Poster: Speech Privacy Attack via Vibrations from Room Objects Leveraging a Phased-MIMO Radar

Cong Shi, Tianfang Zhang, Zhaoyi Xu, Shuping Li, Yichao Yuan, Athina Petropulu, Chung Tse Michael Wu, Yingying Chen
Rutgers University, New Brunswick, NJ, USA
{cs1421,tz203,sl1567,yy470,ctm.wu,yingche}@scarletmail.rutgers.edu,{zx111,athinap}@soe.rutgers.edu

ABSTRACT

Speech privacy leakage has long been a public concern. Through speech eavesdropping, an adversary may steal a user's private information or an enterprise's financial/intellectual properties, leading to catastrophic consequences. Existing non-microphone-based eavesdropping attacks rely on physical contact or line-of-sight between the sensor (e.g., a motion sensor or a radar) and the victim sound source. In this poster, we discover a new form of speech eavesdropping attack that senses minor speech-induced vibrations upon common room objects using mmWave. By integrating phased-array and multiple-input and multiple-output (MIMO) on a single mmWave transceiver, our attack can capture and fuse micrometer-level vibrations upon the surfaces of multiple objects to reveal speech content in a remote and non-line-of-sight fashion. We successfully demonstrate such an attack by developing a deep speech recognition scheme grounded on unsupervised domain adaptation. Without prior training on the victim's data, our attack can achieve a high success rate of over 90% in recognizing simple speech content.

KEYWORDS

Speech Privacy Attack, mmWave Sensing, Phased-MIMO

ACM Reference Format:
Cong Shi, Tianfang Zhang, Zhaoyi Xu, Shuping Li, Yichao Yuan, Athina Petropulu, Chung Tse Michael Wu, Yingying Chen. 2022. Poster: Speech Privacy Attack via Vibrations from Room Objects Leveraging a Phased-MIMO Radar. In *The 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22), June 25–July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3498361.3538790

1 INTRODUCTION

Human voice has always been a dominant medium of communication. For both enterprises and individuals, voice communication plays a critical role in various important tasks, such as meetings, phone calls/messages, and bank transactions. The recent advancements of voice-user interfaces even extend the use of voice to human-to-machine interactions. However, the unencrypted nature of voice creates a looming threat of security and privacy leakage.

MobiSys '22, June 25–July 1, 2022, Portland, OR, USA

ACM ISBN 978-1-4503-9185-6/22/06.
https://doi.org/10.1145/3498361.3538790

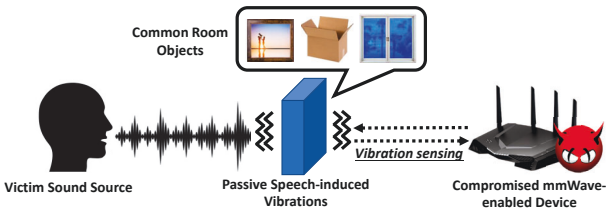


Figure 1: Illustration of the proposed mmWave-based eavesdropping attack exploiting minor passive speech vibrations upon common room objects.

Speech eavesdropping attacks may steal a user's private information (e.g., passwords, social security numbers) or an enterprise's financial and intellectual properties (e.g., transactions, financial reports/blueprint), resulting in high risks of financial and reputation loss.

Exploring the issues of speech eavesdropping has long been a core topic in computer security. Early attempts with tampered microphones can be thwarted by restricting physical access to a space (e.g., using a confidential meeting room). Motion sensors on mobile devices were shown to pick up conductive vibrations caused by the speech playback of a built-in or neighboring external loudspeaker [1]. However, the motion sensors do not generally get impacted by aerial sounds, which precludes their use in natural scenarios with airborne speech. Research studies also revealed the potential of using radio frequency (RF) techniques [2] for speech sensing. However, these approaches are still limited to scenarios with the sound source in the line of sight of the radar sensor.

In this work, we demonstrate a remote and non-line-of-sight speech eavesdropping attack by exploiting *passive speech-induced vibrations* on common room objects (e.g., plastic boxes, glass windows). Speech is carried by sound waves propagating through the air. Upon hitting nearby objects, the sound energy of speech is partially transmitted onto those solid media, causing subtle physical vibrations. As illustrated in Figure 1, our attack lies in sensing such minor vibrations by leveraging widely deployed mmWave-enabled devices (e.g., 802.11ad routers, 5G-enabled devices). Our contributions are summarized as follows:

- To the best of our knowledge, our work is the first attempt to explore mmWave to capture passive speech vibrations for speech eavesdropping, which can turn a common room object into a microphone in a remote fashion.
- We integrate phased-array and MIMO on a single transceiver. Our phased-MIMO radar can steer the mmWave beam to simultaneously sense on multiple vibration sources while utilizing

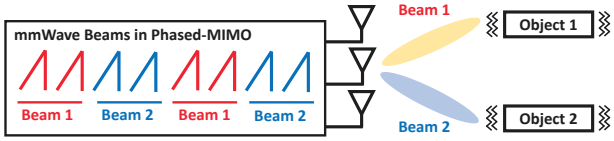


Figure 2: Illustration of the designed phased-MIMO radar capable of sensing and fusing passive speech vibrations on multiple room objects.

all antennas to boost the signal-to-noise ratio (SNR) for speech eavesdropping.

- We design a deep speech recognition scheme based on unsupervised domain adaptation. Our attack can adapt a model built on public audio datasets to recognize speech in passive vibrations, without any prior training.

2 ATTACK DESIGN

Threat Model. We assume a scenario where the victim uses voice communication in a private space/confidential room with restricted access. An adversary tries to breach the security and eavesdrop on the voice communication without physical access to the space/room. The adversary can compromise a mmWave-enabled device by injecting malware and log mmWave signals to capture passive speech vibrations on room objects.

Passive Vibration Sensing via mmWave. Our attack senses the passive vibrations upon room objects based on Frequency-Modulated Continuous-Wave (FMCW) radar, which transmits a sequence of chirp signals sweeping across a fixed bandwidth. The distance between the object and the radar, $R(t)$, is encoded in the received signal $r(t)$:

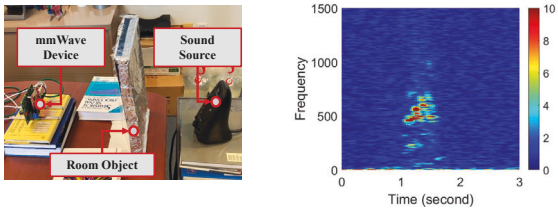
$$r(t) = \alpha exp(j2\pi(2f_c + \beta t)R(t)/c), \quad (1)$$

where f_c is denoted as the carrier frequency, and β is the slope of the chirp. α is an attenuation factor dependent on the propagation distance. We apply dechirp and range-FFT operations on the received signal to determine a particular distance where the object is placed, and extract the time-series signals at this distance:

$$y(t) = \alpha exp(-j4\pi f_c(R(t))/c), \quad (2)$$

where the phase part $exp(\cdot)$ of the time-series signal is associate with the vibrations of the object's surface.

Phased-MIMO Sensing System. Although mmWave sensing can capture displacement of objects, it is still challenging to precisely measure passive vibrations that are orders of magnitude smaller than the wavelength (i.e., micrometer-level), especially under the signal distortions induced by propagation. Therefore, we design a system based on time-division-multiplexing (TDM) phased-MIMO [3], which can significantly enhance the directivity and the sensitivity of mmWave sensing. As shown in Figure 2, for a group of consecutive chirps, the system transmits the FMCW waveform through a subset of the transmitting antennas. The antenna weights are chosen so that a beam, focusing the transmitted power to the direction of the object, is formed. In addition, TDM-MIMO operation enables the formation of a virtual receiving array aperture longer than that of the physical receiving array. The combination of the beamforming can MIMO significantly boosts the SNR, while enabling sensing and fusing vibrations on multiple objects.



(a) Experimental setup (b) Spectrogram of "Zero"

Figure 3: Capturing speech from a loudspeaker by sensing the passive vibrations on an aluminum foil.

Speech Recognition based on Deep Learning. The time-frequency patterns of speech vibrations are dependent on many practical factors, such as the materials of objects, the victim's vocal characteristics, the distance between the object and the sound source, etc. It is unlikely that an adversary can obtain labeled data considering all these impacting factors to build a general speech recognition model. To enable a practical attack, we first leverage public audio datasets to pre-train a bidirectional recurrent neural network (BRNN)-based speech recognition model. Then, the collected vibration data (i.e., without labels) are utilized to adapt the parameters of the BRNN model through adversarial training with a domain discriminator, which transfers the knowledge of speech to the victim's vibration data. The adapted model is then utilized to predict the speech content.

3 PRELIMINARY EVALUATION

As a proof of concept, we implement the proposed TDM phased-MIMO radar on an off-the-shelf TI AWR2243 mmWave device operating within a $76GHz \sim 81GHz$ frequency range. The mmWave device consists of three transmitting and four receiving antennas. An evaluation board TI DCA 1000 is used to acquire raw I/Q signals for speech vibration sensing. We collect audio recordings of digits of $0 \sim 9$ from two participants (i.e., 10 repeats per digit per participant), and then replay the audio using a Logitech Z623 loudspeaker with 85dB volume. An aluminum foil is used as a room object to capture speech vibrations as illustrated in Figure 3 (a), and an example spectrogram of passive speech vibrations of "Zero" is shown in Figure 3 (b). Under this setting, our attack achieves 91.9% accuracy in recognizing the 10 digits.

4 ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation Grants ECCS2033433, CCF1909963, CNS2120396.

REFERENCES

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

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

[3] Zhaoyi Xu, Cong Shi, Tianfang Zhang, Shuping Li, Yichao Yuan, Chung-Tse Michael Wu, Yingying Chen, and Athina Petropulu. 2022. Simultaneous Monitoring of Multiple People's Vital Sign Leveraging a Single Phased-MIMO Radar. *IEEE J-ERM* (2022).