudspeaker so that the targeted *Smartphone B* could replay the audio chirp via its earpiece during the call. During the experiment, we used the mmWave device to eavesdrop on the *Smartphone B*'s earpiece. The recovered audio from each targeted *Smartphone B* is shown in Figure 4. **(1) We found that there were harmonics of 40 Hz (the yellow stripes in the spectrograms) covering from 80Hz to 2kHz.** These harmonic noises dominate the spectrogram of the recovered audio (especially the frequency band below 500 Hz), which would damage the spectrum structure of human voice and interfere with speech intelligibility. In Section 6.4, we will give an analysis of the cause of these harmonic noises and eliminate them. **(2) We also observed that the upper frequency of the recovered audio signals can reach 1.7kHz.** Considering that the non-tonal language like English is dominantly distinguished by formants of vowels and consonants, the bandwidth of 80Hz-1.7kHz has covered F0 formants of all vowels and 62.5% consonants, and F1 formants of 68.8% vowels [9]. The strong low-frequency components of the formants determine the intelligibility of the spoken phonemes [43]. Besides, works have proved that a band of 1.5kHz of human speech contains abundant information related to both the secret (e.g., digit password) and the human speakers' gender and identity [5, 6, 44]. So such bandwidth of recovered audio can cause great threats to human speech.

5.2 Robustness Study of the Side Channel

5.2.1 ***Attack Distance.*** We changed the sensing distance d between the smartphone and the mmWave device and kept other settings unchanged. Figure 5 shows the recovered audio when the sensing distance changes from 2m to 4m. We can observe that when the distance increases to 4m, the recovered speech is completely overwhelmed by the noise which results in poor speech quality and intelligibility. In other words, the sensitivity of the side channel can degrade significantly in a long distance. **Root-cause analysis:** According to $\Delta d = \frac{\lambda \Delta\phi}{4\pi}$ in Section 3, the speech is recovered from the phase of demodulated mmWave signals. Thus, the sensitivity of the side channel can be characterized by the *phase resolution* $\Delta\phi$ of the COTS mmWave sensor. To recover the audio with little distortion, the phase of two demodulated mmWave chirps should satisfy $\phi_1 - \phi_2 > \Delta\phi$, i.e., a smaller $\Delta\phi$ indicates a better sensing resolution. However, the phase resolution is determined by the SNR of received mmWave signals according to $\Delta\phi = \frac{\beta}{\sqrt{SNR}}$ [55], where β is the SNR coefficient for mmWave devices. We denote the attack distance between the mmWave device and the smartphone as d, then the SNR of received signal [39] can be calculated as

$$SNR = \frac{\alpha \lambda^2 G_{Tx} G_{Rx}}{(4\pi)^3 d^4 F}, \tag{2}$$

where α is the coefficient related to the mmWave hardware configuration, λ is the wavelength of transmitted mmWave, G_{Tx} and G_{Rx} are the gain of transmitter and receiver of the mmWave sensor respectively, F is the noise floor of the sensor. According to Eq. 2, we can find that **the increasing distance can cause significant degradation on the SNR of received mmWave, and then worsen the sensing resolution of the COTS mmWave sensor**. To mitigate the distance limitation, an intuitive solution is to increase the gain of the sensor G_{Tx} and G_{Rx} with a more powerful amplifier which requires a customized hardware design. *To deeply understand the side channel posed by the COTS mmWave sensor, we choose a software-based solution (Section 6.4) to decrease the noise floor F of the sensor and boost the sensing resolution of the COTS sensor without hardware changes.*

5.2.2 ***Human Artifacts.*** In a realistic condition, the smartphone is often used in a mobile condition, e.g., a user holds the smartphone to his/her ear and listens to the speech emitted from the earpiece. So there can be involuntary motion when the user holds the smartphone, such as arm moving and body wiggles. To investigate the impact of human artifacts on the side channel, we asked a male to hold a smartphone (GalaxyS20) close to his ear with natural human motion during the call. The spectrogram of recovered audio is shown in Figure 6(b). The area with brighter color indicates a higher signal power. More areas with similar color between two figures indicates higher similarity. **We found that the interfered frequency band of human motion can cover 60-180Hz, indicated by the red ellipse-marked areas.** If not properly mitigated, the speech contained in the phase signal can be overwhelmed and distorted by the human artifacts, which destroy the intelligibility of the recovered speech. Besides, as denoted by the black and red ellipse-marked areas, the power of higher band can degrade due to the domination of the human interference. However, it is nontrivial to suppress such interference because the interfered frequency band overlaps with the fundamental frequency of the human voice (85-255Hz). The components of human voice can also be eliminated

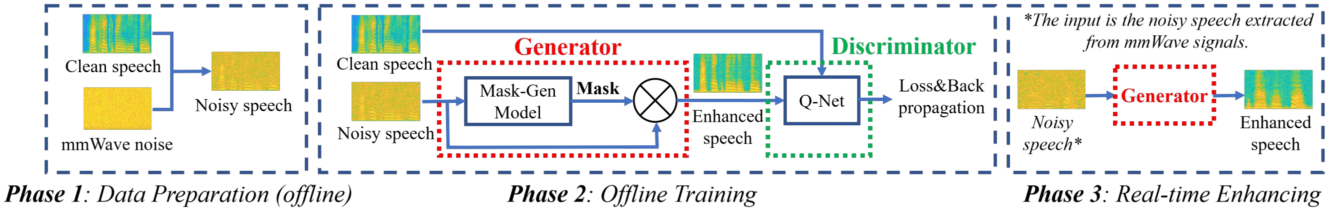

Figure 11: We first generate a training dataset by combining a public voice dataset and mmWave-noise dataset (Phase 1) and then train the Mask-Gen model based on a GAN-based neural network (Phase 2). In the attack phase (Phase 3), we use the *Generator* to denoise noisy speech recovered by mmEve.

when the interfered band is filtered by a digital filter, which will damage the vital fundamental frequency of human speech. So we further investigated the impact of human motion under mmWave sensing. Based on an analysis on the I/Q plane in Section 6.3, we proposed a dynamic clutter suppression method to eliminate motion interference and achieve accurate speech recovery.

6 ATTACK DESIGN

In this section, we introduce our end-to-end attack system **mmEve** for intelligible speech recovery from the targeted smartphone's earpiece. The framework is shown in Figure 7.

6.1 Target Localization and Tracking

When the attacker steers the mmWave beam to the victim coarsely, the distance of the target can be measured according to $d = \frac{cf_0}{2S}$ in Section 3. To ensure stealthy eavesdropping, the attacker should be able to track the target continuously. Note that the attacker is not required to change the position of the mmWave device as long as the targeted victim is within the FoV of the mmWave device. We denote S as the demodulated signal. We first apply range-FFT on every demodulated chirp in S where the peaks in the spectrum indicate detected objects. Due to the fact that the speed of human movement is far less than the fast chirps (10,200 chirps per second), the peak corresponding to the victim, i.e., the location loc_i, can only shift to the adjacent FFT points in the spectrum. By tracking the peak's location in the spectrum [8], we can derive the locations of the victim.

6.2 Static Clutters Suppression

We take the mmWave signals reflected by static objects near the victim as **static clutters.** When the distance of static objects has the same range as the victim to the mmWave device, the static objects can be detected in the same range-bin with the victim. If not suppressed properly, the clutters can degrade the sensing resolution [22] of the side channel, which can be modeled as follows. Here we denote the phasors corresponding to the targeted victim and static objects in the same range-bin as $\overrightarrow{S_t}$ and $\overrightarrow{S_o}$, respectively. The phasor $\overrightarrow{S_r}$ derived from the range-bin of the IF signals can be taken as the superposition of these two phasors, formulated as

$$\overrightarrow{S_r} = \underbrace{A_t e^{j\cdot 2\pi f(\tau_{init}+\Delta\tau_h(t)+\Delta\tau_s(t))}}_{\overrightarrow{S_t}} + \underbrace{\sum\nolimits_i A_i e^{j\cdot 2\pi f_i \tau_i}}_{\overrightarrow{S_o}}, \tag{3}$$

where τ_{init} is chirp delay induced by static parts of human body and nearby objects, $\Delta\tau_h$ is the disturbance of human motion, such as arm moving while holding the smartphone, $\Delta\tau_s$ indicates components induced by the delicate vibration on the smartphone rear.

Based on the above problem formulation, the key goal is to eliminate phasor $\overrightarrow{S_o}$ and estimate the components corresponding to the smartphone's vibration in phasor $\overrightarrow{S_t}$ for speech recovery. An illustration of the impact of static clutters on the vibration measurement is represented in Figure 8. When the smartphone vibrates or the victim moves his/her arm while holding the smartphone during a phone call, the phase ϕ_t of the phasor $\overrightarrow{S_t}$ rotates between A and B on the I/Q plane, where $\phi_t = 2\pi f(\tau_{init} + \Delta\tau_h(t) + \Delta\tau_s(t))$ which contains the speech information. However, due to the existence of the static clutters, i.e., the phasor $\overrightarrow{S_o}$, the actual measured phase ϕ_{meas} is not equal to ϕ_t but ϕ_r as shown in Figure 8. According to $\Delta d = \frac{\lambda\Delta\phi}{4\pi}$ in Section 3, this results in a consequence that the measured displacement Δd is far smaller than the actual value ($\phi_t \geqslant \phi_r$). Based on the above analysis, we can find that if we can estimate the phasor $\overrightarrow{S_o}$ and then eliminate $\overrightarrow{S_o}$ from the samples on the I/Q plane, we can get the ideal measurement of vibration displacement, i.e., $\phi_{meas} = \phi_t$. To achieve this goal, we estimate the center $O'(I_0, Q_0)$ of the circle fitted by the samples on the I/Q plane. Then we get the estimated $\overrightarrow{S_o} = (I_0, Q_0)$. Finally, we further perform a *Coordinate Transformation* from I/Q plane to I'/Q' plane as shown in Figure 8, to get the ideal phase value.

6.3 Motion-resilient Speech Recovery

As analyzed in Section 5.2, the human interference can cause distortion in the recovered speech. The human movement mainly consists of two parts, i.e., the location changes (solved in Section 6.1) and the body movement (such as arm movements and body wiggles while holding a smartphone for calling.) This section focus on solving the impact of the latter. After the static clutters suppression, the processed data samples on the I'/Q' plane can be formulated as

$$\overrightarrow{S_{rr}} = \overrightarrow{S_t} = A_t e^{j\cdot 2\pi f(\tau_{init}+\Delta\tau_h(t)+\Delta\tau_s(t))} \tag{4}$$

To combat the human artifacts, there are two nontrivial problems we need to consider. First, it is likely that the human-motion-induced phase changes $|\phi_h| = |2\pi f\Delta\phi_h| \geqslant \pi$ which will cause *integer ambiguity problem* [47] in the derived phase of $\overrightarrow{S_{rr}}$. Second, the amplitude A_t of the phasor $\overrightarrow{S_{rr}}$ can also change with the human motion. This can pose a great challenge to the circle-fitting introduced in Section

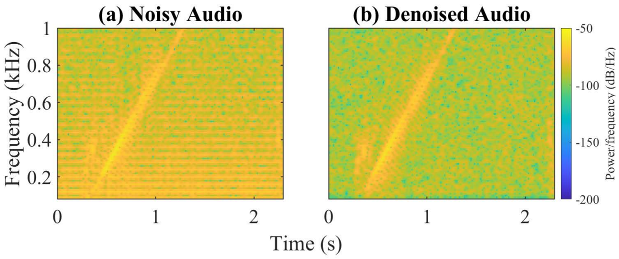

Figure 12: Harmonic noises are eliminated by *Frame-aware Detrend* without damaging audio spectrogram.

6.2 because the resulting trajectory of $\overrightarrow{S_{rr}}$ is not arcs of a circle but helical curves on the I/Q plane as shown in Figure 9(b).

Dynamic Motion Suppression: We model motion suppression as an optimization problem with objective function

$$\mathcal{F}_{obj} = \sum_{n=1}^{N}\left[\sqrt{(I_i - I_0)^2 + (Q_i - Q_0)^2} - R_0\right]^2, \tag{5}$$

where $(I_i, Q_i) \in \mathbb{S}$, $\mathbb{S}$ is the segment of the trajectory of $\overrightarrow{S_{rr}}$. The optimization goal is to find the (I_0, Q_0, R_0) that minimizing the sum of the squared radial deviation. Considering that the speed of human motion can be taken as a constant within a short time, the key idea of the motion suppression is to take every short segment of the trajectory of $\overrightarrow{S_{rr}}$ as an arc of a circle and to find the best circle center for every short segment. We set the length of each segment to 200ms. When the (I_0, Q_0, R_0) for every segment is estimated, the points of the segments are applied to the static clutters suppression as introduced in Section 6.2. The final recovered speech (i.e., the phase information of these translated segments) is derived by connecting the translated segments and applying a *phase unwrapping* [16]. However, a side effect of the segment-based fitting is the induced discontinuity in the junction of two translated successive segments after the phase unwrapping. The discontinuity can induce jitter in the recovered audio, damaging the speech quality and intelligibility. To solve this problem, we propose to apply an *outlier detection and correction* (OutlierDetCo) to the recovered speech signals.

Outlier Detection and Correction: The rationale is that due to the original signal being gradually varied with human motion, we can first apply a window with a size of 1024 on the unwrapped phase and filter out the phase (i.e., the *outliers*) that deviate the median of the window with triple median absolute deviation (MAD) [25]. Then each outlier is replaced by the mean of the phase value in the window. The window slides with no overlaps until all outliers are detected. Finally, the recovered speech is corrected.

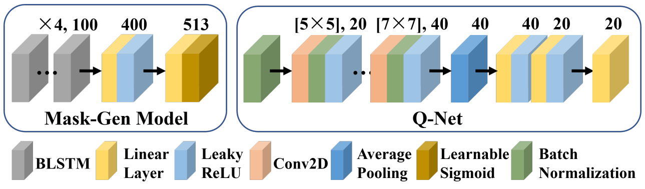

Figure 13: Layers of GAN-based denoising network.

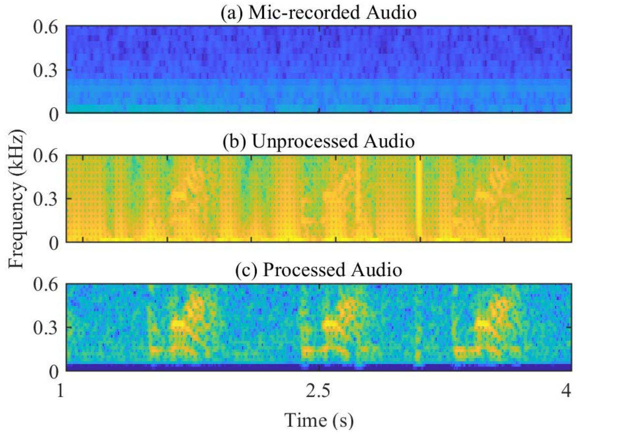

Figure 14: The spectrograms of (a)mic-recorded audio, (b)raw-recovered audio and (c)processed audio by mmEve (Speech: repeating "zero" by three times).

6.4 Speech Enhancement

6.4.1 Frame-aware Detrend. As introduced in Section 5.1, the recovered speech suffers from the harmonic noises of 40Hz which damages the speech intelligibility. Note that this is a nontrivial problem that cannot be solved by a band-pass filter bank as the harmonics overlap with the frequency band of human speech. So We further investigate the phenomena and found that **the harmonic noises result from phase shift of the mmWave sensor**. Specifically, we find that the phase is reset to a specific value at the very beginning of every frame of chirps (the mmWave chirps are transmitted in frame units). Due to the unavoidable phase shift of the mmWave hardware, the derived phases from each frame will drift with time going by, resulting in a step at the junction of two successive frames as shown in Figure 10. According to the *Fourier theorem*, the 40Hz step noise can further induce harmonics of 40Hz. To solve this problem, our key idea is to eliminate the trend of each frame rather than filtering the whole signal. So we apply a *Frame-aware detrend method* based on the polynomial regression [10]. The rationale is to first estimate the low degree polynomial components ($degree = 10$) and then subtract them from the original signal to suppress the trend of phase shift. Figure 12 indicates the effectiveness of this method. We can observe that the harmonics of 40Hz are completely filtered out while the audio components are well retained. After the *Frame-aware detrend* process, a **High-pass Filter** with a cut-off frequency of 80Hz is further applied to eliminate residual low-frequency noise.

6.4.2 Denoising Neural Network. As demonstrated before, the limited SNR of received signals causes poor intelligibility of recovered speech especially when the sensing distance is over 2m. Li et al. [26] proposed a virtual-transceiver solution to improve the sensing range of acoustic sensors. In our studied problem, we focus more on boosting the SNR of the sensor. To solve this problem, we turn to a software-based method to filter out the sensor's self-noise (i.e.,

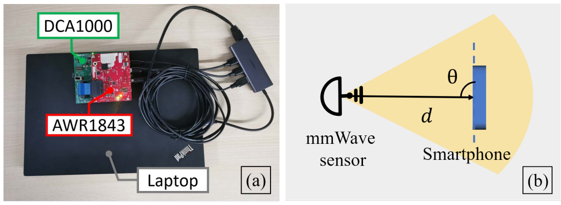

Figure 15: (a) The portable system setup of mmEve. (b) The definition of attack distance d and attack angle θ.

declining the noise floor F). The source of the self-noise can be modeled as a *filter* whose parameters are determined by the whole receiver chain, including antennas and components on chips (e.g, mixers, amplifiers). To combat the noise of the receiver chain, an intuitive solution is to design a corresponding *inverse filter*. However, due to the complexity of the receiving chain, it is impractical to estimate the parameters of the inverse filter by linear filter design methods considering the ubiquitous nonlinearity of the receiving chain components [21, 36, 42]. Considering the nonlinear characteristic of deep neural networks (DNN), we turn to the DNN to estimate the parameters of the *inverse filter* (we call it ***Mask***) which can well characterize both the linearity and nonlinearity of the receiving chain components. The mask can suppress the noise and improve the speech quality and intelligibility, especially when the raw recovered speech by the mmWave device has a poor SNR due to a long attack distance. Existing studies have verified that speech quality and intelligibility can be improved by optimizing objective metrics such as STOI [15]. To improve the speech quality and intelligibility, we include the STOI into the loss function to optimize the network. Furthermore, we apply the generative-adversarial-network (GAN) structure [1] to guarantee the generalization. The speech enhancement contains three phases as shown in Figure 11.

Phase 1: Training Data Preparation. We used a public dataset *VCTK Corpus* [49] and mmWave-noise dataset collected from the mmWave device to generate a training dataset. The mmWave-noise dataset is collected from the mmWave sensor in an open square without moving objects in the background and extracted by the same workflow mentioned before. We resample the mmWave noise from 10.2kHz to 16kHz and combine the two signals, i.e., the public speech signal $s = \{s_1, ..., s_N\}$ and the mmWave noise signal $n = \{n_1, ..., n_N\}$, into noisy speech signals s_n with different SNR according to $s_n = s + \alpha \cdot n$, where $\alpha = \sqrt{E_s/(10^{\frac{snr}{10}} E_n)}$, $E_s = \sum_i s_i^2, E_n = \sum_i n_i^2$, snr is the desired SNR (dB) of synthesized speech which is a random number within $[-9, 9]$.

Phase 2: Offline Training. The *Generator* consists of a *Mask-Gen Model* and a multiply operation. The Mask-Gen Model generates the mask of inputted noisy speech spectrogram and then multiplies the two to generate enhanced speech. The enhanced speech and the corresponding clean speech are fed into the *Discriminator* (Q-Net [14]) to estimate the speech metric score (e.g., STOI) of the enhanced speech. Then the loss is calculated by the mean square error of the estimated and the true scores [15]. The *Generator* and the *Discriminator* are updated alternatively. As shown in Figure 13, the Mask-Gen model has four bidirectional LSTM layers with an input size of 513 and a hidden size of 100 with dropout mechanism (dropout rate=0.1) to avoid over-fitting. The learnable sigmoid is used for frequency-aware compression and improves the performance of speech enhancement [15]. The *Q-Net* contains four two-dimensional convolutional (Conv2D) layers each of which is followed by a Batch Normalization layer and a LeakyReLU layer. The Discriminator output the estimated scores to calculate the loss during training. Detailed parameters are shown in Figure 13.

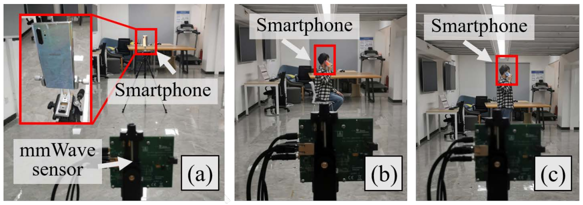

Figure 16: (a) The smartphone is fixed on a tripod. (b) A user sits on a chair holding the smartphone. (c) A user stands/steps back and forth while holding the smartphone.

Phase 3: Real-time Enhancing. Once the offline training finishes, the *Generator* can be deployed on a portable system to enhance the raw recovered speech by mmEve. The raw noisy speech is first resampled to 16kHz and then its spectrogram derived by short-time Fourier transform (STFT) is fed into the Mask-Gen Model to extract a Mask. The spectrogram is further multiplied with the Mask to generate an enhanced spectrogram. Finally, the speech signal is recovered by applying an inverse STFT. An enhanced audio trace with 4m attack distance is shown in Figure 5(d). Figure 14(a)(b)(c) show the spectrograms of mic-recorded, raw-recovered and processed audio by mmEve. We can observe that from a distance of 6m, the microphone cannot record the speech contents but only background noise. By contrast, mmEve can recover speech contents by suppressing the noise and motion interferences in the raw-recovered speech.

7 EVALUATION

7.1 System Setup

The system setup is shown in Figure 15 (a). We use a COTS mmWave sensor AWR1843Boost [20] manufactured by Texas Instruments to transmit and receive mmWave signals. The device is widely applied in automotive and industrial applications. AWR1843Boost has a portable size of $6.5cm \times 8.5cm \times 2.0cm$ and works under a 5V/2.5A power supply. It has an integrated antenna array with 3 Tx and 4 Rx antennas on the same board. It transmits 77-81GHz chirps with a transmitting power of 12dBm. The demodulated chirps are sampled by a general data acquisition board DCA1000EVM and sent to a laptop (ThinkPad T490) for processing. The denoising neural network is implemented with Pytorch and trained offline on a Linux Server with four GeForce RTX 3090 GPUs and then deployed on the laptop for real-time speech enhancement.

7.2 Dataset and Settings

We use the Harvard Sentences in Open Speech Repository (OSR) [13] which is widely used in speech tests, to evaluate mmEve. Note that there is no overlap between the dataset and the VCTK Corpus

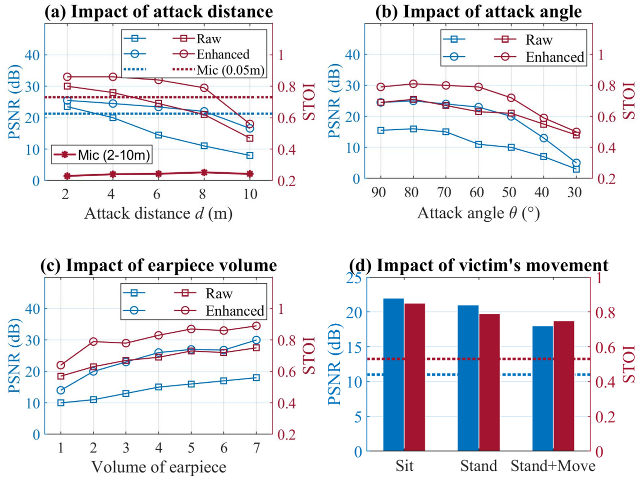

Figure 17: Results of controlled experiments.

(used to train the denoising neural network in Section 6.4) in order to show the generalization of the attack. There are 100 utterances in total including half of male utterances and half of female utterances. We played the utterances via a loudspeaker (50-60dB SPL, the normal SPL of human talking) towards smartphone A to emulate a human talking when smartphone A calls the victim's smartphone B. The communication distance between A and B is larger than 200m. Then we use mmEve to eavesdrop on the smartphone B to recover the speech emitted by B's earpiece. The volume of smartphone B's earpiece mode (Level0-Level7) is set to Level4. The background noise is around 49-50dB SPL measured by a sound level meter from 2m away from the smartphone B. There are 23 different models of smartphones (listed in Figure 18) hired from volunteers included in the experiments. It is ensured the experiments follow the internal review board (IRB) protocol of the host institution.

7.3 Metrics

Speech quality and intelligibility are two significant aspects where the *quality* means how comfortable the speech is, such as clean or noisy) and the *intelligibility* means how comprehensible the speech is for human hearing). In the following pages, we use two metrics to quantify the speech quality and intelligibility, respectively. **1) Peak Signal-to-Noise Ratio (PSNR)** is a commonly used metric to quantify the speech quality [12, 53]. It bounds the human audibility with $0dB$, which means the speech signal with a $PSNR > 0dB$ is audible for human perception [53]. A higher PSNR indicates a better speech quality. **2) Short-Time Objective Intelligibility (STOI)** characterize the intelligibility of human speech with the score within [0,1]. According to Taal's work [48], over 90% of words (Harvard Sentences) can be recognized correctly by humans when STOI>0.7. A higher STOI indicates better intelligibility.

7.4 Controlled Experiment

We performed controlled experiments of different attack distances, angles, and earpiece volumes in a laboratory. The attack distance d and angle θ are defined in Figure 15(b). The smartphones are fixed on a tripod as shown in Figure 16 (a).

7.4.1 Attack Distance. We set the mmWave device towards the targeted smartphone and change the sensing distance from 2m to 10m. The volume of the smartphone's earpiece mode is set to Level 4 (the maximum is Level 7). When the earpiece emits sound, there is no change in the SPL value measured by the sound level meter from 2m away. For a better understanding of the performance of mmEve, we used a professional condenser microphone (Gmtd GM-S801) to collect the sound emitted from the smartphone (GalaxyNote10) earpiece from different distances, i.e., 0.05m, 2m, 4m, 6m, 8m, and 10m. When the microphone is placed near the smartphone's earpiece (0.05m), the PSNR/STOI scores achieve 21.3dB/0.73 as the blue/red dotted lines shown in Figure 17(a) respectively. However, when the distance changes from 2m to 10m, the PSNR of the mic-recorded speech is below -18dB and the STOI score degrades to less than 0.26, in which condition the speech is totally beyond the perception of human hearing. The mic-recorded results indicate that the earpiece-emitting audio over the air has poor intelligibility which guarantees speech confidentiality to some extent. **By contrast**, the mmEve-recovered speech has higher scores of both PSNR and STOI. Although the speech quality (PSNR) and intelligibility (STOI) degrade with the increasing attack distance, the PSNR and STOI scores can be improved to over 20dB and 0.78 via the precessing scheme of mmEve when the attack distance is 8m. When the attack distance increases to 10m, the recovered speech has limited intelligibility (STOI=0.55). However, the eavesdropping performance with an attack distance of 6-8m is enough for a practical attack in the physical world.

7.4.2 Attack Angle. Considering that the orientation of targeted smartphone's rear may change during the phone call due to the victim's movements, we performed controlled experiments to investigate the impact of the smartphone's orientation. The *attack angle* θ related to the smartphone's orientation is defined in Figure 15 (b). We fixed a smartphone on a tripod each time and set the tripod 6m away from the mmWave sensor. We rotated the spindle of the tripod to change the attack angle from 90° to 30°. From Figure 17(b), we can observe that the PSNR and STOI are kept at a high level when the attack angle changes from 90° to 50°, which indicate mmEve can recover speech with high quality and intelligibility when the targeted smartphone changes the orientation within 50°-90°.

7.4.3 Volume. We change the volume of smartphones' earpiece mode from Level 1 to Level 7 and re-conduct the experiments. We set the attack distance to 6m and attack angle within 80°-90°, respectively. The attack results are shown in Figure 17(c). Generally, the PSNR/STOI scores rise with the increasing volume. we can also observe that the PSNR and STOI of recovered speech are respectively larger than 20dB and 0.78 when the volume ranges from Level 2 to Level 7. This indicates that mmEve can cause threats to speech in a wide range of volumes in people's daily life. Considering that a high volume is harmful to human hearing and a low volume can hardly for user perception when using the smartphone, we consider the user takes a medium volume (Level 4) in daily life. So the next experiments take this volume setting unless otherwise specified.

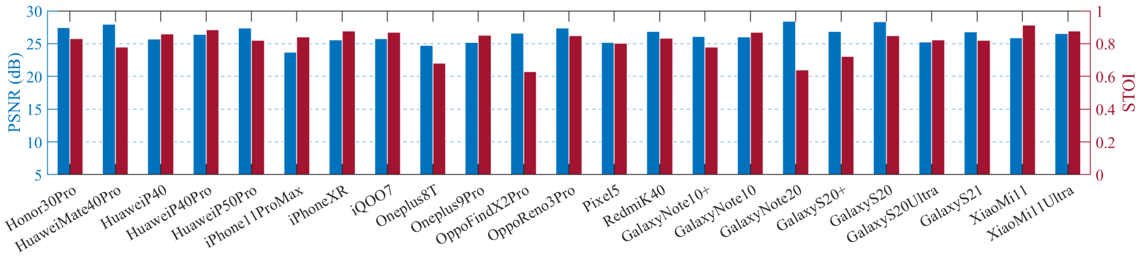

Figure 18: Attack results on 23 smartphones in the Stand+Move condition in Section 7.5 (attack distance: 6m).

7.5 Handhold Experiment

To investigate the performance of mmEve under victim's movements, we asked volunteers to hold the 23 smartphones and evaluate the impact of victims' movements in three conditions as shown in Figure 16(b)(c), i.e., **1) Sit:** sitting on a chair and holding the smartphone with arm wiggles, **2) Stand:** standing and holding the smartphone with arm wiggles, **3) Stand+Move:** moving back and forth while holding the smartphone with arm wiggles. We kept the attack distance to 6m and attack angle within 60°-90°. Figure 17(d) show the results in the three conditions where the dotted lines indicate the PSNR and STOI scores without the processing of mmEve in **Stand+Move** condition. Compared with the former conditions, we find the **Stand+Move** condition poses a greater challenge to mmEve, which causes a lower PSNR (11dB) and STOI (0.53) as the blue and green dotted lines indicate. However, after the clutter suppression and speech enhancement by mmEve, the speech quality (PSNR>18dB) and intelligibility (STOI>0.75) are greatly improved, which enhances attack practicality in the real world. The results of the 23 smartphones are shown in Figure 18. We find that the PSNR/STOI score varies across different models of smartphones. This may result from the structures of the smartphones considering the hardware design and earpiece placement are different and thus cause different levels of information leakage.

7.6 Speech Recognition

The recovered speech can be distorted and unrecognizable by humans. To further investigate the threat, here we performed automatic speech recognition on the recovered speech. (1) Commercial speech recognition: we adopted Amazon Transcribe (a service for speech-to-text transcription) [2]. We uploaded the original and recovered audio files and performed the transcribing. Then we downloaded the speech-to-text results and calculated the word error rate (WER) [54]. (2) Customized models: we also built two speech recognition models with different decoders, denoted as *DeepSpeech_A* (without language model rescoring) and *DeepSpeech_B* (with language model rescoring) [28]. The rescoring mechanism can help to correct misspelling errors. The audio are transformed into spectrograms with short duration (25ms) and fed into a DeepSpeech2-based neural acoustic model [3] with texts for training. Then the output are fed into the decoder to infer final words. We adopted the leave-one-out method to test the two speech recognition models, i.e., for each experiment we used the audio samples from 22 smartphones (Section 7.2) and test with the samples from another different smartphone model. For each tested smartphone, the samples are the recovered audio in Section 7.5. Figure 19 shows the average WER of the leave-one-out-test experiments. We can observe that the WER of the Amazon Transcribe is about 17%-26% while the WER of DeepSpeech_A and DeepSpeech_B are under 9%. This indicates that to further improve the attack performance (e.g., lower WER), the attacker can use the mmWave-recovered speech collected from smartphones to train the model for the recognition of distorted speech. The performance of DeepSpeech_B is better than DeepSpeech_A. The reason is that the rescoring mechanism contribute to correct the misspelled words and thus reduce the WER of speech recognition.

8 CASE STUDY

Experimental Setting: We made phone calls/send voice messages to five volunteers respectively and asked them to hold their smartphones naturally to listen to the speech emitted from their smartphones' earpiece. The phone callers/voice-message senders both include a male and a female who say the following utterances in a conference room via another smartphone, i.e., 1) "your password is zero one two three" (phone call), 2) "today's meeting has been canceled" (voice message). We also placed a microphone near the

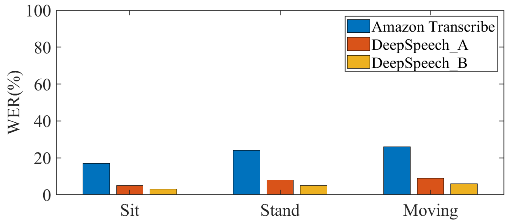

Figure 19: The WER of the speech recognition experiments.

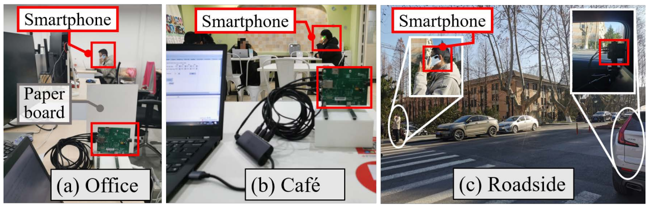

Figure 20: Attack scenarios in the case study.

Table 1: Attack results of the case study.

Smartphone	PSNR/STOI		
	Office	Café	Roadside
GalaxyNote10	25dB/0.83	23dB/0.81	19dB/0.78
GalaxyS20	27dB/0.82	26dB/0.82	21dB/0.76
Oneplus9Pro	21dB/0.79	24dB/0.81	18dB/0.75
HuaweiP40	23dB/0.80	25dB/0.79	14dB/0.76
OppoReno3Pro	26dB/0.81	24dB/0.81	19dB/0.71

phone caller/message sender to record the raw speech. We considered three scenarios as shown in Figure 20. **1) Office:** The victim sits on a chair and holds the smartphone with natural arm movements. An adversarial colleague uses mmEve to eavesdrop on the victim's smartphone. For stealthiness consideration, the attacker blocks the mmWave device with an opaque paperboard (0.5mm thin) which can be easily penetrated by mmWave. The attack distance is about 5m with an attack angle within 70°-90°. **2) Café:** The victim sits on a chair and holds the smartphone close to his/her ear with natural body movements. An adversary uses mmEve to eavesdrop on the victim's smartphone. The attack distance is about 5.5m with attack angle within 70°-90°. **3) Roadside:** The victim stands on the roadside and holds the smartphone with natural body movements. The adversary uses mmEve in a car with attack distance of 8.5m and angle within 60°-90°.

Attack Results: The results are shown in Table 1. mmEve recovers the speech with high PSNR /STOI scores over 21dB/0.79 in the office and café scenarios. There is a slight degradation in the roadside scenario due to the longer distance and tough attack angle during the movements of victims. Overall, mmEve can achieve an average PSNR/STOI score above 18dB/0.75 in all three real-world scenarios. Overall, the results indicate the practicality of mmEve.

9 ROBUSTNESS INVESTIGATION

We investigate the robustness of mmEve with controlled experiments in a laboratory by fixing the targeted smartphone on the tripod (except for Section 9.2). We randomly selected ten sentences from the 100 sentences of OSR and set the attack distance to 6m with an attack angle within 80°-90°. We kept other settings the same as in Section 7.4. Without specific clarifications, the reported results of PSNR and STOI are the mean scores of smartphones. Considering that Android-based smartphones are prevalent due to the open-source feature, we chose the Android-based smartphones to evaluate mmEve in this section (Tested smartphones in Section 9.2-9.5: GalaxyNote10, GalaxyS20, Oneplus9Pro, HuaweiP40, and OppoReno3Pro).

9.1 Smartphone Casing

We investigated the impact of smartphone cases. There are two types of commonly-used cases including hard cases made of plastic and soft cases made of rubber or leather. The attack results are shown in Table 2. We can observe that 1) the plastic case can hardly impact the performance of mmEve because due to the physical connection between the smartphone and the hard case, the case can stand as a solid medium to conduct the vibration of smartphones' surface. 2) mmEve shows resilience to both rubber and leather cases because mmWave can penetrate rubber, leather, paper, clothes, etc., to sense the vibration of smartphone's surface. Overall, mmEve is resilient to commonly-used smartphone cases.

Table 2: Impact of smartphone casing.

Smartphone	PSNR(dB)/STOI			
	No casing	Plastic	Rubber	Leather
GalaxyNote10	27.2/0.94	25.6/0.93	27.0/0.91	26.8/0.91
GalaxyS20	30.2/0.91	27.2/0.89	29.8/0.87	29.5/0.90
OppoReno3Pro	29.1/0.92	25.2/0.87	27.9/0.89	28.5/0.90

9.2 Handhold Habits

We performed experiments to investigate the impact of users' different holding habits. We divided the smartphone rear into three parts coarsely, i.e., the bottom, middle, and upper parts. We asked each volunteer to hold the bottom/middle/ upper parts of smartphones respectively with his/her hands. The attack results are shown in Figure 21. The PSNR score ranges from 22.1dB to 25.3dB and the STOI score changes from 0.75 to 0.82. This indicates that the side channel exists in the whole rear of the smartphone. The coverage of the upper area has a little impact on the PSNR and STOI of recovered speech. The reason is possible that, for the targeted smartphones, the sound source (i.e., the earpiece) locates on the upper part of the motherboard which area tends to have the strongest SPL. Considering that the vibration exists on the whole rear of smartphones, it is difficult to mitigate mmEve unless users cover the whole surface with their hands which brings inconvenience to daily usage.

9.3 CPU Load

We quantitatively evaluated the impact of CPU load by running different numbers of applications on smartphones. There were 15 different applications included in the experiment, such as TikTok, YouTube, and Telegram. Here we defined three statuses of the CPU load, i.e., low (5 apps running), moderate (10 apps running), and high (15 apps running). Besides, there were over 25 WiFi access points in the environment. The results are shown in Figure 21. We can find that the CPU load has no impact on mmEve. That is because the transmitted mmWave by mmEve is in a frequency band of 77-81GHz which is far higher (over 19x) than the operating frequency (about 2-4 GHz) of existing smartphones' CPU. Any wireless signals in the environment out of the band (77-81GHz) will also not be demodulated by mmEve and thus, cause no interference on mmEve.

9.4 Acoustic Noise

We investigate the impact of environmental acoustic noise on mmEve. We played white noise with different SPL in the background via a loudspeaker and performed experiments with other settings the same as in Section 9.3. The loudspeaker is placed on the ground behind the mmWave sensor. The results are shown in Figure 21. We observe that the PSNR and STOI of recovered speech change little when there is acoustic noise of different SPL (55dB/65dB/75dB) in the background. The reason is that mmEve captures the physical vibration of the smartphone's rear for speech recovery and thus hardly by the background noise. This reveals that mmEve is resistant to acoustic noise.

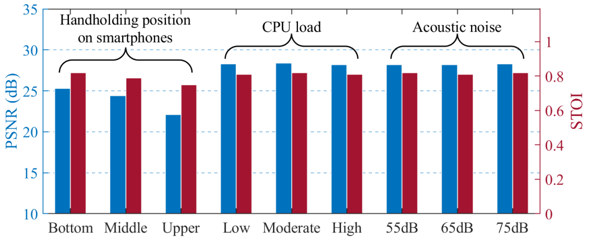

Figure 21: Impact of users' handholding habits, smartphone's CPU load, and environmental acoustic noise.

9.5 Vibration of Inbuilt Motor

To investigate the impact of inbuilt motors on smartphones, we used an application named *VibApp* to control the motor to vibrate in a period of 2s (50% duty cycle) with other settings the same as in Section 9.3. As a control group, we disable the vibration motor and re-perform the experiments. The PSNR and STOI scores without motor vibration are 28.3dB and 0.86. There is a degradation of 5.6dB and 0.13 for the PSNR and STOI respectively when the motor is on. The reason is that the interfering band of the motor (140-180 Hz, GalaxyNote10) overlapped with the fundamental frequency (85-255Hz) of human voice. Based on our observation that the strong-energy vibration of the motor can also be observed in neighboring Range-FFT points due to spectrum leakage[56] while the earpiece-induced delicate vibration is hardly contained, we take the signal derived from the neighboring Range-FFT point as the reference signal and further apply the Normalized Least Mean Square algorithm [23] to eliminate the motor-induced noise. The improvement of PSNR and STOI scores are 3.3dB and 0.09.

9.6 Screen Surface

Here we conduct experiments on the other side of the smartphone, i.e., the screen surface of the smartphone. Specifically, we placed the smartphone (GalaxyNote10) in a fixed condition and a handholding condition, respectively. We followed the same scheme to recover the recovered audio and calculated the PSNR/STOI scores. We find that the score can reach 25.4dB/0.65 and 23.1dB/0.62 for the two conditions. The result indicates that the screen can also leak speech information as the rear surface of the smartphone.

10 COUNTERMEASURES

Detection and Jamming: Considering the transmitted 77-81GHz mmWave by mmEve, an intuitive idea is to detect the malicious signals and disable mmEve by wireless jamming. This is a high-cost strategy. First, to detect the malicious fast chirps, users require both the parameters (e.g., chirp delay and period) of the malicious mmWave sensor and a μs-level synchronization with the malicious device, which is hard to achieve in a real case. Second, mmWave sensors are widely used in automatic driving and industry [30]. Any adversary can purchase such a sensor or compromise a mmWave-equipped system (e.g., automotive vehicles, home monitoring devices) for eavesdropping. Besides, smartphone is widely favored due to its characteristic of mobility and portability. The smartphone user may listen to a phone call or voice messages anywhere, making it high-cost to deploy large numbers of jamming devices.

Shielding: Shielding smartphones with special cases can be effective to mitigate the attack. The case can be designed with vibration-damping materials and wave-absorbing materials. Moreover, smart reflector [35, 46] is a promising technique for mitigation, which can manipulate the phase of reflected mmWave to confuse the attacker.

11 DISCUSSION

Blocking Condition. Objects like metals and human bodies can block mmWave signals. Considering that multipath signals (e.g., reflected mmWave by a wall near the target) have been explored for sensing [45], mmEve is possible to eavesdrop on victim's smartphone via the multipath signals. Thru-wall (e.g., concrete) attack pose greater challenges to mmEve because the penetrating attenuation can be several orders of magnitude greater than line-of-sight sensing, which requires a more specific design.

Attack Distance and Angle. mmEve enlarges the attack distance to 6-8m which is larger than a normal social distance and guarantees the stealthiness. The attack distance can be further enlarged with hardware development. To further improve the performance in tough angles, MVDR algorithm [37] is a promising solution that can strengthen reflected signals from specific directions by estimating a weight factor matrix to increase the SNR of received mmWave.

Distorted Speech. Unvoiced consonants (e.g., /t/) often have lower energy than vowels and have more high-frequency formants far beyond the frequency-response band of the side channel. Thus, the recovered speech containing unvoiced consonants can be somehow distorted. Harmonic extension [31] is promising to compensate the lost formants. But the accuracy of pitch and spectral envelope estimation can have a vital impact on the performance.

Different Sensors. The rationale of boosting the sensing range is to suppress the noise of the sensor's hardware. Considering that the self-noise of sensors are different, the noise pattern in the collected mmWave signals can be different. Thus, to guarantee the performance, the attacker needs to recollect the noise data from the new sensor to train the denoising model.

12 CONCLUSION

In this paper, we achieved a remote and motion-resilient eavesdropping on smartphone earpiece and proposed an end-to-end attack system **mmEve** to recover the speech via a COTS mmWave sensor. We solved technical challenges to boost the eavesdropping performance and achieved practical eavesdropping. We gave countermeasures and emphasized the attack threats due to the increasing number of smartphone users and widely-available attack devices.

ACKNOWLEDGMENTS

The authors would like to thank all anonymous reviewers for their insightful comments on this paper. This paper was in part supported by National Key R&D Program of China (Grant No. 2020AAA0107700), National Natural Science Foundation of China (Grant No. 62032021, 61972348, 61772236, and 62122066), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005), and Research Institute of Cyberspace Governance in Zhejiang University.

REFERENCES

[1] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. 2021. Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights 1, 1 (2021), 100004.

[2] Amazon. 2022. Amazon Transcribe. https://aws.amazon.com/transcribe/

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning. PMLR, 173–182.

[4] S Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 1000–1017.

[5] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2021. 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. 288–299.

[6] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer. In 27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020. The Internet Society.

[7] S. Basak and M. Gowda. 2022. mmSpy: Spying Phone Calls using mmWave Radars. In 2022 2022 IEEE Symposium on Security and Privacy (SP) (SP). IEEE Computer Society, Los Alamitos, CA, USA, 995–1012. https://doi.org/10.1109/SP46214.2022.00058

[8] Alexander Bunkowski, Bertram Bödeker, Stefan Bader, Michael Westhoff, Patric Litterst, and Jörg Ingo Baumbach. 2009. MCC/IMS signals in human breath related to sarcoidosis—results of a feasibility study using an automated peak finding procedure. Journal of Breath Research 3, 4 (2009), 046001.

[9] John Cunnison Catford et al. 1988. A practical introduction to phonetics. Clarendon Press Oxford. 161 pages.

[10] Chi-Lung Cheng and Hans Schneeweiss. 1998. Polynomial regression with errors in the variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60, 1 (1998), 189–199.

[11] Jieun Choi, Hae-Yong Yang, and Dong-Ho Cho. 2020. TEMPEST Comeback: A Realistic Audio Eavesdropping Threat on Mixed-signal SoCs. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 1085–1101.

[12] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Fredo Durand, and William T Freeman. 2014. The visual microphone: Passive recovery of sound from video. (2014).

[13] Philippa Demonte. 2019. HARVARD Speech Corpus—Audio Recording 2019. University of Salford Collection (2019).

[14] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. 2018. Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. arXiv preprint arXiv:1808.05344 (2018).

[15] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. 2021. MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. arXiv preprint arXiv:2104.03538 (2021).

[16] Munther Gdeisat and Francis Lilley. 2011. One-dimensional phase unwrapping problem. signal 4 (2011), 6.

[17] Mordechai Guri, Yosef Solewicz, Andrey Daidakulov, and Yuval Elovici. 2017. SPEAKE (a) R: Turn speakers to microphones for fun and profit. In 11th {USENIX} Workshop on Offensive Technologies ({WOOT} 17).

[18] HStatista. 2021. Smartphone users worldwide 2016-2021. Technical Report. New York, NY, USA.

[19] TeamViewer Inc. 2021. TeamViewer. https://www.teamviewer.com/

[20] Texas Instruments. 2020. AWR1843Boost. https://www.ti.com/lit/ug/spruim4b/spruim4b.pdf?ts=1638342429747

[21] Md Asif Iqbal, Mohammad AZ Al-Khateeb, Lukasz Krzczanowicz, Ian D Phillips, Paul Harper, and Wladek Forysiak. 2019. Linear and nonlinear noise characterisation of dual stage broadband discrete Raman amplifiers. Journal of Lightwave Technology 37, 14 (2019), 3679–3688.

[22] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. mmVib: micrometer-level vibration measurement with mmwave radar. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking. 1–13.

[23] S Kumudini and R Sinha. 2015. Normalized least mean square (Nlms) adaptive filter for noise cancellation. International Journal of Progresses in Engineering, Management, Science and Humanities 1, 1 (2015), 49.

[24] Andrew Kwong, Wenyuan Xu, and Kevin Fu. 2019. Hard Drive of Hearing: Disks that Eavesdrop with a Synthesized Microphone. In 2019 IEEE Symposium on Security and Privacy (SP). 905–919. https://doi.org/10.1109/SP.2019.00008

[25] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of experimental social psychology 49, 4 (2013), 764–766.

[26] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. LASense: Pushing the Limits of Fine-grained Activity Sensing Using Acoustic Signals. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6, 1 (2022), 1–27.

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

[28] Ephrem Tibebe Mekonnen, Alessio Brutti, and Daniele Falavigna. 2022. End-to-End Low Resource Keyword Spotting Through Character Recognition and Beam-Search Re-Scoring. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 8182–8186.

[29] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing speech from gyroscope signals. In 23rd {USENIX} Security Symposium ({USENIX} Security 14). 1053–1067.

[30] Felix Modes. 2021. BMW Automotive sensors: assistance systems' sense organs. https://www.bmw.com/en/innovation/automotive-sensors.html

[31] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE TRANSACTIONS on Information and Systems 99, 7 (2016), 1877–1884.

[32] Ralph P Muscatell. 1984. Laser microphone. The Journal of the Acoustical Society of America 76, 4 (1984), 1284–1284.

[33] Ben Nassi, Yaron Pirutin, Tomer Galor, Yuval Elovici, and Boris Zadov. 2021. Glowworm Attack: Optical TEMPEST Sound Recovery via a Device's Power Indicator LED. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 1900–1914.

[34] Ben Nassi, Yaron Pirutin, Adi Shamir, Yuval Elovici, and Boris Zadov. 2020. Lamphone: Real-time passive sound recovery from light bulb vibrations. Cryptology ePrint Archive (2020).

[35] John Nolan, Kun Qian, and Xinyu Zhang. 2021. RoS: Passive Smart Surface for Roadside-to-Vehicle Communication. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference (Virtual Event, USA) (SIGCOMM '21). Association for Computing Machinery, New York, NY, USA, 165–178. https://doi.org/10.1145/3452296.3472896

[36] Tomofumi Oyama, Hisao Nakashima, Shoichiro Oda, Tomohiro Yamauchi, Zhenning Tao, Takeshi Hoshida, and Jens C Rasmussen. 2014. Robust and efficient receiver-side compensation method for intra-channel nonlinear effects. In OFC 2014. IEEE, 1–3.

[37] Piya Pal and PP Vaidyanathan. 2009. Frequency invariant MVDR beamforming without filters and implementation using MIMO radar. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2081–2084.

[38] Pery Pearson. 1993. Sound Sampling. http://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/I.B.3.a.SoundSampling.html

[39] Sandeep Rao. 2017. Introduction to mmWave sensing: FMCW radars. Texas Instruments (TI) mmWave Training Series (2017).

[40] Wit Rigs. 2019. Samsung Galaxy Note10 Teardown. https://www.youtube.com/watch?v=RfHLb5TagS8

[41] Wit Rigs. 2020. Samsung Galaxy S20 Teardown. https://www.youtube.com/watch?v=zaZsd4Sz83Q

[42] Vittorio Rizzoli, Diego Masotti, and Franco Mastri. 1994. Full nonlinear noise analysis of microwave mixers. In 1994 IEEE MTT-S International Microwave Symposium Digest (Cat. No. 94CH3389-4). IEEE, 961–964.

[43] Nirupam Roy and Romit Roy Choudhury. 2016. Listening through a vibration motor. In Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services. 57–69.

[44] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with your robot vacuum cleaner: eavesdropping via lidar sensors. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems. 354–367.

[45] Elahe Soltanaghaei, Avinash Kalyanaraman, and Kamin Whitehouse. 2017. Peripheral wifi vision: Exploiting multipath reflections for more sensitive human sensing. In Proceedings of the 4th International on Workshop on Physical Analytics. 13–18.

[46] Elahe Soltanaghaei, Akarsh Prabhakara, Artur Balanuta, Matthew Anderson, Jan M. Rabaey, Swarun Kumar, and Anthony Rowe. 2021. Millimetro: MmWave Retro-Reflective Tags for Accurate, Long Range Localization. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom '21). Association for Computing Machinery, New York, NY, USA, 69–82. https://doi.org/10.1145/3447993.3448627

[47] Andrew Sowter. 2003. The derivation of phase integer ambiguity from single InSAR pairs: implications for differential interferometry. In Proceedings of the 11th FIG Symposium on Deformation Measurements, Santorini, Greece. 149–156.

[48] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing 19, 7 (2011), 2125–2136.

[49] Cassia Valentini-Botinhao et al. 2017. Noisy speech database for training speech enhancement algorithms and tts models. (2017).

[50] Chao Wang, Feng Lin, Tiantian Liu, Ziwei Liu, Yijie Shen, Wenyao Xu, and Kui Ren. 2022. mmPhone: Acoustic Eavesdropping on Loudspeakers via mmWave-characterized Piezoelectric Effect. In IEEE INFOCOM 2022 - IEEE Conference on Computer Communications.

[51] Chen-Chia Wang, Sudhir Trivedi, and etc. 2009. High sensitivity pulsed laser vibrometer and its application as a laser microphone. Applied Physics Letters 94, 5 (2009), 051112.

[52] Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani B Srivastava. 2020. UWHear: through-wall extraction and separation of audio vibrations using wireless signals. In SenSys'20. 1–14.

[53] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. ACM, New York, NY, USA, 130–141.

[54] Wikipedia contributors. 2020. Word error rate — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Word_error_rate&oldid=939575741 [Online; accessed 6-August-2022].

[55] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services. 14–26.

[56] Bin Zhang, Min Kong, and Cong Bing WU. 2009. Research of spectrum leakage with window function. Informationization 11 (2009), 10–12.

[57] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services. 301–315.