AmbiEar: mmWave Based Voice Recognition in NLoS Scenarios

JIA ZHANG*, Tsinghua University, China
YINIAN ZHOU*, Tsinghua University, China
RUI XI, Tsinghua University, China
SHUAI LI, Tsinghua University, China
JUNCHEN GUO, Alibaba DAMO Academy, China
YUAN HE†, Tsinghua University, China

Millimeter wave (mmWave) based sensing is a significant technique that enables innovative smart applications, e.g., voice recognition. The existing works in this area require direct sensing of the human's near-throat region and consequently have limited applicability in non-line-of-sight (NLoS) scenarios. This paper proposes AmbiEar, the first mmWave based voice recognition approach applicable in NLoS scenarios. AmbiEar is based on the insight that the human's voice causes correlated vibrations of the surrounding objects, regardless of the human's position and posture. Therefore, AmbiEar regards the surrounding objects as ears that can perceive sound and realizes indirect sensing of the human's voice by sensing the vibration of the surrounding objects. By incorporating the designs like common component extraction, signal superimposition, and encoder-decoder network, AmbiEar tackles the challenges induced by low-SNR and distorted signals. We implement AmbiEar on a commercial mmWave radar and evaluate its performance under different settings. The experimental results show that AmbiEar has a word recognition accuracy of 87.21% in NLoS scenarios and reduces the recognition error by 35.1%, compared to the direct sensing approach.

CCS Concepts: • **Computer systems organization → Embedded and cyber-physical systems**; • **Hardware → Sensor applications and deployments**.

Additional Key Words and Phrases: Wireless Sensing, Millimeter Wave, Voice Recognition

ACM Reference Format:
Jia Zhang, Yinian Zhou, Rui Xi, Shuai Li, Junchen Guo, and Yuan He. 2022. AmbiEar: mmWave Based Voice Recognition in NLoS Scenarios. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 151 (September 2022), 25 pages. https://doi.org/10.1145/3550320

1 INTRODUCTION

Millimeter wave (mmWave) based sensing is a significant technique that can sense the object by collecting and analyzing the reflected mmWave signals. Owing to its short wavelength and high spatial resolution, mmWave as a sensing medium has attracted a large body of research attention in the last few years. The applications range from micro-movement measurement [12, 17], motion and activity sensing [19, 50], material identification [51], to environmental sensing [4, 22], etc.

*Both authors contributed equally to this research.
†Corresponding author

Authors' addresses: Jia Zhang, j-zhang19@mails.tsinghua.edu.cn, Tsinghua University, Beijing, China; Yinian Zhou, zhouyn19@mails.tsinghua.edu.cn, Tsinghua University, Beijing, China; Rui Xi, Tsinghua University, Beijing, China, ruix.ryan@gmail.com; Shuai Li, Tsinghua University, Beijing, China, lis20@mails.tsinghua.edu.cn; Junchen Guo, Alibaba DAMO Academy, Hangzhou, China, juncguo@gmail.com; Yuan He, Tsinghua University, Beijing, China, heyuan@mail.tsinghua.edu.cn.

2474-9567/2022/9-ART151
https://doi.org/10.1145/3550320

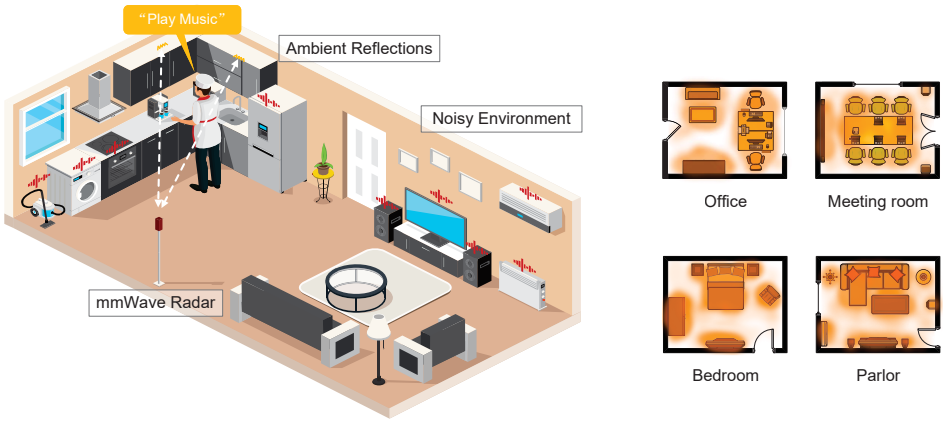


Fig. 1. AmbiEar leverages the mmWave signal reflected from ambient reflections to facilitate voice recognition in NLoS scenarios.

mmWave based sensing has great potential in supporting smart human-computer interaction, e.g., voice recognition. Conventional solutions [26, 34–36, 41] use microphone(s) to collect sound and then analyze the human's voice contained in the sound. Their performance will significantly degrade in noisy environments [1]. As may be our familiar experience: when a machine (a TV set, a washing machine, a sweeping robot, or a vacuum cleaner) is working, the voice recognition function of a smart speaker [46–49] often fails, because the voice signal perceived by the microphone has a limited signal-to-noise ratio (SNR) [31, 32].

Due to the extremely high frequency and high spatial resolution of mmWave signals, mmWave based voice recognition has been proposed to resist noisy environments. By extracting only the voice-related reflected mmWave signals from the near-throat region in the radar's field of view [21, 24, 28, 52], voice recognition can be realized by analyzing those reflected signals. For example, WaveEar [52] directs the mmWave signals towards the near-throat region to sense the vocal vibration and recover the voice.

However, mmWave based voice sensing often fails when the human moves or a blockage exists between the radar and the target. The reason is that the voice-related reflected signals from the near-throat region cannot be obtained due to the extremely high frequency and the weak diffraction characteristics of mmWave. Therefore, the existing approaches encounter serious problems in non-line-of-sight (NLoS) and dynamic scenarios. As shown in Fig. 2, in an ideal scenario, the fixed throat can be easily located and tracked as long as the line-of-sight (LoS) path between the throat and the radar exists. Whereas, in a practical scenario, the LoS path of the throat is hard to find or even does not exist due to the dynamics of the human's position and posture and the blockage. As a result, the applicability of the existing direct sensing approaches is far from satisfactory in the real world.

In order to make mmWave based voice recognition indeed applicable in practice, we turn to explore an indirect sensing approach to complement existing methods. Fig. 1 plots a typical scenario. There are often a number of objects in the living and working environments. We observe an interesting fact: *Sound propagates as a mechanical wave. The human's voice causes vibrations of the surrounding objects, which contain common components that are highly correlated with the human's voice.* Inspired by this insight, we propose **AmbiEar**, a mmWave based indirect sensing approach for voice recognition. AmbiEar converts the surrounding objects into ambient "ears" and uses a mmWave radar to perceive the voice-related signals from the "ears" for voice recognition. AmbiEar is designed for but not limited to NLoS scenarios. Since it utilizes surrounding objects for voice recognition, AmbiEar also works well in LoS scenarios.

To put this idea into practice, one will encounter critical challenges associated with processing the low-quality voice-related signals. First, since the location of the human is unknown in advance and dynamic, it is difficult to

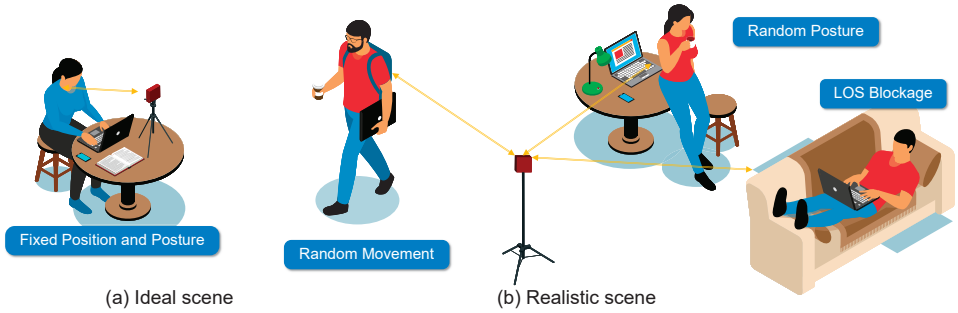


Fig. 2. It is unrealistic to assume that mmWave radars can accurately locate the human's throat in reality.

accurately identify the objects around the human. Second, the surrounding object's vibration is much weaker than the vibration of the throat. According to our measurement, the amplitude of the object's vibration is as weak as 3-5um. Such weak vibration signals are likely to be buried by electromagnetic noise in the view of a mmWave radar. The environment acoustic noise around the object further reduces the signal-to-noise ratio (SNR). In this paper, we distinguish these two types of noises by using the terms "environment acoustic noise" and "electromagnetic noise". How to obtain signals of sufficiently good quality is a daunting task. Last but not least, due to the sound attenuation in the air and the energy loss in the transformation from sound to vibration, the vibration signal is intrinsically distorted, compared to the original voice signal. Directly feeding such semantically incomplete signals into a voice recognition procedure will yield erroneous results.

We address the above challenges with a progressive scheme: First, AmbiEar distinguishes a human and the surrounding objects according to their different dynamics in the reflected signals. Then in the critical second step, AmbiEar extracts and combines the common components of vibration signals from multiple objects. Since the vibration signals of different objects contain the same voice-related components, enhancing their common components can effectively improve the voice-related signal's SNR. By further incorporating an end-to-end network to extract the voice-related features in the vibration signals, AmbiEar effectively deals with the signal distortion problem and achieves accurate voice recognition.

Our contributions can be summarized as follows:

- We propose the concept of mmWave based indirect sensing and convert the surrounding objects into "ears" that help us perceive. To the best of our knowledge, AmbiEar is the first-of-its-kind approach for mmWave based voice recognition in NLoS scenarios.
- The design of AmbiEar effectively tackles a series of technical challenges. We particularly pay attention to the problem of how to utilize the low-SNR and semantically incomplete vibration signals for voice recognition. The tailored design of AmbiEar includes four main parts, namely surrounding detection, common component extraction, signal superimposition, and voice recognition.
- We implement AmbiEar on the commercial device (TI IWR1642 board) and conduct extensive experiments under various settings. The results demonstrate that AmbiEar achieves a word recognition accuracy of 87.21% and a character recognition accuracy of 88.66%. AmbiEar significantly improves the applicability of mmWave based voice recognition in practice.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the background and preliminaries of our work. The theoretical model is introduced in Section 4. Section 5 elaborates on our design. Section 6 presents the implementation and evaluation results. Section 7 discusses some practical problems. We conclude this work in Section 8.

2 RELATED WORK

In this section, we first introduce the works about mmWave based subtle displacement measurement, which is the cornerstone of achieving voice recognition. Then we discuss some works related to mmWave based voice sensing and their limitation in NLoS scenarios. After that, some other works related to mmWave based human sensing are briefly introduced. Finally, we discuss some sensing works based on environmental reflection and sound sensing works in NLoS scenarios.

With the extremely high frequency, mmWave sensing can achieve millimeter-level or even sub-millimeter-level subtle displacement measurement. mTrack [45] utilizes a signal-phase-based model to achieve millimeter-level tracking accuracy. mmVib [17] introduces a multi-signal consolidation model to achieve sub-millimeter vibration measurement error. These works demonstrate the ability of mmWave to measure subtle displacement, which can be used for voice sensing.

The recent mmWave based voice sensing works mostly recognize and analyze the voice by sensing the vibration of the sound source. WaveEar [52] directs the mmWave signals towards the near-throat region to sense the vocal vibration and recover the voice. Vocalprint [21] makes use of mmWave sensing to preserve fine-grained vocal biometric properties and realize voice authentication. RadioMic [28] presents training-free approaches for robust sound detection and high-fidelity sound recovery based on tiny vibrations of sound sources. Wavoice [24] utilizes multi-modal signals (mmWave signals and audio signals) fusion and achieves accurate voice recognition. All of them sense the voice by analyzing the vibration signals of the sound source, which requires the LoS path between the sound source (e.g., the human's throat) and the radar. When the LoS path disappears as the human's position and posture change, these works cannot work well.

This problem also exists in other mmWave based human sensing works. Whether locating and tracking the human's position and posture [19, 23, 50, 54] or sensing the human's vital signals such as breathing and seismocardiogram [13, 53], it is necessary to sense the human body through the LoS path.

There have been some works that utilize reflections in the environment to sense the object or achieve sound sensing in NLoS scenarios. GWaltz [12] combines the coherent observations from multipath reflections to restore the 2D orbit of the target. The 2D rotor orbit is measured by analyzing the extra information extracted from the NLoS path between the target and the radar. However, AmbiEar analyzes the target's change by sensing the reflector itself instead of the NLoS path between the target and the radar. On the other hand, LidarPhone [37] uses a lidar sensor to eavesdrop on the privacy information from the vibrations induced on nearby objects. However, as it can only sense a single object's vibration due to the directivity of the laser, the signal quality will deteriorate and its performance may decrease significantly in noisy environments. In contrast, AmbiEar can effectively resist the impact of environment noise via extracting the common components of multiple objects' vibrations. VisualMic [6] utilizes high-speed video to analyze the subtle vibration of an object and recover the sound around it. The subtle vibration signals are derived from phase variations in the complex steerable pyramid. Such a method requires well-lighting conditions and hours of complex calculations. In contrast, our system can work in any lighting conditions and does not require a huge computing cost. ART [44] eavesdrops on loudspeaker sound by analyzing the subtle disturbance that the sound causes to the radio signal with a prototype based on software-defined radio (SDR). Although it can eavesdrop in NLoS scenarios, it still needs to directly sense the sound source by transmitting carrier signal to penetrate walls. In contrast, AmbiEar uses commercial mmWave radar to sense surrounding reflectors rather than directly sensing the sound source. RadioMic [28] also claims to be able to recover sound in NLoS scenarios. It is able to work in scenarios where there are soundproof materials such as a double glass layer between the sound source and the radar. The radar can still receive and analyze the reflected signal from the sound source. However, AmbiEar can recover sound from surrounding reflectors when the reflected signal of the sound source cannot be received.

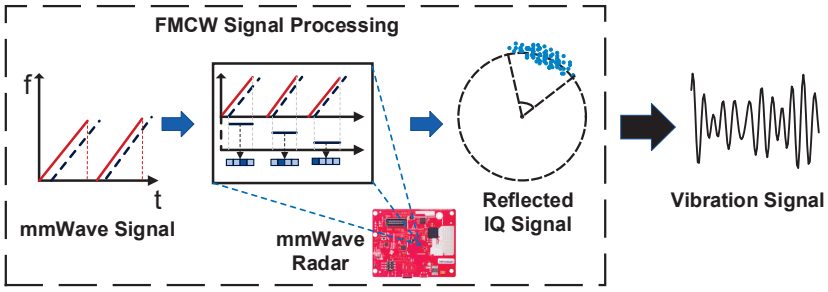


Fig. 3. The process of mmWave based vibration measurement

Compared with the existing works on mmWave based voice sensing, AmbiEar particularly addresses the challenges of voice sensing in NLoS scenarios, which is a missing piece in the literate. Although there are a few works that also utilize environmental reflections, the idea behind is essentially different from what we present in AmbiEar. Specifically, Gwaltz exploits the NLoS paths of mmWave propagation to construct a multi-view measurement of the same target. What the radar senses in Gwaltz is still the vibrating target itself rather than the surrounding reflectors. LidarPhone is based on lidar sensing and can only sense one object at a time. VisualMic utilizes video analysis and requires a huge computing cost. Different from LidarPhone and VisualMic, AmbiEar exploits a mmWave radar's ability to simultaneously sense and analyze multiple targets. It is able to extract the coherent information from multiple objects' vibrations and improve the signal's SNR.

3 BACKGROUND AND PRELIMINARIES

In this section, we first introduce the background knowledge about mmWave based vibration measurement, which is the cornerstone of our voice sensing. Then we discuss our preliminary studies and experimental results. On the one hand, these preliminary studies illustrate the limitation of voice recognition by directly sensing the near-throat region. On the other hand, they demonstrate the feasibility of extracting voice indirectly from the surrounding objects, which inspires us to propose our solutions in this paper.

3.1 mmWave Based Vibration Measurement

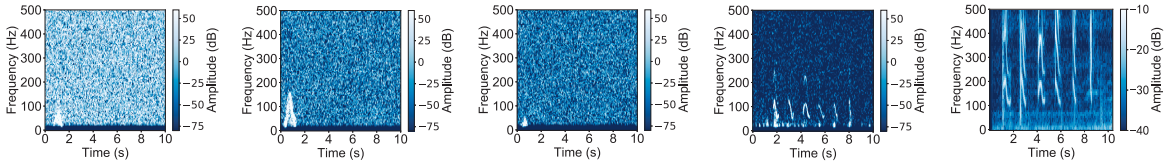
As shown in Fig. 3, the mmWave radar periodically sends *frequency-modulated continuous wave* (FMCW) signals to measure the distance of the target. The frequency difference between the transmitted signal and the received signal corresponds to the signal propagation time and can be used to calculate the propagation distance. By mixing the transmitted signal and the received signal, the *beat frequency signal* $s(t)$ can be obtained as:

$$s(t) = \alpha \exp[j4\pi(f_c + Kt)R(t)/c] \tag{1}$$

where α represents the path loss. f_c and K are the chirp starting frequency and the chip slope of the FMCW signal, respectively. To separate Rx signal components reflected from different ranges, we perform a Range-FFT [23] operation on the samples of $s(t)$ within a chirp to separate the reflected signals from different distances. Then in a certain range bin, these samples can form a new reflected signal $S(t)$. The reflected signal can be calculated as:

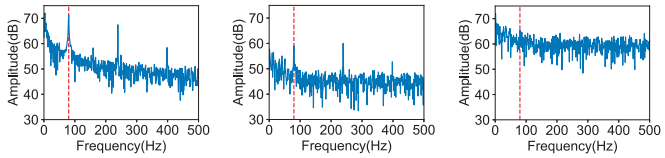
$$s(t) \xrightarrow[\text{at a range bin}]{\text{Range-FFT}} S(t) = \alpha \exp[j4\pi f_c R(t)/c] \tag{2}$$

where $R(t)$ represents the distance between the radar and the object. When the object vibrates, the distance $R(t)$ can be rewritten as $R(t) = R_0 + x(t)$, where R_0 represents the radar-object distance and $x(t)$ represents the object vibration displacement.



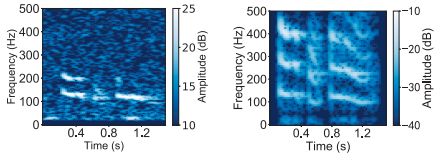
(a) head (b) back (c) arms (d) throat (e) voice

Fig. 4. The STFT results of the different body parts' reflected signals and the original voice signal



(a) 30cm (b) 60cm (c) 90cm

Fig. 5. The Fourier Transform results of the vibration signals from different distances



(a) The vibration signal (b) The voice signal

Fig. 6. The similarity between the vibration signal and the voice signal

Based on Eq. 2, the signal $S(t)$ manifests a circular arc in the IQ domain. The arc's phase change indicates the radar-object distance change, which can be obtained as:

$$x(n) = \frac{c}{4\pi f_c} unwrap(\phi_n) - R_0, n \in [1, N] \tag{3}$$

where ϕ_n represents the n-th sample's phase and $x(n)$ is the radar-object distance change. The signal processing is shown in Fig. 3.

3.2 Limitation of Extracting Voice Directly

We have conducted some preliminary studies about extracting voice directly from the human body. In our experiments, a volunteer is instructed to sit still and say "one, two, three, four, five, six". A commercial mmWave radar (TI IWR1642) is set in front of the volunteer's throat, head, back, and arms. Fig. 4 shows the short-time Fourier transform (STFT) results of these reflected signals and the original voice signal recorded by a microphone.

The results show that only when the throat is directly in front of the radar can the voice-related information be extracted. However, the human's throat is a tiny target that is hard to locate and track. Thus, it is challenging to extract human voices exactly from the human throat.

3.3 Feasibility of Extracting Voice Indirectly

In addition, we conduct some other preliminary studies, which help us discover the feasibility of extracting voice indirectly. We first use the single-frequency sound signal at normal volume to observe the effective range of sound-vibration transformation, and then we explore the similarity between the vibration signal and the voice signal.

3.3.1 The Effective Range of Sound-vibration Transformation. In these experiments, we use a speaker to play a 70dB 80Hz single frequency sound signal. We set a 150g 20cm×20cm iron box at 30cm, 60cm, and 90cm away from the speaker. Such a heavy iron box is chosen to explore the impact of sound on objects in actual scenes. We use an eddy-current sensor to detect the vibration of the iron box. Fig. 5 shows the Fourier transform results of the iron box's vibration signals at different distances. When the distance between the iron box and the speaker is

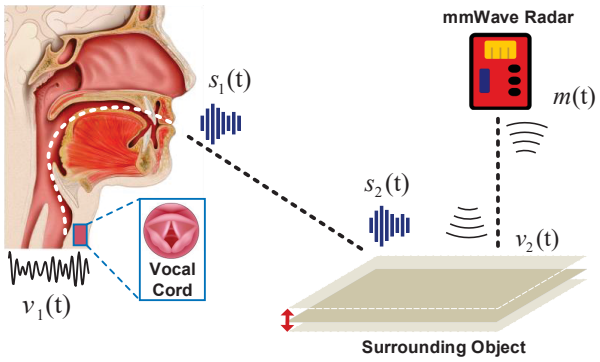


Fig. 7. The theoretical sensing model

within 60cm, the vibration of the iron box contains an obvious 80Hz component. When the iron box is 90cm away, there is still a small energy peak at 80Hz, but it is hard to distinguish. Other obvious peaks in the figures of 30cm and 60cm are mainly harmonics with multiples of 80Hz. These results show that the normal volume sound only causes the objects within a certain range (e.g., 1m) to vibrate in actual scenes. This observation is helpful in overcoming the influence of environment noise. We will explain it later.

3.3.2 Similarity between the Vibration Signal and the Voice Signal. We further verify that the complex voice signal can cause similar vibration to the original voice signal. We set the iron box 60cm away from a volunteer. The volunteer is instructed to speak randomly. We use a microphone to record the original voice signal and the eddy current sensor to record the vibration of the iron box. Fig. 6 shows the STFT results of the original voice signal and the iron box's vibration signal. We find that the two signals contains certain consistency. However, the vibration signal has obvious spectrum distortion, especially in the high-frequency components. The reason is that the human voice attenuates when it propagates in the air and losses when it is transformed into a vibration signal. The high-frequency components are more obviously affected by these two factors.

These experimental results show that it is feasible to sense the voice signal from the vibrations of the surrounding objects. However, the vibration signal has obvious spectrum distortion and poor signal quality. If such a signal is directly used as input, the voice recognition result will be degraded. The problem of spectrum distortion and poor signal quality must be solved to sense the voice signal clearly. We further explain it in the following section.

4 THEORETICAL MODEL

In this section, we introduce the theoretical model of using the surrounding objects to sense the voice signal. Based on this model, there are three factors that affect the quality of the received signal that should be considered.

The whole process is shown in Fig. 7. When a human speaks, the vocal cord first generates the vibration signal $v_1(t)$. According to the source-filter model [8], the voice from the mouth $s_1(t)$ can be represented as a combination of the vocal cord's vibration and a linear acoustic filter formed by the vocal tract.

The previous works, such as WaveEar [52] and VocalPrint [21], both sense the voice signals by analyzing the vocal cord's vibration signals. However, the vocal cord's vibration signals don't need to be considered in our design as the vibrations of the surrounding objects are directly excited by the voice signal. When the voice signal reaches the object, it can be represented as $s_2(t)$ which includes attenuation and phase change brought by the air channel.

The sound around the object will cause the object to vibrate. Considering that the material, shape, and other characteristics of the object will affect the vibration signal excited by the sound, the vibration signal $v_2(t)$ can be represented as a combination of the sound around the object and the conversion process, which is related to the

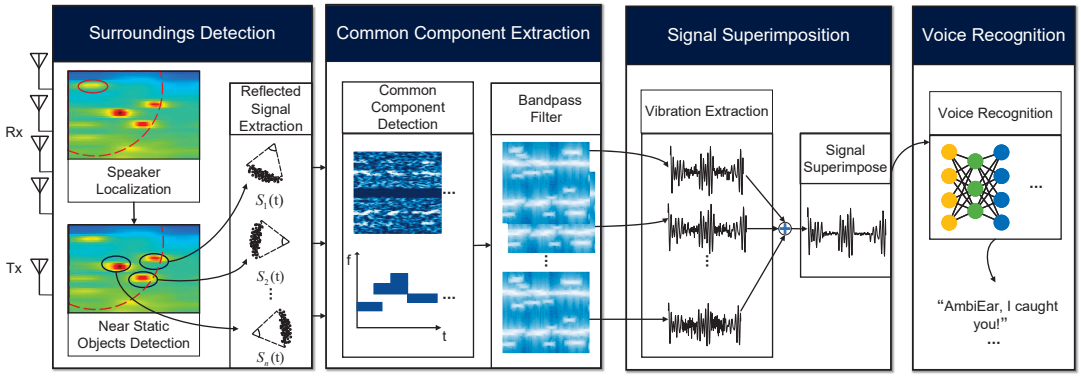


Fig. 8. The oveview of AmbiEar

object's properties. Note that the sound around the object includes the voice signal $s_2(t)$ and the environment acoustic noise. The environment acoustic noise also affects the object's vibration.

Finally, when the vibration signal is sensed by the mmWave radar, the received signal $m(t)$ is the superposition of the vibration-related reflected signal and the electromagnetic noise.

We can find that the mmWave voice is strongly related to the voice signal. In our theoretical model, there are three factors that affect the quality of the received signal, namely the spectrum distortion between the voice signal and the vibration signal, the environmental acoustic noise around the object and the electromagnetic noise in mmWave signal processing. These factors work together to affect the signal quality. Considering the noisy environment, the degradation of the signal quality is more serious and must be addressed.

5 SYSTEM DESIGN

Based on the above observations and analysis, we design our system, AmbiEar, which achieves accurate voice recognition in NLoS scenarios without restricting the human's position and posture. To counter the reduction in the signal quality caused by the three factors mentioned above, we design three modules, namely common component extraction, signal superimposition, and voice recognition. Since AmbiEar only observes the surrounding objects' vibrations and does not need to locate the human's throat, our system can work in mobile scenes compared to previous works.

5.1 Overview

The overview of AmbiEar is shown in Fig. 8. AmbiEar consists of four parts, namely surrounding detection, common component extraction, signal superimposition, and voice recognition. We introduce each module below:

- *Surroundings Detection.* Firstly, AmbiEar scans the environment to obtain information about all the objects. Then the 2D CFAR algorithm and the DBSCAN algorithm are applied to obtain the locations of the objects. We further distinguish the human's trajectory from the others using a maximum likelihood estimation algorithm. Finally, AmbiEar finds the static objects surrounding the human within a specific range. In this way, the **dynamics** of the human body can be handled.
- *Common Component Extraction.* After locating these static objects, AmbiEar extracts the mmWave signals reflected from them by the mmWave radar. Then AmbiEar uses a modified MVDR algorithm and a group of fine-grained bandpass filters to extract their common frequency components. Considering that the vibration signals of the surrounding objects contain the same voice-related components, their common components can be used to effectively resist the impact of **environment acoustic noise** and **electromagnetic noise**.

- *Signal Superimposition.* After improving the SNR of each reflected signal separately, AmbiEar extracts these vibration signals from the reflected signals and superimposes them to further resist the impact of **environment acoustic noise**. Then the time-frequency spectrograms of the superimposed vibration signals are generated for voice recognition.
- *Voice Recognition.* Finally, We use a customized encoder-decoder network to output the corresponding semantics from the generated time-frequency spectrograms, which has obvious **spectrum distortion** and is semantically incomplete compared to the voice signal.

5.2 Surroundings Detection

AmbiEar first scans the environment to obtain information about all the objects, including their positions and reflected signal strength. Then it continuously tracks the objects' positions and selects the human's position based on the variance of the trajectories. After that, AmbiEar can select the surrounding objects in a specific range as the **reflectors** for the next step.

AmbiEar first periodically scans the environment and obtains the signal strengths of all positions in the field of view, namely the range-angle spectrum. It can be obtained by applying classic range FFT and receiver beamforming algorithms [39].

Then the 2D Constant false alarm rate (CFAR) algorithm [33] is applied to the range-angle spectrum to detect these rang-angle bins with objects, as shown in Fig. 9. CFAR is a standard adaptive algorithm used to detect targets against environment noise. After estimating the noise level by convolving the CFAR window with the signal strengths, the bins with energy higher than the noise level will be retained and considered as the bins with objects. According to our experience, the detection result is best when the values of guard cell and training cell are both set to 2.

After obtaining the 2D CFAR result, AmbiEar applies the DBSCAN algorithm [9] to cluster the CFAR result. DBSCAN is a classic clustering algorithm that does not assume the number and shape of clusters, which is suitable for our scenario. Each cluster center represents an object in the field of view, which can be represented as:

$$O_j :< D_j, A_j, S_j > \tag{4}$$

where O_j represents the j-th object. D_j, A_j, and S_j is the distance between the j-th object and the radar, the angle between the j-th object and the radar, and the reflected signal strength, respectively.

Considering that the human's trajectory has more changes compared to that of static objects, AmbiEar can determine the human's position based on the variance of these trajectories. Specifically, AmbiEar first calculates the Jaccard Similarity Coefficients[20] between every pair of object's positions in two adjacent scanning result P_i and P_{i+1}:

$$J_{j,k} = \frac{\left|C_{i,j} \cap C_{i+1,k}\right|}{\left|C_{i,j} \cup C_{i+1,k}\right|},\ (1 \le j \le N_i, 1 \le k \le N_{i+1}) \tag{5}$$

where $C_{i,j}$ represents the j-th cluster in i-th scanning result P_i. N_i and N_{i+1} are the number of clustering results in P_i and P_{i+1}, respectively. The Jaccard Similarity Coefficient is a statistic used for gauging the similarity and diversity of sample sets. Here we use it to measure the similarity of the clusters' positions in two adjacent scanning results. Then we construct a bipartite graph to obtain the trajectories. The two vertex sets of the bipartite graph are the clusters of the two scan results, and the weights of the edges between the vertexes are $J_{j,k}$. By such modeling, we transform the tracking problem into the optimal match problem. We solve it by the classic Kuhn-Munkres algorithm[27], which can find the matching with the largest sum of the weights. Such matching can maximize the position similarity of the matched clusters. If a cluster pair is included in the result of the optimal matching problem, it will be considered as the same object's trajectory. After that, we choose the trajectory with the largest variance as the human's trajectory.

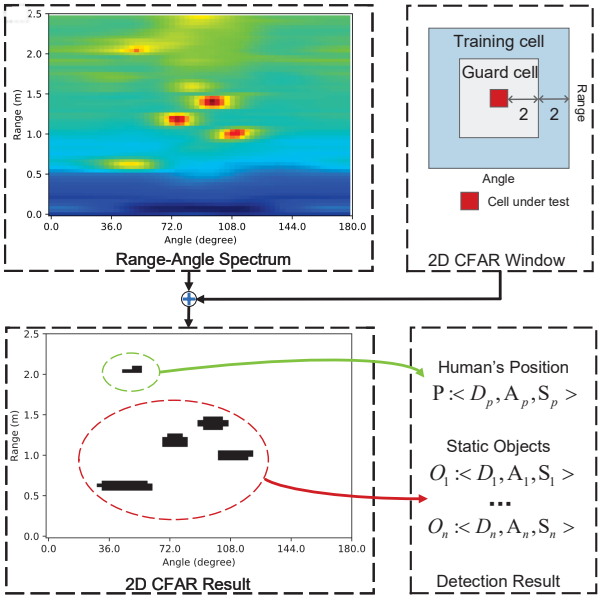


Fig. 9. The process of surrounding detection

After tracking the human's position, AmbiEar tries to find the static objects in a specific range (e.g., 1m) around the human's position. These objects are selected as the reflectors for further signal extraction. Considering the limited movement speed of the human, we set the update frequency of surrounding detection to 0.5s. In this way, AmbiEar can handle the dynamics of the human body well.

5.3 Common Component Extraction

In this part, AmbiEar first extracts the reflected signals from these selected reflectors. Then the common components of these signals are extracted to resist the low SNR.

After selecting the reflectors, AmbiEar applies the receiver beamforming algorithm to extract reflected signals from these static objects. Since the human's voice is concentrated in a certain frequency spectrum, the reflected signal of the static objects will have the same spectral distribution characteristic as the voice. Therefore, we traverse all the bins covered by each static object to find the bin with the highest energy ratio in the certain spectrum and obtain its reflected signal. If the energy ratio is less than our empirical threshold, we consider that the static object has too little vibration caused by voice or it has other vibrations. In this way, we ignore the static object and its reflected signal.

However, these reflected signals cannot be directly used to measure the reflectors' vibrations. For example, the amplitudes of the vibrations in our preliminary study are about $2 \sim 5$um, and the corresponding phase changes in the reflected signals are about $0.003 \sim 0.008$ rad. Such small phase changes caused by the weak vibrations are easily submerged in electromagnetic noise. If the reflection signals are observed in the IQ domain, the arcs formed by these reflected signals are easily converted into clumps with the impact of electromagnetic noise, as shown in Fig. 10. If we directly extract the vibration signal, the submerged arcs cannot be distinguished, and the calculated phase changes will be totally wrong. On the other hand, since we wish our system to work in noisy environments, the reflectors will also be affected by environment acoustic noise. If we directly extract the

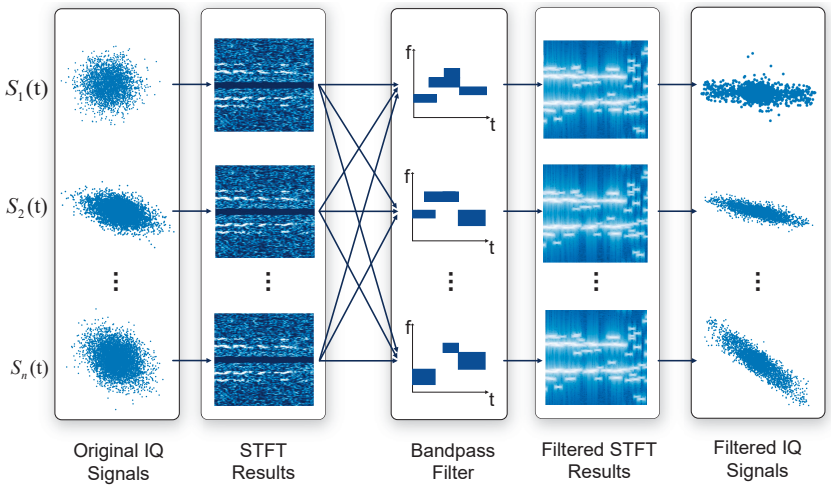


Fig. 10. After filtering out the common components, the IQ signal images change from clumps to arcs.

vibration signal, the vibration signal will contain the vibration components caused by the environment acoustic noise.

To solve this problem, AmbiEar calculates common frequency components of multiple reflected signals, which can be used to improve the SNR of the reflected signals. Our preliminary study shows that the range where sound can excite the surrounding objects' vibrations is limited, so the environment acoustic noise usually only affects parts of the reflectors. Considering that all of the reflectors' vibrations contain the information of the voice, the common components can help us resist the impact of environment acoustic noise. At the same time, this information can guide us to filter out the impact of electromagnetic noise.

Specifically, We design our common component extraction algorithm based on the MVDR algorithm[2], which is used in receiver beamforming tasks and used to analyze whether the signals from different directions are correlated. Its idea and structure are naturally consistent with our requirement of common component extraction. We modify it to analyze whether the signals are correlated in different frequencies.

In the original MVDR algorithm, there is a known unitary matrix $\mathbf{U}$. Each column vector of $\mathbf{U}$ represents a spatial direction. When $\mathbf{U}$ is replaced by the Fourier matrix $\mathbf{F} = \begin{bmatrix}\mathbf{f}_0 & \mathbf{f}_1 & \dots & \mathbf{f}_{K-1}\end{bmatrix}$, the function of the MVDR and the generalized MVDR spectrum in the space domain is replaced by that in the frequency domain. Similar to the original calculation process, we calculate the cross spectrum of two signals $x_i(n)$ and $x_j(n)$ in the k-th frequency bin:

$$S_{i,j}(\mathbf{f}_k) = \mathbf{g}_{i,k}^H \mathbf{R}_{i,j} \mathbf{g}_{j,k} \tag{6}$$

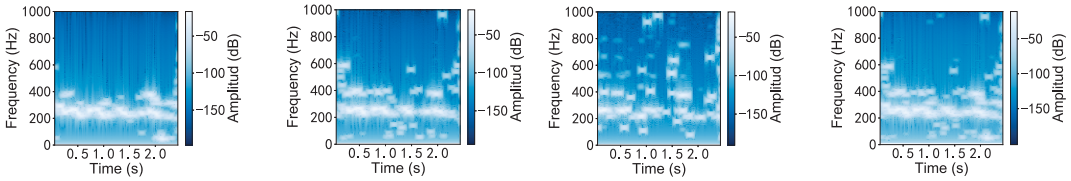
where

$$\mathbf{R}_{i,j} = \mathbf{x}_i(n)\mathbf{x}_j^H(n), \quad \mathbf{g}_{i,k} = \frac{\mathbf{R}_{i,i}^{-1}\mathbf{f}_k}{\mathbf{f}_k^H \mathbf{R}_{i,i}^{-1}\mathbf{f}_k}, \quad i,j \in \{1,2\} \tag{7}$$

and $\mathbf{x}^H$ means the conjugate transpose of $\mathbf{x}$. The correlation of the two signals $x_1(n)$ and $x_2(n)$ in the k-th frequency bin is

$$\gamma_{1,2}^2(\mathbf{f}_k) = \frac{|S_{1,2}(\mathbf{f}_k)|^2}{S_{1,1}(\mathbf{f}_k) S_{2,2}(\mathbf{f}_k)} \tag{8}$$

AmbiEar selects a similar time window length with some voice recognition tasks (e.g., 20ms) and applies the MVDR algorithm to each pair of the reflected signals in every time window. After that, the common components can be determined by comparing each pair's correlation with an empirical threshold. Then AmbiEar applies



(a) The first initial reflection vibration signal (b) The second initial reflection vibration signal (c) The third initial reflection vibration signal (d) The superimposed vibration signal

Fig. 11. The initial vibration signals from different objects and their superimposed vibration signal

fine-grained bandpass filters to all the reflected signals to preserve signals at these correlated frequency bands. As shown in Fig. 10, after filtering the reflected signals, the images of the IQ signals change from clumps to arcs. Such result indicates that our method can effectively filter out the impact of electromagnetic noise.

5.4 Signal Superimposition

After the above steps, the vibration of each reflector can be obtained as the electromagnetic noise has been filtered. To further resist the impact of environment acoustic noise, AmbiEar superimposes them to obtain the superimposed time-frequency spectrograms.

AmbiEar first extracts the vibration signals from the filtered reflected signals. As explained in §3.1, the IQ samples of the reflected signals form an arc-shaped trajectory. The phase changes of the sample points on the arc represent the displacement of the vibration signal. To further resist the low SNR, AmbiEar performs a circle fitting algorithm on these IQ sampling points. Let $\mathbf{P} = \{p_n\}, p_n \in \mathbb{R}^2$ denotes the IQ sampling points. The circle fitting problem can be characterized as calculating a circle center c and a radius r to minimize the sum of the distances between the sampling points and the circle:

$$c^*, r^* = \arg\min_{c,r} \sum_{p_n \in \mathbf{P}} (\|p_n - c\| - r)^2 \tag{9}$$

There are several algorithms to solve such a nonlinear least-squares optimization problem. Here we use the classic Levenberg-Marquardt algorithm [10] for its versatility and effectiveness.

Once the circle center and the radius are determined, the phases of the sampling points on the circle ϕ_n can be determined and can be converted to the vibration signals $x(n)$ according to Eq. 3.

In order to further resist the impact of environment acoustic noise, AmbiEar superimposes these vibration signals into one enhanced signal, as shown in Fig. 11. Considering that different types of objects have different vibration amplitudes when excited by the same sound, we use the nearest vibration signal as a benchmark to calculate the amplification factors of other signals. Those amplified signals that are most similar to the benchmark and the corresponding factors are selected:

$$\lambda_i = \arg\max_{\lambda} xcorr(v_1, \lambda * v_i), 2 \le i \le N \tag{10}$$

where v_1 represents the nearest vibration signal, v_i and λ_i represent the i-th vibration signal and its amplified factor, N is the number of reflected signals and $xcorr$ is the cross-correlation operation. The superimposed signal v_s can be represented as:

$$v_s = v_1 + \sum_{2 \le i \le N} \lambda_i * v_i \tag{11}$$

After obtaining the superimposed signal, AmbiEar performs normalization and STFT on it to obtain the time-frequency spectrogram. In this way, we can get the enhanced signal as input for further voice recognition.

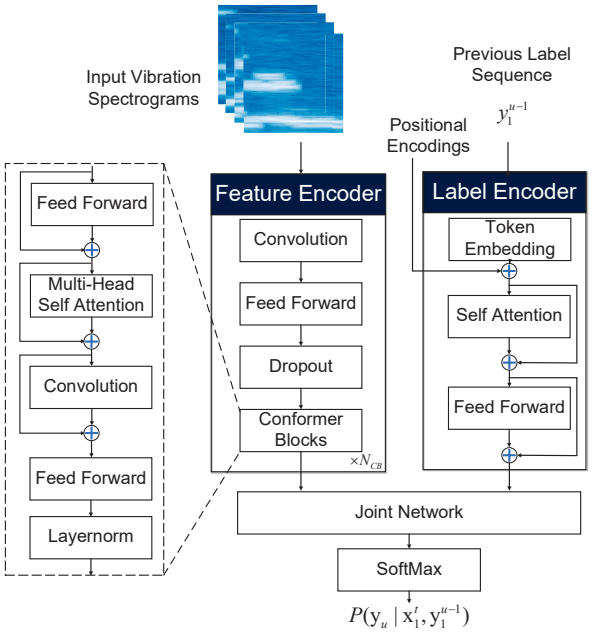


Fig. 12. The architecture of the voice recognition network

5.5 Voice Recognition

In the domain of Automatic Speech Recognition (ASR), a large body of works have demonstrated that it is possible to recover speech information from time-frequency spectrograms [43][29][7]. Through the previous steps, AmbiEar obtains the time-frequency spectrograms of enhanced vibration signals caused by the user's voice. The final task of AmbiEar is to identify the semantic information contained in such vibration signals.

Due to the sound attenuation in the air channel and the energy loss in the transformation process from sound to vibration, the vibration signal has obvious spectrum distortion compared with the voice signal. This signal distortion makes the commonly used voice information feature extraction techniques cannot be used directly, such as MFCC[25], wavelet transform[40], etc. Therefore, an end-to-end network is designed to extract voice-related features and decode them into semantic information.

The misalignment of the vibration signal and the semantic information should be considered. First of all, there is no way to match speech fragments with its corresponding text. Secondly, the signal duration for a specific character is variable. For the same person, different characters may have different pronouncing durations. For different people, the same character may also have different durations due to the variable speaking speed. In this paper, we leverage an RNN-T (Recurrent Neural Network Transducer) framework[11] for voice recognition, due to the reason that RNN-T can solve the misalignment of the vibration signal and the semantic information. The model is small and has high accuracy. Moreover, it does not require contextual information so it could support streaming ASR well.

As the vibration signal has obvious spectral distortion compared with the original voice, the features of the recovered voice are concentrated in the low-frequency part while the high-frequency part is seriously distorted. Therefore, we choose the 0-1500Hz part of the time-frequency spectrum as the network's input. The time-frequency spectrum of each voice is rearranged into a matrix of size $300 * l$, where the height of the matrix represents the analysis frequency band and l represents the duration of the voice.

As Fig. 12 depicts, the proposed recognition framework includes three modules, a feature encoder network, a label encoder network, and a joint network. The former is capable of extracting hidden features from the input sequence, and the label encoder computes the corresponding predictive coding. To preserve historical prediction output, the outputs of two encoders are added linearly by the joint network to compute the probability distribution over the sentence piece vocabulary. Next, we will introduce the design of the three modules in detail.

5.5.1 Feature Encoder. We extract two kinds of feature information from the distorted spectrum, namely the global interaction and the local correlations. The former implies the context of the voice, and the latter represents a certain character. The feature encoder first processes the input with a convolution subsampling layer, which shortens the time sequence length and fuses context information. Then a number of conformer blocks are applied to learn the global interaction and efficiently capture the local correlations.

The conformer block combines self-attention and convolution operation to separately learn the global interaction and capture the local correlations. It contains two Feed Forward (FFN) modules, a Multi-Headed Self-Attention (MHSA) module, and a Convolution module. The FFN module is designed to achieve feature conversion and enhance the model's representation ability. The MHSA modules integrate the relative sinusoidal positional encoding scheme to capture the internal structure and representations of the sentence with the semantics and dependencies at different positions, which can resist the impact of variable input lengths. The Convolution module identically encodes the context at each position into higher-level representations.

5.5.2 Label Encoder and Joint Network. Considering that the voice information is dependent on the context, we bring in a label encoder module to embed the previous outputs. First, an embedding layer converts the previously predicted non-blank labels into vector representations. Then several linear layers project the embedding vectors followed by a self-attention layer. Meanwhile, to only access the past states and ensure causality, a mask operation is added to the attention scores.

For the joint network, we only use a fully-connected feed-forward neural network with a single hidden layer and tanh as the activation function for simplicity and efficiency. The outputs of the multi-channel audio encoder and label encoder are concatenated as the inputs of the joint network.

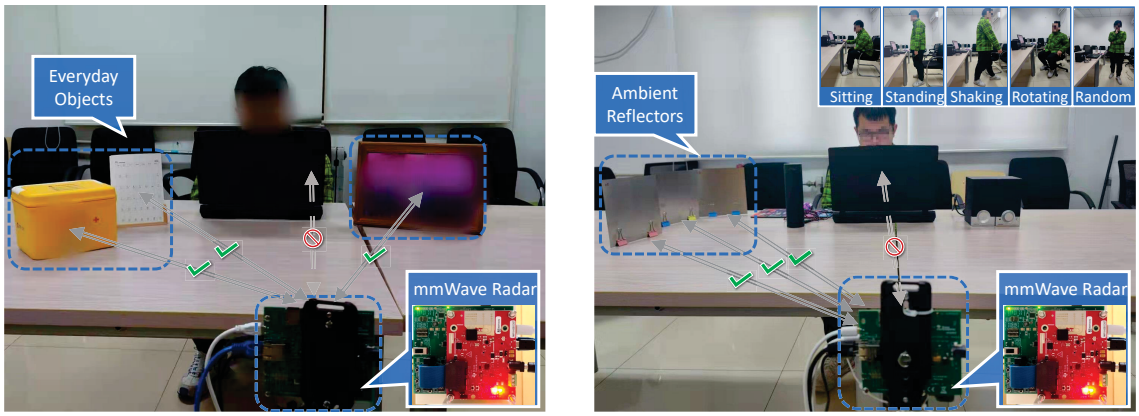
In our implementation, the number of conformer blocks N_{CR} is 4, the embedding dimension is set to 512, and the size of the hidden state in the feed-forward sub-layer is 1024. We train the network with the Adam optimizer [5] with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ and an adaptive learning rate schedule. For regularization, we apply dropout [38] in each residual unit of the conformer, i.e., to the output of each module, before it is added to the module input. We use a rate of $P_{drop} = 0.1$. Variational noise [18] is introduced to the model as a regularization. A l_2 regularization with $1e-6$ weight is added to all the trainable weights in the network.

6 IMPLEMENTATION AND EVALUATION

In this section, we introduce the implementation of AmbiEar and evaluate the performance of our prototype system under different settings.

6.1 System Implementation

We implement AmbiEar based on a commercial mmWave radar Texas Instruments IWR1642 Booster Pack [15]. We use this 2D Radar because AmbiEar only needs the vibration spectrum information of the reflectors surrounding the human body, which has little relationship with the 3D spatial position. There are 2 TX antennas and 4 RX antennas on the radar board. In our implementation, we let one TX transmit FMCW signals starting at 77GHz with 4.0GHz bandwidth, and all RXs receive the reflected signals. The *Ramp End Time* and the *Idle Time* are set to 80us and 20us, respectively. So the duration of a single chirp is 100us. Each frame includes 200 chirps and the frame period is set to 20.1ms due to the extra preparation time for each frame. In this configuration, the chirp sample



(a) (b)

Fig. 13. The experiment scenario

rate can be calculated as $\frac{1s*200}{20.1ms} \approx 9950Hz$. The frequency slope of the FMCW signal is 49.97MHz/us, and the ADC sample rate of the radar is 3430kHz, so the radar's maximum detection range is $\frac{3\times10^8 m/s\times3430kHz}{49.97MHz/us\times2} = 10.30m$. Considering the angle of the radar's FoV is about 120° [15], such FoV is enough to cover our experimental scene. We can further extend the FoV by modifying the radar's configuration or adding more radars. The raw data from the radar are captured by a TI DCA1000EVM data acquisition board [16] which can guarantee high speed and real-time transmission.

To ensure the generality of AmbiEar, we use a public voice data set, TensorFlow Speech Recognition Challenge Data Set (**TSRC**)[3], to generate our training set. **TSRC** includes 65,000 one-second long utterances of 30 short words by thousands of people. We use a commercial speaker to play twenty people's voices at a volume of 75dB. We use random 70% of the vibration signals of the reflectors surrounding the speaker as the training set, and the rest 30% as validation set. Six volunteers of different gender and age are instructed to repeat the short words in **TSRC** at the same volume. The corresponding vibration signals of the reflectors are extracted to generate the testing set. The training set includes voice-induced vibration signals in different experimental settings, including distances between people and radar, distances between people and reflector, reflector thicknesses and environmental noise intensities. We first collect the test data and verify our system performance under the same experimental settings as the training set. Then we further validate the robustness of our system in experimental settings with different environmental noise types, body orientations and body motions. In this way, we can avoid the excessive dependence of the network on environmental settings and increase the universality of our system. We collect more than 4000 seconds signals under different settings. All the experiments are IRB-approved, and all data are anonymized.

The experiment scenario is shown in Fig.13: The radar is placed on one side of the desk, and the volunteer/speaker is on the other side. When the volunteer moves or sits, the LoS path between the volunteer and the radar may not exist. To explore the impact of different factors on AmbiEar's performance, We first use the iron plates mentioned before to conduct some experiments. To further verify the generality of AmbiEar, we conduct some experiments in everyday scenes, placing some everyday objects (medicine box, calendar, photo album) around the volunteer as reflectors.

Scene	Method	WER		
		LoS	NLoS	Agg.
Meeting Room	AmbiEar	15.01%	16.19%	15.60%
	Direct Sensing	5.44%	95.92%	50.68%
Dormitory	AmbiEar	15.19%	16.58%	15.88%
	Direct Sensing	5.44%	96.37%	50.91%

Fig. 14. The comparison of AmbiEar and direct sensing approach

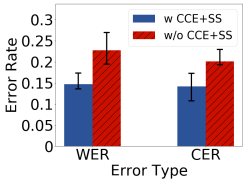


Fig. 15. Ablation study on signal enhancement

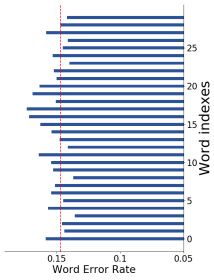


Fig. 16. Word error rates of AmbiEar

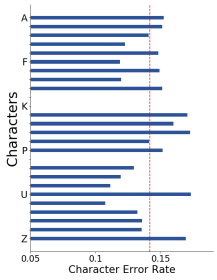


Fig. 17. Character error rates of AmbiEar

6.2 Methodology

We measure the voice recognition accuracy in terms of word and character with two standard metrics: word error rate and character error rate.

6.2.1 Word Error Rate (WER). WER is a common metric to evaluate the performance of a voice recognition system. It calculates the word error by comparing the output words with the reference as follows:

$$WER = \frac{S_w + D_w + I_w}{N_w} \tag{12}$$

where N_w is the number of words in the reference. S_w, D_w, and I_w represent the number of substitutions, the number of deletions, and the number of insertions, respectively. Lower WER indicates higher voice recognition accuracy. Word accuracy can be expressed by $1 - WER$.

6.2.2 Character Error Rate (CER). CER is a common metric to evaluate the performance of a voice recognition system from the perspectives of characters. Its calculation process is similar to WER. The minimum number of operations can be calculated as:

$$CER = \frac{S_c + D_c + I_c}{N_c} \tag{13}$$

where N_c is the number of characters in the reference. S_c, D_c, and I_c represent the number of substitutions, deletions, and insertions, respectively. Lower CER indicates better voice recognition performance. Character accuracy can be expressed by $1 - CER$.

We evaluate the performance of AmbiEar on WER and CER under multiple factors, including (1) distance between people and reflector, (2) distance between radar and reflector, (3) reflector's thickness, (4) environment noise intensity, (5) environment noise types, (6) body orientation, and (7) body motion. The impact of each factor is independently evaluated in the following experiments. In the baseline experiment, the distance between people and reflector is 30cm, the distance between radar and reflector is 1.0m, the reflector's thickness is 0.1mm, the environment noise intensity is 35dB, and the people sit still facing the reflector.

6.3 Overall Performance

We first compare our design with the direct sensing approach. Then we evaluate the overall performance of AmbiEar with WER and CER.

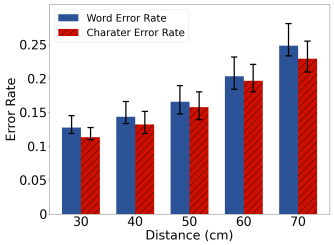


Fig. 18. The impact of distance between people and reflector

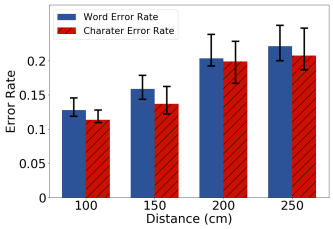


Fig. 19. The impact of distance between radar and reflector

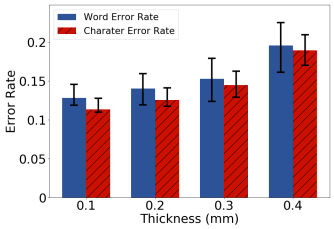


Fig. 20. The impact of reflector thickness

6.3.1 Comparison. We compare our design with existing direct sensing methods. To enrich our experiment scenarios, a volunteer is instructed to repeat the words in TSRC at 16 different positions with two different orientations in a meeting room and a dormitory. The volunteer sits at ten positions and stands at the remaining positions. The volunteer's orientations include one facing the radar (LoS scenario) and one facing away from the radar (NLoS scenarios). The signals from reflectors are extracted as the AmbiEar's data set and are processed by our network to obtain the recognition results.

On the other hand, we have mentioned before that the vocal cord's vibration signals are the input of the direct sensing methods, while AmbiEar utilizes the voice signal as the sensing target. Due to the different sensing targets, we cannot directly conduct comparative experiments. In this way, we refer to the voice recognition results of WaveEar [52] as the direct sensing reference in LoS scenarios. In NLoS scenarios, since there is no relevant experimental result to refer to, we trained a neural network with the same structure as that in §5.5. We use random 70% of the reflected signals from the human body as the training set, and the rest 30% as validation set. As the reflected signals from the human body has little relationship with the human's voice, the recognition result of the trained network is similar to a random guess.

The WERs of the two methods are shown in Fig. 14. We observe that the direct sensing approach has high accuracy in LoS scenarios. In this case, AmbiEar can be an important supplement to the direct sensing approach. We believe that more accurate recognition results can be obtained by fusing the direct sensing results and the indirect sensing results, which are obtained by sensing the vocal cord's vibration signals and the voice signals, respectively. In NLoS scenarios, the performance of direct sensing is severely degraded, while AmbiEar still maintains a stable and high accuracy. AmbiEar achieves an aggregated WER of 15.6% and 15.88%, which is 3.21× and 3.20× lower than that of the direct sensing approach. This result shows that AmbiEar can be used as an effective supplement to the direct sensing approach and achieve accuracy recognition in NLoS scenarios.

6.3.2 Word Error and Character Error. We evaluate the performance of AmbiEar under multiple factors and analyze its overall performance. The error rates of the words involved in the **TSRC** are shown in Fig. 16. The average word error rate is 14.71%. AmbiEar can also recognize other words as our network is designed for voice recognition rather than classification. The character error rates of AmbiEar are shown in Fig. 17. The average character error rate is 14.16%. Since our experimental voice does not contain "J", "K", and "Q", their error rates are missed.

These word error rates and character error rates are different due to their unique pronunciation. Moreover, some characters are not pronounced in certain words, such as the "g" in "eight". The words with such silent characters have higher error rates. Some words containing repeated characters also have higher error rates because repeated characters are recognized only once, such as "tree" containing "ee". We can utilize thesaurus

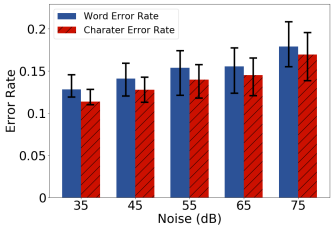


Fig. 21. The impact of environmental noise intensity

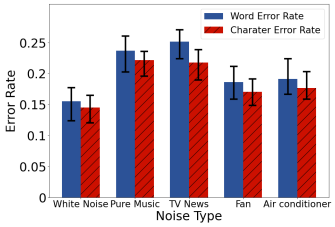


Fig. 22. The impact of environmental noise types

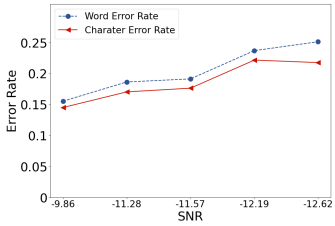


Fig. 23. The relationship between SNR and voice recognition error

and context information to further improve voice recognition accuracy, such as adding a Language Model in our network. However, this is beyond the scope of our work.

6.4 The Impact of Different Factors

6.4.1 The Impact of Distance between People and Reflector. In this experiment, we assess the impact of the distance between the people and the reflector. The distance between the volunteer and the nearest reflector ranges between 30cm and 70cm. Fig. 18 shows the error rates under different distances. With the volunteer moving away from the reflector, the word error rate raises from 12.79% to 24.89%, and the character error rate raises from 11.34% to 22.96%. The reason is that the intensity of the voice signal reaching the reflector is inversely related to the distance between people and reflector. As the voice propagation distance increases, the intensity of the voice signal decreases and the vibrations are weaker. It is more challenging to recognize the voice with weaker vibrations since it has lower SNR. Even so, AmbiEar can maintain a word recognition rate of more than 80% within 60cm with a customized network.

6.4.2 The Impact of Distance between Radar and Reflector. In this experiment, we assess the impact of the distances between the radar and the reflector. The distances between the radar and the nearest reflector range between 1.0m and 2.5m. Fig. 19 shows the error rates under different distances. With the radar moving away, the word error rate raises from 12.79% to 22.13%, and the character error rate raises from 11.34% to 20.77%. The reason is that the SNR of the mmWave reflected signal is related to the distance between radar and reflector. With the increase of the propagation distance of the mmWave signal, the intensity of the reflected signal is attenuated more seriously, and the phase distortion caused by the interference of electromagnetic noise is more obvious. The vibrations of reflectors are more difficult to extract with low-SNR reflected signals. However, with our core signal processing (CCE+SS), AmbiEar can still achieve an accuracy of nearly 80% at a distance of 2.5m.

6.4.3 The Impact of Reflector's Thickness. In this experiment, we assess the impact of the thickness of the reflector. The thickness of the reflectors ranges between 0.1mm and 0.4mm. Fig. 20 shows the error rates under different thicknesses. With the thickness of the reflectors increasing, the word error rate raises from 12.79% to 19.53%, and the character error rate raises from 11.34% to 18.92%. The reason is that when the intensity of the incident voice is the same, the thickness of the reflector affects the amplitude of its vibration induced by the voice. The thicker the reflector, the smaller vibration that can be induced. Since the phase difference of the reflected signal is proportional to the vibration amplitude, smaller phase difference is more likely to be submerged in the electromagnetic noises, causing the recognition errors to rise. However, with the help of our customized network, AmbiEar can achieve an accuracy of more than 80% in the case of 0.4mm thickness.

6.4.4 The Impact of Environmental Noise Intensity. In this experiment, we assess the impact of environmental noise intensity. We use a commercial speaker to play white noise around the reflectors.The speaker is about 1m

away from the nearest reflector and about 1.8m from the farthest reflector. We use a decibel meter to control the noise intensity at the nearest reflector. The noise intensity ranges between 35dB-75dB SPL (measured at 1m). In this setting, the white noise mainly affects a part of the reflectors at close range and has little impact on other reflectors. Fig. 21 shows the error rates under different noise intensities. With the noise level increasing, the word error rate rises from 12.79% to 17.87%, and the character error rate raises from 11.34% to 16.92%. As the environmental noise increases, the SNR of the voice around the reflector decreases. This results in more differences between the time-frequency spectrum of the vibration signal extracted from the reflector and the time-frequency spectrum of the original voice, which leads to an increase in recognition errors. As our core signal processing (CCE+SS) can resist the envrionment noise well, AmbiEar can achieve an accuracy of more than 80% even in the 75dB noise environment.

6.4.5 The Impact of Environmental Noise Types. In this experiment, we assess the impact of environmental noise types. We use a commercial speaker to play different types of sounds as the environmental noise, including white noise, TV news and pure music. The speaker's position is the same as in the previous experiment and the sound intensity at the nearest reflector is controlled at about 65db. We also use a large fan and an air conditioner to generate the environmental noise. Limited by their power, their noise levels at the nearest reflector are 65db and 50db, respectively. Fig. 22 shows the error rates under different noise types. The error rate under the interference of white noise is the lowest, and the error rate under the interference of TV news and pure music is relatively high. This result shows that our system can resist white noise and the noise of home appliances well, but is less resistant to music and TV news. The reason is that the frequencies of music and TV news are more similar to those of the human voice, and they can reduce the voice's SNR more seriously.

We further analyze of the relationship between the SNR of the received signal and the voice recognition error. We calculate the average SNR of the received signal under different noise types and the corresponding error rates. To obtain the SNR of the received signal, we calculate the background noise by subtracting the enhanced signal from the noisy signal, which is calculated directly using the phase change of the reflected radar signal. The corresponding relationship between the SNR and the voice recognition error is shown in Fig. 23. The results show that as the SNR decreases, the recognition error rate tends to increase. The reason is that as the SNR of the received signal decreases, the signal becomes more distorted and harder to recognize.

6.4.6 The Impact of Body Orientation. In this experiment, we assess the impact of body orientation. We record the volunteer facing the reflector as 0° and the volunteer's back to the reflector as 180°. Fig. 24 shows the error rates when the volunteers speak in different orientations. As the body orientation angle becomes larger, the recognition errors increase slightly. The reason is that when the angle between the human body and the reflector increases, the voice propagation path changes, resulting in a slight decrease in the quality of the voice around the reflector. As mentioned before, a lower SNR of the voice means higher recognition errors. Even so, we observe that AmbiEar can achieve high recognition accuracy regardless of the orientation. This result benefits from the fact that AmbiEar extracts voice signals from surrounding reflectors rather than directly sensing the vocal cord.

6.4.7 The Impact of Body Motion. In this experiment, we assess the impact of body motion. The volunteers are instructed to speak while making different motions. The body motions include sitting still, standing up and sitting down, shaking the body back and forth, rotating, and moving freely. Fig. 25 shows the error rates when the volunteers make different motions. It can be observed that the recognition errors differ slightly when the volunteers make different motions. This is also because the voice propagation path changes during the motions, resulting in a change in the quality of the voice around the reflector. However, benefiting from our indirect sensing method, AmbiEar can achieve high accuracy regardless of the volunteers' motion. Such a result shows that AmbiEar is robust to body motion. Based on the above two experiments, we believe that AmbiEar can achieve high accuracy in mobile scenes.

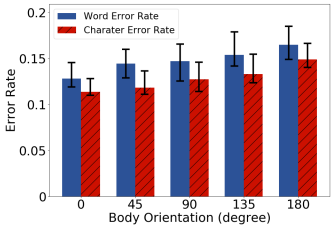


Fig. 24. The impact of body orientation

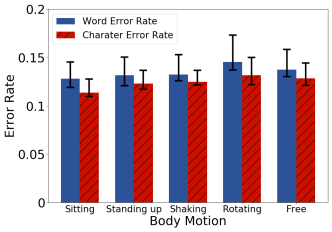


Fig. 25. The impact of body motion

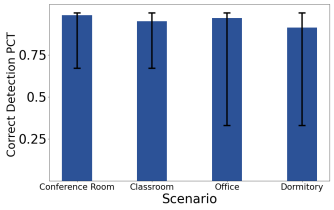


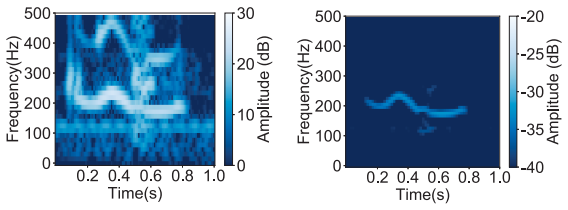
Fig. 26. Ablation study on surrounding detection

6.4.8 The Combined Impact of These Factors. Our experimental results have verified the impact of different factors on our system performance. Indeed, these factors affect the performance of our system simultaneously. Firstly, the body orientation and body motion have less impact on our system performance than other factors. The reason is that they slightly affect the SNR of the voice around the reflectors and AmbiEar can compensate for them by analyzing multiple surrounding reflectors. Secondly, the dominant factors are different in different scenarios. In practical scenarios, when there are many reflectors around the human, such as in the office or the living room, the distance between human and radar and the environmental noise become the main factors affecting recognition errors, because the complex scene limits the radar's placement and other sound sources cause stronger interference to our system. These two factors affect the extracted vibration signal's quality and the SNR of the voice around the reflector, respectively. When there are few reflectors around the people, such as in a conference room, the distance between human and reflector and the reflector's thickness become the main factors affecting recognition errors, because the simple scene may result in a lack of reflectors around the human or the surrounding reflectors to be difficult to vibrate. Both of these factors affect the extracted vibration signal's quality. In general, these factors combine to affect our system performance by affecting the SNR of the voice around the reflectors and the extracted vibration signal's quality. The dominant factors are different in different scenarios while the body factors have less impact on our system.

To mitigate the combined effect, we can choose to avoid the users being too far from the radar or in an environment with few surrounding reflectors. We also recommend that the users keep a certain distance from other sound sources to avoid the effect of noise. To further mitigate the impact of factors on the SNR of the voice, some denoising techniques such as spectral subtraction [42] can be used. By removing the impact of the environment acoustic noise from the voice's time-frequency spectrum, the SNR of the voice signal can be improved. Moreover, in order to improve the quality of the extracted vibration signal, we can explore some signal enhancement methods, such as Tx beamforming, which can concentrate the energy of the transmitted signal near the reflector. In this way, we can obtain an enhanced reflected signal to resist the impact of electromagnetic noises.

6.5 Ablation Study

6.5.1 Surrounding Detection. This section evaluates the performance of the surrounding detection algorithm. We place three iron plates as our target reflectors around people in different scenarios, including a conference room, a classroom, an office and a dormitory. The scene of the conference room is the simplest and the scene of the dormitory is the most complicated. We calculate the correct detection number by comparing the detection results and the actual objects' locations. The evaluation is repeated 100 times in each scene. The average correct detection percentage in these scenarios is 98.67%, 95%, 97% and 91.33%, respectively. The experimental results show that our algorithm can achieve a detection accuracy of more than 90% even in complicated scenarios.



(a) The original voice (b) The recovered voice

Fig. 27. The comparison between the original voice pronounced by a volunteer and the corresponding recovered voice

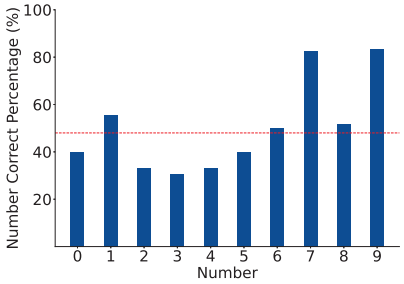


Fig. 28. The average transcription percentage of each number

When the scenario is complex, the reflection points of multiple objects are too close that the DBSCAN algorithm mistakenly treats them as one object, resulting in a decrease in detection performance.

6.5.2 Common Component Extraction and Signal Superimposition. This section evaluates the importance of our core design: common component extraction and signal superimposition (CCE+SS). We generate two different test sets. The first test set consists of the time-frequency spectrograms of the vibration signals extracted via the full version of our design. The second test set consists of the time-frequency spectrograms of the vibration signals of the nearest reflector around the volunteer, which is extracted via the algorithm in §3.1. The error rates of these two test sets are shown in Fig. 15. The word error rate and the character error rate in the first test set are 14.71% and 14.16%, respectively. The word error rate and the character error rate in the second test set are 22.73% and 20.12%, respectively. The results show that our core design, CCE+SS, plays a significant role in improving the signal's SNR and the recognition accuracy.

6.5.3 Voice Recognition. This section evaluates the performance of our network architecture for voice recognition. We compare it with a structure of SENet-CTC that the SENet refers to the network model of [14]. The test sets and the validation sets of these methods are set to be the same. The recognition results show that the average word error rates of our architecture and SENet are 14.71% and 39.48%, respectively, and the average character error rates of our architecture and SENet are 14.16% and 36.3%, respectively. We can find that our network can achieve a better recognition performance. The reason is that our network can effectively extract the global interactions and the local correlations in the time-frequency spectrum than SENet. Besides, Connectionist Temporal Classification (CTC) needs to match speech fragments with their corresponding text, which is very hard to resolve due to the variable speaking speed.

6.6 Case Study

To further verify the generality of AmbiEar, we conduct some experiments in everyday scenes. Some everyday objects (medicine box, calendar, photo album) are placed around the volunteer as reflectors, as shown in Fig. 13. the distance between people and reflector is 40cm and the distance between radar and reflector is 1.0m. The people sit still facing the reflector and the environment noise is 35dB. The average word error rate is 21.17% and the average character rate is 17.92%. Compared with the result of the iron plates under the same configuration, the average word error rate and the average character rate only increased by 5.29% and 3.17%, respectively. The results show that AmbiEar can work effectively in everyday scenes.

6.7 Validation

To further verify the correlation between the extracted time-frequency spectrum and the original voice signal, we conduct a validation experiment. We select five people's voices pronouncing the numbers from "zero" to "nine"

and their corresponding extracted time-frequency spectrums in the baseline experiment. These time-frequency spectrums are directly transformed to the time-domain signals by IFFT, called recovered voices. The original voice's time-frequency spectrum of the number "one" pronounced by a volunteer and the corresponding extracted time-frequency spectrum are shown in Fig. 27(a) and Fig. 27(b), respectively. As we mentioned before, due to the sound attenuation in the air and the energy loss in the transformation from sound to vibration, the vibration signal has obvious spectral distortion compared with the original voice. The features of the recovered voice are concentrated in the low-frequency part while the high-frequency part is seriously distorted. These features of the original voice with small amplitudes are also difficult to be recovered. Nevertheless, the results show that the time-frequency spectrum of the extracted signal is similar with the low-frequency part of the original voice's spectrum. We further invite fifteen volunteers to listen to the recovered voices and transcribe these numbers without hearing the original voices in advance. Each volunteer listens to ten random numbers and the recovered voice of each number is played five times in succession. Each group of numbers is randomly selected and some numbers may appear multiple times. The percentage of correctly transcribed numbers is shown in Fig. 28. We can find that the average percentage of correctly transcribed numbers is 48% and some numbers have an average transcription percentage above 80%. Since the recovered signals still have some spectral distortion, it is challenging to directly identify the recovered signals. However, we find that some numbers, such as "seven" and "nine", can be almost directly distinguished. We believe that such time-frequency spectrums input can help our recognition network perform well.

7 DISCUSSION

In this section, we discuss some practical problems and potential opportunities, including multi-target scenarios, insufficient objects, and multipath effects.

7.1 Multi-target Scenarios

When more than one humans talk together, the interference of their voices must be considered. As the sound can only excite the surrounding objects' vibrations in a limited range, AmbiEar can obtain the corresponding voices directly when the people are separated by a certain range. When two people are very close, AmbiEar can extract the mixed signal of the voices from the reflected signals. Then the classic voice separation algorithms can be applied to obtain the separated signals. In this way, it is possible that the voices from multiple people can be recognized simultaneously with a single radar. We leave this problem to be explored in our future work.

7.2 Insufficient Objects

Due to the occlusion of the human body or the change of the environment, the number of objects found in the surrounding detection module may be insufficient. In this case, The advantages of AmbiEar will be weakened, and the impact of environment acoustic noise will be more obvious. Considering that some objects are blocked by the human body in certain viewing angles, we can choose the optimal deployment position and angle when deploying our radar to solve this problem.

7.3 Multipath Effect

The multipath effect of mmWave signals is much weaker than other wireless signals (such as WiFi, LoRa, etc.) owing to its extremely high frequency. However, some "ghost images" may still appear in our scanning results. These "ghost images" are caused by the reflection of mmWave signals in propagation and can appear in the range-angle bins where there is no object, which affects the results of our surrounding detection module. To solve this problem, we can scan the environment in advance and only reserve the detection results corresponding to the actual objects.

7.4 System Limitation

Although our system performs well in appropriate usage scenarios, including suitable distances, enough reflectors, and tolerable environment acoustic noise, it suffers from some limitations in practice. First, in order to cause the reflectors' vibrations, we should ensure that there are some reflectors that are easy to vibrate and are close to the users. In general, at least three reflectors at different positions are required to extract the voice signal and resist the noise from other positions. Secondly, due to the rapid attenuation of mmWave signals, the distance between people and the radar cannot be too far, otherwise it will cause the reflected signal to be buried in the electromagnetic noise. Finally, the environment acoustic noise that AmbiEar can tolerate is limited. Some noises that are similar to voice in frequency domain can degrade the performance of our system, such as the human voice and TV news close to reflectors. However, our system can resist some environmental interference when the main frequency of the sound is higher than our analysis band (0-1500Hz) or the sound only affects a few reflectors, such as the high-frequency noise from small appliances and the human voice from distant locations.

7.5 Data Collection

Our training set contains voice-induced vibration signals collected in different experimental settings. There are some details of data collection that need to be further considered, such as the volume and type of speakers. The volume of the speaker affects the sound's SNR, while the speaker type affects the details of the sound's time-frequency spectrum. Taking these factors into account when collecting data could further enrich our training set and may improve our system's performance. Meanwhile, some spectral augmentation methods [30] can be applied to further improve the performance of our system.

8 CONCLUSION

In this paper, we present an indirect voice recognition system, namely AmbiEar, to liberate the restrictions on the human's position and posture and achieve accurate voice recognition in NLoS scenarios. We believe that AmbiEar is an important complement to existing direct voice sensing methods. We explore the limitation of extracting voice directly and the possibility of extracting voice indirectly by mmWave sensing. Then we propose our system including a series of modules from surrounding detection to voice recognition. Our system can extract voice-related vibration parts from the objects around the human body and use multiple reflection signals to achieve signal enhancement and voice recognition. Extensive experiments under real-world scenarios show that AmbiEar can effectively infer the voice in NLoS scenarios.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments. This work is supported by the grant No. U21B2007 of National Natural Science Fund of China and the grant No. 2021GQG1002 of the Guoqiang Institute, Tsinghua University.

REFERENCES

[1] Khamis A Al-Karawi, Ahmed H Al-Noori, Francis F Li, Tim Ritchings, et al. 2015. Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance. *International Journal of Information and Electronics Engineering* 5, 6 (2015), 423–427.

[2] Jacob Benesty, Jingdong Chen, and Yiteng Huang. 2005. A generalized MVDR spectrum. *IEEE Signal Processing Letters* 12, 12 (2005), 827–830.

[3] Google Brain. 2017. TensorFlow Speech Recognition Challenge. https://www.kaggle.com/c/tensorflow-speech-recognition-challenge.

[4] Baicheng Chen, Huining Li, Zhengxiong Li, Xingyu Chen, Chenhan Xu, and Wenyao Xu. 2020. ThermoWave: a new paradigm of wireless passive temperature monitoring via mmWave sensing. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[5] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*. PMLR, 933–941.

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

[7] Rohan Doshi, Youzheng Chen, Liyang Jiang, Xia Zhang, Fadi Biadsy, Bhuvana Ramabhadran, Fang Chu, Andrew Rosenberg, and Pedro J Moreno. 2021. Extending Parrotron: An End-to-End, Speech Conversion and Speech Recognition Model for Atypical Speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6988–6992.

[8] Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Févotte. 2010. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE transactions on audio, speech, and language processing* 18, 3 (2010), 564–575.

[9] M. Ester, H. P. Kriegel, Jrg Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *AAAI Press* (1996).

[10] Walter Gander, Gene H Golub, and Rolf Strebel. 1994. Least-squares fitting of circles and ellipses. *BIT Numerical Mathematics* 34, 4 (1994), 558–578.

[11] Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711* (2012).

[12] Junchen Guo, Meng Jin, Yuan He, Weiguo Wang, and Yunhao Liu. 2021. Dancing Waltz with Ghosts: Measuring Sub-mm-level 2D Rotor Orbit with a Single mmWave Radar. (2021).

[13] Unsoo Ha, Salah Assana, and Fadel Adib. 2020. Contactless seismocardiography via deep learning radars. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[15] Texas Instruments Incorporated. 2020. IWR1642: Single-chip 76-GHz to 81-GHz mmWave sensor integrating DSP and MCU. https://www.ti.com/product/IWR1642.

[16] Texas Instruments Incorporated. 2020. Real-time data-capture adapter for radar sensing evaluation module. http://www.ti.com/tool/DCA1000EVM.

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

[18] Kam-Chuen Jim, C Lee Giles, and Bill G Horne. 1996. An analysis of noise in recurrent neural networks: convergence and generalization. *IEEE Transactions on neural networks* 7, 6 (1996), 1424–1438.

[19] Abdelwahed Khamis, Branislav Kusy, Chun Tung Chou, Mary-Louise McLaws, and Wen Hu. 2020. RFWash: a weakly supervised tracking of hand hygiene technique. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 572–584.

[20] Michael Levandowsky and David Winter. 1971. Distance between sets. *Nature* 234, 5323 (1971), 34–35.

[21] 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*. 312–325.

[22] Zhengxiong Li, Baicheng Chen, Zhuolin Yang, Huining Li, Chenhan Xu, Xingyu Chen, Kun Wang, and Wenyao Xu. 2019. Ferrotag: A paper-based mmwave-scannable tagging infrastructure. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 324–337.

[23] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–19.

[24] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, Kui Ren, et al. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th Conference on Embedded Networked Sensor Systems*.

[25] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa Sheeber, and Nicholas Allen. 2009. Content based clinical depression detection in adolescents. In *2009 17th European Signal Processing Conference*. IEEE, 2362–2366.

[26] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083* (2010).

[27] James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics* 5, 1 (1957), 32–38.

[28] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. 2021. RadioMic: Sound Sensing via mmWave Signals. arXiv:2108.03164 [eess.SP]

[29] Ashutosh Pandey and DeLiang Wang. 2019. Exploring deep complex networks for complex spectrogram enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6885–6889.

[30] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[31] K Sreenivasa Rao and Sourjya Sarkar. 2014. *Robust speaker recognition in noisy environments*. Springer.

[32] Patrice Robisson, Thierry Aubin, and Jean-Claude Bremond. 1993. Individuality in the voice of the emperor penguin Aptenodytes forsteri: adaptation to a noisy environment. *Ethology* 94, 4 (1993), 279–290.
[33] Hermann Rohling. 1983. Radar CFAR thresholding in clutter and multiple target situations. *IEEE transactions on aerospace and electronic systems* 4 (1983), 608–621.
[34] Aaron E Rosenberg, Chin-Hui Lee, and Frank K Soong. 1994. Cepstral channel normalization techniques for HMM-based speaker verification. In *Third International Conference on Spoken Language Processing*.
[35] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.
[36] Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* 11, 5 (2007), 561–580.
[37] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with your robot vacuum cleaner: eavesdropping via lidar sensors. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 354–367.
[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
[39] Petre Stoica, Zhisong Wang, and Jian Li. 2002. Robust capon beamforming. In *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.*, Vol. 1. IEEE, 876–880.
[40] Z Tufekci and John N Gowdy. 2000. Feature extraction using discrete wavelet transform for speech recognition. In *Proceedings of the IEEE SoutheastCon 2000.'Preparing for The New Millennium'(Cat. No. 00CH37105)*. IEEE, 116–123.
[41] Diana Van Lancker, Jody Kreiman, and Karen Emmorey. 1985. Familiar voice recognition: patterns and parameters Part I: Recognition of backward voices. *Journal of phonetics* 13, 1 (1985), 19–38.
[42] Saeed V Vaseghi. 2008. *Advanced digital signal processing and noise reduction*. John Wiley & Sons.
[43] Daria Vazhenina and Konstantin Markov. 2020. End-to-end noisy speech recognition using Fourier and Hilbert spectrum features. *Electronics* 9, 7 (2020), 1157.
[44] 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.
[45] Teng Wei and Xinyu Zhang. 2015. mtrack: High-precision passive tracking using millimeter wave radios. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 117–129.
[46] Wikipedia. 2021. Alibaba Tmall Genie. https://en.wikipedia.org/wiki/Tmall_Genie.
[47] Wikipedia. 2021. Amazon Alexa. https://en.wikipedia.org/wiki/Amazon_Alexa.
[48] Wikipedia. 2021. Apple HomePod. https://en.wikipedia.org/wiki/HomePod.
[49] Wikipedia. 2021. Google Nest. https://en.wikipedia.org/wiki/Google_Nest_(smart_speakers).
[50] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. 2020. mmTrack: Passive multi-person localization using commodity millimeter wave radio. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2400–2409.
[51] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. 2020. mSense: Towards Mobile Material Sensing with a Single Millimeter-Wave Radio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–20.
[52] 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.
[53] 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.
[54] Feng Zhang, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2020. mmEye: Super-Resolution Millimeter Wave Imaging. *IEEE Internet of Things Journal* (2020).