2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON) | 979-8-3503-0052-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/SECON58729.2023.10287434

# RemoteGesture: Room-scale Acoustic Gesture Recognition for Multiple Users

Mi Tian[†] , Yannwen Wang[†] , Zheng Wang[†] , Junhua Situ[†]
Xiaoqi Sun[†], Xiaokang Shi[†] , Chenwei Zhang[†] , Jiaxing Shen[‡]
[†]Hunan University, China [‡]Lingnan University, Hong Kong, China
{tianmi, wangyw, wangzheng, situjunhua, sunxiaoqi, shixiaokang, zhangcw}@hnu.edu.cn
jiaxingshen@LN.edu.hk

***Abstract*—As a promising way, controlling smart devices through gestures offers the benefits of non-contact interaction, efficiency and convenience. Previous researches on acoustic-based gesture recognition have mostly focused on near-field gestures within $1$ meter and for a single user only. However, such a near-field sensing scheme is inadequate to meet the growing demands for multi-person human-computer interaction in far-field spaces. In this paper, we present a novel acoustic-based room-scale gesture recognition system that is capable of recognizing gestures simultaneously performed by multi-user. Our approach achieves far-field sensing by examining the relationship between acoustic signal frame length and sensing range, and overcoming a series of practical challenges incurred by far-field sensing. To simultaneously detect and distinguish gestures of multiple persons, we divide the sensing area into multiple beamforming sub-scanning areas and apply binary search to detect multiple users, which allows for an efficient scanning process and facilitates real-time detection. Finally, we conduct a data augmentation scheme to enlarge the training data and apply a lightweight deep learning framework to classify different gestures. Extensive experiments confirm that our system enables multi-user gesture detection and can recognize nine gestures at a distance up to $7$ meters.**

***Index Terms*—acoustic, gesture recognition, far-field sensing, channel impulse response, beamforming**

## I. Introduction

**Motivation.** As a natural, convenient and human-oriented manner, gestures have aroused significant attention in human-computer interaction (HCI) [2], [9], [19]. Recent years have witnessed a massive proliferation of smart devices and applications, facilitating gesture recognition for various usage scenarios, such as VR gaming, metaverse, remote surgery, teaching, smart manufacturing and etc. [4], [6], [7], [20]. Many practical usage scenarios entail far-field sensing of human gestures, usually in a room-scale level, and simultaneous multi-user gesture recognition. Yet most of existing works are designed to fulfill gesture recognition tasks in near-field scenarios ($<1m$) and support only a single user [4], [9], [22]–[24], [29], which restrains the potential usage scenarios in HCI. To bridge this gap, we aim to design a contact-free gesture recognition system that can achieve room-scale level and simultaneous multi-user gestures recognition.

**Prior works and limitations.** Existing works widely apply vision-based method to identify gestures [2], [6], [12] since cameras can capture the continuous motion of gestures by tracking finger joints in real-time. Powerful deep learning techniques are then exploited to accurately classify different gestures. However, vision-based method requires strict lighting conditions and prone to raise privacy concerns.

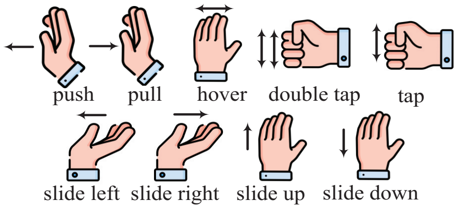


Fig. 1. Nine Gestures in RemoteGesture.

Recent works explore the potential of radio frequency (RF) signals to recognize human gestures based on the intuition that human hand when performing gestures can impact the signal propagation in terms of signal strength and phase, which can be used to infer the gestures [1], [20]. However, RF-based methods either suffer from inaccessibility to channel information for most commercial devices [1] or requiring high-cost specialized hardware (e.g., FMCW, mmWave radar and USRP). More importantly, the singal wavelengths in most mainstream RF techniques are too long ($60mm \sim 320mm$) to capture fine-grained finger movements, which significantly restricts a wide range of public use.

Ultrasonic-based method exploits commodity speakers and microphones to achieve fine-grained gesture recognition [3], [4], [9], [11], [15], [21], [24], [33]. The merits lie in threefold: (1) ultrasonic is able to offer sufficient resolution for capturing finger movement due to its relatively short wavelength; (2) acoustic devices have already been embedded in most of smart devices, which do not incur extra hardware cost; (3) independence of lighting conditions. As human hand will reflect sound waves, by transmitting a known sound wave as a probe, different types of gesture can be inferred using Channel Impulse Response (CIR) extracted from the received sound wave [9], [15], [24], [33].

Although those works enable complex finger movement tracking and achieve fine-grained gesture recognition, the sensing distance of those works are less than $1m$, which requires performing gestures nearby the acoustic devices. Such near-field gesture sensing techniques deliberately discard signal attenuation and interference beyond the scope of target sensing distance and, thus, obtain accurate gesture recognition. RTrack enables room-scale gesture recognition by firstly extracting the angle at which a gesture is performed using 2D MUSIC

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.

979-8-3503-0052-9/23/$31.00 ©2023 IEEE

algorithm and implements a room-scale tracking system [11]. Note that RTrack can effectively expand the sensing range to $4.5m$, yet is still hard to support simultaneous multi-user gesture detection since sound waves reflected from different hands are difficult to separate at the receiver.

**Challenges.** Achieving room-scale level acoustic gesture recognition while simultaneously detecting multi-user gestures is a daunting task. One fundamental challenge is how to expand the effective coverage area of acoustic signals to interpret gestures from meters away. As propagation distance increases, acoustic signal attenuates significantly. As such, subtle gesture pattern modulated by hand in the signal becomes extremely weak, which becomes even more unapparent considering the round trip between hand and acoustic devices. One potential solution is to apply high-end acoustic devices to increase the transmission power, but this may significantly impose cost overhead and reduce its popularity. Furthermore, detecting gestures at a long distance yields greater disturbances caused by the multipath effect, resulting in inconspicuous gesture patterns.

The second challenge that we face is detecting and distinguishing gestures from multiple individuals at long distances. Unlike near-field gesture recognition, simultaneously detecting user gestures is a complex yet essential functionality in far-field gesture recognition systems. Due to the multipath effect, sound waves reflected from different hands superimpose at the microphone, making it difficult to separate the different gesture patterns. One intuitive approach is to scan the entire space continuously and determine the angle at which different people perform their gestures. Beamforming technique can then enhance the signal at this specific angle and impair signals from other directions. However, such a space scanning scheme is time-consuming and does not meet the requirements for real-time gesture recognition.

**Solutions.** In this work, we propose RemoteGesture, an acoustic based gesture recognition system for multi-user recognition in far-field scenarios. RemoteGesture supports multi-user detection and can recognize nine different gestures, as shown in Fig. 1, significantly boosting the sensing distance of gestures to up to 7 meters away from the transceiver. Our approach applies the Zadoff-Chu (ZC) sequence as a channel probing signal to measure the Channel Impulse Response (CIR), which visualizes the impact of gestures on sound waves. The ZC sequence possesses a high auto-correlation property, which enables it to differentiate tiny differences of multipath signals by finding the peaks corresponding to their different delays. More importantly, cross-correlation-based CIR measurements facilitate subtle separation of multipath signals in terms of propagation delay ranges, which greatly complements the high auto-correlation induced probing signals.

In our study, we aim to enable far-field sensing by examining the quantitative relationship between the length of the transmitted sound frame and sensing range. However, we do not simply prolong the length of the emitted frame. Instead, we synthetically consider various practical issues that may affect the system's performance, and conduct a parameter selection scheme for designing the ZC sequence. Moreover, to minimize the scanning time of beamforming and support multi-user gesture detection, we adopt a simple yet effective method inspired by the binary search algorithm. Specifically, we divide the $180°$ area in front of the acoustic devices into multiple subareas and apply binary search to find the exact regions where gestures are performed.

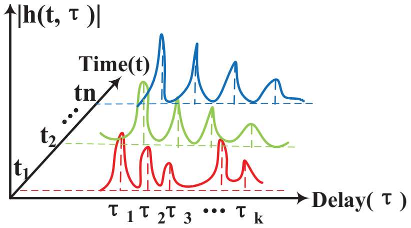


Fig. 2. Channel impulse response.

**Our contribution.** Our holistic system design guarantees that RemoteGesture achieves room-scale gesture recognition and enables multi-user gesture detection and separation. Extensive evaluation manifests that RemoteGesture can detect gestures 7 meters away from acoustic devices and yield a simultaneous detection of maximum 3 users with an accuracy up to $95.13\%$. We summarize our contribution as follows:

- We boost the acoustic sensing range by synthetically correlating the length of the ZC sequence and the sensing distance. Instead of simply using a longer transmission frame, we deeply analyze a bunch of practical issues that affect the acoustic attenuation and conduct a parameter selection scheme to design a proper transmission frame.
- We deliberately apply a simple yet effective approach originated from binary search algorithm to reduce the scanning time of beamforming and achieve a real-time gesture detection for multiple users.
- We design a prototype of RemoteGesture all consisted of cost-effective and commodity acoustic devices. Extensive experiments show that RemoteGesture achieves high gesture recognition accuracy in far-field scenarios and can simultaneously detect gestures from multiple users.

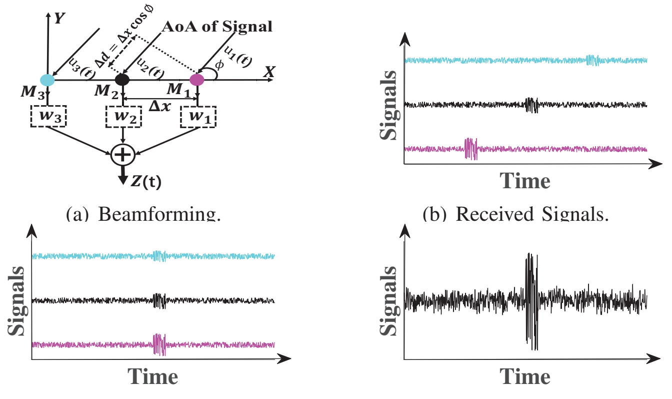
## II. Background

### A. Channel Impulse Response

When a signal propagates in the air, it travels along multiple paths due to the presence of various reflection entities in the environment. As a result, the receiver receives multiple copies of the signal with different delays and attenuation. CIR can separate the multipath signals into bins or taps based on their propagation delays, each of which corresponds to a distinct range of propagation distance, facilitating to pinpoint signals affected by specific target objects. By identifying appropriate bins exclusively affected by the interested targets, we can effectively track the movement of targets by looking into signal fluctuations in both delay and attenuation.

To measure Channel Impulse Response (CIR), a pre-defined frame is transmitted to probe the channel, which is then received by the receiver. For a complex baseband signal, CIR can be measured by calculating the cyclic cross-correlation

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.



(a) Beamforming.

(b) Received Signals.

(c) Received Signals in Delay.

(d) Sum of Delayed Signals.

Fig. 3. Beamforming

between the transmitted baseband signal and the received baseband signal:

$$h[\tau] = \frac{1}{N}\sum_{n=1}^{N} t[n] \cdot r^*[n-\tau] \tag{1}$$

where $h[\tau]$ is the measured CIR, $t[n]$ denotes the transmitted complex baseband signal, $r^*[n]$ is the conjugate version of $r[n]$ at the receiver, and $\tau$ is the multipath delays. $N$ is the length of $t[n]$. As shown in Fig. 2, $h[\tau]$ is a matrix that characterizes how signals with different delays $\tau$ attenuate along time $t$.

### B. Beamforming

In complex multipath channels, directing signal energy towards a specific direction of interest can minimize noise and interference, leading to the amplification of target signals. Beam scanning involves steering sensor beams over a predetermined angular range to detect signals from various directions [8], [17]. This technique is useful in situations where signals are arriving from multiple directions and need to be tracked. In specific, beamforming selectively enhances or suppresses signals that arrive at an array of sensors by controlling their phases and amplitudes [18]. To generate a beam pattern that amplifies signals in a particular direction while reducing noise and interference from other directions, a set of weights is applied to each sensor's signals. These weights can be optimized based on the signal-to-noise ratio (SNR) or a desired radiation pattern.

Beamforming techniques leverage the multipath effect and the geometric layout of the sensor array. The sensors receive successive copies of the transmitted signal with varying delays and strengths. As a result, the signal received in each receiver of the array is a copy of the transmitted signal while exhibiting diverse attenuation and phase shifts that may induce destructive interference of the signal. As shown in Fig. 3(a), assume that a microphone array is composed of three uniformly deployed microphones, denoted as M1, M2, and M3. When sound waves reach the array, each microphone receives identical copies of the signal with different delays (Fig. 3(b)). If we take the signal received by M2 as a reference signal and assign proper weights (phases) to M1 and M3 based on their geometric relationship, the signals can be aligned perfectly (Fig. 3(c)) and the combined signal energy is concentrated effectively, as shown in Fig. 3(d). In other words, by adjusting the phase of the signals received by each sensor, beamforming can create a directional beam that maximizes signal strength in a specific direction while minimizing signal strength in other directions. Therefore, the effective use of beamforming techniques can significantly enhance the quality of sound transmission and reception.

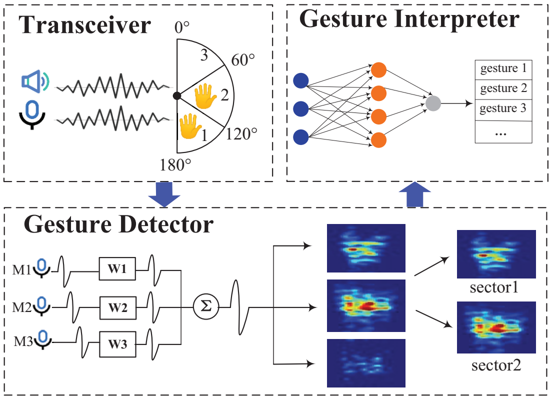


Fig. 4. Overview of system.

## III. System Design

### A. Overview

Fig. 4 depicts the system overview of RemoteGesture. RemoteGesture consists of three modules. In transceiver design, a commodity speaker continuously emits a predefined sound wave to probe the channel, which is then received by a circular microphone array. To receive a reliable signal, the emitted sound wave is carefully designed and transformed to support far-field gesture sensing and alleviate interference in practical environments. Then, the gesture detector will perform a two-stage progressive beam scanning scheme to scan the space where gestures may occur, which can timely detect the direction that is most closely associated with the gesture. With this estimated direction, sound waves reflected from the hand can be remarkably enhanced. Note that our two stage scheme also supports simultaneous detection of gestures performed by multiple users. CIR of the enhanced signal is then measured for visualizing channels impacted by the gestures. In gesture interpreter, we apply a lightweight convolutional neural network to extract features from CIR measurements and perform gesture recognition.

### B. Transceiver Design

The transceiver consists of a collocated commodity speaker and a microphone array. The speaker unceasingly transmits a predefined sound frame, while the microphone array receives those frames. In our system, we apply the ZC sequence as the baseband transmission signal. First, the ZC sequence has been proven to possess higher auto-correlation with narrow side lobe level [15] compared to other mainstream sequences, such as Training Sequence Code, Barker Sequence, and M-sequence. A higher auto-correlation of the transmission frame (sharper main lobe) guarantees easier separation of sound

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.

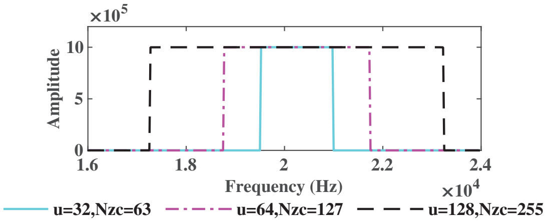


Fig. 5. Bandwidth of three ZC passband sequences.

copies with close delays generated by complex finger movement. Second, the transmitted frame should cover a certain range of frequency band to carry effective information imposed by gestures and mitigate frequency-selective fading. In addition, the transmitted ZC sequence should reside in a frequency band ranging from $18KHz$ to $24KHz$ such that it is inaudible to users [16] and does not pose any disturbance. Typically, the ZC sequence has a constant amplitude covering a range of frequencies, whose bandwidth can be configured with careful parameter selection to satisfy those practical requirements of gesture sensing.

ZC sequence is a complex value sequence, which can be generated as follows:

$$ZC[n] = e^{-j\frac{\pi u n\left(n+c_f+2q\right)}{N_{ZC}}} \tag{2}$$

where $N_{ZC}$ is the length of the sequence, which is usually an odd number and can be configured based on specific requirements. $0 \leqslant n < N_{ZC}$, $0 < u < N_{ZC}$ and $gcd(N_{ZC}, u) = 1$. $c_f = \{N_{ZC}\ mod\ 2\} = 1$ and $q$ is an integer indicating the cyclical shift of Chu sequence by $q$, where Chu sequences is a special case of ZC sequence when $q = 0$. $ZC[n]$ denotes the $n - th$ elements of the ZC sequence. In our design, $c_f = 1$ and we set $q = 0$ to simplify the parameter configuration. Therefore, two parameters $u$ and $N_{ZC}$ jointly determine the ability of ZC sequence carrying information of gestures.

Traditional works use a longer transmitted frame to cover a wider sensing range. With the support of a sampling rate of $fs = 48KHz$ by commodity acoustic devices, we can calculate the sensing distance $d = \frac{N_{ZC} \times c}{2fs}$, where $c = 340m/s$ is the speed of sound propagation in air. In order to enable room-level gesture sensing, which covers a range of approximately $7m$, we target to set $N_{ZC} = 2048$. However, our experiments have verified that simply configuring $N_{ZC}$ to 2048 results in extremely high-pitched screaming of sound, since the frequency of the transmitted ZC sequence leaks into the audible band below $18KHz$.

To meet the inaudibility requirement, we apply interpolation scheme in frequency domain of a short ZC sequence and extend the length of ZC sequence to 2048. Specifically, we first apply Fast Fourier Transform (FFT) to convert the short ZC sequence into the frequency domain and symmetrically pad multiple zeros outside its original frequency components. After zero-padding, we transform the interpolated sequence back into the time domain using Inverse Fast Fourier Transform (IFFT). Such an interpolation scheme in frequency domain results in sharper auto-correlation peak than that in time domain [15]. Note that the bandwidth of the interpolated baseband signal is $B = \frac{f_s \times N_{ZC}}{N'_{ZC}}$, where $N_{ZC}$ and $N'_{ZC}$ denote the sequence length before and after interpolation, respectively, and $f_s$ denotes the signal sampling rate.

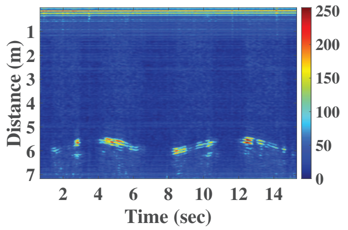


Fig. 6. CIR Measurement for gestures at $6m$.

To transmit channel probe signal in inaudible band, we up-convert the baseband signal to the inaudible band with a carrier wave $f_c$. Specifically, the real and imaginary parts of interpolated baseband ZC sequence are multiplied by $\cos 2\pi f_c t$ and $-\sin 2\pi f_c t$, respectively, and then are summed to form the passband signal. A high-pass filter is applied to exclude any out-of-band noise and interference. Note that the transmitted frame can be saved in audio-compatible files (e.g., WAV, MP3), and can be continuously played by current commercial speakers. At the receiver side, due to asynchronization of speaker and microphone, the first sample of the received ZC frames can be pinpointed using cross-correlation between the transmitted frame and received signal. This allows for the subsequent frames to be aligned due to the fixed frame length. Note that low-frequency components in the environment and other interferences, such as human speech and music, are filtered out by simply applying a high-pass filter. Then we perform down-conversion to the synchronized passband signal by multiplying $\cos 2\pi f_c t$ and $-\sin 2\pi f_c t$ to generate the real and imaginary part, respectively, followed by a low pass filter to remove high-frequency interference.

Fig. 5 shows frequency domain of three passband ZC sequences with different $u$ and $N_{ZC}$ pairs. All ZC sequences are expanded to 2048 samples using our interpolation scheme. We perform up-conversion to obtain passband sequences with a carrier wave at $f_c = 20.25KHz$. The bandwidth for passband ZC sequences with parameters of $(32, 63)$, $(64, 127)$, and $(128, 255)$ are $1.5KHz$, $3KHz$, and $6KHz$, respectively. On one hand, rich multipath in a far-field sensing range aggravates the frequency selective fading (FSF), leading to inadequate channel measurements if a probing signal only covers a $1.5KHz$ bandwidth. On the other hand, although $6KHz$ bandwidth contributes to capture sufficient channel measurements and theoretically satisfies inaudible frequency range, filters applied to up-conversion and down-conversion inevitably incur frequency leakage into the audible band due to imperfect frequency responses. Thus, after comprehensive consideration, we select the ZC sequence with $u = 64$ and $N_{ZC} = 127$, which covers a $3KHz$ frequency band ranging from $18.75KHz$ to $21.75KHz$, enabling a sufficient bandwidth for reliable channel measurement while fully residing in the inaudible band.

We conduct an experiment to validate our parameter selection and interpolation scheme for detecting gestures in the

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.

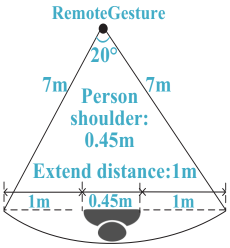


Fig. 7. Beamforming resolution.

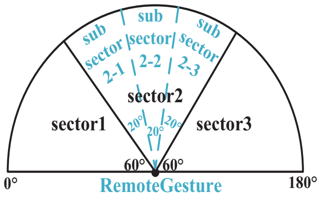


Fig. 8. Two-Stage Beam Scanning.

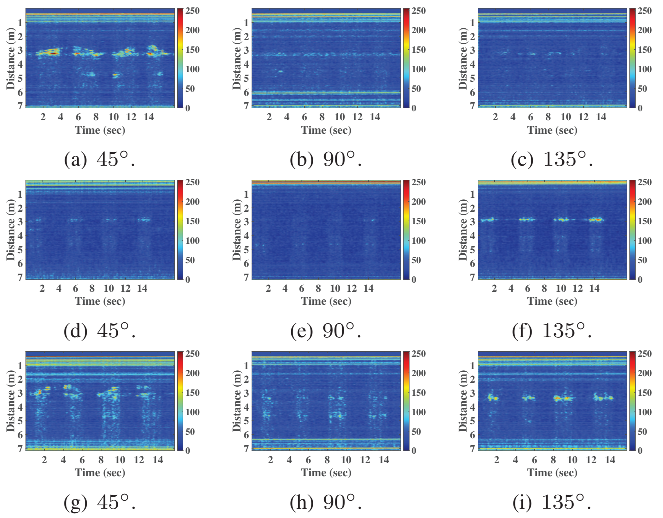
far-field. During the experiment, a speaker and microphone array continuously transmits and receives the designed signal frame using the aforementioned parameters and interpolation scheme. A user is requested to stand 6 meters away from the acoustic devices and perform push and pull twice. Fig. 6 illustrates the measured CIR, which clearly exhibit observable gesture patterns even at a distance of $6m$. The experiment results demonstrate that the chosen parameters and interpolation scheme effectively increase the sensing distance of the acoustic signal and allow reliable channel measurements for the far-field sensing.

### C. Beamforming Scanning

Our objective is to simultaneously detect gestures from multiple users. Traditional beamforming techniques employ continuous scanning of the entire space to identify the directions most associated with the gesture. Based on these directions and geometric layout of sensor array, different weights are assigned to corresponding receivers in the array. By doing so, signals from specific directions where different users perform gestures are enhanced, at the same time signals from other directions are attenuated. Although the space scanning method enables precise angle-of-arrival (AoA) estimation, the scanning time is too long to satisfy the real-time requirements of our gesture recognition system. Intuitively, a higher resolution of scanning results in more precise direction estimation while imposing longer scanning time, and vice versa.

To meet real-time of gesture detection, we design a simple yet effective scheme inspired from binary search algorithm. The core idea is to balance the resolution of scanning and direction estimation precision. To determine the lowest resolution for scanning while still being capable of detecting gestures, we fully exploit the geometric relationship between the human body and acoustic devices during gesture performance. Specifically, we measure the angle of a certain sector area between acoustic devices and two shoulders of the user at a distance of $7m$ as the lowest scanning resolution, as illustrated in Fig. 7. Taking into account practical usage scenarios, we extend a 1 meter distance at each side of the shoulder since multiple users are less likely standing within such a small area in far-field gesture sensing. To mitigate individual stature differences, we ask multiple volunteers, including those of different genders, heights, and figures, and the measured sector angle ranges from $18°$ to $22°$. Therefore, w.l.o.g, we finally configure the lowest resolution for beam scanning to $20°$.

However, such a scanning resolution still challenges the real-time performance for RemoteGesture, as it requires 9 scans to determine the direction. To further reduce the scanning time, we propose a two-stage progressive scanning scheme, as shown in Fig. 8. First, we divide the semicircular region into three $60°$ sector regions and execute beamforming on each sector to determine approximate regions where users are performing gestures (we will explain how to find those regions in detail in Section III-D). In the second stage, we further divide each $60°$ sector region into three $20°$ sub-sectors and perform beam scanning on each sub-sector to pinpoint the exact direction associated with the gesture. It is important to note that beamforming in sub-sectors is only conducted for the parent sector where the gesture is detected in the first stage. Such a scheme enables more than 3 user gestures recognition. Users can be located at random positions rather than being restricted to the assigned sectors. Our approach reduces the scanning complexity from $O(N)$ to $O(logN)$ for detecting a gesture, resulting in reduced scanning time.



(a) 45°. (b) 90°. (c) 135°.

(d) 45°. (e) 90°. (f) 135°.

(g) 45°. (h) 90°. (i) 135°.

Fig. 9. CIR Measurement in different directions.

### D. Channel Estimation

In this section, we will introduce how to determine the region where gestures are performed using CIR measurements. Note that CIR can separate received multipath signals in different delays, providing a basis for detecting gestures performed at different distances. However, the lack of azimuth information in CIR makes it inaccessible to simultaneously detect gestures from multiple users. As a complementary method, our beam scanning scheme enables the measurement of azimuth information. Our key idea is that CIR measured from regions where users perform gestures exhibit a much stronger pattern than other non-gesture directions. Therefore, we first generate the CIR measurements for each selected area and then design an effective method to find the area mostly related to gestures based on CIR amplitude.

Fig. 9 shows CIR measurements when different users perform gestures at the same distance while in different directions to the acoustic devices. In this experiment, we ask two volunteers standing at $3m$ to RemoteGesture while in different directions (i.e., $45°$ and $135°$). First, user1 in $45°$

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.

direction performs push and pull two times while user2 keeps static. Then, user2 in 135° direction performs double tap four times while user1 remains stationary. Finally, user1 performs push & pull and user2 performs double tap simultaneously. Beamforming is executed at 45°, 90° and 135°, respectively.

From Fig. 9 we obtain two key observations based on the measured CIR measurements: (1) matching the beam scanning with the direction in which gesture is performed results in significantly stronger gesture pattern than those in another two sectors (i.e., Fig. 9(a), Fig. 9(f), Fig. 9(g) and Fig. 9(i) are stronger than the others). Additionally, different gestures exhibit distinguishable features in the CIR patterns (i.e., Fig. 9(a) and Fig. 9(g) are push and pull, while Fig. 9(f) and Fig. 9(i) are double-tap); (2) when two users perform gestures simultaneously, beamforming at each user's direction generates different CIR patterns (i.e., Fig. 9(g) and Fig. 9(i)). The other bright paths in the CIR measurement in Fig. 9 are multipath reflections from the far-field space. This experiment demonstrates that our beam scanning and beamforming scheme can effectively distinguish simultaneous multiple gestures even from the same distance. After coarsely finding the gesture region, we further divide the 60° region into three 20° regions and perform beamforming to determine the fine-grained gesture direction.

We apply a simple yet effective method to determine the region where users perform gestures. Our intuition is that CIR amplitude can be significantly enhanced when the regions where users perform gestures precisely match the sectors executed by beamforming. On the contrary, if the gesture regions do not match the beamforming sectors, the CIR amplitude will remain the same. Therefore, we determine the sector as targeted gesture region by finding the largest average CIR amplitude for each candidate sector. Specifically, we first set a threshold for filtering the CIR amplitude in each sector. CIR amplitudes below this threshold can be considered as interference caused by weak multipath copies of the gesture, which are set to zero value, as those amplitudes in yellow rectangle shown in Fig. 10. Next, we measure two indicators that can profile the amplitudes of CIR: (1) the number of non-zero grids and (2) the average value of non-zero grids. The number of non-zero grids quantitatively represents how many reflection paths from hand are successfully captured by CIR, while the average value of non-zero grids demonstrates the overall quality of the hand reflected signals. The gesture regions are pinpointed only if both two indicators satisfy a pre-defined threshold.

CIR amplitudes in red rectangle in Fig. 10 are caused by other parts of human body, such as torso, arm and elbow. We notice that those CIR amplitudes are at the same level as those caused by hands, and cannot be easily removed with a small threshold. However, we discover that movements of torso, arm and elbow are intrinsically involved in a gesture. For example, if a user performs push or pull, in addition to finger movements, the arm and elbow also move simultaneously, which can be regarded as a useful feature for recognizing gestures. Therefore, instead of removing those CIR amplitudes, we synthetically consider them as a part of gestures and reserve such amplitudes for gesture recognition.

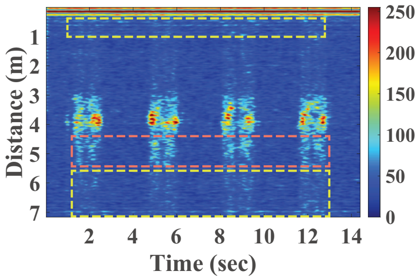


Fig. 10. CIR measurement of four double-tap gestures. Yellow rectangle: multipath caused by weak reflections and red rectangle: multipath caused by human body.

## E. Gesture Recognition

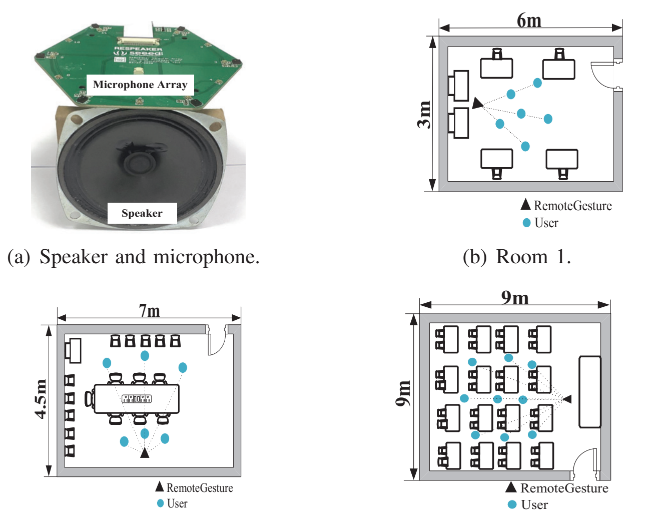
Gesture interpreter extracts features from CIR measurements and classifies them into different gestures. Neural networks have demonstrated excellent performance in image classification, while requiring massive amount of data to achieve high accuracy and strong robustness. To release the heavy burden of manual data collection, we conduct a data augmentation technique similar to the one used in [24] to automatically enlarge the training dataset.

Specifically, CIR measurements of a gesture obtained at varying distances from the transceiver exhibit vertical drift in CIR taps, while different moving speeds of gestures result in horizontal stretching and compression of CIR measurements [24]. These two key observations motivate us to automatically enlarge the gesture training data for different distances and hand moving speeds. Note that our data augmentation technique differs from the previous technique used in RobuCIR. Instead of simply shifting each CIR measurement in all taps, we shift the CIR measurements within a certain range of taps. Restricting the range of drifting taps enables a more effective representation of gestures, since CIR measurements involve similar patterns when gestures are performed at close regions.

In our study, we employ MobileNetV2 as the gesture classifier, given its lightweight neural network structure and low computational overhead that suits mobile devices with limited computing resources [14]. The real-time detection requirement of multi-user gestures in RemoteGesture necessitates fast processing during gesture classification. MobileNetV2 employs lightweight depthwise convolutions to filter features, which reduces computational overheads by approximately one-eighth of traditional convolutional methods while without compromising performance [14].

The input of our classifier is a CIR image with size of $H \times W$, where $H$ is image height and $W$ is image width, respectively. First, 32 convolution kernels, each with a size of $3 \times 3$, normalize the input CIR image. Then, 19 residual bottleneck layers extract the features of the input image. Following this, an average-pooling layer performs global average pooling on the extracted features, resulting in $1 \times 1 \times 1280$ eigenvectors. Finally, the fully connected layer connects the feature vectors and outputs a $1 \times 9$ probability vector, which represents the

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.



(a) Speaker and microphone.

(b) Room 1.

(c) Room 2.

(d) Room 3.

Fig. 11. Commercial speaker and microphone and three different rooms.

classification rank of the nine gestures. The predicted gesture is the one with the highest probability.

## IV. Experiments and Evaluation

### A. Experiment setup

*1) Hardware:* Our transceiver consists of a commercial speaker and a circular microphone array, as shown in Fig. 11(a). We apply google AIY Voice Kit version 2.0 including a speaker with a maximum power of $3W$ and a Raspberry Pi Zero to control the speaker to transmit the acoustic signal frame in the inaudible band continuously. The audio frames are received using the ReSpeaker 6-Mic Circular Array Kit with a signal sampling rate of $48KHz$, which supports a maximum $24KHz$ audio signal.

*2) Data collection:* We select the root ZC sequence with $u = 64$ and $N_{zc} = 127$, and the interpolated sequence length is $N^{'}_{ZC} = 2048$. We invite six volunteers (four males and two females) to perform nine gestures at various distances and angles relative to the RemoteGesture in three rooms with different sizes and layouts. These rooms are sized $6m \times 3m \times 3m$, $7m \times 4.5m \times 3m$ and $9m \times 9m \times 3m$, respectively, as shown in Fig. 11(b)-11(d). In each room, we ask volunteers to stand or sit still in different positions depicted in Fig. 13 and perform gestures at a natural speed. A maximum of three users are allowed to perform gestures simultaneously during the experiment. Each gesture is repeated 20 times at each position, resulting in a total of 2700 gesture samples collected from 15 distinct positions. We implement data augmentation on the real gesture samples with rate = $100\times$. Specifically, we shift each real CIR measurement within 572 taps (corresponding to a $2m$ shifting range, $1m$ above and below), followed by a $0.5\times \sim 1.5\times$ horizontal compression and stretching (representing $0.5s \sim 2s$ for each gesture). The augmented dataset matches the gesture variations covering real-world scenarios.

*3) Model Training:* The gesture classifier is trained and tested using PyTorch on a laptop equipped with a 32 GB

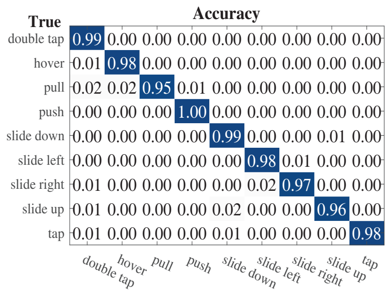


Fig. 12. Overall performance.

memory, a $12th$ Generation Intel Core $i7 - 12700H$ CPU, and an NVIDIA GeForce RTX 3060 Laptop GPU. The dataset is separated with $70\%$ for training and $30\%$ for testing, respectively. We utilize 10-fold cross-validation during model training and testing. The training process is a one-off step and the trained model is $8.8M$, which can be stored offline.

*4) Benchmark:* We compare the performance of our RemoteGesture with the state-of-the-art approach named RobuCIR [24]. As RobuCIR only supports near-field gesture sensing, we fix the distance between hand and speaker-microphone pair to $1m$ while performing gestures at different angles (i.e., $45°$, $90°$, and $135°$). We use the same setting in [24] and apply the configuration above to implement RobuCIR and RemoteGesture, respectively.

### B. Evaluation

*1) Overall system performance:* Fig. 12 shows the overall performance of RemoteGesture across all 9 gestures. For this evaluation, we use a combination of augmented data and real collected data from all three rooms for training and testing. Our RemoteGesture achieves an average recognition accuracy of $97.8\%$ with each gesture exceeding $95\%$ accuracy, even when performed at different distances and angles in multi-user scenarios. With our holistic design, RemoteGesture can recognize gestures simultaneously performed by multi-users at distances up to $7m$, significantly extending the usage scenarios of acoustic-based gesture recognition.

*2) Performance on different distances and angles:* We evaluate the average recognition accuracy for all 9 gestures performed at various distances and angles, as illustrated in Fig. 13. In this experiment, volunteers perform 9 gestures at each position, with each gesture repeated 100 times. At each position, we apply $10\times$ data augmentation on the real gesture samples, in which $70\%$ are used for training and $30\%$ for testing. RemoteGesture achieves an accuracy higher than $94\%$ at all 15 positions. The average accuracy at $5m$ in all directions exhibits only a $0.68\%$ decrease compared to the average accuracy at $1m$, which reveals the high robustness of our RemoteGesture. In addition, although recognition accuracy decreases with increasing distance, RemoteGesture maintains an average accuracy of $95.4\%$ at $7m$, which manifests its capability for far-field sensing. We note that accuracy exceeds $5m$ can be further improved by using high-end acoustic devices with higher Tx/Rx antenna gain.

*3) Performance on multiple users:* To evaluate the ability of RemoteGesture identifying multiple simultaneously performed

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.

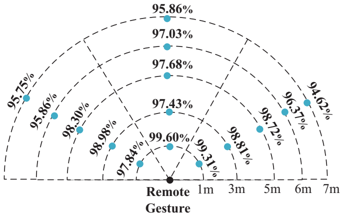


Fig. 13. Performance on different positions.

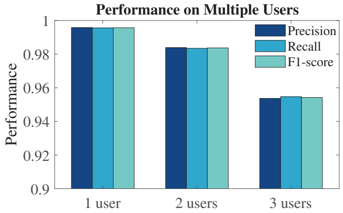


Fig. 14. Performance on multiple users.

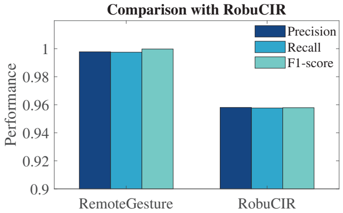


Fig. 15. Comparison with RobuCIR.

TABLE I
RUNNING TIME OF REMOTEGESTURE

| CIR Calculation (ms) | | | Beamforming (ms) | Gesture Recognition (ms) |
|---|---|---|---|---|
| Frame Detection | Down Conversion | CIR | Binary Search | MobileNetV2 |
| 111 | 10.2 | 33 | 200 | 36.9 |

gestures, we conduct an experiment with three volunteers. We ask one, two and three volunteers to stand at different distances (i.e., $> 3m$) and angles from RemoteGesture, respectively. Each volunteer is allowed to randomly perform 3 different gestures at each position with each gestures being repeated 100 times. The gesture samples are divided into $70\%$ for training and $30\%$ for testing, respectively.

Fig. 14 shows the experiment results. RemoteGesture achieves a recognition precision of $98.4\%$ and $95.4\%$ for two and three users simultaneously performing gestures, respectively, which are slightly lower than the single-user scenario. The recall and F1_score all exceed $95\%$. The results demonstrate that our system can effectively detect and distinguish each gesture in multi-user scenarios while maintaining the recognition ability.

*4) Comparison with RobuCIR:* We compare our RemoteGesture with RobuCIR. As depicted in Fig. 15, RemoteGesture achieves an accuracy of $99.7\%$ in recognizing gestures from different angles, surpassing RobuCIR's accuracy of $95.8\%$ in the near-field scenario. This is attributed to our two-stage progressive beam scanning and beamforming scheme, which concentrates the energy of the reflected signal from the hand, thereby enhancing the signal-to-noise ratio of the received signal. As such, more distinguishable patterns can be captured by CIR measurements, which can then be successfully learned by neural network.

*5) Execution Time:* Table I shows the processing time of each step involved in RemoteGesture. Specifically, frame detection is performed only once at the beginning of the received signal, which takes $111ms$. The processing delay for measuring CIR for a gesture is $33ms$. Based on our beamforming scheme, a maximum of six beamforming scannings need to be executed for detecting a gesture, which cost approximate $200ms$ with $33ms$ for each beamforming scanning. The time for identifying a gesture using our MobileNetV2 model is $36.9ms$. Therefore, the overall processing delay takes less than $0.5s$, which satisfies the real-time requirement.

## V. RELATED WORK

### A. Vision Based Gesture Recognition

Powerful cameras, along with the emergence of deep learning technologies, facilitate gesture detection and classification in a contact-free manner using vision-based methods [2], [6], [7], [12], [13]. A CNN-based dense hand pose estimation is proposed, aiming to reconstruct 3D hand shapes and poses from a single RGB image [2]. FOANet [12] shows significantly improved gesture recognition results on two publicly available datasets, which utilizes global channels to process the whole gesture video to look for gross motions while focused channels to detect and process each hand. However, the usage scenarios of these methods are restricted by their reliance on good lighting conditions.

### B. Radio Frequency Based Gesture Recognition

The promising RF sensing technology has been widely applied to human activity recognition [1], [5], [19], [20], [25], [26], [30], [32]. The intuition is that human motion impacts the RF signal propagation in terms of signal strength and phase. RF-Finger [20] deploys a RFID tag array to continuously measure the changes of backscattered signals and recognizes the gestures performed in front of the antenna. EUIGR [32] collects phase and received signal strength from the RFID tags and uses neural networks to learn gestures' features to build a realtime gesture-driven interactive system. WiMU [19] supports simultaneous multiple gesture detection based on virtual sample combinations of different gestures. WiGest [1] realizes gesture recognition nearby mobile devices based on changes in RSS and CSI without requiring any pre-training. However, these approaches either require costly specialized devices or bear low resolution of sensing due to relatively long wavelength of RF signals.

### C. Acoustic Based Gesture Recognition

Currently, speakers and microphones have embedded in almost every smart device, which enables economy and convenience for achieving sensing tasks [3], [4], [9]–[11], [15], [21]–[24], [27]–[29], [31], [33]. Vskin [15] performs fine-grained touch gesture recognition on the surface of mobile devices by measuring amplitude and phase of sound signals. RobuCIR [24] can recognize 15 gestures by measuring CIR and adopting frequency hopping mechanism and data enhancement technology to overcome the problems of frequency selective fading and insufficient training data. PDF [4] proposes a method based on phase difference to extrapolate time delay of FMCW signal to infer the absolute distance, which can recognize tiny movements of fingers. Those works can accurately identify gestures in near-field scenarios (i.e., $<1m$) by intentionally discarding the far-field interference. RTrack [11] implements a room-scale hand motion tracking system

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.

with a range up to $4.5$ meters by using 2D MUSIC (multiple signal classification) algorithm to overcome low signal-to-noise ratio and rich multipath propagation of reflected signals at long distances. However, recent works merely focus on gesture sensing for a single individual. Unlike those works, RemoteGesture achieves room-scale gesture detection (i.e., $7m$) and supports simultaneous gesture recognition for multiple users.

## VI. Conclusion

We present RemoteGesture, a room-scale gesture recognition system based on acoustic signals that can detect gestures simultaneously performed by multiple users. RemoteGesture overcomes a series of practical challenges occurred in far-field sensing scenarios and boosts the sensing range to 7 meters. RemoteGesture supports simultaneous gestures from multiple users by applying beamforming technique. To satisfy real-time gesture detection, we reduce the space scanning time inspired from the binary search algorithm. Extensive experiments demonstrate that RemoteGesture can achieve over $95\%$ recognition accuracy at 7 meters under multi-user scenario.

## Acknowledgment

This work was supported by the National Nature Science Foundation of China under Grant 62102139, the Nature Science Foundation of Hunan Province of China under Grant 2022JJ30168 and 2023JJ20015, and the Fundamental Research Funds for the Central Universities under Grant 531118010612. This research has benefited from financial support of the Hong Kong Institute of Business Studies, Faculty of Business, Lingnan University, Hong Kong Special Administrative Region, China (Yanwen Wang is the corresponding author).

## References

[1] H. Abdelnasser, M. Youssef, and K. A. Harras. Wigest: A ubiquitous wifi-based gesture recognition system. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1472–1480, 2015.

[2] S. Baek, K. Kim, and T. Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1076, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[3] H. Chen, F. Li, and Y. Wang. Echotrack: Acoustic device-free hand tracking on smart phones. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, 2017.

[4] H. Cheng and W. Lou. Push the limit of device-free acoustic sensing on commercial mobile devices. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021.

[5] K. Cui, Y. Wang, Y. Zheng, and J. Han. Shakereader: 'read' uhf rfid using smartphone. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021.

[6] R. Cui, H. Liu, and C. Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1610–1618, 2017.

[7] Leap Motion, Inc. Leap Motion: Mac PC Gesture Controller for Game, Design and More. https://www.leapmotion.com/, 2013.

[8] J. Li and P. Stoica. *Robust adaptive beamforming*. John Wiley & Sons, 2005.

[9] K. Ling, H. Dai, Y. Liu, A. X. Liu, W. Wang, and Q. Gu. Ultragesture: Fine-grained gesture sensing and recognition. *IEEE Transactions on Mobile Computing*, 21(7):2620–2636, 2022.

[10] C. Liu, P. Wang, R. Jiang, and Y. Zhu. Amt: Acoustic multi-target tracking with smartphone mimo system. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021.

[11] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, MobiCom '19, New York, NY, USA, 2019. Association for Computing Machinery.

[12] P. Narayana, J. R. Beveridge, and B. A. Draper. Gesture recognition: Focus on the hands. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5235–5244, 2018.

[13] RoboRealm. Microsoft Kinect. http://www.roborealm.com/help/Microsoft Kinect.php, 2013.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[15] K. Sun, T. Zhao, W. Wang, and L. Xie. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, MobiCom '18, page 591–605, New York, NY, USA, 2018. Association for Computing Machinery.

[16] A. R. Valiente, A. Trinidad, J. R. G. Berrocal, C. G. Górriz, and R. R. Camacho. Extended high-frequency (9–20 khz) audiometry reference thresholds in 645 healthy subjects. *International Journal of Audiology*, 53:531 – 545, 2014.

[17] H. L. Van Trees. *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2002.

[18] B. Van Veen and K. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

[19] R. H. Venkatnarayan, G. Page, and M. Shahzad. Multi-user gesture recognition using wifi. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '18, page 401–413, New York, NY, USA, 2018. Association for Computing Machinery.

[20] C. Wang, J. Liu, Y. Chen, H. Liu, L. Xie, W. Wang, B. He, and S. Lu. Multi - touch in the air: Device-free finger tracking and gesture recognition via cots rfid. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1691–1699, 2018.

[21] P. Wang, R. Jiang, and C. Liu. Amaging: Acoustic hand imaging for self-adaptive gesture recognition. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 80–89, 2022.

[22] T. Wang, D. Zhang, Y. Zheng, T. Gu, X. Zhou, and B. Dorizzi. C-fmcw based contactless respiration detection using acoustic signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4), jan 2018.

[23] W. Wang, A. X. Liu, and K. Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, MobiCom '16, page 82–94, New York, NY, USA, 2016. Association for Computing Machinery.

[24] Y. Wang, J. Shen, and Y. Zheng. Push the limit of acoustic gesture recognition. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pages 566–575, 2020.

[25] Y. Wang and Y. Zheng. Modeling rfid signal reflection for contact-free activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), dec 2018.

[26] Y. Wang and Y. Zheng. Tagbreathe: Monitor breathing with commodity rfid systems. *IEEE Transactions on Mobile Computing*, 19(4):969–981, 2020.

[27] Z. Wang, Y. Wang, M. Tian, and J. Shen. Hearfire: Indoor fire detection via inaudible acoustic sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4), jan 2023.

[28] Q. Yang and Y. Zheng. Model-based head orientation estimation for smart devices. 5(3), sep 2021.

[29] Q. Yang and Y. Zheng. Deepear: Sound localization with binaural microphones. *IEEE Transactions on Mobile Computing*, pages 1–17, 2022.

[30] Y. Yang, J. Cao, and Y. Wang. Robust rfid-based respiration monitoring in dynamic environments. *IEEE Transactions on Mobile Computing*, 22(3):1717–1730, 2023.

[31] Y. Yang, Y. Wang, J. Cao, and J. Chen. Hearliquid: Nonintrusive liquid fraud detection using commodity acoustic devices. *IEEE Internet of Things Journal*, 9(15):13582–13597, 2022.

[32] Y. Yu, D. Wang, R. Zhao, and Q. Zhang. Rfid based real-time recognition of ongoing gesture with adversarial learning. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, SenSys '19, page 298–310, New York, NY, USA, 2019. Association for Computing Machinery.

[33] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '17, page 15–28, New York, NY, USA, 2017. Association for Computing Machinery.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on November 19,2024 at 07:35:41 UTC from IEEE Xplore. Restrictions apply.