Wavesdropper: Through-wall Word Detection of Human Speech via Commercial mmWave Devices

CHAO WANG, Zhejiang University, ZJU-Hangzhou Global Scientific and Tech Innovation Center, China
FENG LIN*, Zhejiang University, ZJU-Hangzhou Global Scientific and Tech Innovation Center, China
ZHONGJIE BA, Zhejiang University, China
FAN ZHANG, Zhejiang University, China
WENYAO XU, State University of New York at Buffalo, United States
KUI REN, Zhejiang University, China

Most existing eavesdropping attacks leverage propagating sound waves for speech retrieval. However, soundproof materials are widely deployed in speech-sensitive scenes (e.g., a meeting room). In this paper, we reveal that human speech protected by an isolated room can be compromised by portable and commercial off-the-shelf mmWave devices. To achieve this goal, we develop *Wavesdropper*, a word detection system that utilizes a mmWave probe to sense the targeted speaker's throat vibration and recover speech contents in the obstructed condition. We proposed a CEEMD-based method to suppress dynamic clutters (e.g., human movements) in the room and a wavelet-based processing method to extract the delicate vocal vibration information from the hybrid signals. To recover speech contents from mmWave signals related to the vocal vibration, we designed a neural network to infer the speech contents. Moreover, we explored word detection on a conversation with multiple (two) probes and reveal that the adversary can detect words on multiple people simultaneously with only one mmWave device. We performed extensive experiments to evaluate the system performance with over 60,000 pronunciations. The experimental results indicate that Wavesdropper can achieve 91.3% accuracy for 57-word recognition on 23 volunteers.

CCS Concepts: • **Security and privacy → Usability in security and privacy**.

Additional Key Words and Phrases: word detection, mmWave sensing, through walls

ACM Reference Format:
Chao Wang, Feng Lin, Zhongjie Ba, Fan Zhang, Wenyao Xu, and Kui Ren. 2022. Wavesdropper: Through-wall Word Detection of Human Speech via Commercial mmWave Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 77 (June 2022), 26 pages. https://doi.org/10.1145/3534592

1 INTRODUCTION

Voice privacy has drawn increasing attention in recent years [1, 2, 21]. Due to the development of computers and mobile devices, speech communication becomes easier and more efficient. However, on the other hand, more attack surfaces are exposed, which results in sensitive information leakage. The security threats to an

[*] Feng Lin is the corresponding author.

Authors' addresses: Chao Wang, Zhejiang University, ZJU-Hangzhou Global Scientific and Tech Innovation Center, Hangzhou, Zhejiang, China, wangchao5001@zju.edu.cn; Feng Lin, Zhejiang University, ZJU-Hangzhou Global Scientific and Tech Innovation Center, Hangzhou, Zhejiang, China, flin@zju.edu.cn; Zhongjie Ba, Zhejiang University, Hangzhou, Zhejiang, China, zhongjieba@zju.edu.cn; Fan Zhang, Zhejiang University, Hangzhou, Zhejiang, China, fanzhang@zju.edu.cn; Wenyao Xu, State University of New York at Buffalo, Buffalo, New York, United States, wenyaoxu@buffalo.edu; Kui Ren, Zhejiang University, Hangzhou, Zhejiang, China, kuiren@zju.edu.cn.

2474-9567/2022/6-ART77 $15.00
https://doi.org/10.1145/3534592

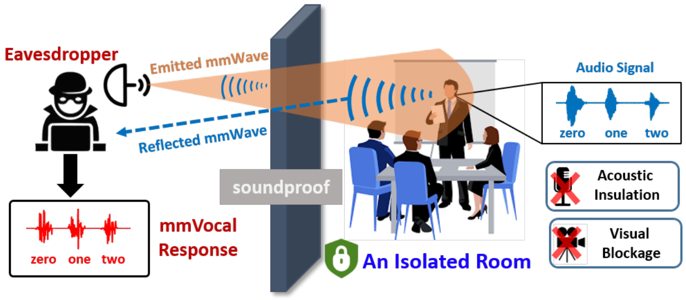


Fig. 1. An attacker can leverage a COTS mmWave probe to launch through-wall word detection of human speech protected in a soundproof scenario.

unsuspecting victim's voice or conversation in a speech-sensitive scene (e.g., a meeting room) can expose private information (e.g., credit card numbers, passwords, and social security numbers). What is worse, the leaked speech contents can cause severe damage to enterprise benefits [4, 36] if sensitive information is involved (e.g., transaction numbers, enterprise decisions). To mitigate acoustic eavesdropping, protecting measures such as soundproofing, are widely adopted in speech-sensitive scenarios (e.g., a conference room, an enterprise office).

Attacks leveraging propagating sound waves [19] and sound wave-induced vibration [35] can fail when the sound wave is constrained by soundproof materials. Although soundproof obstacles can protect the propagating sound waves from leakage, they cannot guarantee the direct leakage from the sound source (e.g., human speakers in this paper). Based on that, we wonder whether the vocal source (i.e., vocal cords) of a human speaker can leak speech information even though the human speaker is protected by a soundproof environment. For example, an adversary may leverage a high-precision device, such as a mmWave probe, to sense the speaker's near-throat skin vibration through the wall and retrieve speech contents remotely. mmWave has been widely adopted in automotive driving [11, 45] and 5G networks [6, 44]. There are many manufacturers of mmWave devices, such as Texas Instruments and NXP Semiconductors, supplying mmWave devices to the public. On one hand, these commercial off-the-shelf (COTS) mmWave devices benefit human life, such as vital sign detection [7, 16, 51], home monitoring [17, 40], and smartphone interaction [22]. On the other hand, an adversary may leverage these widely available and high-precision devices for evil things such as speech eavesdropping. In this paper, we try to investigate whether human speech protected by soundproof measures can be compromised by an outside attacker equipped with a COTS mmWave device. Specifically, as shown in Figure 1, the attacker uses a portable and COTS mmWave probe outside the room to capture the victim's near-throat skin vibration through the wall to recover speech contents, which is resistant to soundproof measures.

However, there are multiple challenges to achieve such a through-wall attack. (1) The attacker has no knowledge about the environment setting due to the blockage of the wall. It is vital to make the attack resilient to environmental changes and locate the speaker. (2) The objects around the speaker (including static and moving objects) can induce clutters, it is important to eliminate the interference and extract speech-related mmWave components from the hybrid reflected signals. Furthermore, as a common situation, the speaker may have motion artifacts (e.g., body wiggles and gestures) during the speech, which can cause speech-irrelevant mmWave echoes and make the extraction of delicate vocal vibration more challenging. (3) Assume that the attacker successfully acquires the mmWave components that contain the speech information, a transition model is required to translate the mmWave signals into speech contents.

To this end, we propose **Wavesdropper**, a through-wall word detection system to retrieve human-rendered speech based on a portable COTS mmWave probe (i.e., TI IWR1642Boost). The proposed Wavesdropper can locate the targeted speaker behind the wall and transmit mmWave to interrogate his/her throat vibration. By applying a spatial-temporal analysis, Wavesdropper can differentiate the speaker from the background and eliminate the impact of background echoes. Afterward, a CEEMD-based clutter suppression is applied to eliminate the dynamic interference in the room (e.g., moving objects). A high-pass filtering and a wavelet-based analysis are further used to filter out the impact of the speaker's motion artifacts (e.g., body wiggle and gestures) and extract clean vocal vibration, namely mmVocal response, containing speech information. Eventually, Wavesdropper segments mmVocal responses into single words automatically and feeds them into a ResNet-based neural network (i.e., *WavesdropNet*) to recover intelligible speech contents. We introduce the proposed system in Section 4 and evaluate the system performance and robustness in Sections 5 and 6. We also conduct an exploration of word detection on multiple speakers in Section 7.

The contributions of our work are as follows:

- We reveal that malicious adversaries can turn widely-available COTS mmWave sensors into eavesdroppers to cause threats to human speech protected by soundproof materials. We investigate the feasibility of using COTS mmWave devices to eavesdrop on human-rendered speech in a through-wall scenario.
- We develop an end-to-end system Wavesdropper for through-wall word detection. Based on our proposed CEEMD-based clutter suppression and wavelet-based signal processing scheme, Wavesdropper can eliminate the clutter interference and extract speech information from reflected hybrid mmWave signals. With a well-designed retrieval model, speech contents can be recovered with high accuracy.
- We conduct experiments to evaluate the system which achieves 91.3% accuracy for 57-word recognition on 23 volunteers. We also test the system robustness under different conditions and give the countermeasures.

2 ATTACK OVERVIEW

2.1 Attack Scenario

We consider a scenario that a victim has a private conversation in an isolated environment (e.g., an enclosed conference room). To ensure the confidentiality of the talk, the victim takes protection measures against attackers, such as sound insulation and visual sheltering. To compromise the privacy and security of the victim's speech, an attacker aims to detect sensitive words in the speech. The attacker can steal the following information by the performed word detection: 1) personal privacy and secret information, e.g., words related to private conversation and passwords; 2) enterprise interest, e.g., words of a confidential meeting that involves enterprise decisions and dates of economic transactions.

2.2 Threat Model

We assume that the victim is under speech protection (soundproof and obstructed environment) as mentioned above. The speech of the victim contains words related to private information. In this paper, we mainly focus on words made up of numbers and hot words, such as passwords, credit card numbers, and social security numbers. The goal of the attacker is to detect and recognize these words to retrieve confidential information. We assume the attacker can acquire the victim's mmWave data when he/she speaks to train the model in advance. This can be achieved by transmitting mmWave to the victim remotely when he/she speaks in some scenarios that do not have protections, such as a public coffee bar. Once the model is trained, it can be used for the following attack. We assume that there is a soundproof and opaque wall between the victim and the attacker. Thus, acoustic-based and visual-based eavesdropping methods will fail. We also assume that the attacker cannot get physical access to the protected room and cannot deploy any eavesdropping devices near the victim.

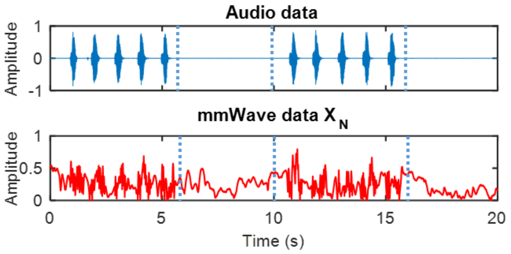


Fig. 2. The volunteer says “one” ten times in front of the probe (the speech takes place in $0 \sim 6s$ and $10 \sim 16s$).

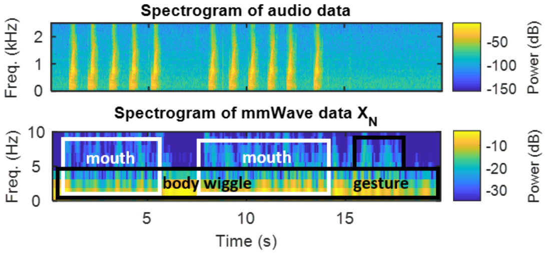


Fig. 3. The frequency of speech-irrelevant actions (black boxes) overlaps with the mouth movement’s (white boxes).

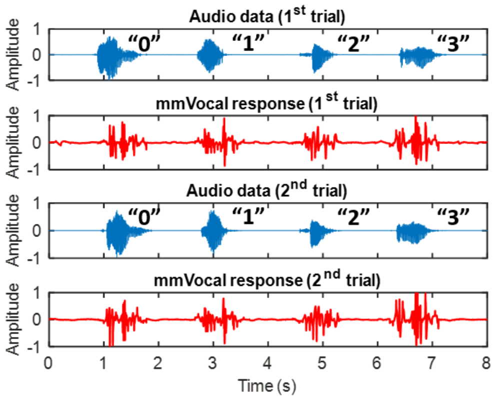


Fig. 4. Two trials of extracted mmWave signals and corresponding audio data when the speaker says “zero”, “one”, “two”, “three”.

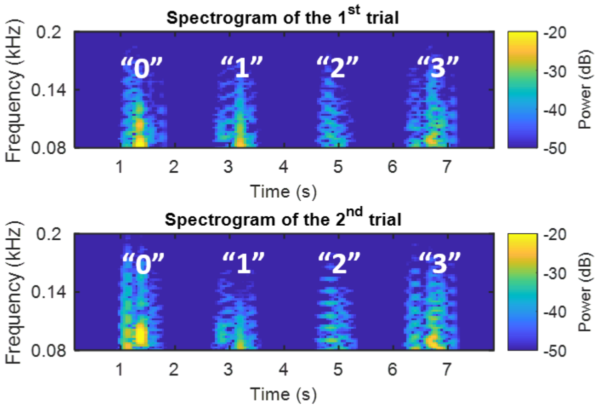


Fig. 5. The mmWave signals of the same word show a high similarity but are different from other words’ in the spectrograms (from 80Hz to 200Hz).

3 PRELIMINARIES STUDY

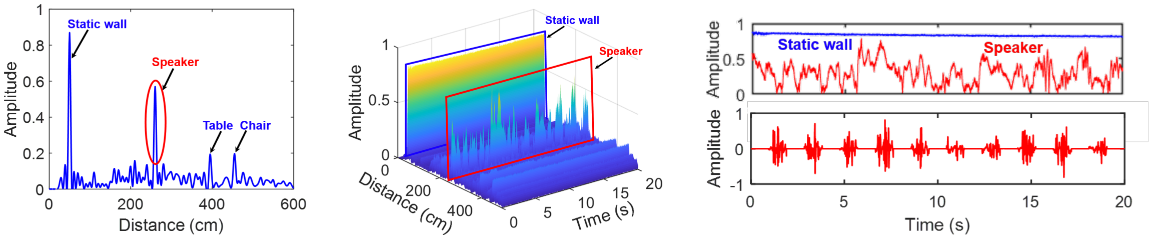
In this section, we present our preliminary results on word detection of human-rendered speech via a COTS mmWave device, i.e., TI IWR1642Boost.

3.1 Mechanism of Voice Generation

When a human speaks, the lungs first produce adequate airflow and air pressure to vibrate the vocal cords. By adjusting the length and tension, vocal cords can fine-tune the pitch and tone of the sound. Then the articulators pronounce and filter the sound to generate human-rendered speech. With the different vibratory frequency of vocal cords and movement of articulators, the human voice is modulated to generate different speech contents. (*We denote the region of near-throat skin as* ***Vocal Region***.) In other words, there is a close relationship between the speech contents and vocal vibration. If an adversary acquires the victim’s vocal vibration with a high-precision sensor, such as a mmWave probe, speech information leakage can happen.

3.2 Eavesdropping with mmWaves

Frequency-Modulated Continuous Wave (FMCW) in the mmWave band is widely adopted in automatic driving and small vibration measurement [24, 28, 31, 49]. The mm-level wavelength of the transmitted signal allows a



(a) The result of range-FFT.

(b) Spatial-temporal analysis.

(c) Human echo and mmVocal response.

Fig. 6. (a) The four peaks indicate four detected objects. We can differentiate the speaker from the backgrounds by applying a spatial-temporal analysis. (b) The variation of the speaker's human echo is far larger than the static backgrounds. (c) The below figure shows the extracted mmVocal response from the red waveform (corresponding to the speaker) in the above figure. Note that the waveforms in the above and below figures are normalized to [-1,1] respectively for better representation.

higher detection accuracy of object movements (e.g., range and velocity) compared with a lower frequency band for sensing. The frequency-modulated transmitted signal is called a chirp which increases linearly with time. With a series of chirps transmitted and received, the mmWave sensor can locate objects with accurate distance and angle estimation. For simplicity, the received waveform can be taken as a delayed replica of the transmitted signal. The receiver down-converts the received signal to the baseband using a local copy of transmitted signals. The frequency shift Δf (i.e., the beat frequency f_b) between the transmitted chirp signal and reflected signal is linearly proportional to the delay Δt, $f_b = \Delta t \cdot S$, $S = B/T$, where B and T are the bandwidth and frequency-modulation sweep time of the transmitted chirp, respectively. Then the baseband signal is sampled by an analog-to-digital converter to generate digital data for further processing. By applying Fast-Fourier-Transform (FFT) to the baseband digital data (which is called a range-FFT [42]), we obtain the range of detected object as $R = c \cdot f_b/(2S)$. Due to the short wavelength of mmWave, the small vibration can be sensed and detected by calculating the phase change $\Delta\varphi$ of the demodulated reflected signal. Given λ as the wavelength of mmWave, the small displacement of vocal vibration Δd can be calculated by $\Delta d = \frac{\lambda \Delta\varphi}{4\pi}$, where $\Delta\varphi$ is the phase change of the demodulated signal. The phase change of demodulated signal is sensitive to the small displacement. For a 77GHz mmWave probe, a 1mm displacement change will cause a phase shift of π in the demodulated signal.

When an attacker uses a mmWave probe to transmit FMCW toward the throat area of a speaker periodically, the near-throat skin vibration will reflect the mmWave that can be captured by the probe, which implies that the reflected mmWave contains speech information. *Hereafter, we define the reflected mmWave signal's components caused by near-throat skin vibration as **mmVocal Response**. In Section 3.3, we demonstrate with our preliminary results that the mmVocal response is closely related to the speech contents.*

3.3 A Feasibility Study

In this part, we investigate the feasibility of detecting words of human-rendered speech by interrogating the speaker's near-throat skin vibration with mmWave. To make things easier, we start with an unobstructed scenario. We ask a volunteer to speak the word "one" in front of a fixed mmWave probe from a distance of 2m. At the same time, we use a microphone to record the audio data as the reference signal. The audio signal and derived mmWave data X_N are shown in Figure 2. We find that when the volunteer speaks ($0 \sim 6s$, $10 \sim 16s$), the mmWave data contains some high-frequency components, which is most likely the *mmVocal response*.

3.3.1 Interrogated Vibration Source. Since the transmitted mmWave is a wave beam with a specific beamwidth (about ±35°), the reflected mmWave can be affected by the near-throat skin vibration, the mouth movement, and other speech-irrelevant artifacts (e.g., body wiggles) of the speaker. *Hereafter, we define the reflective mmWave signal affected by the speaker as* ***Human Echo****, i.e.,* X_N*. To achieve the goal of speech information retrieval, one of the key steps is to extract the mmVocal response from the human echo.* The typical frequency components of human mouth movement and motion artifacts (e.g., body wiggles) are below 12Hz [37] and 10Hz [53], respectively. To investigate the impact of motion artifacts on mouth movement in the spectrogram, we asked the volunteer to shake his body ($0 \sim 20s$) and raise his hands ($15 \sim 18s$) while speaking with mouth movement ($0 \sim 6s$, $8 \sim 14s$). As shown in Figure 3, we observed that the frequency of mouth movement is disturbed by the body wiggle and gesture (especially frequencies below 5Hz). In other words, if we use the mouth movement for speech retrieval, it can be affected by the speech-irrelevant motion artifacts. However, the fundamental frequency of vocal vibration is between 85Hz and 255Hz, higher than the frequency of motion artifacts. To achieve a robust speech retrieval, we apply a high-pass filter and wavelet-based analysis (detailed in Section 4.3) to eliminate the impact of motion artifacts (along with the mouth movement impact) and extract the high-frequency components corresponding to the mmVocal response for speech retrieval.

Next, we perform further experiments to investigate the consistency of *mmVocal response* (i.e., the near-throat vibration) and the speech contents. We ask the volunteer to speak “zero, one, two, three” twice in front of the fixed mmWave probe from a distance of 2m and also use the microphone to record audio data as the reference signal. The two trails of recorded audio and *mmVocal responses* are shown in Figure 4. The short-time Fourier transform of the two *mmVocal responses* are shown in Figure 5. We can observe an obvious discrepancy between different words in the spectrogram (from 80Hz to 200Hz) while the *mmVocal responses* of the same word show high similarity, which indicates that the ***mmVocal response has a unique and persistent relationship with human speech contents***.

3.3.2 Through-wall Attack. As demonstrated in the threat model, the adversary cannot get physical access to the isolated zone and there is an opaque and soundproof wall between the victim and the adversary. Thus, it is crucial to investigate the impact of the wall on the *mmVocal response*. We asked a volunteer to speak the words (i.e., “zero, one, ..., nine”) towards the mmWave probe (2m away). There was a soundproof-glass wall between the mmWave probe and the volunteer. The transmitting power for each channel of the probe is 12.5 dBm. The range-FFT [42] result is shown in Figure 6(a). The peaks represent detected objects by the probe, which indicates the probe can detect objects in a through-wall scenario. However, it fails to tell apart which peak corresponding to the speaker and distinguish the speaker from backgrounds only by the range-FFT result. So we further develop a spatial-temporal analysis approach to solve this problem (detailed in Section 4.2). Figure 6(b) depicts the result of spatial-temporal analysis of the speaker’s position changes with time. Figure 6(c) (above) depicts the mmWave data X_N of the speaker (red curve) and the wall (blue curve). We can observe that the variation of the mmWave data X_N corresponding to the speaker is far larger than the static wall and other static layouts (e.g., table and chair), which helps to differentiate the speaker from backgrounds. To differentiate the speaker from other moving objects, we apply a speaker detection mechanism in the spatial-temporal analysis as introduced in Section 4.2. Figure 6(c) (below) depicts the extracted *mmVocal response*. The results show that *mmVocal response* can still be acquired through the wall with our proposed spatial-temporal analysis.

3.3.3 Challenges of the through-wall Attack. Although the feasibility study shows a promise of the through-wall eavesdropping, there are several challenges we need to overcome:

- How to locate the speaker in the obstructed scenario (e.g., a speaker in an isolated room). (The solution is detailed in Section 4.2.)

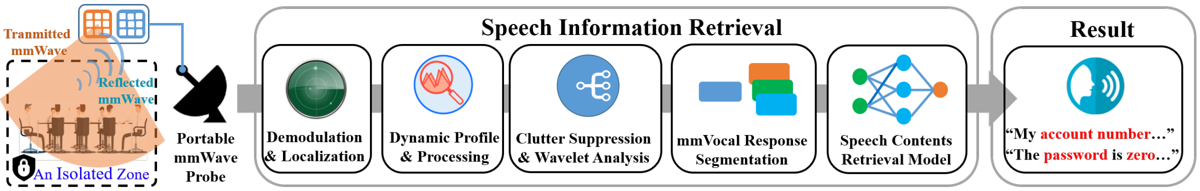

Tranmitted mmWave
Reflected mmWave
An Isolated Zone
Portable mmWave Probe
Speech Information Retrieval
Demodulation & Localization
Dynamic Profile & Processing
Clutter Suppression & Wavelet Analysis
mmVocal Response Segmentation
Speech Contents Retrieval Model
Result
"My account number…"
"The password is zero…"

Fig. 7. The system framework of Wavesdropper.

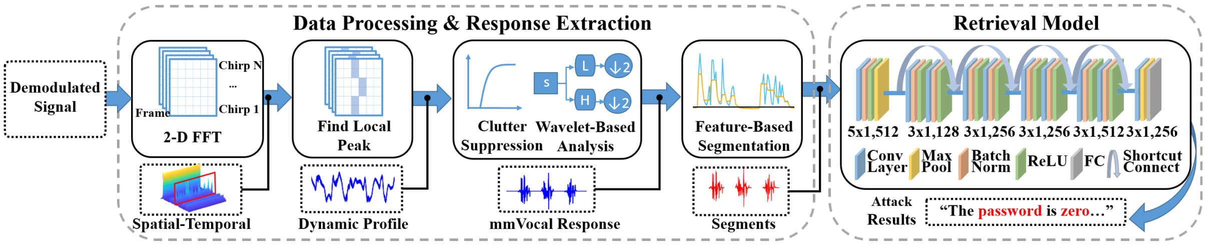


Fig. 8. The flow chart of Wavesdropper.

- How to eliminate the clutters induced by objects in the room, especially moving objects which may saturate the receiver and interfere with extracted mmVocal responses. Based on our observation, the large-scale movements of listeners near the speaker can induce glitch-like noise in the extracted mmWave signals. (The solution is detailed in Section 4.3.)
- How to design a transition model to translate the processed mmWave signals into intelligible speech contents. (The solution is detailed in Section 4.4.)

4 SYSTEM DESIGN

In this section, we detail *Wavesdropper*, a portable and robust system for through-wall word detection of human-rendered speech. The framework of *Wavesdropper* is shown in Figure 7.

4.1 Wavesdropper: A mmWave-based Word Detection System

The flow chart of *Wavesdropper* is shown in Figure 8. Once the raw mmWave is received and demodulated by the probe, the *human echo* corresponding to the speaker is first extracted and filtered to eliminate the clutter interference. Then a wavelet-based analysis is employed to extract the *mmVocal response* which contains speech information. A feature-based segmentation module segments the mmVocal responses into short signals, each of which corresponds to a single word. Finally, the segments are fed into *WavesdropNet*, a ResNet-based speech retrieval model to further extract representative features and recover speech contents.

4.2 Speaker Localization

4.2.1 Locating the Speaker. To retrieve the speech information, the first thing we need to consider is how to locate the speaker (i.e., near-throat skin). The mmWave probe has a 3dB beamwidth of ±35° (denoted as Field-of-view, FoV), out of which the SNR decreases rapidly. Within the FoV, *Wavesdropper* can recover human speech information with high accuracy (refer to Section 6.1). When the target is out of the FoV, we should adjust the mmWave beam to the targeted speaker's vocal region. To achieve this goal, *Wavesdropper* adopts a localization method in radar sensing to guide the beam-steering to boost the eavesdropping performance. By applying 2-D

FFT [42] to the demodulated mmWave data of the four receiving channels, *Wavesdropper* can acquire the distance d and horizontal angle θ of the target to guide the beam-steering. For M demodulated chirps with N samples of each (i.e., $x(i,j)$), we get the samples matrix

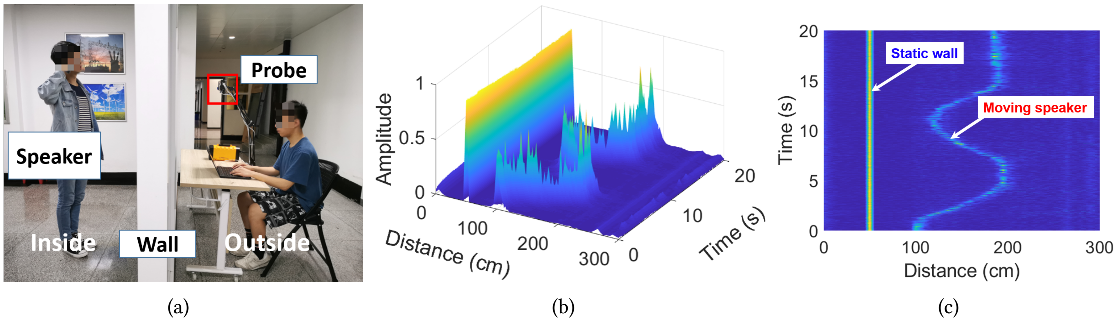
$$X_{M\times N} = [x(i,j)]_{M\times N}, i \in [1,M], j \in [1,N]. \tag{1}$$

The 2-D FFT is achieved by applying FFT to the matrix $X_{M\times N}$ by row and column successively (which is known as range-FFT and doppler-FFT). Then the distance and angle of arrival for a detected object can be calculated as:

$$d = \frac{f \cdot c}{2S}, \theta = \arcsin \frac{\lambda \Delta\phi_{ab}}{2\pi d_{ab}}, \tag{2}$$

where f is the frequency of the local peak in the range-FFT spectrum, c is the speed of light in the vacuum, S is the bandwidth of the transmitted chirp, λ is the wavelength of mmWave, and $\Delta\phi_{ab}$ is the phase difference between receiving antenna a and b.

4.2.2 Spatial-temporal Analysis. Considering that the *human echo* X_N can be used to differentiate the dynamic objects (including the speaker) from other static objects, we calculate the standard deviation of X_N for every detected object. For n detected dynamic objects in the room, we can get n traces of $x(i,j)$. To differentiate the speaker from other dynamic objects, we first apply band-pass filtering (with cut-off frequencies of 85Hz and 255Hz) on the derived mmWave data and then calculate the power spectral density for each trace. Then the one with the highest intensity value will be chosen as the speaker's. The rationale behind this is that the throat vibration contains most of fundamental frequency of human speech ($85 \sim 255Hz$) so a trace with a high power spectral density within $85 \sim 255Hz$ can be the speaker. Then we calculate the horizontal angle θ corresponding to the speaker and steer the beam to the specific direction. The location of the speaker may change when he/she speaks in a realistic scenario (we assume he/she is still within the FoV), so the peak corresponding to the target in the range-FFT can shift with time, which makes it difficult to extract the *mmVocal response.* To solve this problem and make the attack system more robust, we apply a peak-searching algorithm [12] to the spatial-temporal analysis to trace the moving speaker and then extract *mmVocal responses* for speech information retrieval. As shown in Figure 9(a), we asked a volunteer to step back and forth behind the wall (inside the room) and then we collected the mmWave data from outside the room. From Figure 9(b)(c), we can observe the peaks corresponding to the volunteer shift along the distance axis.



(a) (b) (c)

Fig. 9. Spatial-temporal analysis. The shift of peak along the distance axis indicates a moving speaker.

Algorithm 1: CEEMD-based Dynamic Clutter Suppression

```
Input: $s(n)$: Raw mmWave signal, $n = 1, 2, ..., N$
Output: $s_r(n)$: Reconstructed mmWave signal, $n = 1, ..., N$
Initialize the index set of noise-interfered segments $S = \{\}$;
for $0 < k_1 < k_2 < N$ do
    if $\frac{\sum_{k=k_1}^{k_2} |s(k)|^2}{k_2 - k_1 + 1} > \frac{\sum_{n=1}^{N} |s(n)|^2}{N}$ then $(k_1, k_2) \in S$;
$\{IMF_i(n)\} = CEEMD(s(n)), n = 1, ...N, i = 1, ..., I$;
for $i = 1, ..., I$ do
    if $\frac{\sum_{(k_1,k_2) \in S} \sum_{k=k_1}^{k_2} |IMF_i(k)|^2}{\sum_{n=1}^{N} |IMF_i(n)|^2} > \epsilon$ ($\epsilon = 0.9$ empirically) then
        $T_r = \sigma_i \sqrt{2log(N)}$, $\sigma_i$ is the noise variance of $IMF_i(n)$;
        $IMF'_i(n) = \begin{cases} 0 & |IMF_i(n)| \le T_r \\ (2 * sigmoid(IMF_i) - 1)(|IMF_i| - T_r) & |IMF_i(n)| > T_r \end{cases}$
    else
        $IMF'_i(n) = IMF_i(n)$;
$s_r(n) = \sum_{i=1}^{I} IMF'_i(n), n = 1, ..., N$;
return $s_r(n), n = 1, ..., N$
```

4.3 Clutter Suppression

The raw mmWave signal (i.e., *human echo*) can be distorted by human actions and movement that should be eliminated for clean *mmVocal response* extraction. To solve this problem, we develop a CEEMD-based clutter suppression algorithm to eliminate the dynamic clutters mostly induced by dynamic (moving) objects in the environment. Then we apply a high-pass filter and a wavelet-based analysis to mitigate the speaker's body movement according to our preliminary study in Section 3.3 and extract the delicate vocal vibration (i.e., mmVocal response).

4.3.1 Dynamic Clutter Suppression. When the victim speaks, the random body motion of listeners nearby can induce glitch-like noise in the extracted mmWave signals (as the red-elliptic area in Figure 24(a) shows). We find that the induced noise is hard to eliminate by applying a digital filter because the magnitude of the noise spectrum envelop is random. If not properly mitigated, the noise can reduce the system performance on speech recovery. To suppress the dynamic clutters, we develop a CEEMD-based adaptive noise cancellation algorithm (Algorithm 1). The complete ensemble empirical mode decomposition (CEEMD) [43] is used to analyze non-stationary signals. It decomposes the mixed data into several intrinsic mode functions (IMFs) and has a good spectral separation of the modes. Considering that the clutter can have high energy with a short duration, we first apply an energy-based detection to locate the glitch-noise segment. Then we decompose the original mmWave signal into different IMFs. To suppress the interference, an intuitive way is to discard the noise-related IMFs and then reconstruct the signal. However, this method will also abandon part of useful components, which damages the completeness of the mmVocal response and causes a false speech inference. So we proposed a soft-threshold-based method in Algorithm 1 to suppress the noise and reserve useful components for mmVocal response reconstruction. We show an example of dynamic clutter suppression in Figure 24 and evaluate this proposed method in Section 7.1.

4.3.2 High-pass Filtering. As demonstrated in Section 3.3, the speaker's actions (e.g., gestures and body wiggle) can cause low-frequency components in the human echo as shown in Figure 3. Considering that the typical

Algorithm 2: Segment The *mmVocal Response*

```
Input: $s(m)$: mmVocal response, $m = 1, 2, ..., M$
Output: $S_k$:Segments, $k = 1, ..., K$
Divide $s(m)$ into 50ms-length frames $s_i(n), n = 1, 2, ..., N; i = 1, 2, ..., I$;
for $i \in \{1, ..., I\}$ do
    calculate feature values: $E(i)$ and $C(i)$;
Calculate the histograms of $E(i)$ and $C(i)$ and estimate the thresholds $T_E$ and $T_C$;
for $i \in \{1, ..., I\}$ do
    while $E_i > T_E \&\& C_i > T_C$ do
        $s_{i(n)}$ is an active frame;
Merge successive active frame and get segments $S_k, k = 1, ..., K$;
return Segments: $S_k, k = 1, ..., K$
```

frequency of human motion is below 10Hz [53], we design a high-pass Butterworth filter with a cut-off frequency of 80Hz to eliminate the speaker's motion interference. After filtering, we take further steps (i.e., wavelet-based analysis) to extract the clean *mmVocal response* that contains speech information.

4.3.3 Wavelet-based Analysis. Wavelet transform is an effective multi-resolution analysis tool for signal decomposition. Its fine-grained multi-scale analysis on both time and frequency domains is beneficial for speech information extraction. After high-pass filtering and eliminating most human motion interference, the filtered signal *s(t)* becomes a signal with zero-mean and satisfies the following conditions: $\int_{-\infty}^{\infty} s(t)\mathrm{d}t = 0$,where *s(t)* is a waveform. Wavelet transform uses $\psi_{a,b}$ and $\phi_{a,b}$, where $\psi_{a,b} = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a})$ and $\psi_{a,b} = \frac{1}{\sqrt{a}}\phi(\frac{t-b}{a})$, as the mother wavelet function that satisfies the condition of dynamic scaling and shifting, where a and b are the scale and translation parameters accordingly [41]. To get the *mmVocal response* at high frequency, the wavelet-based analysis is achieved as Equation 3:

$$s(t) = A_0 + D_1 + D_2 + D_3 + D_4 + D_5 + D_6, \tag{3}$$

where $A_0 = \frac{1}{C_\phi}\int_{-\infty}^{\infty} F_W(a_0, b)\phi_{a_0,b}\frac{\mathrm{d}b}{\sqrt{a_0}}$ is the approximation part, $D_i = \frac{1}{C_\psi}\int_{-\infty}^{\infty} F_W(a_i, b)\phi_{a_i,b}\frac{\mathrm{d}a}{a_i^2}\frac{\mathrm{d}b}{\sqrt{a_i}}, i = 1, 2, 3, 4, 5, 6$ is the Level i detail part, $F_W(a_i, b)$ is the corresponding coefficient. After the wavelet-based analysis, the *mmVocal response* corresponding to the 4th level detail part is extracted.

4.4 Speech Information Retrieval

When the speech of the targeted speaker contains pre-defined sensitive words, *Wavesdropper* leverages the acquired mmVocal responses to recognize corresponding words and recover sensitive information. However, the *mmVocal response* corresponding to each word is difficult to separate because there is no reference signal to identify the starting and ending points of each pronounced word. To address this problem, we proposed the adaptive segmentation method to acquire the mmVocal responses corresponding to each word. To translate the mmVocal response (mmWave data) to intelligible speech contents, we design a neural network to extract the inner features of mmVocal response which contains speech information, and output semantic information.

4.4.1 Segmentation. To translate the mmVocal responses into intelligible speech, we first need to segment the time sequence into single words and then feed the segments into the speech retrieval model for speech recovery. The rationale of the segmentation is that two successively spoken words have boundaries in both the time and frequency domain. This motivates us to achieve the adaptive segmentation based on the feature of signal energy (time-domain) and spectral centroid (frequency-domain) as introduced in Algorithm 2. *Wavesdropper* first divides the whole *mmVocal response* sequence into M frames with equal length (50ms). Let $s_i(n), n = 1, 2, ..., N$ denote the i-th frame of length N. We call $s_i(n)$ an *active frame* if it is part of *mmVocal response*. For each frame $s_i(n)$, two adopted features (i.e., signal energy $E(i)$ and spectral centroid $C(i)$) are calculated:

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} |s_i(n)|^2, C(i) = \frac{\sum_{k=1}^{N}(k+1)S_i(k)}{\sum_{k=1}^{N} S_i(k)}, \tag{4}$$

where $S_i(k), k = 1, 2, ..., N$ is the discrete Fourier transform (DFT) coefficients of i-th frame. Then we compute histograms H_E and H_C of the two acquired feature sequences. Two adaptive thresholds for each feature sequence are computed:

$$T_E = \frac{WM_{E1} + M_{E2}}{W + 1}, T_C = \frac{WM_{C1} + M_{C2}}{W + 1}, \tag{5}$$

where M_{E1}, M_{E2}, M_{C1}, and M_{C2} represent the first and second local maxima of $E(i)$ and $C(i)$, respectively. Finally, the segments are formed by successive frames whose feature values are larger than both T_E and E_C in Equation 5.

4.4.2 Interpolation & Normalization. Considering that when the same person speaks the same word twice, there is a slight difference in the duration between them, which has a negative impact on the speech contents retrieval. To solve this problem, we interpolate the segments into the same size before feeding them into the retrieval model. In practice, the reflected mmWave signal decays with distance and penetration, which affects the robustness of *Wavesdropper*. To suppress the disturbance of distance and penetrating attenuation, *Wavesdropper* applies normalization to the amplitude of segments. After normalization, the amplitude of all the segments is within [-1,1] and the normalized segments are fed into the speech retrieval model for speech recovery.

4.4.3 Speech Retrieval. Traditional machine learning algorithms, e.g., support vector machine, require expertise to design and extract the features. But the differences among the extracted *mmVocal responses* of some words can be subtle. To tackle this challenge, we design **WavesdropNet**, a residual neural network-based classifier for speech retrieval. The structure of the developed network is shown in Figure 8. Different from ResNet for image classification [20], we choose to take the 1-D *mmVocal response* segments acquired from the 4 channels of mmWave probe as data series with time-domain and channel-domain. The basic residual block adopts a one-dimension convolution kernel and takes the four-channel time series as input [46]. WavesdropNet consists of an initial layer, four residual blocks, and a prediction layer as shown in Figure 8. The initial layer aims to convert four-channel time series $S \in R^{4\times2560}$ into a latent space. The initial layer includes four consecutive operations: a convolution (Conv), a batch normalization (BN), a rectified unit (ReLU), and a max pooling. Each of the four subsequent residual blocks contains two basic residual blocks with a kernel size of 3×1. The residual architecture contains the shortcut connections which help convergence [20]. Let x represent the input of a residual block, the shortcut connections can be formulated as:

$$y = F(x, \{W_i\}) + x, \tag{6}$$

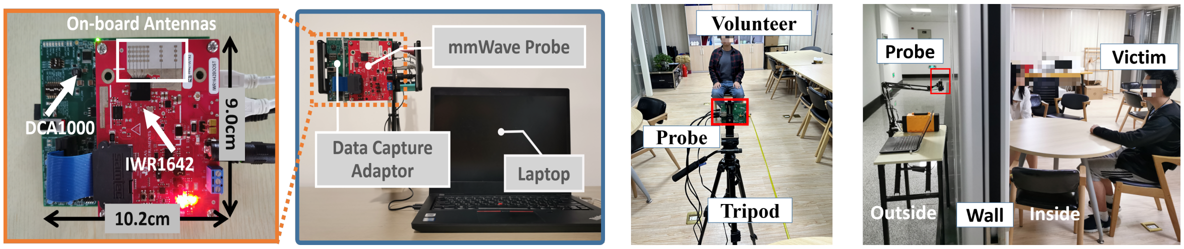
where the function $F(x, \{W_i\})$ represents the residual mapping to be learned, and $F(x, \{W_i\}) + x$ represents element-wise addition. According to He *et al.* [20], the shortcut connections make it easy to optimize the network. In the prediction layer, the output of the residual block is fed into a 3×1 convolution layer, followed by a max pooling operation and a fully-connected layer (FC) for classification. In the training process, we use Cross Entropy Loss [54] as the loss function and choose Adam [25] to optimize network parameters. We implement the retrieval model in Pytorch.

Table 1. Experimental setting.

Section	Experimental setting			
	Num. of participants	Testing distance/orientation[1]	Testing scene (through-wall)	Num. of words
5.3	23	$2.5m/0^\circ$	Conf. room (Soundproof wall)[2]	57
6.1	5	$0.5m \sim 5m/0^\circ$, $2.5m/0^\circ \sim 40^\circ$		
6.2	5	$2.5m/$ $0^\circ \sim 10^\circ$	Conf. room, cafe, and office	
6.3	5	$2.5m/0^\circ$	Glass, sponge, wood, and brick	
6.4-6.6	5	$2.5m/0^\circ$	Conf. room (Soundproof wall)	
7	2	$3.6m/18^\circ$, $3.7m/17^\circ$		

[1] The orientation is defined as the angle deviation of how the speaker's throat facing towards the probe, where 0° means the speaker's throat is facing the probe frontally.

[2] The conference room in this paper is shown in Figure 20(1) and Figure 25. The soundproof wall has a total thickness of 8cm (two 1cm thick glasses and 6cm thick vacuum layer).



(a) Experimental setup of Wavesdropper.

(b) Setting of training data collection.

(c) Setting of testing data collection.

Fig. 10. (a) The setup of Wavesdropper, (b) the setting of training data collection in a conference room, and (c) the testing data collection by penetrating through the soundproofing glass of the conference room.

5 EVALUATION

5.1 Experimental Setup and Datasets

5.1.1 System Setup. The system setup of Wavesdropper is shown in Figure 10(a). Wavesdropper utilizes a COTS mmWave probe IWR1642Boost [3] for mmWave interrogation. The IWR1642Boost is an integrated single-chip mmWave sensor that operates in 77-81 GHz with 4 GHz bandwidth. It employs phased antenna arrays (i.e., 2 transmitting and 4 receiving antennas integrated on a single board) to generate high antenna gains to fight against the attenuation. The antenna array can steer the beam to a specific direction and thus reduce the interference from the background. The probe has an 18mW transmitting power with a 5V/2.5A power supply and a portable size of $10.2cm \times 9.0cm \times 1.5cm$. The received RF gain is 30dB. The raw mmWave data is collected by a data capture board DCA1000EVM and sent to a laptop (ThinkPad T490). We use the laptop to process the raw mmWave data and infer speech contents.

Table 2. Details about the 57 tested words.

Digits	1.zero 2.one 3.two 4.three 5.four 6.five 7.six 8.seven 9.eight 10.nine
Secret&Gender	11.username 12.password 13.telephone 14.account 15.number 16.credit 17.card 18.balance 19.stock 20.price 21.dollar 22.hundred 23.thousand 24.million 25.male 26.female
Time&Date	27.date 28.time 29.morning 30.afternoon 31.evening 32.Monday 33.Tuesday 34.Wednesday 35.Thursday 36.Friday 37.Saturday 38.Sunday
Political&Sensitive	39.government 40.nuclear* 41.agency* 42.military* 43.emergency* 44.force* 45.army*
Other	46.in 47.at 48.on 49.the 50.my 51.his 52.her 53.owner 54.am 55.is 56.are 57.address

[1] Words with * are chosen from sensitive keywords published by the U.S. government [33].

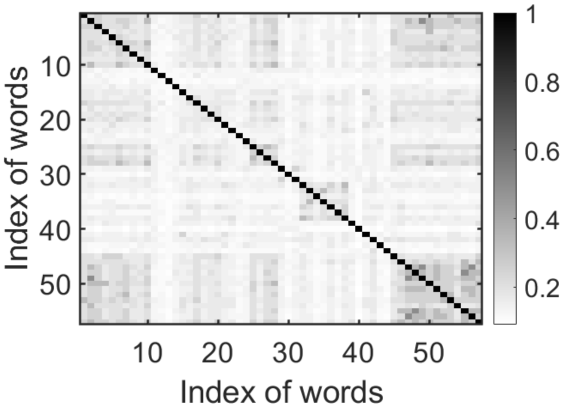


Fig. 11. The Levenshtein Distance matrix of the tested words indicates the similarity among these words.

5.1.2 Tested Words. There are 57 words included in this paper as listed in Table 2. Ten digits (from *zero* to *nine*) are chosen for the consideration that the digits are often related to secrets such as the password. The topic of the hot words relates to people's secret (such as *username, password*) and privacy (such as *male* and *female*), time (such as *morning, Monday*), often-used prepositions (such as *in, at* and *on*), and etc. The tested words include both monosyllabic and polysyllabic words. We use the Levenshtein Distance matrix to quantify the similarity of these words as shown in Figure 11. The Levenshtein Distance [52] between two words is within [0,1] where a higher score indicates a higher similarity.

5.1.3 Training Data Collection. Our experiments involve 23 participants with ages from 19 to 58 years old, including 17 males and 6 females with diverse accents. It is ensured that all the participants follow the host institutional review board (IRB) protocol. In all the experiments, the volunteers were asked to speak in a normal sound pressure level (SPL) within 60-70dBA. We collected the data in a soundproofing conference room as shown in Figure 10(b). We fixed the probe on a tripod. Each volunteer was about 2.5m away from the probe and faced the probe without blockage with the orientation of about 0 degree. We asked the volunteers to keep still when speaking. There was no moving but static objects in the background during the collection. We construct a training dataset containing 45,885 mmVocal response samples of the 57 words from the 23 volunteers (i.e., 1,995 samples from each person). We use the dataset to train the speech retrieval model on a Linux server with two Nvidia GPUs (GeForce GTX 1070). The consuming time for training is about nine minutes.

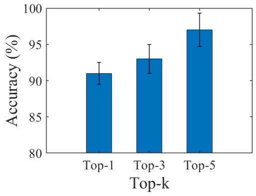


Fig. 12. Overall performance of through-wall attack.

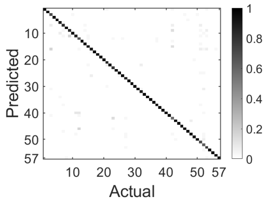


Fig. 13. Confusion matrix of top-1 accuracy.

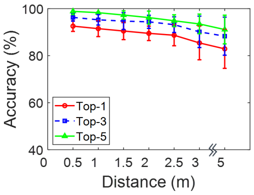


Fig. 14. Distance impact.

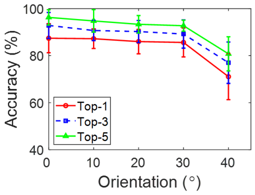


Fig. 15. Orientation impact.

5.2 Evaluation Metrics

5.2.1 Top-k Accuracy. The top-k accuracy is the probability that the correct label is within the top-k classes predicted by the *WavesdropNet*. Specifically, we report the top-1, top-3, and top-5 testing accuracy of the 57-word recognition to clarify the system performance. Without specific clarification, the recognition accuracy is defined as the top-1 accuracy.

5.2.2 mmVocal-Signal-to-noise Ratio (mmVSNR). To facilitate quantitative comparison of *mmVocal response* under different experimental settings, we define the *mmVocal Signal-to-Noise Ratio* as formulated in Equation 7:

$$mmVSNR = 10\log_{10}(\frac{P(s)}{P(n)}) \tag{7}$$

where P is the mean of the summed square magnitude, s is the speech segment in an extracted mmVocal response trace and n is the noise segment in the extracted trace. *Similar to the SNR, a higher mmVSNR indicates a better resilience to speech-irrelevant noise and thus a better system performance for speech retrieval.*

5.3 Overall Performance

We used the collected training data from the 23 volunteers in Section 5.1 to train the model for speech recognition. The model was trained on a Linux server with two Nvidia GPUs (GeForce GTX 1070). We started the training with a learning rate of 10^{-3} and set the batch size to 64. The consuming time for training is about nine minutes. The tested data was collected from the same 23 volunteers mentioned before, i.e., a target-dependent attack which requires prior knowledge about the victim. During the testing phase, each volunteer was asked to sit still in a chair without body motions and speak the 57 words ten times (i.e., 570 samples from each person) in the same soundproof conference room as the data collection. The testing scene is shown in Figure 10(c). The distance of the testing scenario is **2.5m** and the orientation is about 0 degree. We deployed the mmWave probe outside the room to transmit and collect mmWave data. Wavesdropper extracts mmVocal responses from the collected mmWave data and feeds them into the pre-trained model for word recognition. We calculate and show the recognition results in Figure 12 and Figure 13. The top-k accuracies are shown in Figure 12. We can observe that *Wavesdropper* achieves 91.3% average accuracy for 57-word recognition among the 23 speakers. The Top-3 and Top-5 average accuracies are 93.2% and 97.3%, respectively. This indicates the ability of Wavesdropper to retrieve speech contents through the soundproof wall. Figure 13 shows the confusion matrix of the top-1 accuracy of the 57 words. We can observe that the speech retrieval model achieves a high interference accuracy for all the words. Besides, there is no severe bias on word misclassification or confusion among specific classes observed from the result.

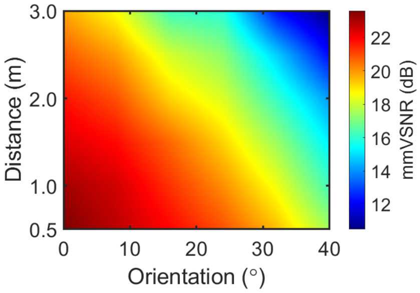


Fig. 16. mmVocal-signal-to-noise ratio

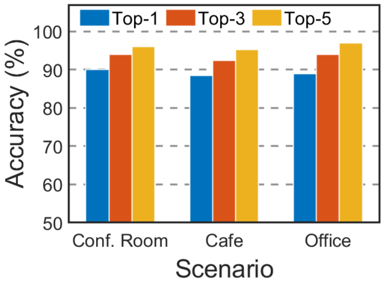


Fig. 17. Impact of environmental change.

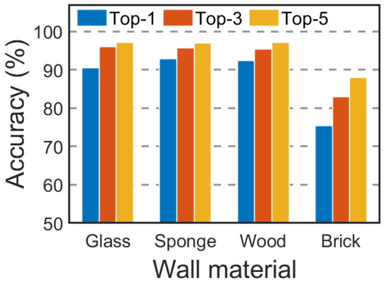


Fig. 18. Impact of blocking wall.

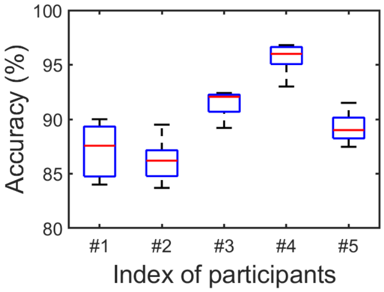


Fig. 19. Impact of body motion.

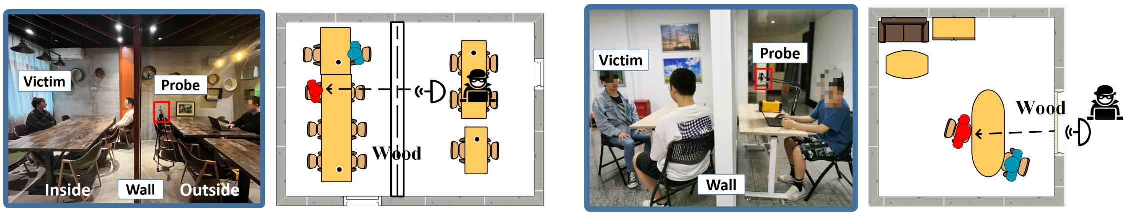
6 COMPLEX ENVIRONMENT ANALYSIS

In this section, we study the robustness of the proposed attack in complex scenarios. The experiments aim to quantify the ability of an attacker who leverages a COTS mmWave probe (i.e., IWR1642Boost) for through-wall word detection of human-rendered speech. Five out of the 23 volunteers are included in the experiments. Experimental settings are detailed in Table 1.

6.1 Impact of Distance and Orientation

In real-world scenarios, the distance and orientation between the probe and targeted victim may change. In this part, we investigate the impact of distance and orientation on the system. We asked the five out of the 23 volunteers to sit still without body motions in the chair and speak the 57 words ten times in the conference room for each experiment setting. We deployed the probe outside the room. The speech retrieval model is trained with the dataset in Section 5.1. When we studied the impact of attack distance (from $0.5m$ to $5m$), we kept the orientation within $0° \sim 10°$. For the orientation ($0° \sim 40°$) evaluation, we set the attack distance as $2.5m$.

Results: As shown in Figure 14, the top-1 inference accuracy of the 57 words is above 83% when the sensing distance varies within 5m. In general, the performance degrades with the increasing distance. Because the power density of the transmitted mmWave beam decreases with increasing distance [38], resulting in a low SNR for the near-throat vibration sensing. This can be further improved by a probe with a more concentrated beam design and higher transmitting power. As shown in Figure 15, the top-1 accuracy is above 86% within 30° but decreases remarkably in the attack orientation of 40°. Because the reflecting surface of the targeted victim's throat area reduces with the attack orientation increasing. Thus, the mmVSNR decreases as shown in Figure 16, resulting in



(a) Cafe

(b) Office

Fig. 20. Experiments in different environments.

the performance degradation in the orientation of 40°. Generally speaking, *Wavesdropper* is robust for speech retrieval within 5m and an orientation less than 40° in a through-wall scenario.

6.2 Impact of Environmental Change

Except for echoes reflected from the victim's near-throat skin, background objects in the room can also reflect mmWave signals (i.e., clutters), which may influence the attack performance. To investigate the impact of environmental changes on the proposed attack system, we performed experiments under three scenarios as shown in Figure 20, i.e., a conference room, a cafe, and an office room. The conference room is the same room for training data collection but with layouts (e.g., positions of chairs and cabinetry) changed. We asked five out of the 23 volunteers to keep still without body motions and speak the 57 words ten times in each environment and collected the mmWave data through the wall. The sensing distance is set to **2.5m** and the orientation is kept within 0 ∼ 10°. Then we used the model trained with the dataset in Section 5.1 to recover speech contents.

Results: As shown in Figure 17, the top-1 accuracies under three scenes are above 88% with little performance fluctuation (about 1.3%) across different scenes. This is because compared with just feeding the raw mmWave data into the retrieval model (which contains environment-dependent information), *Wavesdropper* first differentiates the speaker from the background by applying the spatial-temporal analysis and clutter suppression, and then extracts the intrinsic vocal vibration information of the speaker (Section 4). This can eliminate speech-irrelevant echoes at most and thus improve the robustness of the eavesdropping system under different scenarios.

6.3 Impact of Blocking Wall

It is unavoidable that the blocking wall will cause attenuation on the mmWave. We observe that the attenuation caused by common soundproof wall materials (e.g., soundproof glass, sound-absorbing sponge, and wood) are slightly different. To mitigate the impact of attenuation, we apply the amplitude normalization to the extracted mmVocal response as introduced in Section 4.4. In this part, we evaluate *Wavesdropper* with different wall materials, i.e., soundproof glass (a conference room wall), sound-absorbing sponge (a customized wall), wood (a wooden folding wall), and brick (an office wall). For each experimental setting, we asked the five out of the 23 volunteers to keep still without body motions and speak the 57 words ten times in the room and used the mmWave probe to collect the mmWave data through the wall with the attack distance of **2.5m** and orientation of 0°. The speech recovery model is trained with the dataset in Section 5.1.

Results: The results are shown in Figure 18. We can observe that for each common soundproof material (i.e., glass, sponge, and wood), the top-1 accuracy is above 90% which indicates that the proposed attack is resilient to different kinds of soundproof obstacles. But for the brick wall (about 12*cm* thick), the average top-1 accuracy is only 75.3%. The degradation of the performance results from the large attenuation caused by the brick wall, and

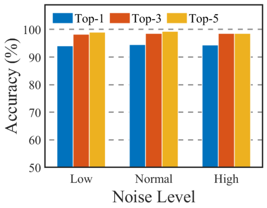


Fig. 21. Impact of background noise.

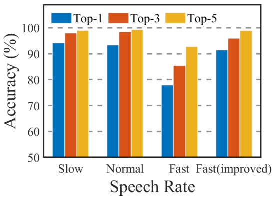


Fig. 22. Impact of speech rate.

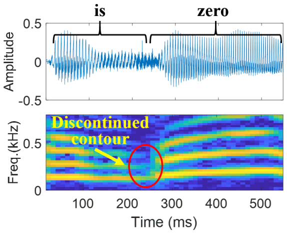


Fig. 23. Discontinued contours separates coherent words in the time-frequency domain.

the transmitting power is limited for the COTS mmWave sensor. However, the experiment validates the ability of COTS mmWave probes to penetrate common soundproof materials and cause threats to in-room speech.

6.4 Impact of Body Motion

Considering that speakers are not still but with small body wiggles or gestures when they speak, we evaluate the impact of speaker's body motion on *Wavesdropper*. We asked the five out of the 23 volunteers to sit on a chair and say the 57 words ten times. When speaking, each volunteer was asked to wave their hands slowly in front of the chest, which is beyond the range-bin of the throat. The attacker launched the word detection outside the conference room (the same room as in Section 5.1) through the soundproof wall and used the dataset in Section 5.1 to train the speech retrieval model. The attack distance is set as **2.5m** and the orientation is 0°. The detailed experimental setting is shown in Table 1.

Results: The recognition accuracy of the 57 words are shown in Figure 19. The top-1 average accuracy of the five volunteers varies from 84% to 96%. The low accuracy of volunteer #2 (84%) is most likely due to his waving hands blocking the transmitted mmWave towards the throat area and thus degrading the attack performance. As analyzed in Section 3.3, the frequency of human motion artifacts is far below 80Hz so the high-pass filter with a cut-off frequency of 80Hz can effectively eliminate most of the impact of speaker's gental gestures and body wiggle.

6.5 Impact of Background Acoustic Noise

We asked the five out of the 23 volunteers to keep still without body motions and speak the 57 words ten times in the soundproof room and evaluate the performance of *Wavesdropper* across three different levels of real-world acoustic noise: low (noise of the air conditioner), normal (daily conversation played by a loudspeaker), and high (loud music played by a loudspeaker). The sensing distance is **2.5m** and orientation is 0°. According to the experiment results in Figure 21, we find that the top-1 recognition accuracy is stable at 93% without noticeable disturbance. The results indicate that the background noise of these three levels has little impact on the performance of *Wavesdropper*.

6.6 Impact of Speech Rate

In Section 5 and 6, we asked the participants to speak in their normal way. To investigate the impact of speech rate, we asked the five out of the 23 volunteers to keep still without body motions and speak the 57 words ten times with different speech rates, i.e., low (0.6 words/sec on average), normal (2.1 words/sec on average), and fast (2.9 words/sec on average) speed. We kept other experimental settings the same as Section 6.5. The sensing distance is **2.5m** and the orientation is 0°. The recognition accuracy is shown in Figure 22. The top-1 accuracy is above

93% when the speaker speaks in a low and normal speed. When the speech rate increases to the fast speed, the performance degrades to 78%. We find that some segments under the fast-speed condition have a long duration, which indicates the segment possibly contains two or more words rather than a single word due to imperfect segmentation. The reason for the connected words is that the signal energy in the segmentation algorithm (Section 4.3) is calculated within non-overlap windows. If there is little interval between two successively pronounced words in the time domain, the signal energy changes little across successive windows, which wrongly indicates a single pronounced word. Another observation is that connected words occur more often when two successive words both have the same or similar phonemes causing coherent pronunciation, e.g., “is zero” (/ɪz ˈzɪəɾəʊ/).

Although coherent pronunciation is not always the case, we propose a method based on the discontinued contour of coherent pronunciation in joint time-frequency domain to tackle this problem. The method is inspired by the fact that, even though there is little interval observed between audio signals of connected words in the time domain (as shown in the top figure in Figure 23), the connected words can be separated in the joint time-frequency domain, i.e., frequency contours of connected words (yellow stripes in the below figure of Figure 23) are discontinuous in the joint time-frequency domain (as indicated by the red circle in Figure 23). Specifically, we first apply short-time Fourier transform with a sliding window to covert the input signal to the joint time-frequency domain. Then we derive contours of the frequency components of the speech signal in the joint time-frequency domain using an f_0 estimation algorithm [32]. The algorithm returns a binary sequence consisting of “1” and “0” where “1” and “0” indicate the high energy and low energy part of the contours, respectively. Then we calculate the differential sequence of which the peak indicates the changing point of the contours. Finally, connected words are separated by the changing points. The recognition result with the improved segmentation is shown in Figure 22, i.e., Fast (improved), which indicates a 14% performance improvement on the recognition accuracy.

7 EXPLORATION: WORD DETECTION ON MULTIPLE SPEAKERS

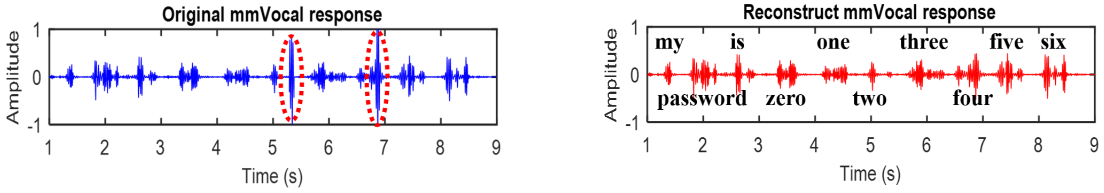
In previous sections, we have evaluated the system under complex scenarios. In this part, we explore the ability of an attacker/attackers to detect words of multiple (two) victims simultaneously in a controlled environment. Two out of the 23 volunteers are involved in the experiments.

7.1 Word Detection with Multiple Probes

To eavesdrop on multiple targets in a conference room, an intuitive way is using multiple probes targeting on different speakers. However, we find that there are two further problems we need to address. First, probes with the same operating frequency band have a probability of causing mutual interference, which should be analyzed and mitigated if there is any severe interference. Second, we observed that large-scale motions (e.g., shaking body from side to side) of the listener can influence the mmWave reflected from the speaker’s throat and reduce the attack performance. Next, we give the mutual interference analysis and evaluate the dynamic clutter suppression method (detailed in Section 4.3) for the listener’s body movement.

7.1.1 Mutual Interference Analysis. We assume the attacker uses two same devices with the same parameter settings (e.g., chirp slope, chirp bandwidth). In such a condition, there is a possibility that Probe B receives and demodulates the chirp signal transmitted by Probe A, and thus, the mutual interference can happen. As mentioned above, the interference requires that Probe B’s chirp is demodulated by Probe A. This means a strict timing condition that two probes transmit chirps almost simultaneously. The probability of such interference can be calculated using the chirp max-delay (t_d) and chirp repeat periodicity (t_r), and the number of probes present in the scene (N_p) [39], as shown in Equation 8:

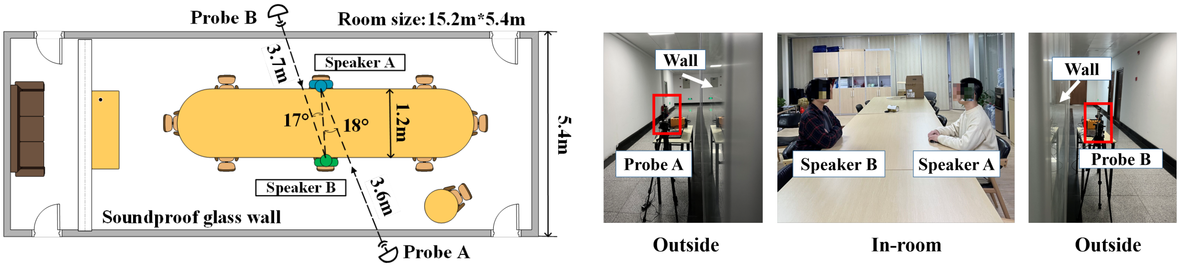
$$P_{interference} = 1 - (1 - \frac{t_d}{t_r})^{N_p - 1} \tag{8}$$



(a)

(b)

Fig. 24. (a) Original and (b) reconstructed mmWave data when the speaker says "my password is zero...six".



(a)

(b)

Fig. 25. The (a) sketch map and (b) pictures of eavesdropping with multiple probes in a conference room with multiple probes to show the in-room and out-room settings.

The device (IWR1642Boost) in our experiments has a max-delay of 0.33µs [3] and we set the chirp repeat periodicity as 250µs in all the experiments. When the attacker uses two probes ($N_p = 2$) for eavesdropping, the probability of mutual interference is 0.13%, which is neglectable, so we do not take further exploration about this in the following pages.

7.1.2 Listener-introduced Noise Cancellation. When participant A speaks, we find that reflected mmWave from A can be influenced by participant B's large-scale motion which induces glitch-like noise in the extracted mmVocal response (the red-elliptic area in Figure 24(a)). By applying the dynamic clutter suppression in Section 4.3, the impact of listener's motion can be mitigated as Figure 24(b) indicates.

7.1.3 Result. As shown in Figure 25, we asked the two volunteers to sit face to face and speak the 57 words ten times alternatively. When one speaks, the other wiggles his body from side to side at random times. We used two mmWave probes outside the conference room to launch the attack through the soundproof-glass wall and used the model trained with the dataset in Section 5.1 to recover their speech contents. Among the total 136 interfered samples by the listener's body motion, only 11 samples are correctly recognized without the dynamic clutter suppression. However, the number increases to 129 when the proposed clutter suppression is applied. The average accuracies for speakers A and B are 90.2% and 88.5%, respectively. The results indicate that the proposed dynamic clutter suppression (Algorithm 1) effectively mitigates the listener-induced interference and improves the system performance.

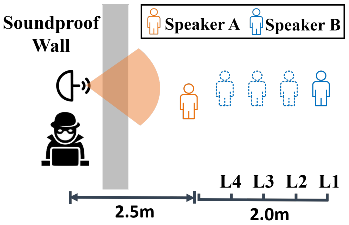


Fig. 26. Eavesdropping on multiple targets with a single probe.

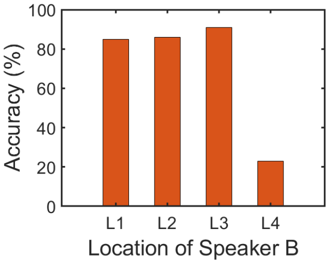


Fig. 27. The Top-1 accuracy of multi-person word recognition.

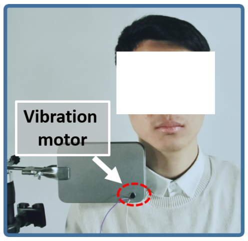


Fig. 28. A deliberate vibration source near the speaker's throat.

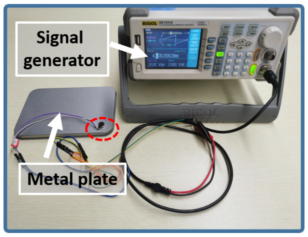


Fig. 29. The metal plate attached with a motor driven by a signal generator.

7.2 Word Detection with a Single Probe

Except for the previous multi-probe solution, we also exploit the attacker's ability to recover speech contents of multiple targets with only a single mmWave probe.

7.2.1 Key Insight. As introduced in Section 4.2, the samples of the mmVocal response are derived from FFT results of successively demodulated chirps, i.e., an N-point mmVocal response is derived from N demodulated chirps. In other words, a demodulated chirp can be taken as a sampling of the speaker's vocal vibration. One of the key characteristics of the demodulated chirp is that its FFT result can indicate multiple detected targets. If there are multiple speakers present, the demodulated chirp can be taken as the simultaneous sampling of multiple targets where different FFT points correspond to different speakers. This paves the way for eavesdropping on multiple targets with a single probe.

7.2.2 Experimental Validation. We asked two volunteers to sit still in chairs, face the probe and speak in the conference room and deployed a single probe outside the room for through-wall eavesdropping. During the experiment, we find that the distance between the two speakers can influence the speech retrieval performance. We change the distance between speaker A and B as shown in Figure 26. For each distance setting, we asked the two volunteers to simultaneously speak the 57 words ten times and used the model trained with the dataset in Section 5.1 for speech retrieval. The average top-1 accuracy of the two speakers is shown in Figure 27. We can observe that when the speakers have a distance larger than $0.5m$, Wavesdropper can retrieve the speech of both speakers with top-1 accuracy above 85%. But the performance degrades to 23% when the speakers are close to each other ($0.5m$). This is because the spatial-temporal analysis (Section 4.2) is unable to distinguish the two speakers in the space due to the distance resolution if they are too close to each other physically (note that the

Table 3. Comparison with previous work.

Work	Experimental target	Training data from the target	Resistant to acoustic noise	Performance		
				Words	Accuracy	Distance
Kwong *et al.*[26]	Loudspeaker	No	No	-	-	0.25m
Michalevsky *et al.*[30]	Loudspeaker	Yes	No	11 digits	17%(-)	-
Ba *et al.*[9]	Loudspeaker	Yes	No	36 words	55% (-)	-
Teng *et al.*[48]	Loudspeaker	No	Yes	10 digits	99% (through-wall)	4-11m
Wang *et al.*[47]	Human	Yes	Yes	33 words	18% (through-wall)	2-7m
Ours	Human	Yes	Yes	57 words	91% (through-wall)	2.5m

distance is smaller than a normal social distance). The attack results indicate that Wavesdropper can recover speech contents of two targets simultaneously with only a single probe but the performance can be influenced by the distance between the two targets. Generally speaking, the eavesdropping system works well when the distance between two speakers is larger than $0.5m$.

8 COUNTERMEASURES

In this section, we discuss four potential countermeasures and perform experiments to evaluate the last defense method, i.e., a deliberate vibration source near the throat. **Passive defense:** (1) Considering that Wavesdropper leverage mmWave signals for eavesdropping, electromagnetic shielding (e.g., a Faraday cage) can reduce the coupling of electromagnetic fields and thus block the transmitted mmWave signal. Besides, (2) wave-absorbing materials can also reduce the reflected mmWave signals to defend against the proposed attack passively. **Active defense:** The mutual interference between two probes can be significantly small (Section 7.1) when the transmitted chirps of the two probes have the same slope. But when their chirp slopes are different, sweeping interference [39] can happen. Therefore, (3) an active countermeasure is that the speaker can use a mmWave probe to transmit chirps with a random slope to interfere with the adversarial probe. However, this defense method requires prior knowledge about the working frequency of the attack device. (4) Another active defense is to confuse the attacker's retrieval model by placing a delicate vibration source near the throat. The additional source vibrates in the same frequency band as the vocal vibration. Next, we perform experiments to validate the effectiveness of this method.

We placed a metal plate attached with a vibration motor close to the speaker's throat area as shown in Figure 28. We used a signal generator (RIGOL DG3121A, shown in Figure 29) to drive the motor to vibrate in the form of frequency modulation covering the bandwidth $80 \sim 260Hz$ with a period of 4ms and asked the volunteer to speak the 57 words ten times within the conference room. The remaining experimental setting is the same as Section 6.4. We find that Wavesdropper achieves a poor recognition accuracy of 1.9% and the mmVSNR is only 0.31dB when the vibration motor works. Because the vibration source is too close to the speaker's throat physically and Wavesdropper cannot tell the two vibration sources apart. The extracted mmVocal response is overwhelmed by the disturbance from the vibrating metal plate, which results in a low mmVSNR and a poor recognition accuracy.

9 DISCUSSION AND FUTURE WORK

9.1 Attenuation and Interference

Attenuation is a tough problem for wireless sensing [47, 48], especially in a long-range and through-wall scene. As the results in Sections 5.3 and 6 indicate, mmWave can easily penetrate soundproof materials for eavesdropping but has a limited penetrating performance on the brick. In addition, the moisture material covering the throat area may have a certain impact on the performance due to the absorption. Limited by the COST mmWave devices (with a transmitting power of 12.5dBm) in this paper, this problem can be solved by adopting powerful antenna arrays with larger transmitting power and advanced noise reduction techniques [15] in the future design of COTS mmWave devices. Considering the growing amount of mmWave devices in human life, the signals in 77-81GHz is possibly demodulated by the malicious mmWave sensor and thus, cause interferences to the system. Based on this, a jamming mitigation can be applied to defense the attack but requires the parameters of the malicious mmWave sensor, such as the chirp rate, the operating band, the duty cycle, etc.

9.2 Improved Eavesdropping on Multiple Targets and Motion Interference

In Section 7.2, we explore the feasibility of eavesdropping on multiple victims with a single mmWave probe. Although the experimental results validate the feasibility to some extent, the result is not satisfying when the two targets were close to each other physically, which results from the mixed mmVocal responses of the two targets. Considering that the probe has an on-board antenna array with 4 Rx of which each Rx can derive a trace of mmVocal response, a potential method is to adopt the blind source separation [14] to separate mmVocal responses of the two targets from the mixed one. The current attack shows resilience to gentle movements of the speaker. However, the drastic motion of the speaker (e.g., turning the head and standing up/down suddenly) may induce noise covering the fundamental frequency band of human voice (85-255Hz) in which case the performance can be degraded largely. A potential solution is to use self-similarity matrices [18] to eliminate these human artifacts and recover the clean mmVocal response.

10 RELATED WORK

Recent research reveals that attackers can use motion sensors [8, 9, 30], radio frequency signals [48], and even hard disk drives [26] to eavesdrop on machine-rendered speech. These works leveraged the conductive vibration, wireless vibrometry, and air vibration caused by a loudspeaker for speech retrieval. Kwong *et al.* [26] revealed that hard disk drives can record loudspeaker audio with a sound pressure level (SPL) of 85dBA, which is louder than most normal conversations as they demonstrated. Michalevsky *et al.* [30] and Ba *et al.* [9] eavesdropped with limited accuracy using motion sensors built in smartphones. Teng *et al.* [48] leveraged the wireless vibrometry caused by a static loudspeaker to recover machine-rendered speech. Our work mainly focuses on retrieving human-rendered speech. For the human-rendered speech retrieval, Wang *et al.* [47] leveraged the multipath effect of WiFi to capture speakers' mouth movement but they have a strong assumption that the victim is totally static and speaks standing by the WiFi device, which is often not the case. Our work considers a more practical condition where the environment is dynamic (e.g., body movement). In our work, mmWave has a higher directivity with a concentrated beam to focus on specific targets for eavesdropping. The proposed spatial-temporal analysis and dynamic clutter suppression method can eliminate the background clutters and make the system resilient to environment changes (evaluated in Section 6). Our speech retrieval model can recognize words of human speech with an accuracy of 91% which is higher than Wang's work. We summarize the compared previous works on eavesdropping and our proposed *Wavesdropper* in Table 3. Wavesdropper achieves satisfying performance within the orientation range of 0°-30°. Considering previous works do not quantitively evaluate the orientation, we do not list the performance in the table for comparison. To retrieve human-rendered speech, the attacker may also leverage visual side channels, such as lipreading [13, 23] and sound wave-caused vibration on surrounding objects [35]. However, these visual-based approaches do not work in a through-wall (opaque) scenario. Xu *et al.* [50] developed a customized mmWave device for noise-resistant speech sensing. Our work achieves eavesdropping leveraging a widely-available COTS mmWave device that is not specially designed for speech sensing. Xu's work requires both mmWave and audio signals to train the model but our method solely relies on the mmWave signals to infer the speech. Besides, Xu's work mainly focuses on the line-of-sight sensing condition without blockage. Our work aims to achieve the through-wall word detection to retrieve sensitive information and faces different challenges in the obstructed condition, i.e., speaker localization and suppressing environment clutters. Through-wall information retrieval is a common and challenging research topic. Li *et al.* [29] use a customized mmWave probe to acquire the liquid crystal state of the victim's screen and infer screen contents. Adib *et al.* [5] use radio frequency to capture human figures through walls. Zhao *et al.* [55] use wireless and visual signals to estimate human pose behind walls. Our work focuses on acoustic side-channel and faces different challenges. Banerjee *et al.* [10] investigate the feasibility of using wireless links to predict the moving direction of the user behind the wall. Nandakumar *et al.* [34] propose to use a smartphone and a loudspeaker to form an active sonar system to recognize human motion through the wall. Both of them study coarse sensing of human activities and aim to compromise activity privacy. our work focuses on speech retrieval by sensing more delicate vocal vibration of human being.

Limitation of Wavesdropper: 1) Due to the propagating and penetrating attention of mmWave and the COTS hardware limitation, Wavesdropper has a distance of around 2-3m and orientation range of 0°-30° to achieve satisfying performance compared with previous work [47, 48]. Besides, materials like metal or human tissue containing moisture (e.g., hands) can block the transmitted mmWave and thus, cause great performance degradation on the system. 2) Wavesdropper requires pre-collected data from the victim to train a victim-specific model for word inference compared with previous work [26, 48], i.e., a target-dependent attack. Due to the discrepancy in pronunciation habits and vocal physiological structures [27], *WavesdropNet* learns speaker-dependent features, which can only achieve satisfying performance when with the training data from the targeted speaker. 3) *Wavesdropper* eavesdrops on speech contents (word classification) which is a subset of vocabulary compared with previous work that can recover audible speech [9, 26, 48]. This limitation is mainly determined by the mechanism of speech production. Intelligible speech is produced by the collaboration of the vocal cords and articulators (e.g., the tongue and the palate). With the different vibratory frequency of vocal cords and movement of articulators, the human voice is modulated to generate different speech contents. Wavesdropper is hard to capture the movement of the inner articulators in human body. Thus, voice formants produced by the inner articulators cannot be recovered, which makes Wavesdropper difficult to recover audible voice.

11 CONCLUSION

In this paper, we reveal a new speech threat posed by widely-available COTS mmWave devices. An adversary can break through the soundproof protection and detect words of human-rendered speech using the portable COTS mmWave probe. We first investigated the relationship between the vocal vibration and reflected mmWave signals (i.e., mmVocal response) and find that the threat still exists in a through-wall condition. Then we solve the challenges in the obstructed condition and proposed *Wavesdropper*, an end-to-end word detection system to compromise speech privacy and security in a soundproofing environment. The experiments on 23 volunteers indicates that Wavesdropper can achieve 91% accuracy for 57-word recognition in a through-wall scenario. The results of extensive experiments show the system robustness in complex conditions, such as environment changes and different soundproof materials. Considering the increasing applications of widely-available COTS mmWave devices, we hope the new acoustic side-channel attack can raise attention of related researchers and the public.

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 62172359), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005), and Research Institute of Cyberspace Governance in Zhejiang University.

REFERENCES

[1] [n. d.]. Apple Siri Eavesdropping Puts Millions Of Users At Risk. [online]. https://www.forbes.com/sites/kateoflahertyuk/2019/07/28/apple-siri-eavesdropping-puts-millions-of-users-at-risk/#1dd38829a530 Accessed on March 26, 2021.

[2] [n. d.]. The Group Dynamics of Smartphone Eavesdropping. [online]. https://www.securitymagazine.com/articles/90577-the-group-dynamics-of-smartphone-eavesdropping, Accessed on March 26, 2021.

[3] [n. d.]. IWR1642Boost. [online]. https://www.ti.com/tool/IWR1642BOOST, Accessed on March 26, 2021.

[4] [n. d.]. Why Enterprises Still Have to Worry about Eavesdropping. [online]. https://blog.trendmicro.com/why-enterprises-still-have-to-worry-about-eavesdropping/, Accessed on March 26, 2021.

[5] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[6] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. 2016. Next generation 5G wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 18, 3 (2016), 1617–1655.

[7] Adeel Ahmad, June Chul Roh, Dan Wang, and Aish Dubey. 2018. Vital signs monitoring of multiple people using a FMCW millimeter-wave sensor. In *2018 IEEE Radar Conference (RadarConf18)*. IEEE, 1450–1455.

[8] 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.

[9] 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 the Network and Distributed Systems Security (NDSS) Symposium*. 23–26.

[10] Arijit Banerjee, Dustin Maas, Maurizio Bocca, Neal Patwari, and Sneha Kasera. 2014. Violating privacy through walls by passive monitoring of radio windows. In *Proceedings of the 2014 ACM conference on Security and privacy in wireless & mobile networks*. 69–80.

[11] Mattia Brambilla, Monica Nicoli, Sergio Savaresi, and Umberto Spagnolini. 2019. Inertial sensor aided mmWave beam tracking to support cooperative autonomous driving. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 1–6.

[12] A Bunkowski, B Bödeker, S Bader, M Westhoff, P Litterst, and J I 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 (sep 2009), 046001.

[13] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3444–3453.

[14] Pierre Comon and Christian Jutten. 2010. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

[15] Khadidja Ghribi, Mohamed Djendi, and Daoued Berkani. 2016. A wavelet-based forward BSS algorithm for acoustic noise reduction and speech enhancement. *Applied Acoustics* 105 (2016), 55–66.

[16] Jian Gong, Xinyu Zhang, Kaixin Lin, Ju Ren, Yaoxue Zhang, and Wenxun Qiu. 2021. RF Vital Sign Sensing under Free Body Movement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–22.

[17] Tianbo Gu, Zheng Fang, Zhicheng Yang, Pengfei Hu, and Prasant Mohapatra. 2019. Mmsense: Multi-person detection and identification via mmwave sensing. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*. 45–50.

[18] Unsoo Ha, Sohrab Madani, and Fadel Adib. 2021. WiStress: Contactless Stress Monitoring Using Wireless Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 103 (sep 2021), 37 pages. https://doi.org/10.1145/3478121

[19] Jun Han, Albert Jin Chung, and Patrick Tague. 2017. Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 181–192.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Yongzhong He, Xuejun Yang, Binghui Hu, and Wei Wang. 2019. Dynamic privacy leakage analysis of Android third-party libraries. *Journal of Information Security and Applications* 46 (2019), 259–270.

[22] Jerry Hildenbrand. [n. d.]. Here's how the Pixel 4's Soli radar works and why Motion Sense has so much potential. [online]. https://www.androidcentral.com/how-does-googles-soli-chip-work, Accessed on March 26, 2021.

[23] Denis Ivanko, Alexey Karpov, Dmitry Ryumin, Irina Kipyatkova, Anton Saveliev, Victor Budkov, Dmitriy Ivanko, and Miloš Železnỳ. 2017. Using a high-speed video camera for robust audio-visual speech recognition in acoustically noisy conditions. In *International Conference on Speech and Computer*. Springer, 757–766.

[24] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. *MmVib: Micrometer-Level Vibration Measurement with Mmwave Radar*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3372224.3419202

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] 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)*. IEEE, 905–919.

[27] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th*

Conference on Embedded Networked Sensor Systems.

[28] Yingsong Li, Zelong Shao, Xiangkun Zhang, and Jingshan Jiang. 2018. Vibrations monitoring for highway bridge using mm-Wave radar. In *2018 IEEE Asia-Pacific Conference on Antennas and Propagation (APCAP)*. IEEE, 501–502.

[29] Zhengxiong Li, Fenglong Ma, Aditya Singh Rathore, Zhuolin Yang, Baicheng Chen, Lu Su, and Wenyao Xu. 2020. Wavespy: Remote and through-wall screen attack via mmwave sensing. In *2020 IEEE Symposium on Security and Privacy (SP)*. 49–64.

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

[31] Jochen Moll, Kaspar Bechtel, Bernd Hils, and Viktor Krozer. 2014. Mechanical vibration sensing for structural health monitoring using a millimeter-wave doppler radar sensor.

[32] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. 2009. fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. *journal of the audio engineering society* (february 2009).

[33] Timothy I Murphy. [n. d.]. Analyst's Desktop Binder. [online]. https://www.hsdl.org/?view&did=710020 Accessed on March 26, 2021.

[34] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[35] Ben Nassi, Yaron Pirutin, Adi Shamir, Yuval Elovici, and Boris Zadov. 2020. Lamphone: Real-Time Passive Sound Recovery from Light Bulb Vibrations. *BlackHat USA* (2020).

[36] K O'flaherty. 2019. Amazon Staff Are Listening To Alexa Conversations—Here's What To Do'. *Forbes* (2019).

[37] Hyojin Park, Christoph Kayser, Gregor Thut, and Joachim Gross. 2016. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife* 5 (2016), e14521.

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

[39] S. Rao and A. V. Mani. 2020. Interference Characterization in FMCW radars. In *2020 IEEE Radar Conference (RadarConf20)*. 1–6. https://doi.org/10.1109/RadarConf2043947.2020.9266283

[40] Linda Senigagliesi, Gianluca Ciattaglia, and Ennio Gambi. 2020. Contactless Walking Recognition based on mmWave RADAR. In *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–4.

[41] Satish Sinha, Partha S Routh, Phil D Anno, and John P Castagna. 2005. Spectral decomposition of seismic data with continuous-wavelet transform. *Geophysics* 70, 6 (2005), P19–P25.

[42] M. Song, J. Lim, and D. Shin. 2014. The velocity and range detection using the 2D-FFT scheme for automotive radars. In *2014 4th IEEE International Conference on Network Infrastructure and Digital Content*. 507–510.

[43] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin. 2011. A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4144–4147. https://doi.org/10.1109/ICASSP.2011.5947265

[44] Vutha Va, Takayuki Shimizu, Gaurav Bansal, and Robert W Heath Jr. 2016. Millimeter wave vehicular communications: A survey. *Foundations and Trends® in Networking* 10, 1 (2016), 1–118.

[45] Pradeep Verma, Vivek Sheel Shakya, Divya Sharma, Usha Chauhan, and Abhilash Gaur. 2020. MM-Wave Radar Application for Autonomous Vehicles. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 556–559.

[46] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. 2019. Joint Activity Recognition and Indoor Localization With WiFi Fingerprints. *IEEE Access* 7 (2019), 80058–80068.

[47] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. 2016. We can hear you with Wi-Fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920.

[48] 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*. 130–141.

[49] Henk Wymeersch, Gonzalo Seco-Granados, Giuseppe Destino, Davide Dardari, and Fredrik Tufvesson. 2017. 5G mmWave positioning for vehicular networks. *IEEE Wireless Communications* 24, 6 (2017), 80–86.

[50] 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.

[51] Zhicheng Yang, Parth H Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring vital signs using millimeter wave. In *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*. 211–220.

[52] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.

[53] Hansong Zeng and Yi Zhao. 2011. Sensing movement: Microsensors for body motion measurement. *Sensors* 11, 1 (2011), 638–660.

[54] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*. 8778–8788.

[55] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7356–7365.