mmEar: Push the Limit of COTS mmWave Eavesdropping on Headphones

Xiangyu Xu^y, Yu Cherl, Zhen Ling , Li Lu^z, Junzhou Lub and Xinwen Fu ^ySoutheast University, China, Emailxy-xu, yu-chen, zhenling, jlug@seu.edu.cn ^zZhejiang University, China, Email: li.lu@zju.edu.cn [{] University of Massachusetts Lowell, Lowell, MA, USA, Email: xinween@uml.edu

Abstract-Recent years have witnessed a surge of headphonessound-induced vibrations. Optical sensors, including lasers[2], comfortable having con dential communication wearing headwe present an end-to-end eavesdropping systemmEar, which from previous works that realize eavesdropping by sensing speech-induced vibrations with reasonable amplitude,mmEar focuses on capturing the extremely faint vibrations with a low this end, we propose a faint vibration emphasis (FVE) method yet threaten the headphone scenarios. that models and amplies the mmWave responses to speechinduced vibrations on the In-phase and Quadrature (IQ) plane, followed by a deep denoising network to further improve the and setups, we propose a cGAN model with a pretrain- netune mmEar with extensive experiments on different headphones and earphones and nd that most of them can be compromised by the proposed attack for speech recovery.

mmWave sensing

I. INTRODUCTION

(including in-ear headphones) usage in works and communi- high-speed cameras[3], electro-optical sensors[4] and lidars[5] cations. Because of the privacy-preserve property, people feel could capture the vibrations for voice eavesdropping, under phones and pay little attention to speech leakage. In this paper, the assumption of static targets and no occlusion. With the surge of mobile devices, the built-in motion sensors are shows the feasibility of launching an eavesdropping attack on exploited to infer the speech[6], [7], [8], [9], [10], under the headphones leveraging a commercial mmWave radar. Different assumption of accessible motion sensor data. More recently, radio frequency (RF), such as Wi-Fi[11], [12] and UWB[13], has been explored for contactless eavesdropping, but not robust signal-to-noise ratio (SNR) on the surface of headphones. Toward to noise. Due to their constraints, all these attacks could not

With the development of 5G and IoT, researchers nd that mmWave signal could capture small vibrations[14] and SNR. To achieve practical eavesdropping on various headphones thus enable eavesdropping on loudspeakers[15], sound waveinduced objects[16] and smartphones[17], [18]. These apscheme, boosting the generalization ability and robustness of the proaches work well with enough strength and signal-toattack by generating high-quality synthesis data. We evaluate noise ratio (SNR) of the sound-induced vibrations. However, eavesdropping on headphones with mmWave signal is much more challenging as most sound-induced vibrations are wak-Index Terms-side-channel attack, headphone eavesdropping, ened and absorbed within the headphone structure. To tackle this problem and enable headphone eavesdropping with a commercial-of-the-shelf (COTS) mmWave radar, we face the following challenges. First, the sound-induced vibrations on

The global headphone market size (including in-ear headphone surface are too weak that a COTS mmWave radar could barely capture their pattern, and the SNR of phones) is valued at more that billion in 2022 and is expected to grow at a compound annual growth rate of ptured sound-induced vibrations is too low that the noises 12.6% from 2023 to 2030[1], showing an unstoppable trendominate the in uences on mmWave signals. Second, there are of increasing headphone usage in the daily lives for peoplerious headphones with different structures and materials, and around the world. Besides the usage of entertainment, nowae setups in real-world scenarios can also be different, making days more and more people choose to wear headphones and working, including phoning, attending online conferences, another extremely dif cult and costly.

listening to voice messages. Compared to other alternatives this work, we address the above challenges and propose (e.g., loudspeakers), headphones limit the transmission minEar a practical mmWave-based headphone eavesdropping sound close to the human ear, greatly reducing sound leakage eme using single COTS mmWave radar. To capture the In that case, headphone users usually pay little attention faint speech-induced vibrations on the headphone surface, we the threats of eavesdropping, making room for potential side popose a faint vibration emphasis (FVE) method that amplies the vibration-induced phase changes of re ected mmWave channel attacks.

Several side-channel attacks on loudspeakers leveraging on the In-phase and Quadrature (IQ) plane, followed by non-acoustic sensors have been revealed by previous studies. The key insight is that when a loudspeaker generates sound induced vibrations, so that the sound inforwaves, it induces the physical vibration of itself and the mation carried by the vibrations could be partially revealed surrounding objects. Instead of directly recording the sound, the mmWave signals. To further improve the quality and side-channel eavesdropping attacks focus on capturing intelligibility of the extracted sound information, we propose a cGAN model with a pretrain- netune scheme, which leverages

* Corresponding author: Prof. Zhen Ling of Southeast University, Chinathe physical mechanism of vibration generation on headphones

Fig. 1: A typical headphone structure and related sound/vibrations.

to generate large-amount synthesis pretraining data for various

headphones and setups, followed by a netune step with small

amount real-world collected data to strengthen the robustness. 2: Range-FFT on fast-time samples is needed to locate of the model, and nally recover the speech on the headphotime target bin and phase variations of slow-time samples is and accomplish the eavesdropping. Overall, our contributionsed to capture the tiny displacement. are summarized as follows:

Our work reveals a practical non-contact eavesdropping

attack on headphones leveraging single COTS mmWave radar, which demonstrates the alarming threat of user In practice, to measure the distance, a Fast Fourier Trans-privacy leakage from widely used headphones. privacy leakage from widely used headphones. We propose FVE, a method for capturing faint speechwhich improves the perceived granularity for mmWave signals, corresponding to a discretized distance. For a COTS mmWave radars and extends their eavesdropping scenarios radar with 4GHz bandwidth for FMCW chirps, the distance We design a cGAN model with a dedicated pretrainradars and extends their eavesdropping scenarios. netune scheme and a synthesis data generation approach to ensure that our attack could recover intelligible speech_

from different headphones and various setups with a small For more ne-grained vibration sensing, the phase variations of the IF signal from the target range bin are needed. As deamount of real-world collected data. picted in Fig. 2, the samples in every chirp in a target range bin proposed attack on different headphones and setups demonstrating that most of them can be compromised time samples as $= \frac{2}{d}$, where denotes the wavelength by the proposed attack for speech recovery.

II. BACKGROUND AND THREAT MODEL

A. Headphone Vibraition

of the mmWave chirp signal. Given the mm-level wavelength for mmWave radars, um-level displacement changes could be detected[17], which could capture the faint speech-induced vibrations from the headphonedsnot considering the real-

Fig. 1 depicts the typical structure of a headphone and tworld noises related sound/vibrations when playing an audio signal. The

sound wave is generated from the driver and most signars Threat Model

propagate through the ear pad to human ears, providing highScenarios: The scenario could be that a victim is making quality sound to users. Meanwhile, there are some soupdone calls or participating in online conferences using headsignals and sound-induced vibration signals within the heaphones, which happens every day in the subwayé, cafc. phone. Part of these signals transmits to the housing part of Meanwhile, an attacker co-located with the victim leverages a headset, suffering great attenuation, and nally inducing fai@OTS mmWave radar sensor (which could be hidden in bags) vibrations on the surface of the headphone. With a normal transmit mmWave signals toward the surface of the target sound volume setup, such pulses are too tiny to be observed adphones to perform eavesdropping. The goal of the attacker

B. mmWave FMCW Radar

is to recover the speech content played in the headphone. Assumptions: We assume that there are no physical block-

The Frequency-Modulated Continuous Waveform (FMCWages between the victim's headphones and the mmWave radar mmWave radar works by transmitting chirp signals whose that the mmWave signals can be transmitted directly toward frequency increases linearly over time. Then the radar cothe headphone of the victim. We do not assume the attacker lects signals re ected from re ectors and mixes the received stalls any malware or attaches any customized hardware on signal with the transmitting signal to generate the intermediate victim's device. And we do not assume the adversary frequency (IF) signal, whose frequency is proportional to the prior knowledge of the victim's headphone type and time the chirp signal travels. Thus, the distance between the attack scenario context information

Fig. 3: The comparison between spectrogram of clean speech (a), speech recovered from the phone (b), and the speech captured from the headphone (c).

III. FEASIBLE STUDY

recent work[21] proposes a multi-signal consolidation (MSC) model to guide the extraction of tiny machine vibrations, that

Fig. 4: Phase variation tting on the IQ plane.

could not help in extremely low SNR cases.

We rst explore the feasibility of capturing the patterns of According to our studiesextremely faint sound-induced speech-induced vibrations on the headphone surfaces in real rations with low SNR have made eavesdropping on world scenarios. As described in Section. II, phase variation adphones a new challenging task for COTS mmWave of the mmWave IF signal can be utilized for tiny vibration adars, necessitating the design of more re ned signal prodetection. Based on this principle, preliminary experimenteessing approaches to capturing the speech patterns. were conducted with a COTS TI AWR1642 mmWave radar

IV. SYSTEM OVERVIEW and an HP GH10GS headphone. The distance between the headphone and the mmWave radar is kept constablent. We design and implement a mmWave-based headphone During experiments, audio clips sourced from the "Harvardavesdropping systemmetar With the system, an attacker Sentences" in the Harvard Speech Corpus (HSC)[19] atould leverage a COTS mmWave radar to capture the faint played on the headphone. Meanwhile, the mmWave radarface vibrations induced by the speech playing on the target captures the speech-induced vibrations on the headphone headphone and recover the corresponding speech information, face as phase variations following the operations describes the attack roadmap shown in Fig. 5. For constructing in Section. II, and the extracted phase variations are themmEar the architecture contains two modules: transformed into the corresponding spectrogram using Short-A. Headphone Speech-vibration Capturing Module

For comparison, an iPhone 13 with earpiece mode on isThis module aims to extract the faint speech-induced vitested under the same setup, since previous works[17], [168]ations on the surface of the headphone and enhance the have explored the feasibility of eavesdropping on smartphones/rresponding speech patterns. After the basic mmWave signal earpieces using mmWave radars. The results are shownpincessing procedurmmEardetermines the relative distance Fig. 3. The sound-induced vibrations from smartphones' earetween the radar and the target headphone by range bin pieces could be roughly captured. While for headphones, dselection. Then, a faint vibration emphasis (FVE) approach to the weaker vibration and lower SNR, most patterns are lost proposed to amplify the speech-induced vibrations aiming leaving only limited and vague patterns on the spectrogramtimproving their intelligibility, followed by a deep denoising network to enhance the corresponding speech patterns. that could hardly represent the speech information.

To further improve the perceived granularity, we explore the phase variations of mmWave signals on vibration detection. Speci cally, when plotting the phase variation samples To further improve the quality and intelligibility of the on the In-phase and Quadrature (IQ) plane, theoretical speech signal on various headphomes Ear dethese samples could form arc-shape trajectories whose lengigins a Conditional Generative Adversarial Network (cGAN) is proportional to vibration amplitudes[20]. Considering the model[22] with a pretrain-netune scheme. Concretely, folbackground noises, the arc trajectory could be shifted on the physical mechanism of sound-induced vibration on IQ plane and the linear relation between the arc length aheadphones, a synthesis data generation method is proposed the vibration amplitudes may not be kept. Fig. 4 illustrates stimulate mmWave responses for various headphones and the vibration-induced phase variations on the IQ plane. It cartups, and form a large dataset for pretraining. Then a small be seen that with relatively large vibration amplitudes and highmount of real-world data is collected to netune the model, SNR (such as vibrations on loudspeakers), the circle related ther improving the robustness of the attack.

to the arc trajectory could be t and the linear relation could V. HEADPHONE SPEECHVIBRATION CAPTURING be regained by translating the circle center to the origin of IQ plane. On the other hand, with relatively small vibration. Headphone Localization

amplitude and low SNR (such as vibrations on headphoneTo capture the speech-induced vibrations from the headsurfaces), the sample trajectories lose the arc shape, making ne surfacemmEar rst localizes the target headphone, the circle tting almost impossible, which prevents high which involves a range-FFT to separate the space into multiple quality pattern capturing with mmWave signals. Although bins, each representing a different distance from the mmWave Fig. 5: Overview of mmEar architecture.

Fig. 6: Procedure of the FVEFig. 7: STOI heatmap on the Fig. 8: Illustration of the FVE-ampli ed sample and networkapproach. IQ plane. enhanced sample.

radar. Therefore, the headphone localization problem becomes Step 3: In the circle- tting process, the nal step is to a range bin selection problem. Note from Fig. 3 that for the correct range bin, there are detectable low-frequency (< 1000Hz) patterns of phase variations on the spectrogram, which do not hold for other bins. Based on the observation, mmEar calculates the variance of phase variations for within the frequency between100Hz and 500Hz for each range bin and selects the largest one as the target.

B. Faint Vibration Emphasis

After range bin selection, the phase variations of the IF mmWave signal of the target range bin can be calculated to capture the vibrations. However, as discussed in Section.III, the sound-induced vibrations on headphones are too faint and with low SNR, making existing circle- ttingbased approaches[20][21] fail on the IQ plane.

To capture the faint headphone vibrations with low SNR, we propose a faint vibration emphasis (FVE) approach to amplify the phase variations on the IQ plane to improve intelligibility. As shown in Fig. 6, the procedure of FVE containsteps:

Step 1: Because of extremely small amplitude and low SNR, the phase variation trajectory looks more like a short thick line segment instead of an arc. Therefore, we t the line segment leveraging the least square tting. Step 2: Although we could not conduct circle- tting because of the loss of arc shape for phase variation trajectories, given the tted line segment obtained from Step 1, it could be observed that the center of the expected circle should lie on the perpendicular bisector of the line segment. SommEar calculates the corresponding perpendicular bisector in this step.

move the center of the tted circle to the origin of the IQ plane. In our case, we got a line across the center of the circle from Step 2 instead. So we translate the line to make it across the origin of the IQ plane and move the phase variation trajectories accordingly. Step 4: Finally, to amplify the phase variations on the

IQ plane, mmEar further translate the phase variation trajectories in Step 3 along with its perpendicular bisector toward the origin of the IQ plane, until the anglein Fig.6 reaches a pre-determined valuewhich is set as 30° in mmEar

Fig. 9: The process of synthesis data generatiominEar

generating physical acoustic signals. Then, the physical acoustic signals induce vibrations on the surface of smartphones, which is modulated by Td [(k7c(y)-24[(gene(7ed)-4ransforma)-303(

Fig. 10: Spectrograms of digital speech sample(a), synthesis sample (b) and mmWave-captured real-world sample (c).

C. Speech Pattern Enhancement

It could be observed from Fig. 8 that FVE signi cantly ampli es the speech patterns und £000Hz, but the speech patterns higher than 1000Hz are still buried in noise. To address the problem, we apply a deep denoising network model for speech pattern enhancement [24] to improve the SNR in the high-frequency range where a feed-forward neural network with many levels of non-linearities is adopted. The corresponding result is shown in Fig. 8c. We could observe that the speech-related pattern is enhanced in the high-frequency component after applying the model.

VI. HEADPHONE EAVESDROPPINGENHANCEMENT

To further improve the intelligibility of the extracted speech signal, as well as extend the attack scope to various headphones and setupmmEardesigns a cGAN model[22] with a pretrain- netune scheme to realize the headphone eavesdropping enhancement, which includes a synthesis data generation method and a cGAN enhancing process.

A. Synthesis Training Data Generation

The goal is to create a large amount of data that has similar signal patterns with real-world-collected mmWave samples when sensing the headphone vibrations. Moreover, the generated data should cover the in uences of different headphones and setups. Toward this end, the process is presented in Fig. 9. We model the physical process from a digital speech sample to the mmWave responses by oustic vibration response calculation and mmWave noise distribution analysismd design the data generation procedure accordingly.

Acoustic-vibration Response Calculation. mmExempts by capturing the speech-induced vibrations on the headphone surface. Thus, to generate synthesis data, we rst model the responses from digital acoustic signals to headphone vibrations in Fig. 9, It can be seen that digital speech signal is rst in uenced by the frequency response of the headphone drivers,

Fig. 12: Experimental setups of mEar

The cGAN model is rst pre-trained using synthesis data Fig. 11: Structure of cGAN model for speech enhancements input. After that, we leverage the real-world mmWavecaptured signal to netune the pre-trained model, further

 f_d , we have collected the frequency response information proving the robustness of the model. Finally, we obtain a of 47 different headphones from their documentation andell-functioned generator model that takes the output specrandomly select one of them for each generated sample. Whitegram of Headphone Speech-vibration Capturing Modate for f_h , we determine the responses according to a previous ut and obtains the recovered speech with high quality and study[27]. After that, we further multiply the damping factor intelligibility for eavesdropping.

randomly generated from (0; 1) to simulate the signal attenu-

ation. Afterward, the mmWave noises are added. To simulate

various SNRs from different setups, we multiply the noise b. System Setup

a normally distributed random numberranging between 0.1 The system setup is shown in Fig. 12. We utilize a and 10. Finally, for each frequency component of the curreGOTS mmWave radar, Texas Instruments AWR1642BOOST, sample, we add noise that follows a normal distribution with transmitting and receiving mmWave signals. The dedifferent mean and variance for real and imaginary parts. modulated chirp signals are sampled by the data acquisition

The synthesis data generated through the aforementionbed and DCA1000EVM and sent to a desktop (HP Pavilion14process is depicted in Fig. 10b. It can be observed that tote1004TX) for processing. The cGAN model is implemented synthesis data exhibits a high degree of similarity to thesing TensorFlow and is trained of ine on a desktop NUC with mmWave-captured real-world data (Fig. 10c). Through thetel-1260P CPU and 16GB RAM memory. Subsequently, it processes mmEarcould generate large-amount synthesis data deployed on the desktop for real-time speech enhancement. for pretraining the cGAN model to obtain robustness for the utilized headphones include HP GH10GS, HP H320GS, various headphones and setups. Lenovo L7, Sony XM3, ATH-M30x, and earphones include

B. cGAN Enhancing

Lenovo L7, Sony XM3, ATH-M30x, and earphones include Airpods Pro3 and FreeBuds SE 2. Except for Section. VII-L, the headphones are placed on a human head model as de-

The synthesis training data is then utilized to pre-traipicted in Fig. 12a. The results are average performance across the cGAN model[22], the architecture of which is shown indifferent headphones and earphones except for Section. VII-I. Fig. 11. Such an adversary model could be used for detailed

information generation[28]. Particularly, this model includes. Dataset and Data Collection

a conditional generator and a discriminator. The generatorWe conduct evaluations using the widely adopted Harvard takes the spectrogram generated from the headphone spescheech Repository (HSC)[19] and Open Speech Repository vibration capturing module as its input and outputs the re ne(@SR)[34] datasets. HSC consists 720 sentences spoken by spectrogram. The discriminator takes two kinds of inputs ne person while OSR consists sentences from different speak-i.e., a re ned spectrogram output by the generator, and thes. To demonstratemEats generalization capability, we corresponding audio spectrogram, and outputs the discrimelect the initial 640 sentences from HSC as training set and nation results. With the cGAN structure, the generator ansemtences from OSR as testing set.

discriminator are trained iteratively, targeting a well-trained Pre-train Dataset. The pre-train data is generated using the generator that could generate high-quality speech spectmethod presented in Section. VI-A. We select the initial 640 grams for eavesdropping.

For the construction details, the generator consists **ge**nerate 100 samples for each sentence with different generfour fully connected layers with LeakyReLU[29] and Batchation parameters, forming a dataset with 64000 sentences. Normalization[30] in each layer, while the discriminator con- Fine-tune Dataset.The ne-tune data are collected in realsists of three fully connected layers with LeakyReLU and orld scenarios. We play the given 640 speech samples from Dropout Normalization[31] with an output rate of 0.4, and the HSC dataset through the headphones while simultaneously it utilizes the sigmoid function[32] as its activation functioncapturing the vibration signals using mmWave radar. The loss of the generator is Mean Square Error (MSE) and the Testing Dataset. In various experimental settings, we colloss of the discriminator is binary cross entropy. The optimizer to sentences from the OSR dataset to assess the system's we use is Adam[33].

VII. EVALUATION





Fig. 14: Impact of incident

angle of radar.

Fig. 13: Impact of attack distance.

C. Metrics

Short-Time Objective Intelligibility (STOI): The STOI[23] measures the level of intelligibility of the processed speech by quantifying the similarity between the original speech and processed speech. The STOI value is within [0,1], where higher values indicate better speech intelligibility, an STOI greater than 0.7 generally represents high intelligibility.

Perceptual Evaluation of Speech Quality (PESQ): The PESQ[35] is a quantitative algorithm employed for objectively evaluating the perceived quality of speech signals after conducting transmission or processing. The PESQ is defined in the range of [-0.5,4.5], with increasing values indicating improvement in speech quality.

D. Different Attack Distances

We direct the mmWave radar toward the target headphone and vary the distance between the radar and the headphone from 20cm to 200cm. The result is shown in Fig. 13. Without enhancement, the STOI/PESQ scores of the speech are 0.50/1.08 at a proximity of 20cm, while the enhanced speech achieves scores of 0.71/1.46. Within the range of 20cm to 100cm, showing significant improvements in recovered speech quality and intelligibility.

E. Different Incident Angles of the mmWave Radar

Considering the practical attack scenario where the attacker may not ensure precise alignment of the radar with the headphone, we conduct a controlled experiment to investigate the impact of radar misalignment. The distance between the headphones and the radar is set as 50cm. We vary the radar position to introduce an angle between the radar's line-ofsight and the headphone, ranging from 0° to 30°. The result is presented in Fig. 14. We can observe a general trend that the STOI and PESQ decrease with increasing angle of the mmWave radar. Moreover, the STOI and PESQ remain stable when the radar angle is within 10°, showing a moderate tolerance on mmWave radar angles.

F. Different Incident Angles of Headphones

In a practical eavesdropping scenario, the attackers may not always have the opportunity to transmit mmWave signals perpendicular to the target headphone. Therefore, we also investigate the impact of headphone misalignment. Placing the headphones at a fixed distance of 50cm from the radar, we tilt the headphone body by angles from 0° to 30° to explore their





Fig. 15: Impact of incident angle of headphone.

Fig. 16: Impact of threshold



Φ.

Fig. 17: Impact of different volumes.

effects on the system's attack performance. The experimental results, shown in Fig. 15, indicate that when the angle is less than 20°, *mmEar* could realize high-quality eavesdropping by significantly improving both STOI and PESQ. This suggests that *mmEar* can still eavesdrop on the speech content played on the headphones even when the headphone body is misaligned to a certain extent.

G. Impact of threshold Φ of samples

Section. V-B defines a key parameter Φ in our fiant vibration emphasize (FVE) method, which directly influences the quality of captured speech-induced vibrations. We investigated the impact of Φ on the intelligibility of the raw captured vibrations. Concretely, we vary the parameter Φ from 15° to 90° and show the results in Fig. 16. From the figure, it can be observed that when the Φ is set within the range from 30° to 45°, the intelligibility is relatively higher than others, considering the original parameter before FVE is small, the results validate the effectiveness of FVE.

H. Different Volumes of Headphones

We investigated the influence of headphone volume levels by manipulating the volume settings during playback. The headphone volume set at 100%, 80%, 60%, 40%, and 20% of the maximum volume. The result is presented in Fig. 17. It can be observed that when the headphone volume exceeds 60%, the STOI and PESQ for the enhanced audio reach 0.6/1.2. Conversely, when the headphone volume falls below 40%, there is a sharp decline in system performance. This phenomenon arises from the insufficient audio amplitude at lower volume levels, preventing the headphone diaphragm from generating a sufficiently large vibration, thereby impeding the radar's ability to capture meaningful vibrational information.

I. Different Headphones

We conduct experiments with five headphones (HP GH10GS, HP H320GS, Lenovo L7, Sony XM3, ATH-M30x)



Fig. 18: Different headphones and earphones.

and two earphones (Airpods Pro3 and FreeBuds SE 2), with a distance of 50cm between the headphones/earphones and the radar. The STOI of the different headphones after enhancement is shown in Fig. 18. It indicates that the headphones exhibit higher STOI and PESQ, all-surpassing 0.71/1.35, with Sony XM3 and ATH-M30x over 0.77/1.46. In contrast, earphones demonstrate a slightly lower STOI and PESQ but still exceed 0.64/1.18. This suggests that the system is feasible for both headphones and earphones, but more feasible for headphones that have larger reflection areas.

J. Module Performance

We investigated the impact of the mmEar modules on the system's performance. For each module, we computed metrics for audio processed solely through each module (denoted as *Module 1* and *Module 2*), and through both modules for comparison. The results are depicted in Fig. 19. The experimental findings demonstrate that the audio processed through both *Module 1* and *Module 2* exhibit an STOI improvement of 0.08 compared to audio processed solely through *Module 1* and a 0.06 improvement compared to audio processed solely through *Module 1* and a 0.06 improvement compared to audio processed solely through *Module 2* are enhanced by 0.24 and 0.21 compared to those processed solely through *Module 1* and solely through *Module 2*, respectively. This indicates that each module of the system significantly influences the final output.

K. Noise Influence

To investigate the impact of background noise on system performance, we conducted experiments under different background white noise environments simulating 50 dB, 60 dB, and 70 dB noise levels, representing quiet office, meeting room, and noisy street scenarios respectively. The distance between the headphone and the radar was set to 50cm with a 0° angle. The experimental results are presented in Fig. 20. From the figure, it can be observed that background noise has minimal effect on the system performance, as both STOI and PESQ values fluctuate around 0.72/1.33.

L. Human Influence

We investigated the impact of various human actions on system performance when wearing headphones. The radar was positioned facing the headphones at a distance of 50cm from the person. As shown in Fig. 12b and Fig. 12c, three different actions of the person were explored, including sitting, standing, and pacing, where sitting and standing involved minimal movement while pacing encompassed moderate movement. The experimental results are illustrated in Fig. 21. From the figure, it can be observed that when the person is sitting and standing, the distance between the person and the radar remains unchanged, and the system still achieves favorable performance, with STOI and PESQ values approximately at 0.66/1.28. However, when the person is pacing, the distance/angle with respect to the headphone and the radar changes, causing a slight decrease in the system performance, but still available to capture speech information.

M. Subjective Scores for Intelligibility.

We recruited 6 volunteers (3 males and 3 females, ages 20-41) to assess the intelligibility of 60 sentences randomly selected from the collected audio under the aforementioned environments, both in their original form (raw) and after enhancement. The volunteers scored intelligibility on a scale of 0 to 10, with higher scores indicating higher speech intelligibility. The average scores given by the 6 volunteers are presented in Fig. 22. In the "sit" and "stand" environments, the raw scores are below 2, showing that most sentences recovered from raw data can not be understood, while the enhanced scores are around 7.2, indicating that a large part of recovered sentences of *mmEar* could be understood by the volunteers. For the "pace" environment, scores for raw vibrations and mmEar recovered sentences drop to 0.57/6.5, showing that *mmEar* turns the nearly incomprehensible vibrations into intelligible speech for eavesdropping.

VIII. RELATED WORKS

With the surge of mobile computing and Internet-of-things, eavesdropping technology has developed beyond traditional hidden microphones to a broader range of sensors and signals.

A. Eavesdropping with Optical Signals

The laser microphone[2] is a well-known method to eavesdrop with a laser beam, which captures the tiny vibrations of objects induced by sound waves in the victim's room. Similarly, a high-speed camera could also capture sound-induced vibrations and become a visual microphone for eavesdropping[3], so as electro-optical sensors[4] and lidars[5]. Although these optical-based eavesdropping could be launched long distances, they require a static target and a direct path with no occlusion, which limits their scopes.

B. Eavesdropping with Motion Sensors

With the capability to sense vibrations, motion sensors on smartphones are exploited to capture the sound-induced vibrations and perform eavesdropping[6], [7], [8], [9], [10]. Such attack could be achieved with single gyroscope[6] or accelerometer[7], [8], [9], or a combination of multiple motion sensors (accelerometers, gyroscope, and geophone) to realize eavesdropping under limited sampling rates[10]. Besides smartphones, the feasibility of motion sensor-based eavesdropping is proved on VR/AR devices[36], earphones[37], and



Fig. 19: Impact of modules. Fig. 20: Impact of noise.

general IoT devices[38]. However, the assumption of these attacks is the existence of a physical connection between the motion sensor and the sound source, making them not feasible in contactless eavesdropping scenarios.

C. Eavesdropping with RF Signals

Radio frequency (RF) has been explored for contactless eavesdropping in real-world scenarios. As the most widelydeployed RF signal, Wi-Fi shows the feasibility of eavesdropping by leveraging CSI changes[11] and MIMO properties[12] with respect to acoustics, and UWB could achieve better performance with a larger bandwidth[13]. However, these approaches suffer much from multi-path transmission and are thus not robust enough in real-world scenarios. Most recently, the mmWave signal has been exploited for eavesdropping. MILLIEAR[15][39] achieves eavesdropping by extracting phase changes from the reflected mmWave signals targeting the loudspeakers. Along with the path, mmEcho[16] and mmPhone[40] focus on eavesdropping from other objects near the sound source, releasing the requirement of directly targeting the user or the loudspeakers. Meanwhile, mmSpy[17] and mmEve[18] exploit the feasibility of contactless smartphone calls eavesdropping by sensing the speech-induced vibrations on the smartphone's earpiece, showing practical threats in real-world scenarios. However, for the phone call scenarios, people are getting used to wearing headphones or earbuds to prevent privacy leakage, and this work is targeting the more challenging case of headphone eavesdropping with mmWave signals.

D. Other Eavesdropping Approaches

There are other eavesdropping approaches exploiting the side-channel leakage of sound. Motors[41] and hard drives[42] are leveraged for side-channel eavesdropping attacks. Also, magnetic side-channel signals leaked by a microspeaker[43][44] are also utilized for eavesdropping. Different from these methods that require specially designed sensors or scenarios, *mmEar* could launch eavesdropping attacks with COTS mmWave sensors in various real-world scenarios.

IX. DISCUSSION

Defense: The fundamental principle of *mmEar* is to leverage mmWave signals to capture the faint speech-induced vibrations on the headphone surface for eavesdropping. So the defense could be hiding or interfering with the vibration pattern. For the former, more acoustic and vibration-absorbing





Fig. 21: Impact of human actions.

Fig. 22: Subjective scores of human actions.

materials could be added to the headphones to make the vibration weaker than the physical sensing limit of mmWave signals. For the latter, an active vibration signal can be generated by a headphone on its surfaces, which could interfere with the pattern of speech-induced vibrations.

Automatic Speech Recognition: Because the target of *mmEar* is to eavesdrop, The system evaluation is based on scoring the intelligibility of speech using metrics such as STOI and PESQ, as well as human-understanding assessments. We notice that automatic speech recognition (ASR) has already been widely used and intend to involve ASR for more extensive validations in the future.

X. CONCLUSION

This paper reveals an eavesdropping attack on headphones by capturing the faint speech-induced vibrations on headphone surfaces, and proposes a practical headphone eavesdropping system, *mmEar*, leveraging single COTS mmWave radar. *mmEar* explores the feasibility of mmWave-based speech recovery from extremely faint sound-induced vibrations with a low SNR, and proposes a range of techniques to achieve a robust attack with high generalization ability. Extensive experiments validate the feasibility of the attack, demonstrating the alarming threat of user privacy leakage from widely-used headphones.

ACKNOWLEDGMENTS

This research was supported in part by National Natural Science Foundation of China Grant Nos. 62022024, 61972088, 62232004, 62202099, and 62102354, by US National Science Foundation (NSF) Awards 1931871, 1915780, and US Department of Energy (DOE) Award DE-EE0009152, by Jiangsu Provincial Natural Science Foundation of China Grant No. BK20220806, Jiangsu Provincial Key R&D Programs Grant Nos. BE2021729, BE2022680, BE2022065-5, Jiangsu Provincial Key Laboratory of Network and Information Security Grant No. BM2003201, Key Laboratory of Computer Network and Information Integration of Ministry of Education of China Grant No. 93K-9, and Collaborative Innovation Center of Novel Software Technology and Industrialization. Any opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

- [1] G. V. RESERACH, "Earphones and headphones market size, share & trends analysis report. [online]. available:," 2023. [Online]. Available: https://www.grandviewresearch.com/industryanalysis/earphone-and-headphone-market/toc
- [2] R. P. Muscatell, "Laser microphone," <u>The Journal of the Acoustical Society of America</u>, vol. 76, no. 4, pp. 1284–1284, 1984.
- [3] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," <u>ACM Trans. Graph.</u>, vol. 33, no. 4, jul 2014.
- [4] B. Nassi, Y. Pirutin, R. Swisa, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Passive sound recovery from a desk lamp's light bulb vibrations," in Proc. USENIX Security'22. USENIX Association, 2022, pp. 4401–4417.
- [5] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: Eavesdropping via lidar sensors," in <u>Proc.</u> SenSys'20. ACM, 2020, pp. 354–367.
- [6] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in Proc. USENIX Security'14. USENIX Association, 2014, pp. 1053–1067.
- [7] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in <u>2018 IEEE Symposium on</u> <u>Security and Privacy (SP)</u>. IEEE, 2018, pp. 1000–1017.
- [8] 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 <u>Proc. WiSec'21</u>. ACM, 2021, pp. 288–299.
- [9] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learningbased practical smartphone eavesdropping with built-in accelerometer." in <u>Proc. NDSS'20</u>, vol. 2020, 2020, pp. 1–18.
- [10] Y. Liang, Y. Qin, Q. Li, X. Yan, L. Huangfu, S. Samtani, B. Guo, and Z. Yu, "An escalated eavesdropping attack on mobile devices via low-resolution vibration signals," <u>IEEE Transactions on Dependable and Secure Computing</u>, vol. 20, no. 4, pp. 3037–3050, 2023.
- [11] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" <u>IEEE Transactions on Mobile Computing</u>, vol. 15, no. 11, pp. 2907–2920, 2016.
- [12] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in <u>Proc. MobiCom'15</u>. ACM, 2015, pp. 130–141.
- [13] 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 <u>Proc. SenSys'20</u>. ACM, 2020, pp. 1–14.
- [14] X. Xu, J. Yu, C. Ma, Y. Ren, H. Liu, Y. Zhu, Y.-C. Chen, and F. Tang, "mmecg: Monitoring human cardiac cycle in driving environments leveraging millimeter wave," in <u>Proc. INFOCOM'22</u>. IEEE, 2022, pp. 90–99.
- [15] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in <u>Proc. IEEE INFOCOM'22</u>. IEEE, 2022, pp. 11–20.
- [16] P. Hu, W. Li, R. Spolaor, and X. Cheng, "mmecho: A mmwave-based acoustic eavesdropping method," in <u>IEEE Symposium on Security and</u> Privacy (SP). IEEE, 2023, pp. 1840–1856.
- [17] S. Basak and M. Gowda, "mmspy: Spying phone calls using mmwave radars," in <u>2022 IEEE Symposium on Security and Privacy (SP)</u>. IEEE, 2022, pp. 1211–1228.
- [18] C. Wang, F. Lin, T. Liu, K. Zheng, Z. Wang, Z. Li, M.-C. Huang, W. Xu, and K. Ren, "mmeve: Eavesdropping on smartphone's earpiece via cots mmwave device," in <u>Proc. MobiCom'22</u>. ACM, 2022, pp. 338–351.
- [19] P. Demonte, "Harvard speech corpus-audio recording 2019," <u>University</u> of Salford Collection, 2019.
- [20] I. Mikhelson, S. Bakhtiari, T. W. Elmer, and A. V. Sahakian, "Remote sensing of heart rate and patterns of respiration on a stationary subject using 94-ghz millimeter-wave interferometry," <u>IEEE Trans. Biomed.</u> <u>Eng.</u>, vol. 58, no. 6, pp. 1671–1677, 2011.
- [21] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, "mmvib: micrometer-level vibration measurement with mmwave radar," in <u>Proc.</u> <u>MoboCom'20</u>. ACM, 2020, pp. 1–13.
- [22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
- [23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "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.

- [24] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," <u>IEEE/ACM Transactions</u> on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2015.
- [25] K. AN, "Sulla determinazione empirica di una legge didistribuzione," Giorn Dell'inst Ital Degli Att, vol. 4, pp. 89–91, 1933.
- [26] N. V. Smirnov, "Approximate laws of distribution of random variables from empirical data," <u>Uspekhi Matematicheskikh Nauk</u>, no. 10, pp. 179– 206, 1944.
- [27] K. Carillo, F. Sgard, and O. Doutres, "Numerical study of the broadband vibro-acoustic response of an earmuff," <u>Applied Acoustics</u>, vol. 134, pp. 25–33, 2018.
- [28] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "Breathlistener: Finegrained breathing monitoring in driving environments utilizing acoustic signals," in Proc. MobiSys'19. ACM, 2019, pp. 54–66.
- [29] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in <u>Proc. ICASSP'13</u>. IEEE, 2013, pp. 1234–1238.
- [30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in <u>Proc. ICML'15</u>. JMLR.org, 2015, pp. 448–456.
- [31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," <u>The journal of machine learning research</u>, vol. 15, no. 1, pp. 1929–1958, 2014.
- [32] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," <u>The bulletin of mathematical biophysics</u>, vol. 5, pp. 115–133, 1943.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [34] V. Troubleshooter, "The open speech repository," 2010.
- [35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in <u>IEEE international conference on acoustics, speech, and signal processing</u>, vol. 2. IEEE, 2001, pp. 749–752.
- [36] C. Shi, X. Xu, T. Zhang, P. Walker, Y. Wu, J. Liu, N. Saxena, Y. Chen, and J. Yu, "Face-mic: Inferring live speech and speaker identity via subtle facial dynamics captured by ar/vr motion sensors," in <u>Proc.</u> MobiCom'21. ACM, 2021, pp. 478–490.
- [37] Y. Cao, F. Li, H. Chen, X. Liu, C. Duan, and Y. Wang, "I can hear you without a microphone: Live speech eavesdropping from earphone motion sensors," in Proc. INFOCOM'23. IEEE, 2023, pp. 1–10.
- [38] J. Han, A. J. Chung, and P. Tague, "PitchIn: Eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in <u>Proc.</u> <u>IPSN'17</u>. ACM, 2017, pp. 181–192.
- [39] P. Hu, W. Li, Y. Ma, P. S. Santhalingam, P. Pathak, H. Li, H. Zhang, G. Zhang, X. Cheng, and P. Mohapatra, "Towards unconstrained vocabulary eavesdropping with mmwave radar using gan," <u>IEEE Transactions on Mobile Computing</u>, 2022.
 [40] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu,
- [40] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu, and K. Ren, "mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect," in <u>Proc. IEEE</u> INFOCOM'22. IEEE, 2022, pp. 820–829.
- [41] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in Proc. MobiSys'16. ACM, 2016, pp. 57–69.
- [42] A. Kwong, W. Xu, and K. Fu, "Hard drive of hearing: Disks that eavesdrop with a synthesized microphone," in <u>IEEE symposium on</u> <u>security and privacy (SP)</u>. IEEE, 2019, pp. 905–919.
- [43] J. Choi, H.-Y. Yang, and D.-H. Cho, "Tempest comeback: A realistic audio eavesdropping threat on mixed-signal socs," in <u>Proc. CCS'20</u>. ACM, 2020, pp. 1085–1101.
- [44] Q. Liao, Y. Huang, Y. Huang, Y. Zhong, H. Jin, and K. Wu, "Magear: Eavesdropping via audio recovery using magnetic side channel," in <u>Proc.</u> MobiSys'22. ACM, 2022, pp. 371–383.