WaveEar: Exploring a mmWave-based Noise-resistant Speech Sensing for Voice-User Interface

Chenhan Xu[1], Zhengxiong Li[1], Hanbin Zhang[1], Aditya Singh Rathore[1], Huining Li[1,2], Chen Song[1], Kun Wang[3], Wenyao Xu[1]
[1]University at Buffalo, the State University of New York, Buffalo, New York, USA
[2]Nanjing University of Posts and Telecommunications
[3]University of California, Los Angeles, California, USA
{chenhanx,zhengxio,hanbinzh,asrathor,csong5,wenyaoxu}@buffalo.edu,huinli@outlook.com,wangk@ucla.edu

ABSTRACT

Voice-user interface (VUI) has become an integral component in modern personal devices (*e.g.,* smartphones, voice assistant) by fundamentally evolving the information sharing between the user and device. Acoustic sensing for VUI is designed to sense all acoustic objects; however, the existing VUI mechanism can only offer low-quality speech sensing. This is due to the audible and inaudible interference from complex ambient noise that limits the performance of VUI by causing denial-of-service (DoS) of user requests. Therefore, it is of paramount importance to enable noise-resistant speech sensing in VUI for executing critical tasks with superior efficiency and precision in robust environments. To this end, we investigate the feasibility of employing radio-frequency signals, such as millimeter wave (mmWave) for sensing the noise-resistant voice of an individual. We first perform an in-depth study behind the rationale of voice generation and resulting vocal vibrations. From the obtained insights, we present *WaveEar*, an end-to-end noise-resistant speech sensing system. *WaveEar* comprises a low-cost mmWave probe to localize the position of the speaker among multiple people and direct the mmWave signals towards the near-throat region of the speaker for sensing his/her vocal vibrations. The received signal, containing the speech information, is fed to our novel deep neural network for recovering the voice through exhaustive extraction. Our experimental evaluation under real-world scenarios with 21 participants shows the effectiveness of *WaveEar* to precisely infer the noise-resistant voice and enable a pervasive VUI in modern electronic devices.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Speech recognition**; • **Hardware** → **Signal processing systems**; *Sensor applications and deployments.*

Corresponding Contact: wenyaoxu@buffalo.edu; wangk@ucla.edu.

MobiSys '19, June 17–21, 2019, Seoul, Republic of Korea

ACM ISBN 978-1-4503-6661-8/19/06…$15.00
https://doi.org/10.1145/3307334.3326073

KEYWORDS

Voice-user interface; speech recognition; mmWave; neural network

ACM Reference Format:
Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, & Kun Wang, Wenyao Xu. 2019. WaveEar: Exploring a mmWave-based Noise-resistant Speech Sensing for Voice-User Interface. In *the 17th Annual Int'l Conference on Mobile Systems, Applications, and Services (MobiSys '19), June 17–21, 2019, Seoul, Republic of Korea.* ACM, NY, NY, USA, 13 pages. https://doi.org/10.1145/3307334.3326073

1 INTRODUCTION

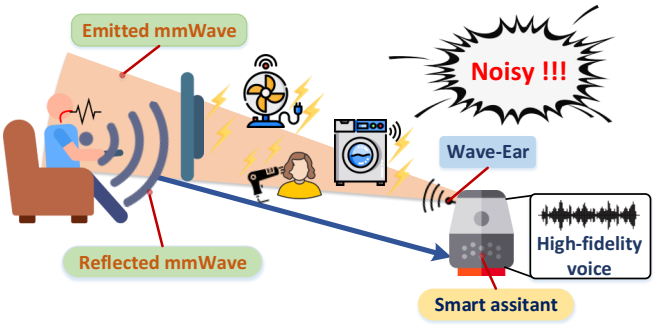

Figure 1: *WaveEar* leverages the reflected mmWave signal from vocal vibrations to facilitate human-device interaction in noisy environments.

With the prevalence of smart devices and home automation, voice has become a popular medium of communication in the Internet of Things (IoT) environment through voice-user interfaces (VUIs). Through voice interactions, VUIs facilitate various daily tasks with superior efficiency such as private communication (*e.g.,* phone calls, messages, emails), information sharing (*e.g.,* voice search) and even monetary transactions [1]. Reports reveal that VUI is striving to become the primary user interface for smart homes and connected lifestyle [2–4]. The voice-enabled smart home market is expected to reach $1.3 billion by 2022 [5].

Although VUI is considered as the most promising interface, a fundamental problem needs to be addressed before it can completely replace the physical user interaction. The existing VUIs rely on an acoustic-based approach of using a low-cost microphone or array of microphones for sensing the voice signals of the user [3, 6].

However, these sensors are highly sensitive to the ambient noise and unable to provide high-quality reproduction of original sound, also known as high-fidelity voice. The primary consequence of this limitation is audible interference during voice transmission in daily environments (*e.g.,* airports, public transportation) with more severe outcomes leading to denial-of-service (DoS) during voice search and breach of VUI security through inaudible voice commands [7, 8].

Mitigating the influence of ambient noise during voice sensing has a long and rich history in literature and is a core topic in the acoustic inference community. The classic solution is active noise cancellation (ANC) which leverages additional microphones to generate the anti-noise signals and nullify the noise from the environment [9]. However, when there are multiple noise sources in the environment, ANC requires multiple microphones (one for each noise channel) to cancel the noise interference [10]. Recent studies explore the feasibility of employing non-acoustic sensors, *e.g.,* electromagnetic motion sensors [11, 12] and bone-conduction microphones [13] for recovering the speech signals by analyzing the skin-vibrations. Nevertheless, these sensors require prolonged contact, thereby obstructing the daily activities of the user and frequently causing skin irritations. While non-contact lip reading utilizes the visual motion of the lip to infer the phoneme, *i.e.,* the minimum unit in speech, it requires a highly directional front camera to record the lip motion which is impractical in real-world scenarios [14]. As a result, how to acquire noise-resistant voice in an unobtrusive, robust and cost-efficient manner to enable pervasive VUI in electronic devices remains an unsolved challenge.

In this paper, we take an exploratory step and unveil the opportunity of leveraging radio-frequency (RF) signals for acquiring the noise-resistant voice of an individual. We observe that when a person speaks, the vibrations of the vocal cord containing the acoustic information are propagated through the body tissue and can be measured on the skin's surface. Motivated by this observation, we focus on analyzing the unique distribution of RF signals caused by the minute vibrations near the throat region. Due to its short wavelength and strong reflectance properties, high-frequency RF signals such as millimeter wave (mmWave), can effectively infer the vocal vibrations and eventually recover the voice information. The objective of this paper is to shift from the conventional microphone-based voice recording that has been utilized for decades to a new mmWave sensing of noise-resistant voice. This will facilitate three vital advantages for voice sensing in real-world applications:

1) **Noise-resistant:** Due to directional beamforming of mmWave signals, the proposed system will be insensitive to the ambient noise even in robust environments and under flexible human dynamics (*e.g.,* movement).
2) **Informative:** The high-frequency mmWave allows inference of comprehensive speech information regardless of the language, pitch or pronunciation of the user.
3) **Secure:** The mmWave sensing provides superior security for VUI due to its resilience to inaudible white noise or hidden voice commands.

Naturally, to enable a pervasive VUI with these advantages, two major challenges will need to be addressed. 1) Unlike the omni-directional sensing in microphones, mmWave has a highly directional beamforming. How can we design our system to support non-invasive voice sensing when the user begins her speaking in an unfixed position? 2) How to reconstruct noise-resistant voice from the unstable mmWave signals which comprise not only the vocal vibrations but also other interference from environmental reflections (*e.g.,* objects, multiple users) and human motion artifacts?

To this end, we propose our system, *WaveEar*, to facilitate remote sensing of noise-resistant voice in robust environments as shown in Figure 1. We prototype and develop a portable (11.8 cm × 4.5 cm × 1.8 cm), light-weight (45.4 *g*), and low-cost (less than $100) 24GHz mmWave probe that is able to localize the legitimate speaker among multiple users and non-invasively sense high-resolution vocal vibrations. Considering the fact that deep neural networks have demonstrated immense capability to learn the mappings between two different domains, we design and implement a novel network architecture, namely Wave-voice Net to recover noise-resistant voice from the received mmWave signal. The voice recovery procedure contains three steps. First, we enhance the time-and-frequency domain information by transforming the reflected 1-D time-domain signal to 2-D spectrogram. Afterward, a residual architecture based neural network is employed to learn and establish the mapping between the mmWave spectrogram and voice spectrogram. Finally, we adopt a phase reconstruction algorithm to recover the voice. As a first exploration study for achieving noise-resistant speech sensing via mmWave, we select a set of metrics, *e.g.*, Mel-Cepstral Distortion (MCD), to better evaluate our system. Using these metrics, we perform a comprehensive evaluation with 21 participants to validate the effectiveness of *WaveEar* in inferring the noise-resistant voice under real-world scenarios.

Summary: The contributions of our work are three-fold:

- We develop *WaveEar*, a noise-resistant vocal sensing system that leverages the disturbance of the mmWave signal caused by near-throat skin vibration to recover the human voice.
- We design a Frequency-Modulated Continuous-Wave (FMCW) based mmWave probe which is capable of localizing the speaker among multiple people and sensing the minute near-throat vocal vibrations. A novel deep neural network, *i.e.,* Wave-voice Net is proposed to reconstruct the noise-resistant human voice from the probe signal.
- We evaluate the robustness of *WaveEar* under the interference from distinct ambient noises and other factors such as sensing distance, motion artifacts and emotional state of the user. Our results demonstrate that the system achieves superior performance and provides a foundation for enabling a pervasive VUI in modern electronic devices.

2 SPEECH SENSING VIA MMWAVE: CONCEPT AND PRELIMINARIES

In this section, we present the background on physiology of the human voice and rationale behind obtaining noise-resistant speech through sensing the near-throat vocal vibration. We also perform a feasibility study to prove this concept.

2.1 The Principle of Human Voice

The human voice is generated from an integrated effort of three physiological organs, *i.e.,* lungs, vocal cords, and articulators. During the exhalation process, the lungs create sufficient airflow that

passes over the vocal cords. This air pressure causes vocal cord vibrations and produces audible pulses, commonly known as voiced and unvoiced sound. The articulatory organs (*e.g.,* tongue, palate, thyroid, uvula) further articulate and filter the sound to strengthen or weaken it while the larynx regulates the tension on vocal cords to calibrate the pitch and tone. The vocal cords and articulatory organs are adept at producing a highly comprehensive array of sounds containing profound information.

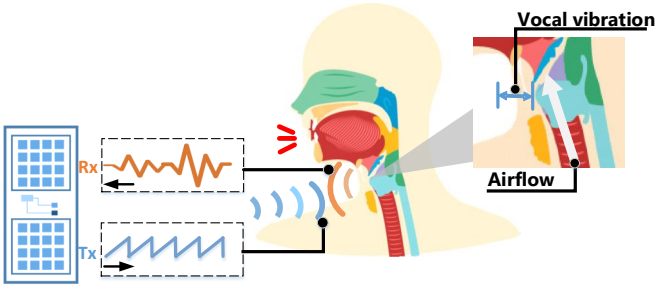

Figure 2: The vocal vibration distorts the skin-reflected mmWave signal. The distortion has a strong correlation with human speech.

2.2 mmWave Sensing on Vocal Vibration

Based on the knowledge of vocal vibration, we hypothesize that an FMCW probe can be used to sense the propagated vocal vibration on the skin's surface around the throat and mouth region. This region is in proximity to the human articulatory system to guarantee the strong correlation between vibration and human voice. As illustrated in Figure 2, the FMCW probe transmits the saw-tooth wave in the frequency domain to human throat periodically. When the person is speaking, the airflow goes through the articulatory system, issuing the vocal vibration and voice signal. The vocal vibration disturbs the skin-reflect mmWave, which will be sensed by the mmWave probe.

Due to the short wavelength of the mmWave, even a small displacement at the micron-level can be detected by calculating the phase change of a few degrees. In particular, given the λ as the length of transmitted wave, the minute displacement d of the skin can be calculated as:

$$d = \frac{\lambda \Delta\varphi}{4\pi}, \tag{1}$$

where $\Delta\varphi$ is the phase shift resolution. It is related with the probe's signal noise ratio (SNR) as:

$$\Delta\varphi^2 = 1/(2 \cdot \text{SNR}). \tag{2}$$

A typical value of the SNR is in the range of 15dB-20dB. We observe from Eq. (1) that a shorter wavelength of transmitted wave results in a higher resolution of the skin minute displacement. Therefore, taking the phase noise into consideration, we choose high-frequency mmWave at 24GHz as the transmitted wave to reach the displacement resolution at around 1 mm.

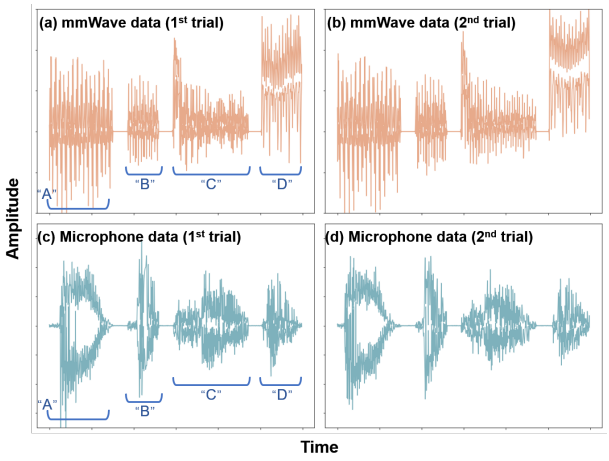

Figure 3: The envelope of time-domain reflected mmWave signal (a), (b) and corresponding speech signal (c), (d) from a subject saying "A,B,C,D" twice. The reflected signal keeps a persistent and unique correlation with human speech.

2.3 A Feasibility Study

To validate the relationship between the received mmWave signal and the speech information, we conduct this proof-of-concept. In our experiment, we ask one subject to speak "A, B, C, D" two times in a normal tone. The mmWave probe is fixed in front of the subject and pointed to her throat from a distance of 40cm. Figure 3 shows the envelope of the time-domain reflected mmWave signals received by the probe when the subject is speaking together with the corresponding voice signals. We observe that the mmWave signals for the same character show a high similarity, while there is an obvious discrepancy between the signals from different characters. Based on this observation, we prove that the sensed mmWave signal has a persistent and unique relationship with the speech information, which matches our claim in Sections 2.1 and 2.2.

3 SYSTEM OVERVIEW

In this paper, we propose *WaveEar*, a noise-resistant speech sensing system. The end-to-end system overview is shown in Figure 4.

***WaveEar* Hardware:** A new mmWave probe with an audio interface is designed to remotely and accurately acquire the human vocal vibration for voice reconstruction. Specifically, the probe first locates the human vocal vibration by the throat localization module. Then, it transmits the continuous wave and processes the reflected signal. After that, the received data are sent to the device for voice reconstruction via the line-in audio card converter.

***WaveEar* Software:** The Wave-voice Net on the device side performs voice reconstruction. Upon receiving data, the data representation module first transforms the data to enhance the temporal and spatial voice information explicitly. Then, the represented spectrogram is fed to the deep network, and our encode-decode architecture achieves the spectrogram transformation. The final voice recover is achieved by recovering the implicit phase information from the spectrogram.

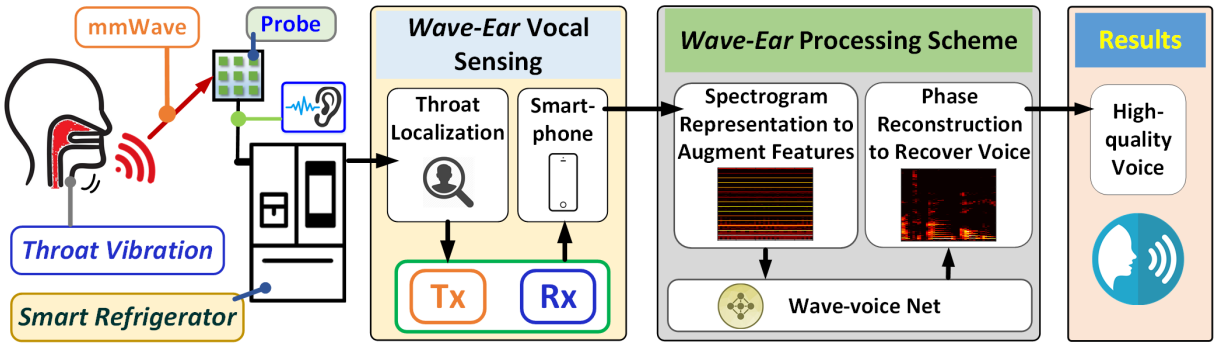

Figure 4: The system architecture for *WaveEar* mainly consisting of a mmWave sesing module in the front-end to locate and sense the vocal vibrations and Wave-voice Net in the back-end to reconstruct the high-quality voice even in the presence of ambient noise.

4 WAVEEAR VOCAL SENSING

In this section, we first introduce the design of the mmWave probe, which includes a three-board structure. Then, we present the vocal vibration locating algorithm that can facilitate spatial vocal sensing in detail.

4.1 mmWave Probe Design

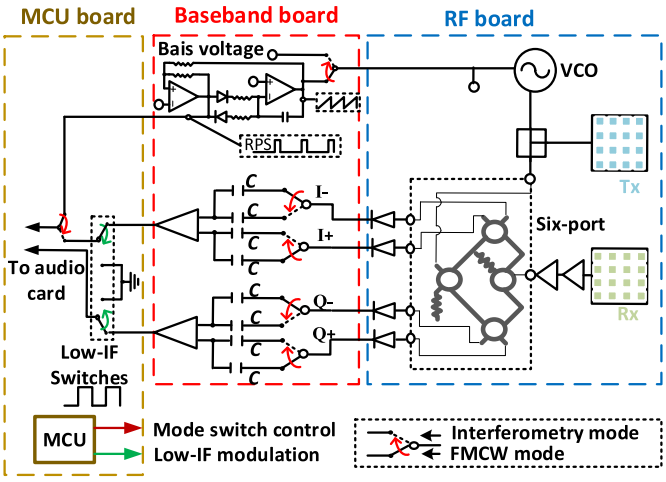

Figure 5: The hardware design for the mmWave probe which comprises MCU, Baseband and RF board to infer the vocal vibrations from mmWave signals.

Figure 5 shows the schematic design of the mmWave probe. It contains three parts, i.e., the RF board, the baseband board, and the Microprocessor Control Unit (MCU) board.

The RF board achieves the transmission and reception of the frequency-modulated RF signal. Because the waveform is continuous, we employ two separate RF channels to minimize coupling between Tx and Rx. The RF front-end employs the phased antenna arrays for two reasons. First, mmWave attenuates rapidly out to a few meters, and antenna array can help generate high antenna gains to fight with this attenuation. Second, the antenna array can steer the beam to a specific direction and thereby reduce the interference from the background.

The baseband board achieves the signal generation and modulation. It employs an operational-amplifier based circuit to generate an analog sawtooth signal and a reference pulse sequence (RPS) which is locked to the sawtooth ramp voltage signal. Afterward, the analog sawtooth signal controls a free-running voltage-controlled oscillator (VCO) to generate the desired frequency-modulated RF signal. The power splitter then divides the output signal of VCO into two parts. One part is fed to the Tx antenna, and the other part is fed to the mixer to help demodulation of the received signal.

After being amplified by the baseband amplifiers, the received signals are modulated with the low-IF digital local oscillator. Then, the two-channel modulated signals are sampled by the audio card.

Table 1: Time-Domain Features for Throat Localization.

Name	Description
Mean Value	$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x(i)$
Standard Deviation	$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x(i)-\overline{x})^2}$
Skewness	$\gamma = \frac{1}{N}\sum_{i=1}^{N}(\frac{x(i)-\overline{x}}{\sigma})^3$
Kurtosis	$\beta = \frac{1}{N}\sum_{i=1}^{N}(\frac{x(i)-\overline{x}}{\sigma})^4 - 3$
RMS Amplitude	$\lambda = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x(i)^2}$

4.2 Throat Localization

Although the above-illustrated probe design facilitates vibration sensing, the WaveEar cannot achieve high-resolution sensing without the localization of throat vibration. As aforementioned, the mmWave probe employs a phased directional array with a 3 dB beamwidth of 12°, out of which antenna gain attenuates rapidly. When the probe misses the throat, the system achieves a low antenna gain at the side lobe and thereby brings a low SNR. According to the Eq. (1) and Eq. (2), low SNR significantly hurts the sensing resolution and therefore impairs the performance of speech recovery.

Our objective is to ensure the mmWave probe can always achieve the strongest SNR by steering beam to the location of the vocal vibration. To achieve this goal, we propose our solution based on the fact that the FMCW probe is sensitive to distance and displacement. Considering the pattern of throat vibration is significantly different from the background (*e.g.,* static objects or slow movement), we employ a classifier to differentiate the background and throat vibration according to the different echo signals which our FMCW probe receives. First of all, the mmWave probe steers beam in all directions to sense the environment. Afterward, a feature detector extracts time domain features (see Table 1), and a classifier is employed to differentiate the throat vibration from backgrounds. These procedures can be completed within 5 ms through digital beamforming [15], a widely used technique in the mmWave systems. Moreover, adaptive beam training protocols could further reduce the localization time with 2 ms [16] (including the 1-ms inference time cost by a classifier such as decision tree). Through throat localization, *WaveEar* achieves the highest resolution in vibration sensing.

5 WAVEEAR PROCESSING SCHEME

In this section, we introduce the voice recovery methodology. Due to the collaboration among all the vocal organs, the information contained in the voice is more than just vocal vibration. Therefore, the *WaveEar* leverages a deep-neural-network based processing scheme to make an end-to-end inference from reflected mmWave to voice. The scheme contains the spectrogram representation module for feature enhancement, the Wave-voice Net for spectrogram transformation, and the phase reconstruction algorithm to recover real voice from the spectrogram.

5.1 Spectrogram Representation to Augment Features

Given the mmWave signal, intuition is to learn the mapping between mmWave and voice signals in the time domain. However, in the signal from the time domain, the frequency-information containing acoustic characteristics is implicit. Deep learning may be a choice to learn the implicit information by cascading more layers, but tremendous extra parameters make training extremely time-consuming and prone to failure. Thereby, we utilize spectrogram as the inner presentation of mmWave and voice signals in Wave-voice Net. The mmWave signal captured by the audio interface will be segmented into a series of small windows before the data representation, which guarantees an appropriate time-domain resolution for each spectrogram. After segmentation, 1-D signals in these small windows will be transformed to 3-D spectrograms by

$$\mathrm{SP}\{x(t)\}(m,\omega) = \left| \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \right|^2, \tag{3}$$

where the Hanning window is chosen to prevent the contained human voice information from the high-frequency interference and spectral leakage:

$$w[n] = \frac{1 + \cos 2\pi \cdot \frac{n}{N-1}}{2}. \tag{4}$$

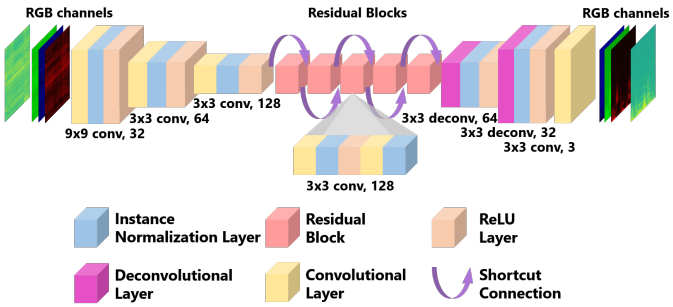

Figure 6: The structure of Wave-voice Net based on encoders and decoders to perform exhaustive extraction of speech information from voice and mmWave spectrogram.

The 2-D spectrogram presents more correlations and patterns, such as formant and sound intensity, which can be easily extracted by a deep learning model. We detail the novel Wave-voice Net in the next subsection.

5.2 Wave-voice Net

Our objective is to solve the image transformation problem between the mmWave spectrogram and the voice spectrogram. As observed in Figure 6, the transformation between mmWave signal and voice requires changing large parts of the image to appear coherently. Therefore, it is advantageous for each pixel in the output to have a large effective receptive field in the input. Consequently, we design our neural network to include an encoder and a decoder.

Encoder. The Wave-voice Net learns the mapping between the mmWave spectrogram and voice spectrogram. Without encoding, each additional 3×3 convolutional layer increases the effective receptive field size by 2. Instead, after encoding by a factor (*e.g.,* D), each 3×3 convolution layer increases its effective receptive field size by 2D. Given the same number of convolutional layers, the encoding achieves larger effective receptive fields. Specifically, our encoder contains three 3×3 convolutional layers. The first layer contains 32 filters, the second layer contains 64 filters, and the third one contains 128 filters. Each of them employs a stride of 2 to achieve encoding.

Residual Blocks. As aforementioned in [17], neural networks need multiple layers to learn features at various levels of abstraction, which benefits the representation. Therefore, we are motivated to add extra weighted layers besides the encoder and decoder. We introduce the residual blocks in the encoder and decoder block. The reason is that the residual architecture contains the shortcut connections which help convergence. The shortcut connections can be formulated as:

$$y = \mathcal{F}(x, \{W_i\}) + x, \tag{5}$$

where $\mathcal{F}(x, \{W_i\})$ represents the multiple convolutional layers, and $\mathcal{F}(x, \{W_i\}) + x$ represents element-wise addition. According to He *et al.* [18], shortcut connections make it easy for the network to learn the identity function. This is an appealing property for transformation between the mmWave signal and voice since in most cases the output of the voice spectrogram shares structures with

the input mmWave probe spectrogram. Specifically, our network contains five residual blocks, each of which contains two 3×3 convolutional layers.

Decoder and Constraint. The decoder contains two 3×3 convolutional layers with a stride of 1/2 to enable upsampling. To facilitate the convergence, we employ the mean square error (MSE) to constrain the loss l between the predictions and the ground truths, which can be formulated as follows:

$$L = \{l_1, \ldots, l_N\}^\top, \quad l_n = (x_n - y_n)^2, \tag{6}$$

where N is the batch size, x is the input and the y is the target. Then the batch loss ℓ can be calculated as:

$$\ell(x, y) = mean(L), \tag{7}$$

Wave-voice Net then minimizes the loss function $\ell(x, y)$ through back propagation to recover the voice from the the mmWave signal. As shown in Figure 4, it achieves an end-to-end transformation from a $3 \times 256 \times 256$ mmWave spectrogram to a $3 \times 256 \times 256$ voice spectrogram.

5.3 Phase Identification to Reconstruct Voice

The restored spectrogram contains amplitude and phase information of the voice that is necessary to determine the voice. However, the phase information is implicit in the spectrogram. To tackle this problem and reconstruct the final high-fidelity voice, we apply spectrogram inversion. In particular, we aim to estimate the phases from the spectrogram and thereby reconstruct the time-domain high-fidelity voice signal.

Phase Initialization. In the magnitude spectrum, we identify the bins that represent peaks by comparing the magnitude of each bin j with that of neighbors, $j+1$ and $j-1$. We consider bin j as a peak if $|X(mS, \omega_j)| > |X(mS, \omega_{j-1})|$ and $|X(mS, \omega_j)| > |X(mS, \omega_{j+1})|$. Here m is the index of the frame, S refers to the synthesis step size, which is one quarter of the frame length N, and $\omega_j = \frac{2\pi j}{N}$ presents the frequency of bin j. Afterward, we employ a phase accumulator ϕ [19] to represent the phase value for bin j at frame m,

$$\phi_{m,j} = \phi_{(m-1),j} + S\omega_j. \tag{8}$$

As the phases at the peaks have been determined, the phases at the remaining bins can be determined depending on the Pi-phase alternation strategy [20].

Griffin-Lim Reconstruction. Starting with the initial phase ϕ_m estimated above, we then employ the G&L algorithm [21] to iteratively renew the estimation:

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(n - mS) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \hat{X}^i(mS, \omega) e^{-j\omega n} \, \mathrm{d}\omega}{\sum_{m=-\infty}^{\infty} w^2(n - mS)}, \tag{9}$$

where $\hat{X}^i(mS, \omega)$ is the Short-Time Fourier Transform (STFT) of $x^i(n)$ with the following magnitude constraint, which can be described as:

$$\hat{X}^i(mS, \omega) = X^i(mS, \omega) \frac{|X(mS, \omega)|}{|X^i(mS, \omega)|}, \tag{10}$$

where $|X(mS, \omega)|$ is the Short-Time Fourier Transform Magnitude (STFTM) of the original signal $x(n)$. According to the results from [21], the G&L algorithm can reduce the Signal-to-Error Ratio (SER) through each iteration.

6 SYSTEM IMPLEMENTATION AND EVALUATION SETUP

6.1 mmWave Probe Implementation

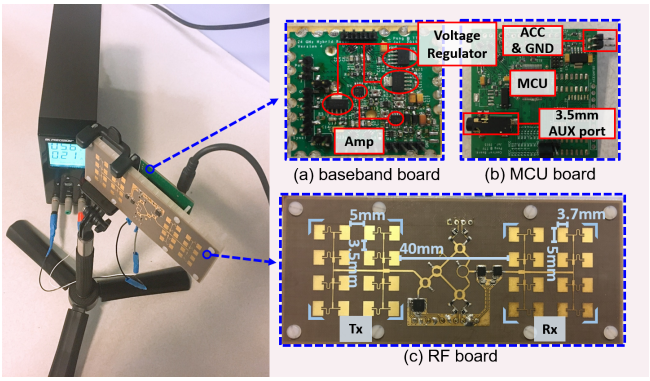

Figure 7: The implementation of 24GHz mmWave probe follows the three-board architecture: (a) a down-frequency baseband board (b) an MCU board and (c) a radio-frequency (RF) Tx/Rx board.

Figure 7 shows the implementation of the designed mmWave probe. The implementation consists of the RF board, the baseband board, and the MCU board. The implementation of the RF board adopts glass microfiber reinforced polytetrafluroethylene composites (Rogers RT/duroid 5880) as the substrate, which provides the thickness lower to 0.245 mm (0.0096 in). The size of the microstrip patch antenna is 5 mm × 3.7 mm, and the gap between two neighbor antennas is 5 mm in the x-axis and 3.5 mm in the y-axis. Both Tx and Rx consist of 16 antennas following the 4×4 layout. The distance between Tx and Rx of the probe is 40 mm.

We integrate the baseband board into the FR4 substrate, which supports the RF board. On the Tx side, three voltage regulators help stabilize the power to generate the mmWave as our design. On the Rx side, high-speed difference amplifiers are used to restrain the noise. The amplified signal then flows to the MCU board, where the embedded MCU is MSP430F2610. The on-chip 12-bit ADC of the MCU converts the amplified signal to the digital signal and feeds it to the 3.5 mm AUX port. We place the power interface in the MCU board, and the MCU board cascades the baseband board to power the entire probe.

Integrating all of these components, the size of the probe reaches 11.8 cm (4.65 in) × 4.5 cm (4.65 in) × 1.5 cm (0.59 in) and the total weights is 45.4 g. We use a 5.9 V power supply so that all ICs work at their nominal voltage. As a result, the total power consumption is 1.23W.

We adopt the time interval (dwell time) of 1000 milliseconds for throat localization, and we achieve an average accuracy of 96.7%. We observe that as the dwell time increases, the classification accuracy is promoted rapidly when the dwell time is less than 1000 milliseconds long. If we continue to increase the dwell time, the classification accuracy becomes progressively saturated.

6.2 Wave-voice Net Implementation

We implement the proposed Wave-Net in Pytorch. To avoid loss of the voice information, we adopt the reflection padding with $padding\ size = kernel\ size/2$ in all the convolutional layers. Moreover, the size of reflection padding guarantees a fixed feature map size under convolution stride = 1, thereby supporting our sequential residual block design of Wave-voice Net. We use the nearest upsampling in the decoder to retain time correlation of the restored voice spectrogram. Adam optimizer with an initial learning rate of 0.001 is used to train the Wave-voice Net. Additionally, we use a batch size of 32 and reduce the learning rate to 10% in every 128 epochs during the training process to fine tune the Wave-voice Net.

6.3 Dataset

Preparation: We conduct extensive experiments to prove the voice reconstructing capability of the proposed *WaveEar*. Figure 8 shows the experimental setup. A subject sits down in a normal position. We place the implemented mmWave probe in front of the subject and align it to the subject's throat. The mmWave probe connects to a 5.9V power supply, and the working current is 0.21A. A microphone is placed close to the probe to record the human voice. We attach another microphone to the subject's collar to collect near-throat voice as the reference signal. We deploy three laptops with Windows 10 and the open-source audio software Audacity to record the signals from the mmWave probe and the microphones, respectively. The three laptops are configured to start the recording at the same time so that the signals are aligned. The training and inference processes are performed by a workstation equipped with an Intel Xeon CPU E5-1620 v4 @ 3.50 GHz and an Nvidia Titan Xp 12 GB.

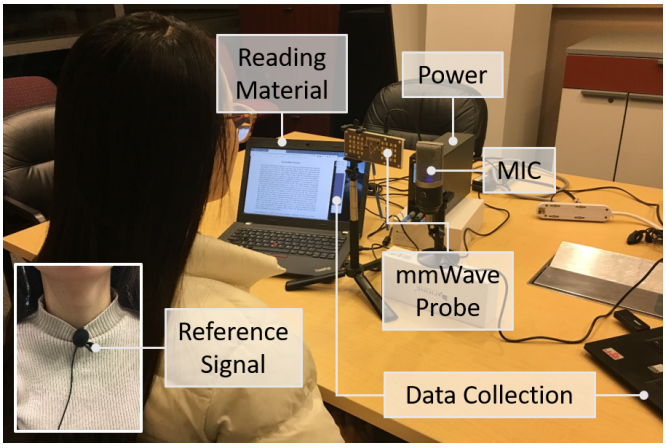

Figure 8: Experimental setup for speech sensing. A subject is sitting 0.5 meter away from the adjacent mmWave probe and microphone (MIC). We attach another microphone on subjects' collar to collect near-throat reference signals.

Participants: Our experiments involve 21 participants including 11 males and 10 females. The participants are recruited by emails including both graduate students and undergraduate students. These participants include both native and non-native English speakers with ages from 21 to 35 years old. We explicitly tell the participants that the purpose of the experiments is to perform voice reconstruction.

Data Collection: In our experiments, we choose five reading materials which cover all the phonemes in English and being used in many phonetic tests as the reading test benchmark. Table 2 shows the five materials and their statistics information. We denote the words per sentence, characters per word, and the average reading speed (words per second) of each material by "W/S", "C/W", and "Speed", respectively. During the experiments, all 21 subjects read the designated materials three times under the controlled quiet environment. We apply one-second segmentation and randomly choose the samples from two male subjects and two female subjects as the test data. Thus, there are 35394 samples used for training and 8328 samples for testing. Note that we have evaluated the system with different experimental setups and data partition for the varying ambient sounds and robustness analysis which will be described in these sections in particular.

Table 2: The reading materials used in the experiments.

Reading materials (Abbr.)	Length	W/S	C/W	Speed
The Rainbow Passage (Rainbow)	330	17.3	4.3	1.90
The Grandfather Passage (Grandfather)	132	16.5	4.2	1.97
Comma Gets a Cure (Comma)	375	17.7	4.2	1.98
The North Wind and the Sun (North)	113	22.6	4.2	2.05
Arthur the Rat (Arthur)	590	14.6	3.8	2.80

6.4 Evaluation Metrics

As a novel voice sensing system, *WaveEar* is compared with the conventional acoustic voice sensing system in the evaluation. Moreover, we also adapt objective quality assessments to evaluate the reconstructed voice. The evaluation metrics include:

Speaking-Signal-to-Noise Ratio (SSNR). *WaveEar* is a novel vocal sensing system that directly senses the vocal vibration but not the sound pressure. The typical reference signal for the SNR evaluation of the acoustic systems is the standard 1 kHz, 94 dB Sound Pressure Level (SPL) signal generated by the test instrument. However, such typical reference signal is not suitable for the proposed system. Therefore, we use the voice signal sensed by *WaveEar* as the reference signal and name the new measurement as Speaking-Signal-to-Noise Ratio (SSNR). Similar to the SNR, a higher SSNR indicates the system has a higher capacity for extracting the voice rather than the ground noise while sensing the voice.

Sensitivity. Sensitivity is measured by the same standard reference signal with SNR and can be formulated as follows:

$$20 \times \log_{10}\left(\frac{Sensitivity_{mV/Pa}}{Output_{AREF}}\right), \tag{11}$$

where $Output_{AREF}$ is the $1000mV/Pa$ ($1V/Pa$) reference output ratio. We replace the reference signal with the voice signal for the same reason as we propose the SSNR. Higher sensitivity provides a better resistance to the signal attenuation.

Word Error Rate (WER). In speech recognition, WER measures the difference between the reference and the recovered words (also named as hypothesis words). Specifically, the WER is formulated

as follows:

$$WER = \frac{S + D + I}{S + D + C}, \tag{12}$$

where S, D, I and C represent the number of substitutions, deletions, insertions, and correct words, respectively. We compare the WER of the reconstructed voices with that of the reference signal collected from the attached microphone based on the Google Cloud Speech-to-Text API [22]. The output is from the "Default model" and language is set to "English". Besides, we remove all the punctuation in the output text and disable the Speaker diarization. A lower WER indicates that the voice reconstructed by *WaveEar* is considered closer to the reference signal in the Speech-to-Text application.

Mel-Cepstral Distortion. MCD is a psychoacoustically motivated measure, which can evaluate the voice considering the human acoustic system preference. The MCD of the k-th frame is denoted by:

$$MCD = \sqrt{\sum_{i=1}^{13} [MC_X(i,k) - MC_Y(i,k)]^2}, \tag{13}$$

where the Mel-frequency cepstral coefficients (MFCCs) is calculated as:

$$MC_X(i,k) = \sum_{i=1}^{20} X_{k,n} \cos\left[i(n - \frac{1}{2})\frac{\pi}{20}\right], \; i \in [1, M] \tag{14}$$

and

$$X_{k,n} = \log_{10}\left(\sum_{M} |X(k,m)|^2 \cdot w_n m\right). \tag{15}$$

We use the audio signal from the attached microphone as the reference signal. A lower MCD indicates that the voice sensed by the system is more close to the original voice that humans hear.

7 PERFORMANCE ANALYSIS

In this section, we analyze the results of our extensive experiments, aiming to address the following questions: 1) the overall voice reconstruction performance of *WaveEar*; 2) performance comparison for various noise types; 3) robustness of the *WaveEar*.

7.1 WaveEar Performance

We evaluate the ability of *WaveEar* to reconstruct human voice in the controlled lab environment.

To give an intuitive awareness of the voice reconstruction performance of the *WaveEar*, we illustrate the spectrograms of the microphone-detected signals and the reconstructed voice from the Wave-voice Net in Figure 9. The overall shapes of the spectrograms are similar, indicating that the general variation of the human voice is reserved. The reason is we embedded a number of convolution layers into the Wave-voice Net, which has a strong ability to capture the high-level feature of the spectrogram. In addition, there is almost no difference in the color depth between the microphone's and the reconstructed voice signals. This observation shows that the energy and the magnitude of the human voice on a different frequency are not distorted with the help of the shortcut connection in residual blocks. The shortcut connection makes the input a part of the output, preventing the *WaveEar* from the data vanishing problem. Much reserved detail in low-frequency also confirms the excellent performance of the proposed *WaveEar*. We annotated these details, *i.e.*, formants, of the voice with their corresponding words in both the spectrograms of the microphone's and the reconstructed voice signal. Formants are distinctive frequency components of the human voice, which help humans detect the source of the voice. Having general awareness of the performance, we deepen our performance analysis by measuring *WaveEar* from different aspects using the four metrics aforementioned.

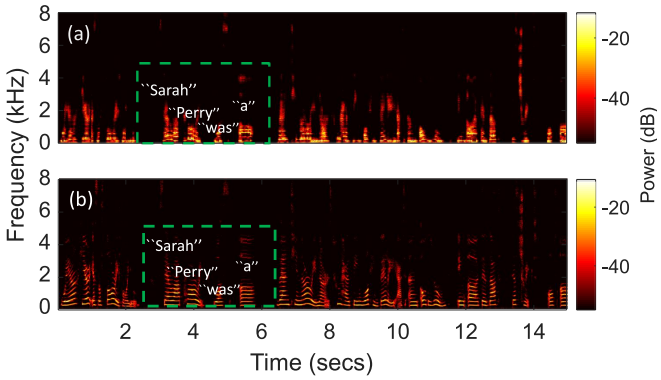

Figure 9: Spectrograms of (a) microphone-detect signal and (b) reconstructed voice signal. The vital voice information (refer to the annotated formants) are reconstructed by the Wave-Net. (Training loss = 0.00604357, Test loss = 0.00599584)

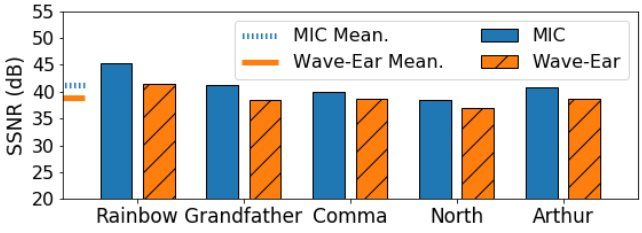

Figure 10: The SSNR of *WaveEar* with different reading materials.

Then, we further evaluate the *WaveEar* performance by the proposed three metrics. Figure 10 shows the average SSNR of the *WaveEar* and the microphone over the test set. In this experiment, subjects keep silent for 5 minutes when they sit in the chair so that we can capture the ground noise from both microphone and *WaveEar*. We observed that both *WaveEar* and microphone provide high SSNR during recording of human speech in the controlled environment. Specifically, the average SSNR of *WaveEar* is 38.85 dB while that of the microphone is 41.18 dB, which implies that *WaveEar* only introduces minute ground noise into the sensed voice. The results are because the mmWave probe is able to recognize and locate the near-throat vocal vibration leveraging the proposed Algorithm 1. When subjects keep silent, *WaveEar* will pause the wave-voice inference and stay mute.

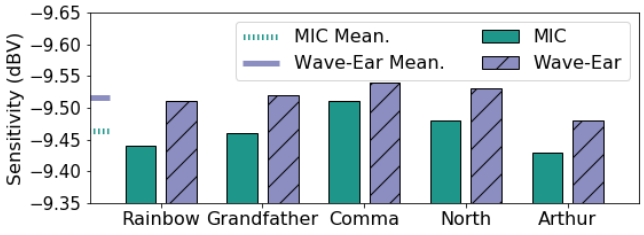

Figure 11: The sensitivity of *WaveEar* with different reading materials.

We use the reference signal from the attached microphone to calculate the sensitivity. The results are illustrated in Figure 11. For each reading material, the sensitivities of *WaveEar* and microphone have no significant difference. The mean sensitivities of *WaveEar* and microphone are -9.516 dBV and -9.464 dBV, respectively. Even mmWave attenuates more rapidly than sound waves, *WaveEar* senses the voice in a different way, *i.e.*, it learns the map from the reflected mmWave signal to speech voice directly, which compensates the sensitivity decreasing caused by the signal attenuation. Moreover, we evaluate the MCD of signals from *WaveEar* and microphone. As shown in Figure 12, the reported mean MCD of *WaveEar* is lower than 1.5, slightly higher than that of the microphone. These results imply that *WaveEar* is able to sense the voice as clear as the microphone.

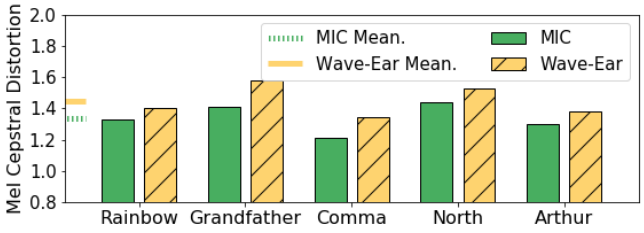

Figure 12: The MCD of *WaveEar* with different reading materials.

Finally, we leverage the WER to show the performance of voice reconstruction from the voice-to-text perspective. Figure 13 shows that the WER of *WaveEar* over the five reading materials is lower than 6%. Because Google Cloud Speech-to-Text not only leverages the voice itself but also uses the speech context to infer the text, a sequential word error happens only when a word is not reconstructed perfectly. Even *WaveEar* reports a low WER on Google Cloud Speech-to-Text, a joint optimization between *WaveEar* and the speech-to-text system would further enhance the speech-to-text application experience. In conclusion, our results demonstrate that a noise-resistant speech sensing can be achieved by *WaveEar*.

7.2 Resistance Test of Ambient Noise

The ambient environment can introduce different noises or even interfere with the probe hardware operation. In this subsection, we evaluate voice reconstruction performance across three different

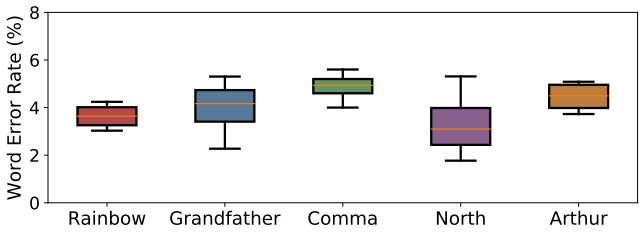

Figure 13: The WER of *WaveEar* with different reading materials.

types of real-world noises with different spectral characteristics: pop music, presentation, and water flow. In the experiments, one subject wears a noise-isolation headset and reads "The north wind and the sun" three times while a loudspeaker plays the three types of noises. The received mmWave signals are fed to the model trained on the training set to reconstruct the speech voice.

Pop music: Pop music consists of sounds coming from instruments and the singer. Figure 14(a) and (b) show the signals recorded by microphone and reconstructed by *WaveEar* in spectrogram. In the spectrogram of the microphone's signal, the beats of the music and the harmonics of the song are evident and mixed with the speech voice. In contrast, the spectrogram of the *WaveEar* outputs is clear with the speech voice reserved.

Presentation: As shown in Figure 14(c) and (d), the presentation voice has a large overlap with the reading voice of the subject. Besides, the echo of the presentation is recorded by microphone, messing up the signal. Signal reconstructed by *WaveEar* does not contain the frequency components of the presentation voice and the echo.

Water flow: The frequency of water flow noise is typically at 1-2KHz, which locates in the frequency range of human voice. Figure 14(e) demonstrates that the water flow noise forms a shallow horizon frequency band, which overlaps with the frequency components of speech voice in the spectrogram of the signal recorded by microphone. Compared to it, the spectrogram of *WaveEar*'s output does not contain this frequency band.

With all the observations above, we draw the conclusion that the proposed *WaveEar* is immune from the ambient noise.

7.3 Robustness Analysis

Impact of Distance and Orientation: In practical scenarios like smart homes, users should be able to walk around with *WaveEar* according to their accessibility, without placing it at a fixed location. However, such a convenient practice will result in variant distance and orientation between the users' throat and the mmWave probe. Therefore, we are motivated to analyze the performance of *WaveEar* under different distances and orientations. Specifically, the mmWave probe is placed in different orientations (from 0° to 90°) and different distances (from 0.5 m to 2 m) to the subject's throat when subjects read "Comma Gets a Cure" material. Figure 15 shows the measurement results. The MCD value remains low (less than 1.45) when the sensing distance is within 1.5 m. As the sensing distance exceeds 1.8 m, MCD value increases faster. As for the orientation,

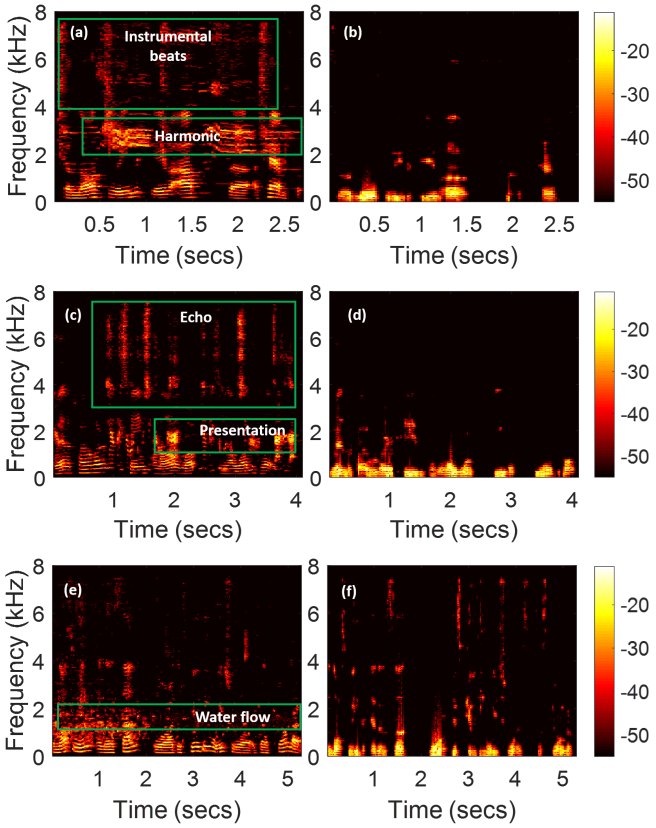

Figure 14: Spectrograms of microphone-detected signal (left) and WaveEar-sensed signal (right) with three different noises: music, presentation, and water flow. Noises cannot break into *WaveEar*.

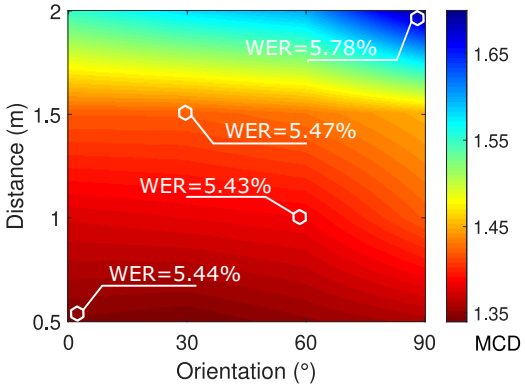

Figure 15: MCD varies under different probe orientation and sensing distance.

the MCD value slightly changes when the orientation varies. As the distance and orientation increase, WER stabilizes at around 5.5%. We can conclude that *WaveEar* is robust to the orientation changes,

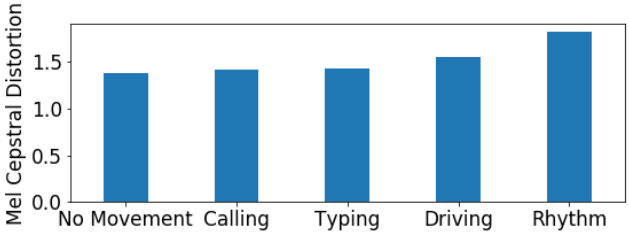

Figure 16: The average MCD comparison among different body motions.

and it can display superior performance within 1.5m sensing distance. Thereby, *WaveEar* can support robust and convenient speech sensing in real practice.

Impact of Body Motion: For comfortability, users always want to communicate with the *WaveEar* when needed, without putting down the things in their hands, such as driving. Therefore, we are curious about how well the *WaveEar* can be utilized for effective speech sensing under the effect of body motion. Specifically, we conduct a set of experiments to show the impact of body motion activities to the *WaveEar* performance, ranging from small to large-range motions. Twenty-one subjects are involved in this experiment and each one is required to perform four groups of body motion activities when reading the first three sentences of "Arthur the Rat" material, including making phone calls, typing, mimic driving, and rhythmical movement with beats. The overall performance on MCD when reading "Arthur the Rat" material is used as a control group without body motions. The mmWave probe is placed in front of the subject's body, where the distance and orientation between the subject and the probe are set at 0.5 m and 0°. With the current experiment setting, we examine the average MCD value of 21 subjects among different body motion activities. The results of the comparison are shown in Figure 16. We can observe that when subjects are making phone calls or typing during reading, the average MCD value almost keeps unchanged compared with the control group. While subjects perform mimic driving activities during reading, the average MCD value slightly increases. As for the large-motions such as rhythmical movement with the beats, the MCD value gains an obvious increment. The performance shows that small-range body motion artifacts can be removed via setting parameters of Wave-voice Net. Since our system is evaluated based on a training set containing a limited number of young subjects, some large-range body motions cannot be completely eliminated. A larger number of subjects with a more diverse background will help training the *WaveEar* system to become more robust.

Impact of Emotional State: A user's emotional state is changeable and one of its external expressions is speaking volume and speaking speed. Thereby, it is interesting to verify whether the user's emotional state can affect the *WaveEar* performance. According to [23], music can evoke emotions, 21 subjects are arranged to listen to blues, classic music, pop music, and heavy metal music via earphones, respectively. When reading "The Rainbow Passage", collect subjects' vocal vibration signals in different emotional states. We observe that the average time reading "The Rainbow Passage"

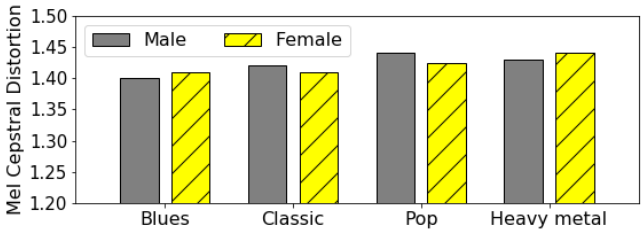

Figure 17: The average MCD comparison of men and women among different emotional states.

under blues is 157s while the average time under pop music is 151s, and even more than 90% subjects spend more reading time under blues. These preliminary results indicate that subjects' emotional states are changeable in our experiment settings. The mmWave probe is also positioned at a distance of 0.5m and an orientation of 0° facing toward the subject. Since Simon [24] proposed that women are more emotional than men, we measure the average MCD values of both men and women under different emotional states. Figure 17 shows the MCD comparison results. We observe that the average MCD values are almost stable when reading under different types of music, and the average MCD values of men and women differ a little. These results show that the *WaveEar* performance is robust to users' different emotional states regardless of a subject's sex.

8 IMPLICATIONS

8.1 Implications on Smart Device Architecture

Based on the real-world development process, this is an upcoming goal to market the mmWave-capable smartphones or smart devices.

Smart Device Architecture in 5G and Beyond: Due to its high frequency and board bands, mmWave can facilitate next-generation high-speed communication, i.e., 5G. Ericsson reports that the number of smart devices accessing 5G networks will reach one billion over the next 5 years. Currently, the first-generation smartphone equipping 5G baseband is introduced to the market. Therefore, we take the smartphone exploration as one illustration, as this end other devices have a similar mechanism. Compared to the model smartphone, the architecture of 5G smartphone differs mainly in the hardware, including the new phased array antenna, the extra beamforming module [25]. First, the embedded phased array antennas are separated to different parts of the entire phone body, which enables beamforming required by 5G, thereby providing spatial multiplexing. This leads to the non-physical eSIM design that provides more space inside the phone body. Second, the size of the extra beamforming model is well controlled below $10 \times 10 \times 1.5$ mm with the help of the emerging commercial solutions. Therefore, more beamforming chipsets are embedded on the back housing of the handset to support higher order MIMO. Given these considerations, we next discuss the compatibility of *WaveEar* with contemporary devices.

Compatibility with Smart Devices: On the hardware side, as aforementioned, the schemes of the 5G smartphone have comprised the phased array and beamforming module, which are the key components required by *WaveEar* front-end. On the software side, the wave-voice spectrogram and spectrogram inversion are easy to implement leveraging the high-speed DSP in the smartphone. Although Wave-voice Net is deployed on the server side and built with Pytorch, the ONNX format and TensorFlow mobile framework can deploy the Wave-voice Net to mobile devices, thereby training a more personalized model and further improving fidelity. Moreover, more and more AI chips have been applied to smart device products [26], which naturally meets the needs for Wave-voice Net. To sum, we can conclude that our system *WaveEar* is compatible or can be updated quickly with contemporary devices or smartphones.

8.2 Implications on Privacy

We consider the potential privacy leakage via mmWave side-channel and the related permission management.

mmWave-voice Side-channel: mmWave-voice can be a new side-channel that has the potential to be utilized by malicious applications or attackers. Once the malicious application controls the mmWave module of the smart device in the background, it can discover the voice of the nearby victim and extract sensitive information from the audio signal, although the user does not give any audio permission to the malicious application. Even smart devices without microphones are able to sense the voice. This implies the user should be aware of the related permission management as illustrated in the next section.

Permission Management: Our system empowers the ability of IoT devices (*e.g.*, smart refrigerator) to interact with the user through physical interaction capabilities. Although such physical interactions of IoT devices could bring significant convenience to end users, they also could be potentially exploited by attackers to jeopardize IoT environments. The access control policies can be generated automatically by machine learning techniques or from users' inputs. Thus, our system could achieve runtime interaction control without modifying existing applications.

8.3 Implications on Cocktail Party Effect

In real practice, it is not uncommon for multiple people to be present in the background of the legitimate user, hereafter Alice, who is speaking into the *WaveEar*. If those people are silent, our system can precisely localize Alice's position and record her vocal vibrations as previously shown in Section 4.2. However, the situation becomes more challenging when the people are also conversing which may cause *WaveEar* to incorrectly select the target speaker and attempt to record their vocal vibrations instead of Alice. To address this challenge, parameters such as sensing distance or body motion, which are unique to an individual, can be additionally leveraged to enhance the performance of localizing the target user's position. We plan to apply such techniques in our future versions of *WaveEar*.

9 DISCUSSION

Attenuation. Though Section 7.3 has shown that *WaveEar* can achieve low MCD and WER within the distance of 2m, the attenuation along the sensing distance due to the millimeter-level wavelength or caused by an object's block is still a major concern in the long-distance application of mmWave-based *WaveEar*. This

is because the attenuation will result in a low SNR thereby decreasing the sensing quality. To further improve the sensing distance of *WaveEar*, advanced noise cancellation technology [27] is needed to increase the SNR. Specifically, empirical mode decomposition is applied to the received low SNR mmWave signal to decompose it into oscillatory components. Then, the mutual information entropy is calculated on each component and a threshold is applied to filter out the noise components.

Human-factor distortion. The large body motion and human orientation are the two major factors that would distort the reflected mmWave signal thereby decreasing the performance of voice reconstruction. For large body motion, while the *WaveEar* shows the resistance to the body motion as we show in Section 7.3, it could be vulnerable when large body motion presents because the large motion will shift the reflected signal in the frequency domain. Thus, the signal alignment [28] could be involved in the signal pre-processing stage to compensate the distortion. For human orientation, recent research [29] reveals that the vocal vibration near the throat, neck, and even ear can be used for human authentication, indicating that the vocal vibration from different near-throat areas also contains rich voice information for reconstruction. Based on this exploration, one potential extension of *WaveEar* is to reconstruct the voice even with the hand or jaw occlusion.

Environmental independence. While the experiments of *WaveEar* do not appear to be overfitting, the reflected mmWave signal may carry latent information that is related to the environments where the experiments are conducted. Therefore, when *WaveEar* is applied to different environments, such as outdoors and conference rooms, removing these environmental specific information will help *WaveEar* to further enhance its performance. In particular, domain adversarial training [30] is a promising approach that could be integrated into *WaveEar* to filter out the latent environmental related information. This approach leverages a discriminator that tries to distinguish environments to indicate and suppress the latent environmental information, thereby facilitating the environmental independent voice reconstruction.

10 RELATED WORK

Speech Sensing: Speech sensing has a long and rich research history. Holzrichter *et al.* [12] used the glottal electromagnetic motion sensor (GEMS) to measure the movement of posterior tracheal wall during voiced speech. Many researchers also used sensors to detect electrical signals from various muscles related to speech production, such as electromyograms (EMG) sensor [31], and electroglottalgrams sensor [32]. In addition, many silent speech interfaces (SSIs) exist, which mainly work for sensing bone structure vibration (*e.g.,* bone-conduction sensors [33]) and skin vibration (*e.g.,* throat microphone [34], skin-conduction accelerometers [35], physiological microphone [36]). Recently, quite a few researchers studied more novel approaches for voice sensing. Zhang *et al.* [37] re-used the smartphone as a Doppler radar to sense a user's articulatory gestures. Ding *et al.* [38] adopted Kinect sensor for voice sensing. Roy *et al.* [39] demonstrated that vibration motor embedded in phones and wearables can be used as a sound sensor. However, most of sensors and interfaces need to have close contact with users or locate in a fixed place, which may cause user discomfort.

mmWave Sensing: In recent years, the development of millimeter wave (mmWave) radar makes it possible to sense biometrics and detect chemicals [40, 41]. Lin *et al.* [42] developed a smart DC-coupled continuous-wave (CW) radar to sense the high-resolution cardiac motion in a non-contact and continuous manner. Lien *et al.* [43] used mmWave to sense gestures for low power human-computer interaction. Lin *et al.* [44] proposed a sleep monitoring system, where a Doppler radar sensor is used to collect respiration data. Liu *et al.* [45] developed a novel gas sensor system based on a mmWave radar operating at 60GHz, which supports higher sensitivity and lower cost. Vocal vibration sensing by mmWave, as a new remote sensing modality attracts more attention in the academic community [46]. However, vocal vibration is a part of human speech which is the result of the collaboration of all vocal organs. To facilitate modern VUI, most of which is based on voice recognition, there is still a gap from vocal vibration to voice. *WaveEar*, as a mmWave based speech sensing system, is inspired by these previous studies but different from them in its noise-resistance and robustness.[1] *WaveEar* lays the first stone for Wave-voice interaction.

11 CONCLUSION

In this paper, we proposed a noise-resistant speech sensing system named *WaveEar*, to facilitate the human voice sensing for the widespread Voice-user interface (VUI) in a noisy environment. First, we explore the feasibility of acquiring high-quality voice by mmWave. Then, we proposed a low-cost 24 GHz mmWave probe with phased array, based on which we developed a throat localization approach. After that, the sensed signal containing speech information flows to the novel deep neural Wave-voice Net to reconstruct the speech via exhaustive extraction. Extensive experiments under real-world scenarios with 21 participants show the effectiveness of *WaveEar* to precisely infer the speech in noisy environments.

ACKNOWLEDGMENTS

We thank our shepherd, Dr. Romit Roy Choudhury, and all anonymous reviewers for their insightful comments on this paper. This work was supported by the National Science Foundation under grant No. 1718375 and the National Science Foundation of China under grant No. 61872195.

REFERENCES

[1] S. Chung and I. Rhee, "vtrack: virtual trackpad interface using mm-level sound source localization for mobile interaction," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct.* ACM, 2016, pp. 41–44.

[2] "Whitepaper: Top 10 consumer iot trends in 2017." [Online]. Available: https://www.parksassociates.com/whitepapers/top10-2017

[3] Y.-C. Tung, D. Bui, and K. G. Shin, "Cross-platform support for rapid development of mobile acoustic sensing applications," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services.* ACM, 2018, pp. 455–467.

[4] Y.-C. Tung and K. G. Shin, "Expansion of human-phone interface by sensing structure-borne sound propagation," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services.* ACM, 2016, pp. 277–289.

[5] *Smart Home Devices Market Forecast to Be Growing Globally at 31% Annual Clip.* https://www.securitysales.com/research/smart-home-devices-market-forecast/, Accessed: 2018-10-12.

[1] It is worth to mention that Wei *et al.* [47] also leveraged the vibration as a side-channel to eavesdrop a target loudspeaker by Wi-Fi signal.

[6] G. Laput, E. Brockmeyer, S. E. Hudson, and C. Harrison, "Acoustruments: Passive, acoustically-driven, interactive controls for handheld devices," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 2161–2170.

[7] Y. Meng, Z. Wang, W. Zhang, P. Wu, H. Zhu, X. Liang, and Y. Liu, "Wivo: Enhancing the security of voice control system via wireless signal in iot environment," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2018, pp. 81–90.

[8] X. Lei, G.-H. Tu, A. X. Liu, C.-Y. Li, and T. Xie, "The insecurity of home digital voice assistants-vulnerabilities, attacks and countermeasures," in *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2018, pp. 1–9.

[9] S. M. Kuo, K. Kuo, and W. S. Gan, "Active noise control: Open problems and challenges," in *Green Circuits and Systems (ICGCS), 2010 International Conference on*. IEEE, 2010, pp. 164–169.

[10] S. Shen, N. Roy, J. Guan, H. Hassanieh, and R. R. Choudhury, "Mute: Bringing iot to noise cancellation," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '18. New York, NY, USA: ACM, 2018, pp. 282–296. [Online]. Available: http://doi.acm.org/10.1145/3230543.3230550

[11] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in *2018 IEEE Symposium on Security and Privacy (SP). Vol. 00*, 2018, pp. 116–133.

[12] J. Holzrichter, G. Burnett, L. Ng, and W. Lea, "Speech articulator measurements using low power em-wave sensors," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 622–625, 1998.

[13] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air-and bone-conductive integrated microphones for robust speech detection and enhancement," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 249–254.

[14] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild." in *CVPR*, 2017, pp. 3444–3453.

[15] P. White and G. L. Reil, "Millimeter-wave beamforming: Antenna array design choices & characterization white paper."

[16] J. Kim and A. F. Molisch, "Fast millimeter-wave beam training with receive beamforming," *Journal of Communications and Networks*, vol. 16, no. 5, pp. 512–522, Oct 2014.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness business," in *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*. IEEE, 1997, pp. 4–pp.

[20] M. Puckette, "Phase-locked vocoder," in *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*. IEEE, 1995, pp. 222–225.

[21] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[22] "Cloud speech-to-text documentation," 2019, google cloud.

[23] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and brain sciences*, vol. 31, no. 5, pp. 559–575, 2008.

[24] R. W. Simon and L. E. Nath, "Gender and emotion in the united states: Do men and women differ in self-reports of feelings and expressive behavior?" *American journal of sociology*, vol. 109, no. 5, pp. 1137–1176, 2004.

[25] Y. Huo, X. Dong, and W. Xu, "5g cellular user equipment: From theory to practical hardware design," *arXiv preprint arXiv:1704.02540*, 2017.

[26] *Smartphone Artificial Intelligence (AI) Chip Market Poised for Growth*. https://www.strategyanalytics.com/strategy-analytics/blogs/components/handset-components/handset-components/2018/08/10/smartphone-artificial-intelligence-(ai)-chip-market-poised-for-growth, Accessed: 2018-10-09.

[27] F. Chen, S. Li, C. Li, M. Liu, Z. Li, H. Xue, X. Jing, and J. Wang, "A novel method for speech acquisition and enhancement by 94 ghz millimeter-wave sensor," *Sensors*, vol. 16, no. 1, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/1/50

[28] J. M. Muñoz-Ferreras and F. Pérez-Martínez, "Subinteger range-bin alignment method for isar imaging of noncooperative targets," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, p. 438615, 2010.

[29] R. Liu, C. Cornelius, R. Rawassizadeh, R. Peterson, and D. Kotz, "Vocal resonance: Using internal body voice for wearable authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 19, 2018.

[30] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: ACM, 2018, pp. 289–304. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241548

[31] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 201–216, 2001.

[32] I. R. Titze, B. H. Story, G. C. Burnett, J. F. Holzrichter, L. C. Ng, and W. A. Lea, "Comparison between electroglottography and electromagnetic glottography," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 581–588, 2000.

[33] S. Stenfelt, T. Wild, N. Hato, and R. L. Goode, "Factors contributing to bone conduction: The outer ear," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 902–913, 2003.

[34] B. E. Acker-Mills, A. J. Houtsma, and W. A. Ahroon, "Speech intelligibility in noise using throat and acoustic microphones," *Aviation, space, and environmental medicine*, vol. 77, no. 1, pp. 26–31, 2006.

[35] Ü. Ç. Akargün and E. Erzin, "Estimation of acoustic microphone vocal tract parameters from throat microphone recordings," in *In-Vehicle Corpus and Signal Processing for Driver Behavior*. Springer, 2009, pp. 161–169.

[36] S. A. Patil and J. H. Hansen, "The physiological microphone (pmic): A competitive alternative for speaker assessment in stress detection and speaker verification," *Speech Communication*, vol. 52, no. 4, pp. 327–340, 2010.

[37] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 57–71.

[38] I.-J. Ding, C.-M. Ruan, and J.-Y. Shi, "Operational control of the multimedia player of smart phones using a kinect voice-sensing scheme," in *Proc. Int. Conf. Inf. Eng. Mech. Mater.* Citeseer, 2015, pp. 194–198.

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

[40] C. Li, Z. Peng, T. Huang, T. Fan, F. Wang, T. Horng, J. M. noz Ferreras, R. Gómez-García, L. Ran, and J. Lin, "A review on recent progress of portable short-range noncontact microwave radar systems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 5, pp. 1692–1706, May 2017.

[41] J. M. oz Ferreras, Z. Peng, R. Gómez-García, and C. Li, "Review on advanced short-range multimode continuous-wave radar architectures for healthcare applications," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 1, no. 1, pp. 14–25, June 2017.

[42] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 2017, pp. 315–328.

[43] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 142, 2016.

[44] F. Lin, Y. Zhuang, C. Song, A. Wang, Y. Li, C. Gu, C. Li, and W. Xu, "Sleepsense: A noncontact and cost-effective sleep monitoring system," *IEEE transactions on biomedical circuits and systems*, vol. 11, no. 1, pp. 189–202, 2017.

[45] S. Liu, G. Shaker, S. Safavi-Naeini, and J. M. Chong, "Low-cost gas sensors utilizing mm-wave radars," in *Antennas and Propagation & USNC/URSI National Radio Science Meeting, 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1853–1854.

[46] C. Lin, S. Chang, C. Chang, and C. Lin, "Microwave human vocal vibration signal detection based on doppler radar technology," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 8, pp. 2299–2306, Aug 2010.

[47] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: ACM, 2015, pp. 130–141. [Online]. Available: http://doi.acm.org/10.1145/2789168.2790119