Multiuser Wireless Hand Gesture Recognition by Spatial Beamforming

Rui Peng, *Student Member, IEEE*, Yafei Tian, *Member, IEEE*, and Shengqian Han, *Member, IEEE*

***Abstract*—Hand gesture recognition through wireless sensing is a new method of human-machine interaction, and is an important research direction for next generation integrated sensing and communication systems. Recently, although there are plenty of works on wireless sensing and hand gesture recognition, most of them are applied in single user scenario. Since the recognition algorithm is mainly realized by analyzing the variation of wireless propagation channel, the simultaneous movement of multiple people will cause superposition of the dynamic channel responses and cause interference to each other. This paper proposes to use spatial beamforming to alleviate the mutual influence of multiuser hand gestures. Since the hands of multiple users reflect the same signal, the channel responses are usually not orthogonal even if the users are well separated in different locations, and there is no reference signals to help distinguishing the channel response reflected by different users. We propose a preamble gesture scheme to estimate the spatial channel of dynamic reflections under the impact of strong phase noise, and use the Doppler variation feature to verify the channel response belongs to hand movement. We have a thorough analysis of the inter-user interference impacted by the transmitter-user-receiver geometry relations, user movement speeds and directions. The interference suppression effect is demonstrated by prototype experiments under real LOS and NLOS scenarios exploiting LTE signals.**

***Index terms*—** Beamforming, channel state information, gesture recognition, multiuser, wireless sensing.

I. Introduction

Device-free wireless sensing (DFWS) is a promising technology which attracts much attention recently since it can capture the electromagnetic disturbance caused by human movement and does not need the user to carry any physical device. The in-air propagated radio-frequency (RF) signal not only carries information data, but also involves with the surrounding environment during its transmission [1]. From transmitter to receiver, the electromagnetic wave might goes through many times of reflections and diffractions, by the static or moving objects like walls, furnitures, and humans. With human body movements, the propagation channel changes, and the wireless sensing system aims to extract such changes and further recognizes human motions [2].

Taking advantage of this capability, DFWS can be widely used in many scenarios in daily lives, such as indoor human activity detections including localization [3], respiration detection [4], and gait recognition [5]. Except for traditional indoor scenes, DFWS also shows great potential on smart cars [6] like driver's behaviour detection [7], driver authentication [8] and vehicle speed estimation [9]. Among these applications, hand gesture recognition is suitable for both indoor and in-car circumstances, which enables human to interact with intelligent electric appliances or automotive multimedia. Hand gesture recognition is challenging compared with body movement detection since hands have much less impact on wireless propagation.

To capture the environment changes, multiple kinds of channel measurement information are exploited, for example, received signal strength (RSS) [10], [11], time-of-flight (ToF) [12] and channel state information (CSI). Among these measurement information, CSI is most widely used because CSI can provide precise amplitude and phase variations of wireless channels [13].

According to the electromagnetic wave propagation theory, if the moving object can be viewed as a point target, there is a deterministic relation between the CSI variation and movement trajectory, given the positions of the transmitter and receiver. In [14], the concept of Fresnel zone is first introduced to reveals the effect of human position and orientation to CSI fluctuations. In [15], the Doppler shift is calculated from CSI and nine kinds of body movement can be recognized. In [16], both amplitude and phase of CSI are used to detect moving and stationary human. But the relation becomes complicated if we take into account situations that one activity may involves several body parts. There are researches focusing on deep learning methods in these situations, using multi-layer perception (MLP) [17], [18], convolutional neural network (CNN) [19]–[21], or recurrent neural network (RNN) [22], [23]. Such methods can achieve good performance in single user scenario by capturing the hidden information it learned from CSI [24].

However, there are few studies regarding the human activity detection problem in multiuser scenario. If two or more people move simultaneously, human activity effects are mixed up, leading to strong inter-user interference (IUI) [25]. With different dynamic path lengths, the activities of different people can be separated by different ToFs. This method is usually implemented by specialized system since it requires large bandwidth to achieve higher ToF resolution. For example, DeepBreath [26] and Witrack2.0 [27] use frequency-modulated continuous wave (FMCW) signals with 1.5 GHz and 1.69 GHz bandwidth, respectively. However, specialized system is hard to be widespreadly used. Recently, some works focus on multi-

This work was supported by the National Natural Science Foundation of China under Grants 61971023 and 61871015.

The authors are with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, P. R. China (e-mail: pengrui@buaa.edu.cn; ytian@buaa.edu.cn; sqhan@buaa.edu.cn).

target recognition with WiFi systems. WiMU [28] generates virtual samples for all possible single gesture combinations and matches with real multi-gesture samples. This method cannot associate gesture to users since it does not acquire any user location information. MultiTrack [29] dynamically switches 24 non-continuous channel in 5 GHz WiFi band and the total frequency spanning achieves 600 MHz, thus it can separate users by ToF. But in real communication system, it is not viable to continually switch the channel during transmission.

Except for the ToF, different reflection paths usually have different angles of arrival (AoA) and angles of departure (AoD). Some works propose to separate users by AoAs or AoDs [4] [30], but indoor multipath deteriorates angle estimation performance, especially in non line-of-sight (NLOS) scenario. In conventional wireless communications, multiuser beamforming can be applied directly based on the channel response instead of the angle information. But the difference is that it is hard to obtain accurate channel response here. In wireless communications, the signal is strong, and there are specially designed orthogonal reference sequences to estimate the channel responses of different users. In DFWS, the dynamic reflection is weak, and multiple users reflect the same signal. The channel responses from different users are usually not orthogonal even that the users are well separated in space, and there is no reference signals to help distinguishing them.

Furthermore, CSI calibration is another challenge. Raw CSI extracted from device is polluted by carrier frequency offset (CFO), sampling timing offset (STO) and phase noise. Such interferences impose a time-variant random phase to real CSI, which will cause power leakage from the static path to the dynamic path. The detection performance will thus be severely deteriorated as the phase variation is the most important information for human motion detection [13]. A common used method for single user detection is CSI ratio model [31], since CFO, STO and phase noise is generated by oscillator vibration, and neighbored antennas share the same CFO, STO and phase noise. Set the CSI from one antenna as reference, by dividing operation, the random phase is removed and the CSI ratio can be used to extract movement information. Nevertheless, if the dynamic part is relatively strong compared to the static part, there will be severe distortion in the ratio.

In previous works, the CSI was often acquired by WiFi signals since there are commodity devices like Intel WiFi link 5300 wireless NIC and it is easy to extract CSI from the drivers. But if not for research, WiFi signal is burst and randomly arrived, it may has rare data packets to get enough samples of CSI, especially when the network service is not busy [32]. On the contrary, Long Term Evolution (LTE) signal is always on the air, and is well covered in most inhabited areas. Its cell-specific reference signal (CRS) is periodically transmitted in every slot (0.5 ms), and we can acquire CSI in a constant rate as high as 2 kHz.

In this work, we develop a practical multiuser hand gesture recognition system based on LTE signal. We separate users and suppress phase noise in spatial domain. To estimate the spatial channel of the weak dynamic paths under the strong interference of static paths, we use a two-layer method to find optimal beam steering in the null space of static channels. We design a preamble gesture which has distinct pattern in Doppler-time domains, to insure the channel response belongs to the expected user and is not interfered by other movements. Since there is inter-user interference when two users perform gestures simultaneously, we have a thorough analysis of the interference impacted by the transmitter-user-receiver geometry relations, and user movement speeds and directions. The inter-user interference and phase noise effect are then suppressed in a unified framework. The gestures of each user are recognized by Doppler shift analysis after beamforming. To verify the system performance, we build a prototype system to receive LTE signals and perform beamforming and gesture recognition in real-time. The interference suppression capability and gesture recognition accuracies are evaluated under LOS and NLOS scenarios.

Our main contributions are listed as follows:

1) We design a preamble gesture to estimate the spatial channel of each user in multiuser scenarios. By analyzing the impact of phase noise, we provide a two-layer method to precisely extract the channel response of weak dynamic paths, where the first layer is used to eliminate the leakage of static paths and the second layer is to maximize the power of dynamic paths.
2) To analyze the inter-user interference, we derive the effect of hand movement speeds and directions of two users, on the eigenvalues and eigenvectors of the spatial covariance matrix. We propose an active motion number detection scheme, and design corresponding beamforming methods to suppress inter-user interference and phase noise.
3) We build a prototype system to receive LTE signals and conduct gesture recognition experiments in real-time, and experiment results show that our system can achieve good SINR and high accuracy rate in different scenarios.

The rest of the paper is organized as follows. Section II introduces the LTE related background, channel model and analyzes the impact of phase noise. In Section III, we discuss the preprocessing step to separate dynamic paths and static paths. Section IV presents the preamble gesture design, and the spatial channel estimation method. In Section V, inter-user interference analysis, multiuser beamforming and gesture recognition schemes are provided. In Section VI, we conduct experiments on the prototype and evaluate the performance and implementation complexity of the system. Finally, the conclusion is drawn in Section VII.

II. System Model

A. LTE Related Background

LTE is a complex system, and we only introduce the basic downlink frame format and some reference signals that we used in the signal processing flow. The complete description of LTE signal can be seen in 3GPP Specification TS 36.211 [33].

In LTE signal, a system frame with duration 10 ms is divided into 10 subframes and each subframe contains 14 or 12 symbols with normal cyclic prefix (CP) or extended CP,

respectively. In frequency domain, the subcarrier spacing is 15 kHz, and each resource block consists of 12 subcarriers. For maximum, each cell supports 20 MHz bandwidth for data transmission, where 1200 subcarriers are used, corresponding to 18 MHz effective bandwidth.

For synchronization purpose, the predefined primary synchronization signal (PSS) and secondary synchronization signal (SSS) are broadcasted in every half frame with fixed time-frequency positions. In addition, CRS is configured for fine synchronization and channel estimation, as shown in Fig.1. LTE supports multiple-input multiple-output (MIMO) transmission mode, and maximum four Tx ports CRS can be employed. According to different number of transmit ports, CRS occupies different number of time and frequency resources. For the signals transmitted from port 0 and 1, CRS occupies four symbols in each subframe and is uniformly distributed in frequency domain with six subcarriers interval. For the signals transmitted from port 2 and 3, only two symbols in each subframe are occupied by CRS. CRS sequence is defined as a pseudo-random sequence initialized by the cell-ID $N_{\mathrm{ID}}^{\mathrm{cell}}$. In this work, we use frequency division duplexing (FDD) mode so that channel estimations can be acquired in every subframes.

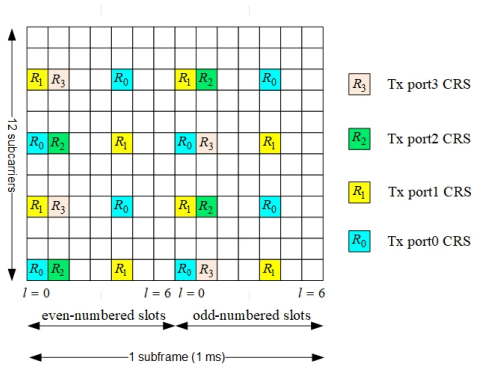


Fig. 1. The symbols and subcarriers occupied by CRS in a resource block, where $N_{\mathrm{ID}}^{\mathrm{cell}} = 0$ and normal CP is used.

B. Channel Model

We consider a MIMO time-variant wideband channel, where the transmitter has N_t antennas and the receiver is equipped with N_r antennas. Ideally, CSI can be modeled as the summation of channel responses of static paths and dynamic paths, for time t and angular frequency ω, as

$$\mathbf{H}(\omega,t) = \mathbf{H}_{\mathrm{s}}(\omega) + \mathbf{H}_{\mathrm{d}}(\omega,t), \tag{1}$$

where $\mathbf{H}_{\mathrm{s}}(\omega) \in \mathbb{C}^{N_r \times N_t}$ is the static channel response, caused by LOS path and other paths reflected from static objects, and $\mathbf{H}_{\mathrm{d}}(\omega,t) \in \mathbb{C}^{N_r \times N_t}$ denotes the time-variant dynamic channel response caused by moving reflections. In Eq. (1), we can further expand the dynamic channel response as a multiplication of the responses in time domain and in space domain,

$$\mathbf{H}_{\mathrm{d}}(\omega,t) = \sum_{i=1}^{M} a_i(t) e^{-j\omega\frac{d_i(t)}{c}} \mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{r},i}) \mathbf{a}_{\mathrm{T}}^{H}(\theta_{\mathrm{t},i}), \tag{2}$$

where M is the number of dynamic reflection paths, c is light speed, $a_i(t)$, $d_i(t)$, $\theta_{\mathrm{r},i}$, and $\theta_{\mathrm{t},i}$ are the channel gain, path length, AoA, and AoD of path i, respectively. The vector $\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{r},i}) \in \mathbb{C}^{N_r \times 1}$, $\mathbf{a}_{\mathrm{T}}(\theta_{\mathrm{t},i}) \in \mathbb{C}^{N_t \times 1}$ denotes array response on AoA $\theta_{\mathrm{r},i}$ and AoD $\theta_{\mathrm{t},i}$, which is also related with the array manifold and antenna spacing. Since the hand movement distance is small compared with the user-to-receiver and user-to-transmitter distance, the AoA and AoD of the reflection path can be assumed as a constant.

The channel gain $a_i(t)$ depends on the radar cross-section (RCS) of the reflection object and the signal propagation path length [34]. The path length $d_i(t)$ usually varies in dozens of centimeters for hand movement, but the caused phase variation can be significant, since the phase lag changes 2π as the propagation path increases one wavelength λ, i.e.,

$$e^{-j\omega\frac{d_i(t)}{c}} = e^{-j2\pi\frac{d_i(t)}{\lambda}}. \tag{3}$$

Furthermore, the effect of RF imperfection cannot be ignored. Except for the additive white Gaussian noise introduced in the front-end, oscillator vibration introduces CFO, STO, and phase noise. Even after tracking and compensation, there are still residual errors of CFO and STO, and we can incorporate them together into the phase noise. Phase noise is a kind of multiplicative noise, which will affect both the static and dynamic channel responses, i.e.,

$$\mathbf{H}(\omega,t) = e^{-j\theta_{\mathrm{n}}(t)} \left[\mathbf{H}_{\mathrm{s}}(\omega) + \mathbf{H}_{\mathrm{d}}(\omega,t)\right] + \mathbf{N}(\omega,t), \tag{4}$$

where $\theta_{\mathrm{n}}(t)$ is the phase noise, and $\mathbf{N}(\omega,t) \in \mathbb{C}^{N_r \times N_t}$ denotes the white Gaussian noise.

Here we have assumed that phase noise is a scalar item, since usually in one receiver the multiple RF chains will share the local oscillator or be synchronized through some phase locking mechanism. As the phase noise is time-variant and multiplicative, and the static channel response is much stronger than the dynamic part, phase noise will cause power leakage from the static part to the dynamic part. To give a straightforward explanation, we expand Eq.(4) to Eq.(5), where the final step approximation comes from the fact that $|\theta_{\mathrm{n}}(t)| \ll 1$, and $\cos\theta_{\mathrm{n}}(t) \approx 1$, $\sin\theta_{\mathrm{n}}(t) \approx \theta_{\mathrm{n}}(t)$. Influenced by phase noise, all the terms except $\mathbf{H}_{\mathrm{s}}(\omega)$ are time-variant. If we define $\hat{\mathbf{H}}_{\mathrm{d}}(\omega,t)$ as the summation of dynamic part channel response, then

$$\mathbf{H}(\omega,t) = \mathbf{H}_{\mathrm{s}}(\omega) + \hat{\mathbf{H}}_{\mathrm{d}}(\omega,t), \tag{6}$$

where

$$\begin{aligned}\hat{\mathbf{H}}_{\mathrm{d}}(\omega,t) = \mathbf{H}_{\mathrm{d}}(\omega,t) - j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{s}}(\omega) - j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{d}}(\omega,t) \\ +\mathbf{N}(\omega,t),\end{aligned} \tag{7}$$

the item $j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{d}}(\omega,t)$ is negligible since dynamic path reflected by hand is much smaller than static path and also $|\theta_{\mathrm{n}}(t)| \ll 1$. Thus, the dynamic part channel response becomes

$$\hat{\mathbf{H}}_{\mathrm{d}}(\omega,t) \approx \mathbf{H}_{\mathrm{d}}(\omega,t) - j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{s}}(\omega) + \mathbf{N}(\omega,t), \tag{8}$$

the strength of $j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{s}}(\omega)$ is comparable to that of $\mathbf{H}_{\mathrm{d}}(\omega,t)$. According to the experiments, actually $j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{s}}(\omega)$ is usually much stronger than $\mathbf{H}_{\mathrm{d}}(\omega,t)$, and causes severe contamination to the dynamic channel response. We can see that the impact

$$\begin{aligned}\mathbf{H}(\omega,t) =& e^{-j\theta_{\mathrm{n}}(t)}\left[\mathbf{H}_{\mathrm{s}}(\omega)+\mathbf{H}_{\mathrm{d}}(\omega,t)\right]+\mathbf{N}(\omega,t)\\ =& \cos\theta_{\mathrm{n}}(t)\left[\mathbf{H}_{\mathrm{s}}(\omega)+\mathbf{H}_{\mathrm{d}}(\omega,t)\right]-j\sin\theta_{\mathrm{n}}(t)\left[\mathbf{H}_{\mathrm{s}}(\omega)+\mathbf{H}_{\mathrm{d}}(\omega,t)\right]+\mathbf{N}(\omega,t)\\ \approx& \mathbf{H}_{\mathrm{s}}(\omega)+\mathbf{H}_{\mathrm{d}}(\omega,t)-j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{s}}(\omega)-j\theta_{\mathrm{n}}(t)\mathbf{H}_{\mathrm{d}}(\omega,t)+\mathbf{N}(\omega,t).\end{aligned} \tag{5}$$

depends on the channel gain ratio between the static paths and the dynamic paths, but is not related to the SNR. That means, we cannot reduce the phase noise impact by simply increasing the transmit power.

III. Preprocessing

For LTE signal, after cell search and time-frequency synchronization by PSS and SSS. We can generate local CRS sequence and get the position of received CRS. If full 20 MHz bandwidth is used, there are $N_{\mathrm{CRS}} = 200$ subcarriers occupied by CRS in each symbol. Then we can estimate the channel response for each antenna port by, for example, the least square (LS) algorithm.

Channel estimation is usually implemented in frequency domain, and will be used in frequency domain for channel equalization and decoding. To improve the SNR of estimated channel response, inverse Fourier transform (IFFT) and time windowing are often used, where the frequency response is first transformed to time domain, in which noise samples beyond the multipath delay spread are removed, and then transformed back to frequency domain. But for the gesture recognition task, the required information is involved in the dynamic channel response in time domain, thus it is not necessary to transform back to frequency domain again.

In indoor environments, the distance differences between LOS path and strong reflection paths are usually smaller than 15 meters, which is the distance resolution of LTE signal with 20 MHz bandwidth. So the channel energy is highly concentrated in the strongest path, which is defined as main path. Since the hand movement is usually not far away from the receiver, we consider that the dynamic path is not separable with the main path. By extracting the main path in time domain, the SNR of dynamic path can be improved. This assumption does not hold if the delay difference between the dynamic path and the strongest path is longer than the signal resolution interval, say in outdoor environments or employing signals with much higher bandwidth. But the idea still holds, in those cases, that we can extract the time domain path with the most amplitude variation.

In this work, we use a $N_{\mathrm{IFFT}} = 256$ points IFFT with 200 CRS subcarriers and 56 zero padding. By extracting the strongest path, we obtain one samples in a subframe for each Tx-Rx pair, so the sampling interval T_s is 1 ms. This time domain channel is represented as $\hat{\mathbf{H}}(t) \in \mathbb{C}^{N_r \times N_t}$. In the rest of this paper, all operations are implemented in time domain, and we use $\hat{\mathbf{h}}(t) \in \mathbb{C}^{N_r N_t \times 1}$ to stand for the channel estimation result of the main path, which is a vectorized form of $\hat{\mathbf{H}}(t)$.

After initial synchronization, the residual CFO and STO still need to be tracked. CFO will induce phase fluctuation for the channel response, and STO will induce main path drift and dramatic SNR drop. The phase tracking and delay tracking are thus operated in two different loops. Timing drift is gradual and non-integer, and is thus estimated and compensated in frequency domain [35].

The static and dynamic channel response can be separated by long and short term smoothing [34], which is proved to be an effective and simple method. We can get the static channel response as

$$\hat{\mathbf{h}}_{\mathrm{s}}(t) = F(\hat{\mathbf{h}}(t), L_{\mathrm{long}}), \tag{9}$$

where $F(\cdot)$ is a smoothing filter with window length L_{long}. Although static path should be constant, there is still some slow environment and circuit changes. Such change has impact on dynamic path extraction because dynamic path is much smaller than static paths, thus the long term smoothing is used to track such slow channel changes and also smooth the fast changed channel. The dynamic channel is the rest part of time domain channel, i.e.,

$$\hat{\mathbf{h}}_{\mathrm{d}}(t) = F(\hat{\mathbf{h}}(t), L_{\mathrm{short}}) - \hat{\mathbf{h}}_{\mathrm{s}}(t), \tag{10}$$

where the short term smoothing aims to reduce the white noise. The selection of L_{long} and L_{short} depends on the expected target movement speed. In this work, we pick $L_{\mathrm{long}} = 101$ and $L_{\mathrm{short}} = 21$, taking into consideration the channel variation periods of static paths and normal hand movement speed. If we hope to track a slow movement behavior, like respiration, we can select larger L_{long} and L_{short}.

IV. Spatial Channel Estimation

To recognize gestures, we need to estimate the channel response of dynamic paths, i.e., the reflection paths of transmitter-user-receiver, not like in wireless communications where we estimate the channel response of transmitter to receiver. Since the reflection of hand is weak, the amplitude of dynamic path is usually far below that of other static paths. According to Eq.(8), the phase noise will cause power leakage of static paths, and severely deteriorate the estimation of dynamic paths.

Furthermore, we know from Eq.(2) that the dynamic channel response is the superposition of reflection paths from multiple users. Different from wireless communications, there is no orthogonal reference sequence to help distinguish different users. When two users perform gestures simultaneously, the channel responses between them may have strong correlation, and there is no method to separate them precisely. Therefore, we propose to use preamble gesture method to solve this problem. Before formal gesture recognition, each user perform a preamble gesture independently. The preamble gesture has a special variation pattern in Doppler-time domains. If two users perform preamble gestures separately in different time, we can clearly see the pattern, so that we can make sure that

it is an effective channel estimation. Otherwise, if two users perform preamble gestures simultaneously, or there are other interference activities when one user performs, the interference will cause distortion on this pattern, and we can find the collision and ask the user to perform again.

When one user performs the preamble gesture, the main interference is the power leakage of static paths, we will use a two-layer method to estimate the spatial channel. The Doppler variation pattern should be verified after spatial beamforming. Thus, in this section we will first introduce the two-layer spatial channel estimation method, and then introduce the preamble gesture design and verification method.

A. Two-Layer Channel Estimation

After the separation of static and dynamic paths, according to Eq.(8), the filtered dynamic path encounters strong power leakage from the static path caused by phase noise. In time domain, it can be written as

$$\hat{\mathbf{h}}_{\mathrm{d}}(t) \approx \mathbf{h}_{\mathrm{d}}(t) - j\theta_{\mathrm{n}}(t)\mathbf{h}_{\mathrm{s}} + \mathbf{n}(t), \tag{11}$$

where $\mathbf{h}_{\mathrm{s}} \in \mathbb{C}^{N_r N_t \times 1}$ represents the static channel, and $\mathbf{n}(t)$ is the residual noise with zero mean and covariance matrix $N_0\mathbf{I}_{N_r N_t}$. Although the dynamic channel in time domain changes fast, its array response in spatial domain changes slowly, since the relative position of hand and antennas has small variation in a gesture period. Thus we can inhibit the impact of phase noise and refine the dynamic channel response through spatial domain processing.

In a K_1 points time window, the spatial covariance matrix of the estimated dynamic path is

$$\begin{aligned}\hat{\mathbf{R}}_{\mathrm{d}}(t) &= \frac{1}{K_1}\sum_{k=t-K_1+1}^{t}\hat{\mathbf{h}}_{\mathrm{d}}(k)\hat{\mathbf{h}}_{\mathrm{d}}^{H}(k) \\ &= \frac{1}{K_1}\sum_{k=t-K_1+1}^{t}\big[\mathbf{h}_{\mathrm{d}}(k)\mathbf{h}_{\mathrm{d}}^{H}(k) - j\theta_{\mathrm{n}}(k)\mathbf{h}_{\mathrm{s}}\mathbf{h}_{\mathrm{d}}^{H}(k) + \\ &\quad j\theta_{\mathrm{n}}(k)\mathbf{h}_{\mathrm{d}}(k)\mathbf{h}_{\mathrm{s}}^{H} + |\theta_{\mathrm{n}}(k)|^2\mathbf{h}_{\mathrm{s}}\mathbf{h}_{\mathrm{s}}^{H}\big] + N_0\mathbf{I}. \end{aligned}\tag{12}$$

Since the phase noise $\theta_{\mathrm{n}}(t)$ is a random variable with zero mean, and $\mathbf{h}_{\mathrm{d}}(t)$ and $\theta_{\mathrm{n}}(t)\mathbf{h}_{\mathrm{s}}$ are independent, the time average of their cross term can be assumed as 0. Eq.(12) is then simplified as

$$\begin{aligned}\hat{\mathbf{R}}_{\mathrm{d}}(t) = \frac{1}{K_1}\sum_{k=t-K_1+1}^{t}&\big[\mathbf{h}_{\mathrm{d}}(k)\mathbf{h}_{\mathrm{d}}^{H}(k) + \\ &|\theta_{\mathrm{n}}(k)|^2\mathbf{h}_{\mathrm{s}}\mathbf{h}_{\mathrm{s}}^{H}\big] + N_0\mathbf{I}. \end{aligned}\tag{13}$$

In case of one hand movement, the spatial channel only has minor change during each gesture, by eigenvalue decomposition of $\hat{\mathbf{R}}_{\mathrm{d}}(t)$, there should be two dominant eigenvalues. From experiment, we have observed that the power of static path leakage is much larger than the power of dynamic path. Thus the largest eigenvalue is mainly induced by the second term in Eq.(13), and the strength of dynamic path is reflected by the second largest eigenvalue. We can set a threshold and compare the second eigenvalue with the threshold to judge whether there is a gesture or not. To illustrate the impact of phase noise, an example of $\hat{\mathbf{h}}(t)$, $\hat{\mathbf{h}}_{\mathrm{d}}(t)$ and eigenvalues of $\hat{\mathbf{R}}_{\mathrm{d}}(t)$ is shown in Fig. 2(a), (b), and (c).

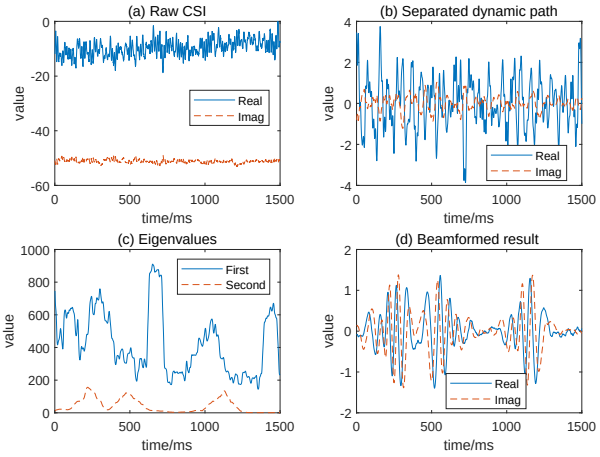


Fig. 2. An example of CSI waveforms and eigenvalues when an approaching-departing-approaching gesture is performed with $N_r = 4$, $N_t = 1$. (a) Waveform of the raw CSI in one antenna. (b) Waveform of the separated dynamic path after long and short term smoothing, where the dynamic path is submerged in static path leakage. (c) When a gesture is performed, there are two large eigenvalues, where the first one indicates the power of static path leakage and the second one indicates the power of dynamic path. (d) After beamforming, the influence of phase noise is removed and we can see a clear waveform of the dynamic channel response.

Obviously, suppress the static channel in spatial domain can eliminate the effect of phase noise. We can first look for the null space of the static channel, and then project the contaminated dynamic channel response into this null space.

The null space of static channel $\mathbf{h}_{\mathrm{s}}$ can be calculated from its spatial covariance matrix $\hat{\mathbf{R}}_{\mathrm{s}}(t)$, i.e.,

$$\hat{\mathbf{R}}_{\mathrm{s}}(t) = \frac{1}{K_2}\sum_{k=t-K_2+1}^{t}\hat{\mathbf{h}}_{\mathrm{s}}(k)\hat{\mathbf{h}}_{\mathrm{s}}^{H}(k). \tag{14}$$

Since the static channel is not completely static, the covariance matrix $\hat{\mathbf{R}}_{\mathrm{s}}(t)$ is still changing with time. But we can use a longer integration period K_2 for $\hat{\mathbf{R}}_{\mathrm{s}}(t)$ than for $\hat{\mathbf{R}}_{\mathrm{d}}(t)$. For example, in this work, we use every 1000 samples to update $\hat{\mathbf{R}}_{\mathrm{s}}(t)$ which counts the slow variation of the static channel in 1 second, and use the last 100 samples to calculate $\hat{\mathbf{R}}_{\mathrm{d}}(t)$, which can reflect the fast variation of the dynamic channel in 100 milliseconds.

With $\hat{\mathbf{R}}_{\mathrm{s}}(t)$, we can get the signal subspace and null space of the static channel. Since $\hat{\mathbf{R}}_{\mathrm{s}}(t)$ is nearly a rank-1 matrix, by eigenvalue decomposition, its signal subspace corresponds to the eigenvector with the largest eigenvalue, and its null space consists of the remaining $N_r N_t - 1$ eigenvectors. Denote the null space matrix as $\mathbf{W}_0 \in \mathbb{C}^{N_r N_t \times (N_r N_t - 1)}$, the projection of the dynamic channel is

$$\hat{\mathbf{h}}_{\mathrm{d1}}(t) = \mathbf{W}_0^{H}\hat{\mathbf{h}}_{\mathrm{d}}(t) \approx \mathbf{W}_0^{H}\left[\mathbf{h}_{\mathrm{d}}(t) + \mathbf{n}(t)\right]. \tag{15}$$

We can see that, after projection, the impact of static channel leakage is removed.

So we use $\mathbf{W}_0$ as the first layer weighting matrix, and we need to search the direction of the dynamic path in the second

layer. To find the subspace where the power of the dynamic path is concentrated on, we need construct the covariance matrix of $\hat{\mathbf{h}}_{\mathrm{d1}}(t)$ and choose the eigenvector corresponding to the largest eigenvalue.

The dynamic path only exists when a gesture is performed, which can be recognized by the eigenvalue variation. Setting an eigenvalue threshold λ_{th}, the time period involving hand movement can be filtered out as

$$T_{\mathrm{motion}} = \left\{ t | \hat{\lambda}_{\mathrm{d},2}(t) > \lambda_{\mathrm{th}} \right\}, \tag{16}$$

where $\hat{\lambda}_{\mathrm{d},i}(t)$ denotes the i^{th} biggest eigenvalue of $\hat{\mathbf{R}}_{\mathrm{d}}(t)$. The threshold λ_{th} can be set to a multiple of the noise variance, say $10N_0$.

With T_{motion}, we extract corresponding channel response set by

$$\mathbf{H}_{\mathrm{d}} = concat \left\{ \hat{\mathbf{h}}_{\mathrm{d}}(t) | t \in T_{\mathrm{motion}} \right\}, \tag{17}$$

and

$$\mathbf{H}_{\mathrm{d1}} = concat \left\{ \hat{\mathbf{h}}_{\mathrm{d1}}(t) | t \in T_{\mathrm{motion}} \right\}, \tag{18}$$

where $\mathbf{H}_{\mathrm{d}} \in \mathbb{C}^{N_r N_t \times L_{\mathrm{d}}}$, $\mathbf{H}_{\mathrm{d1}} \in \mathbb{C}^{(N_r N_t - 1) \times L_{\mathrm{d}}}$, L_{d} is the number of sampling points in time period T_{motion}, $concat$ is the concatenation operation.

The covariance matrix of $\hat{\mathbf{h}}_{\mathrm{d1}}(t)$ is then calculated as $\mathbf{H}_{\mathrm{d1}}\mathbf{H}_{\mathrm{d1}}^H$, and we can use its first eigenvector as the second layer subspace vector $\mathbf{w}_1 \in \mathbb{C}^{(N_r N_t - 1) \times 1}$. The overall spatial channel is the production of the first and second layer matrix/vector,

$$\mathbf{h}_{\mathrm{u}} = \mathbf{W}_0 \mathbf{w}_1. \tag{19}$$

Since $\mathbf{W}_0$ and $\mathbf{w}_1$ are generated by eigenvalue decomposition, $\mathbf{h}_{\mathrm{u}}$ has constant norm. If only for single user gesture recognition, we can directly use this estimated spatial channel to combine the MIMO channel, so the refined dynamic channel response is beamformed as

$$\hat{h_{\mathrm{d}}}(t) = \mathbf{h}_{\mathrm{u}}^H \hat{\mathbf{h}}_{\mathrm{d}}(t). \tag{20}$$

With this two-layer channel estimation and beamforming method, we suppress the leakage of static channel in the first layer, and concentrate the power of dynamic channel in the second layer. After beamforming, the CSI waveform of the dynamic path can be significantly improved, as can be seen in Fig. 2(d).

B. Preamble Gesture and Channel Verification

As shown in Fig. 3, the preamble gesture is designed as two consecutive approaching-departing or departing-approaching motions. For each approaching or departing motion, there are two stages of movements where the first stage is acceleration and the second stage is deceleration. The speed of hand starts from zero, achieves maximum at the midpoint, and slows down to zero again. So in the approaching stage, the Doppler shift will first increase to maximum, and then decrease to zero; in the departing stage, the Doppler shift will first decrease to negative maximum, and then return to zero. As shown in Fig. 4, the Doppler shift introduced by preamble gesture will behave as a two-period sinusoid curve. From the physical implication we know that this is a special variation pattern, and it is hard to be mimicked by other interference movement. Using this feature, we can calculate the Doppler shift and then fit it with sinusoid curve to verify the preamble gesture.

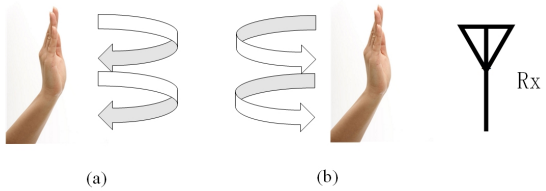


Fig. 3. Preamble gesture.

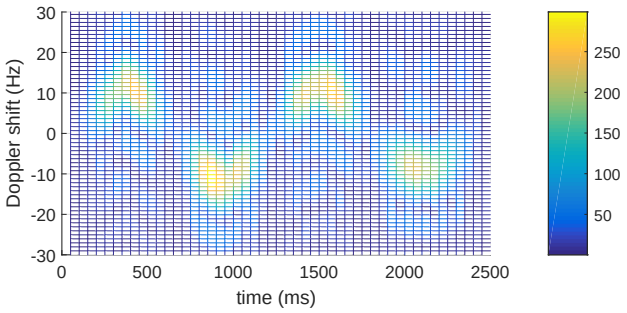


Fig. 4. A Doppler shift spectrum of preamble gesture.

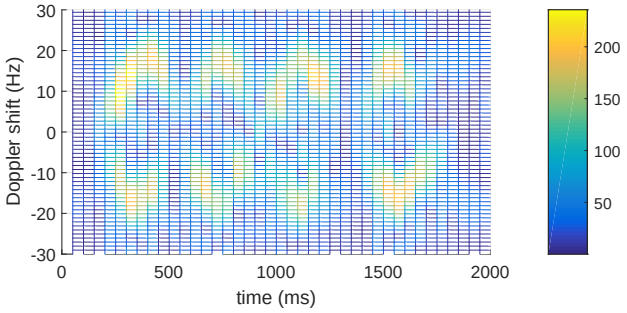


Fig. 5. A distorted Doppler shift spectrum.

In details, we divide the beamformed dynamic channel response $\hat{h_{\mathrm{d}}}(t)$ to short windows with length K_3. There is overlapping between each windows, and the sliding step is $K_3/2$. In each window, we use discrete Fourier transform (DFT) with resolution 1 Hz to calculate the Doppler shift spectrum. We were not using FFT because the frequency resolution of FFT is only $1/(K_3 T_s)$, i.e., 10 Hz for $K_3 = 100$. To reduce the computation complexity of DFT, we only calculate the low frequency part in range of -30 to 30 Hz since the LTE carrier frequency is around 2 GHz and the hand movement speed varies from 0.5 m/s to 1.5 m/s.

After obtaining the Doppler shift spectrum, we can find the frequency shift $\hat{f_D}(n)$ with maximal amplitude at each short window. Assume the n^{th} window starts from sampling point t_n, then

$$\hat{f_D}(n) = \underset{f}{\mathrm{argmax}}\ \mathrm{DFT}\{[\hat{h_{\mathrm{d}}}(t_n), ..., \hat{h_{\mathrm{d}}}(t_n + K_3)]\}. \tag{21}$$

We can first judge whether the sign of $\hat{f_D}(n)$ is consistent with the hand moving direction. Then we will fit $\hat{f_D}(n)$ to a sinusoid curve, and calculate the root-mean-square error (RMSE) of the fitting. If the RMSE is less than a threshold, for example 5 Hz, we can confirm that it is an effective preamble gesture and the spatial channel estimation of user i is thus obtained as $\mathbf{h}_{\mathrm{u},i}$. For example, in Fig. 4, the RMSE of sinusoid fitting is 1.23 Hz. On the contrast, as in Fig. 5, when two users perform preamble gestures simultaneously, there will be large distortion on the Doppler shift variation pattern, where the RMSE of sinusoid fitting is 11.43 Hz. In this case, we will abort the estimation result of this time, and ask the user perform the preamble gesture again.

V. Multiuser Gesture Recognition

In multiuser scenarios, each user keeps its position and performs gesture without any time restriction. When two users perform simultaneously, there will be overlapping on the dynamic channel responses, and will cause inter-user interference. The severity of interference depends on the positions, moving speeds, and moving directions of two users. We hope to distinguish different users and separate their dynamic channel responses in spatial domain. For this purpose, in each time window, we will detect the number of active users, and employ different beamforming schemes according to different number of users. To have explicit analysis results, we only consider two users in this work, but the scheme can be expanded to accommodate more users.

A. Inter-user Interference

A straight forward impact of the inter-user interference is the eigenvalue distribution of the spatial covariance matrix. The number of large eigenvalues is directly related with the number of active users. But the dynamic channel responses of different users are correlated, and higher correlation enlarges the eigenvalue dispersion and makes the users indistinguishable. The correlation depends on many factors including the geometry relation of the transmitter, receiver and users, the hand moving speeds and directions, the spatial channel of reflected signals, etc. We will have a theoretical analysis on the correlation characteristics and eigenvalue distribution due to these factors.

Using the same assumption of Eq.(2), the spatial channel response of the dynamic path for user i is

$$\mathbf{h}_{\mathrm{d},i}(t) = a_i(t)e^{-j2\pi\frac{d_i(t)}{\lambda}}\mathbf{a}(\theta_{\mathrm{r},i},\theta_{\mathrm{t},i}), \tag{22}$$

where $\mathbf{a}(\theta_{\mathrm{r},i},\theta_{\mathrm{t},i}) \in \mathbb{C}^{N_rN_t\times 1}$ denotes vectorized spatial channel response $\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{r},i})\mathbf{a}_{\mathrm{T}}^H(\theta_{\mathrm{t},i})$.

In a two-user scenario, without considering the impact of phase noise, the covariance matrix of the dynamic channel response is represented in Eq.(23). $\overline{h}_{\mathrm{d},1}(t)$ is the conjugate of $h_{\mathrm{d},1}(t)$. We define $\lambda_{\mathrm{d},i}(t)$ as the i^{th} biggest eigenvalue in $\mathbf{R}_{\mathrm{d}}(t)$, and define the eigenvalue ratio as

$$r_{\mathrm{d}}(t) = \frac{\lambda_{\mathrm{d},2}(t)}{\lambda_{\mathrm{d},1}(t)}. \tag{24}$$

To make the second user detectable, $\lambda_{\mathrm{d},2}(t)$ should be notably larger than the noise power. But in real applications, only considering this relationship is not enough. During one gesture period, the arm actually moves with the hand, and the head and body may also have adjoint swings. There may be secondary reflections in the environment and user's spatial channel may have minor change during the hand movement. These imperfections might contribute to a second eigenvalue larger than the noise power. Hence we need to constrain $r_{\mathrm{d}}(t)$ as well, to ensure that $\lambda_{\mathrm{d},2}(t)$ is not less than a given portion of $\lambda_{\mathrm{d},1}(t)$, and there is a real second user.

For convenience, we define the middle part matrix of Eq.(23) as $\mathbf{R}_{\text{time}}(t)$ and define $\rho_{\text{time}}(t) = E\left\{h_{\mathrm{d},1}(t)\overline{h}_{\mathrm{d},2}(t)\right\}$. Assume that the amplitude term $a_i(t)$ in Eq.(22) keeps invariant in a short time window K_1, and its value equals one, thus

$$\mathbf{R}_{\text{time}}(t) = \begin{bmatrix} 1 & \rho_{\text{time}}(t) \\ \overline{\rho}_{\text{time}}(t) & 1 \end{bmatrix}. \tag{25}$$

The ratio of the second and the first eigenvalues of $\mathbf{R}_{\text{time}}(t)$ can be derived as

$$r_{\text{time}}(t) = \frac{\lambda_{\text{time},2}(t)}{\lambda_{\text{time},1}(t)} = \frac{1-|\rho_{\text{time}}(t)|}{1+|\rho_{\text{time}}(t)|}. \tag{26}$$

The time-domain correlation $\rho_{\text{time}}(t)$ can be expressed by the phase variations of two dynamic paths,

$$\begin{aligned}\rho_{\text{time}}(t) &= \frac{1}{K_1}\sum_{k=t-K_1+1}^{t} h_{\mathrm{d},1}(k)\overline{h}_{\mathrm{d},2}(k) \\ &= \frac{1}{K_1}\sum_{k=t-K_1+1}^{t} e^{-j2\pi\frac{d_1(k)-d_2(k)}{\lambda}}.\end{aligned} \tag{27}$$

We can see that the correlation value is determined by the dynamic path length difference $d_1(k)-d_2(k)$, time window length K_1 and the wavelength λ.

To further investigate this issue, we build a geometry model for LOS scenario as in Fig. 6. Points A and B denote the transmitter and receiver, respectively. In a time period K_1T_s, user i moves his hand from point C to point D with a constant speed v_i, and we define the approaching direction as positive. Since $CD \ll CB$, the length of dynamic path changes approximately

$$\begin{aligned}\Delta d_i(t) &= AC + BC - AD - BD \\ &\approx CD(\cos\alpha_i + \cos\beta_i) \\ &= v_iK_1T_{\mathrm{s}}(\cos\alpha_i + \cos\beta_i).\end{aligned} \tag{28}$$

It can be seen that $\Delta d_i(t)$ depends on the moving speed v_i and two angles α_i and β_i. These two angles are determined by moving direction and geometry relation among the transmitter, receiver and user.

Furthermore, define the equivalent speed difference of two users as

$$\Delta v = v_1(\cos\alpha_1 + \cos\beta_1) - v_2(\cos\alpha_2 + \cos\beta_2), \tag{29}$$

the modulus of $\rho_{\text{time}}(t)$ can be simplified as

$$|\rho_{\text{time}}(t)| = \frac{1}{K_1}\left|\sum_{k=0}^{K_1-1} e^{-j\frac{2\pi kT_{\mathrm{s}}\Delta v}{\lambda}}\right| \approx \left|\mathrm{sinc}\frac{K_1T_{\mathrm{s}}\Delta v}{\lambda}\right|. \tag{30}$$

$$\mathbf{R}_{\mathrm{d}}(t) = E\left\{\mathbf{h}_{\mathrm{d}}(t)\mathbf{h}_{\mathrm{d}}^{H}(t)\right\} = \begin{bmatrix}\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) & \mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})\end{bmatrix} \begin{bmatrix} E\left\{|h_{\mathrm{d},1}(t)|^2\right\} & E\left\{h_{\mathrm{d},1}(t)\overline{h}_{\mathrm{d},2}(t)\right\} \\ E\left\{\overline{h}_{\mathrm{d},1}(t)h_{\mathrm{d},2}(t)\right\} & E\left\{|h_{\mathrm{d},2}(t)|^2\right\}\end{bmatrix} \begin{bmatrix}\mathbf{a}^{H}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) \\ \mathbf{a}^{H}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})\end{bmatrix} \quad (23)$$

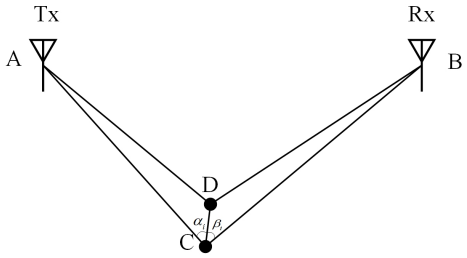


Fig. 6. Geometry model for gesture recognition in LOS scenario.

With given parameters K_1, T_{s} and λ, the time-domain correlation is a sinc function of Δ_v. It has maximal value 1 when $\Delta_v = 0$ and minimal value 0 when $\Delta_v = n\frac{\lambda}{K_1 T_{\mathrm{s}}}$, where n is any non-zero integer. The example curves of $\rho_{\mathrm{time}}(t)$ and $r_{\mathrm{time}}(t)$ are shown in Fig.7.

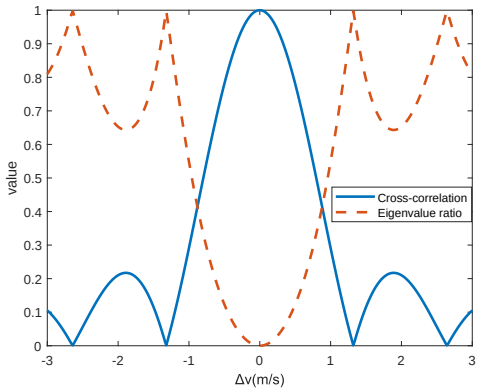


Fig. 7. Curves of $\rho_{\mathrm{time}}(t)$ and $r_{\mathrm{time}}(t)$ with variable Δv, RF frequency 2.27GHz, $K_1 = 100$.

In real applications, the carrier frequency is often fixed and thus λ is fixed. A larger K_1 can help to reduce the mainlobe width. Statistically, a simple gesture like approaching or departing only lasts 200-400 ms. So we choose the observing window parameter $K_1 = 100$. For example, in an LTE system, the carrier frequency is 2.27 GHz, and the corresponding wavelength λ is 0.133 meter. The mainlobe width of $\rho_{\mathrm{time}}(t)$ is $\lambda/(K_1 T_{\mathrm{s}}) = 1.33$ m/s, i.e., the first zero-crossing point of Δv is 1.33 m/s. In 5G millimeter band, the wavelength can be as small as several millimeters, and $|\rho_{\mathrm{time}}(t)|$ will have a very narrow mainlobe.

Considering that the typical hand movement speed is about 1 m/s, according to the given parameters of λ, T_{s} and K_1, there are two cases for $|\Delta v|$. The first is that two users move in opposite directions, which has a relative large $|\Delta v|$. With the effect of α_i and β_i, $|\Delta v|$ may vary from 2 m/s to 4 m/s, and it will fall outside the mainlobe of $|\rho_{\mathrm{time}}(t)|$. The second is that two users move in the same direction. In this case, $|\Delta v|$ is as smaller as 0-1 m/s, and it will fall within the mainlobe of $|\rho_{\mathrm{time}}(t)|$.

As can be seen from Eq.(23), except the time domain correlation characteristic $\rho_{\mathrm{time}}(t)$, spatial channel responses $\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1})$ and $\mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})$ also have influence to the covariance matrix $\mathbf{R}_{\mathrm{d}}(t)$ and its eigenvalue ratio $r_{\mathrm{d}}(t)$ when they are not orthogonal. To further derive the joint effect of the time domain and spatial domain correlations, let us first execute the singular value decomposition (SVD) on the spatial channel matrix as Eq.(31), where $\mathbf{U}$ is a unitary matrix, and $\rho_a = \mathbf{a}^H(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1})\mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})$ is the inner product of two array responses with different AoAs and AoDs. The value of ρ_a reflects the spatial correlation characteristic of two dynamic channels.

Thereafter, $\mathbf{R}_{\mathrm{d}}(t)$ can be represented as in Eq.(32), and we can derive closed-form expressions of the eigenvalues as in Eq.(33), where ϕ is the phase difference between ρ_a and $\rho_{\mathrm{time}}(t)$.

From these explicit expressions, we can see that:

1) If any of ρ_a and $\rho_{\mathrm{time}}(t)$ equals to 1, $\lambda_{\mathrm{d},2}(t)$ is 0. That means, if there is full correlation no matter in spatial domain or in time domain, the eigenspace of two users collapses to dimension one. When $\rho_{\mathrm{time}}(t) \approx 1$, $\mathbf{R}_{\mathrm{d}}(t)$ approximates to a rank-1 matrix, i.e.,

$$\mathbf{R}_{\mathrm{d}}(t) \approx \left[\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) + \mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})\right] \cdot \left[\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) + \mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})\right]^{H}. \quad (34)$$

In this case, we can get a new dominant eigenvector

$$\mathbf{u} = \frac{\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) + \mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})}{||\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) + \mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})||_2}. \quad (35)$$

It is like a single user case, but the equivalent spatial channel response changes.

2) If ρ_a equals to 0, $r_{\mathrm{d}}(t)$ is the same with $r_{\mathrm{time}}(t)$. That means, if the array responses of two users are orthogonal, the eigenvalue ratio only depends on the time-domain correlation characteristic.

Compared with $\rho_{\mathrm{time}}(t)$, ρ_a is more controllable since it is only determined by user position, and it is nearly constant during hand movement. We provide an intuitive example in a widely used $N_r \times 1$ uniform linear array (ULA) deployment, where the antenna spacing is set as half of the wavelength. The array response is formulated as

$$\mathbf{a}_{\mathrm{ULA}}(\theta) = \sqrt{\frac{1}{N_r}}\left[1, e^{-j\pi\cos\theta}, ..., e^{-j\pi(N_r-1)\cos\theta}\right]^{T}. \quad (36)$$

To change ρ_a, we fix one user in direction $\theta_{\mathrm{r},1} = 90°$ and move the other user in range $\theta_{\mathrm{r},2} \in [0°, 180°]$. The eigenvalue ratio results are shown in Fig. 8, where $\theta_{\mathrm{r},2}$ and Δv are independent variables. These results also prove that to get a bigger eigenvalue ratio, $|\rho_a|$ and $|\rho_{\mathrm{time}}(t)|$ should be as small as possible. An outlier is $\theta_{\mathrm{r},2} = 98.4°$ and $\Delta v = -0.373$. In this point, $\rho_a = -\rho_{\mathrm{time}}(t)$, so $r_{\mathrm{d}}(t) = 1$.

$$\begin{bmatrix}\mathbf{a}(\theta_{\mathrm{r},1},\theta_{\mathrm{t},1}) & \mathbf{a}(\theta_{\mathrm{r},2},\theta_{\mathrm{t},2})\end{bmatrix} = \mathbf{U}\begin{bmatrix}\sqrt{1+|\rho_a|} & 0\\ 0 & \sqrt{1-|\rho_a|}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{\rho_a}{\sqrt{2}|\rho_a|} & -\frac{\rho_a}{\sqrt{2}|\rho_a|}\end{bmatrix}^H \tag{31}$$

$$\mathbf{R}_{\mathrm{d}}(t) = \mathbf{U}\begin{bmatrix}(1+|\rho_a|)(2+\frac{\overline{\rho}_a\rho_{\mathrm{time}}(t)}{|\rho_a|}+\frac{\rho_a\overline{\rho}_{\mathrm{time}}(t)}{|\rho_a|}) & \sqrt{1-|\rho_a|^2}(\frac{-\overline{\rho}_a\rho_{\mathrm{time}}(t)}{|\rho_a|}+\frac{\rho_a\overline{\rho}_{\mathrm{time}}(t)}{|\rho_a|}) & \mathbf{0}\\ \sqrt{1-|\rho_a|^2}(\frac{\overline{\rho}_a\rho_{\mathrm{time}}(t)}{|\rho_a|}-\frac{\rho_a\overline{\rho}_{\mathrm{time}}(t)}{|\rho_a|}) & (1-|\rho_a|)(2-\frac{\overline{\rho}_a\rho_{\mathrm{time}}(t)}{|\rho_a|}-\frac{\rho_a\overline{\rho}_{\mathrm{time}}(t)}{|\rho_a|}) & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{U}^H \tag{32}$$

$$\begin{aligned}\lambda_{\mathrm{d},1}(t) &= 1+|\rho_a||\rho_{\mathrm{time}}(t)|\cos\phi+\sqrt{|\rho_a|^2+|\rho_{\mathrm{time}}(t)|^2-|\rho_a|^2|\rho_{\mathrm{time}}(t)|^2\sin^2\phi+2|\rho_a||\rho_{\mathrm{time}}(t)|\cos\phi}\\ \lambda_{\mathrm{d},2}(t) &= 1+|\rho_a||\rho_{\mathrm{time}}(t)|\cos\phi-\sqrt{|\rho_a|^2+|\rho_{\mathrm{time}}(t)|^2-|\rho_a|^2|\rho_{\mathrm{time}}(t)|^2\sin^2\phi+2|\rho_a||\rho_{\mathrm{time}}(t)|\cos\phi}\end{aligned} \tag{33}$$

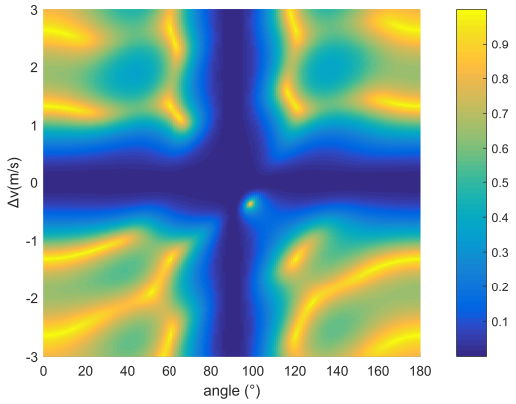


Fig. 8. Eigenvalue ratios with independent variables θ_2 and Δv. In this example, $N_t = 1$, AoA for user 1 is 90°, RF frequency is 2.27GHz, antenna spacing is half of the wavelength, and $K_1 = 100$.

B. Active User Number Detection

Since each user may perform gesture in any time, in each time window with dynamic path, there may be one or two gestures performed. If there is only one user, we only need to inhibit the phase noise impact using the single user beamforming scheme in Eq.(20). If there are two users, we need to inhibit both the phase noise impact and the inter-user interference. So we should first detect the number of active users in each time window.

As aforementioned, there is static path leakage contained in the observed dynamic channel response $\hat{\mathbf{h}}_{\mathrm{d}}(t)$, whose spatial covariance matrix is shown in Eq.(13). The first and second eigenvalues $\lambda_{\mathrm{d},1}(t)$ and $\lambda_{\mathrm{d},2}(t)$ of $\mathbf{R}_{\mathrm{d}}(t)$ actually correspond to the second and third eigenvalues $\hat{\lambda}_{\mathrm{d},2}(t)$ and $\hat{\lambda}_{\mathrm{d},3}(t)$ of $\hat{\mathbf{R}}_{\mathrm{d}}(t)$. In the implementation, we will set an eigenvalue ratio threshold r_{th} and an absolute threshold $\lambda_{\mathrm{th},2}$. For example, $r_{\mathrm{th}} = 0.1$ and $\lambda_{\mathrm{th},2} = \lambda_{\mathrm{th}}/2$, to distinguish real gestures from the noise and disturbance.

When both thresholds are satisfied, we can directly judge that there are two active users. Otherwise, we will further compare the spatial correlation coefficients. From the inter-user interference analysis in Section V-A, we know that when the time-domain correlation of two dynamic paths is high, the eigenspaces of two users collapse to one and we can only observe a very small eigenvalue ratio. In this case, we will first calculate the beamforming vector $\mathbf{w}_{\mathrm{d}}$ by Eq.(19) as if there is only one active user. Then, calculate the correlation coefficients between $\mathbf{w}_{\mathrm{d}}$ and three possible channels

$$\rho_i = |\mathbf{w}_{\mathrm{d}}^H\mathbf{h}_{\mathrm{u},i}|, \tag{37}$$

and

$$\rho_{\mathrm{sum}} = \frac{|\mathbf{w}_{\mathrm{d}}^H(\mathbf{h}_{\mathrm{u},1}+\mathbf{h}_{\mathrm{u},2})|}{\|\mathbf{h}_{\mathrm{u},1}+\mathbf{h}_{\mathrm{u},2}\|_2}, \tag{38}$$

where ρ_i stands for the correlation with user i, and ρ_{sum} stands for the correlation with the superposed channels. If ρ_{sum} is larger than any of ρ_i, we can determine that there are actually two users moving simultaneously.

The complete procedure is summarized in Table I.

TABLE I
USER NUMBER DETECTION

1. Obtain the spatial channel $\mathbf{h}_{\mathrm{u},i}$ of user i using preamble gesture;
2. Calculate $\hat{\mathbf{R}}_{\mathrm{d}}(t)$, $\hat{\lambda}_{\mathrm{d},2}(t)$ and $\hat{\lambda}_{\mathrm{d},3}(t)$;
3. **if** $\frac{\hat{\lambda}_{\mathrm{d},3}(t)}{\hat{\lambda}_{\mathrm{d},2}(t)} > r_{\mathrm{th}}$ **and** $\hat{\lambda}_{\mathrm{d},3}(t) > \lambda_{\mathrm{th},2}$
   There are two active users
   elseif $\hat{\lambda}_{\mathrm{d},2}(t) > \lambda_{\mathrm{th}}$
   Calculate $\mathbf{w}_{\mathrm{d}}$ and correlation coefficients ρ_1, ρ_2, ρ_{sum}
   if $\rho_{\mathrm{sum}} > \max(\rho_1, \rho_2)$
   There are two active users
   else
   There is one active user

C. Beamforming

The beamforming scheme in multiuser case relies on the active user number in each time window. In time windows with only one active user, we still use the single-user two-layer beamforming scheme introduced in Section IV. In time windows with two active users, we should suppress the inter-user interference and separate the dynamic path of each user.

A simple method to do this is using zero-forcing (ZF) beamforming [36], where the beamforming vectors are acquired through pseudo-inverse of the channel matrix.

To mitigate the inter-user interference and meanwhile suppress the impact of phase noise, the constructed channel matrix should not only involve the dynamic channel of each user, but also involve the static channel,

$$\mathbf{H}_{\text{con}} = \left[\mathbf{h}_{\text{u},1} \;\; \mathbf{h}_{\text{u},2} \;\; \hat{\mathbf{h}}_{\text{s}}(t)\right], \tag{39}$$

where $\hat{\mathbf{h}}_{\text{s}}(t)$ is the dominant eigenvector of $\mathbf{R}_{\text{s}}(t)$ and stands for the eigenspace of static channel. Through pseudo-inverse of this matrix, we obtain the beamforming vectors as

$$\left[\mathbf{w}_{\text{zf},1} \;\; \mathbf{w}_{\text{zf},2} \;\; \mathbf{w}_{\text{zf},3}\right]^H = pinv(\mathbf{H}_{\text{con}}), \tag{40}$$

where $pinv$ denotes the Moore-Penrose inverse operation. In Eq.(40), $\mathbf{w}_{\text{zf},1}$ and $\mathbf{w}_{\text{zf},2}$ are ZF beamforming vectors for user 1 and 2 respectively, and $\mathbf{w}_{\text{zf},3}$ is not used. The beamforming vector of each user is orthogonal to the spatial channel of other user, and orthogonal to the static channel as well.

The ZF beamforming result for user i can then be written as

$$\hat{h}_{\text{d},i}(t) = \mathbf{w}_{\text{zf},i}^H \hat{\mathbf{h}}_{\text{d}}(t), \tag{41}$$

where $\hat{\mathbf{h}}_{\text{d}}(t)$ is the original dynamic channel response as defined in Eq.(10), a mixture of dynamic paths of all users and the leakage from the static path. We finally can acquire a clean scalar channel response for the dynamic path of each user.

In Fig. 9, we provide three examples of the beamforming results when two users perform gestures simultaneously. In the first row of Fig. 9, we plot the second and the third eigenvalues of $\hat{\mathbf{R}}_{\text{d}}(t)$. As illustrated in Fig. 9(a), when two users perform gestures with opposite directions, we can observe two evident eigenvalues. If two users perform in the same direction, sometimes we can also find the third eigenvalues as in Fig. 9(d), but this value is relatively small. Occasionally, as in Fig. 9(g), there may be only one dominant eigenvalue. However, in all these cases, by ZF beamforming we can successfully separate two dynamic paths, as can be seen from the waveforms demonstrated in each column.

For two users with close distance, there will be strong interference between users, which affects the performance of gesture recognition. In LOS channel, the angular resolution depends on the number of antennas and antenna spacing. For linear arrays with half wavelength spacing, the angular resolution is about π/N, where N is the antenna number. In NLOS channel, due to multipath scattering, the channel response may have low correlation even that the two users are relatively close, thus with the same array the spatial resolution is usually better than in LOS case. In any cases, increasing antenna number is always effective to achieve better beamforming performance.

D. Gesture Recognition

In this work, to focus on the spatial domain processing and inter-user interference suppression, we only use one receiver and identify the gesture movement in one dimension, i.e., identify the movement direction of approaching towards or departing from the receiver. To recognize more complex gesture, we need more receivers deployed in different locations to form a triangular geometry relation.

The beamforming result $\hat{h}_{\text{d},i}(t)$ is used to calculated the Doppler shift of the dynamic path, as elaborated in Section IV-B. Positive and negative Doppler shifts are corresponding to approaching and departing gestures, respectively. For one user, the difference of hand moving speed only affects the possible values of Doppler shift, and does not affect the sign. For two or more users, however, their difference of velocities do affect the inter-user interference, as we have clarified in Section V-A.

Although we can calculate DFT at 1 Hz interval, the actual resolution actually depends on the width of time window (which is 100 ms in our experiment). That means the frequency response calculated by DFT still has a main lobe with zero-to-zero width 20 Hz. Thus if the Doppler shift is too small, for example only a few Hz, it has chance to make wrong decision on the sign. The small Doppler shift usually happens in the transition point. Thus, we will not confirm the transition unless we observe that the Doppler shift changed its sign for several consecutive time windows. To distinguish the combination gesture like approaching-departing, we set the maximal interval of separate motions as 0.5 second. Beyond this interval, it will be treated as two gestures.

The moving distance of the hand is a quite robust parameter. Given the moving speed, the moving distance only affect the signal lasting time. The provided gesture recognition algorithm is not sensitive to the lasting time, it only looks for the transition point of the Doppler shift sign.

In most cases, the gestures of two users are not synchronized but partially overlapped. We divide the whole waveform into short windows, and the motion number detection, beamforming and Doppler shift estimation are processed independently for each window. So there is no influence of the start timing difference.

VI. Evaluations

A. Experiments Environment and Configuration

To verify the gesture recognition system, we conduct some experiments with single user and multiuser. To make the test condition under control, i.e., no other movement in the environment, we designed a prototype system to transmit and receive LTE signal. It is noteworthy that currently there is no off-the-shelf device to extract CSI for LTE signal. This forces us to build a system by software radio platform to estimate and extract the CSI for validating our method. If LTE-based wireless sensing is applied on mobile phones, CSI acquisition will not be a problem, the baseband module will continually estimate CSI when the terminal is in connected state.

The transmitter is implemented on AD-FMCOMMS2, which is an evaluation board of the 2×2 RF transceiver chip AD9361. AD9361 is a widely-used RF chip in LTE base stations, it has transmission frequency range from 47 MHz to 6 GHz. We transmit 20MHz bandwidth signal with carrier frequency 2.27 GHz, and the maximum transmit power is 10

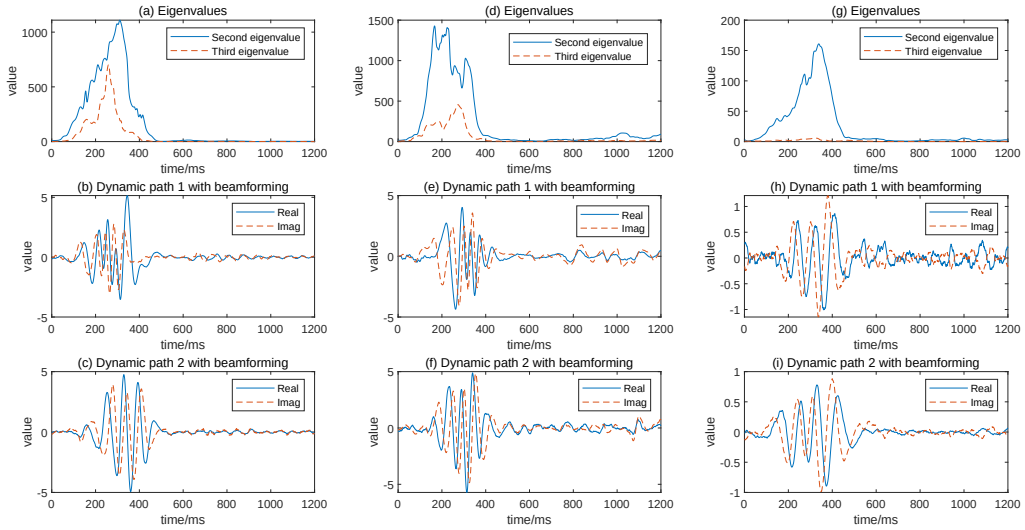


Fig. 9. Examples of multiuser separation by ZF beamforming: (a)(b)(c) user1 departing, user2 approaching; (d)(e)(f) user1 approaching, user2 approaching; (g)(h)(i) user1 departing, user2 departing.

dBmW. For comparisons, the maximum transmit powers of macro-BS and micro-BS in commercial networks are typically 46 dBmW and 30 dBmW, respectively. But the signals received from outdoor base stations generally suffer from penetration loss. We have measured that the received power of our own signal is comparable or even weaker than the signal from commercial network.

In Rx side, we use a software radio platform YunSDR Y550s, which has four channels of RF links and enables us to perform beamforming with four receiving antennas. Antennas are deployed as a ULA with half wavelength spacing. Four-antenna is also a common configuration in cell phones nowadays. Y550s is equipped with two chips of AD9371, and can support 100MHz bandwidth, while in our prototype system only 20MHz is used. The software radio platform implements down-conversion and sampling, and all baseband processing are implemented in the host computer with an Intel Core i7-8700 CPU working at 3.20 GHz. There is an optic fiber link between Y550s and the host computer, and the transmission rate is 4 Gbps for four channels IQ sampling with 30.72 MHz sampling rate.

In order to improve the processing speed, a self-developed C/C++ program is implemented to do LTE physical layer time-frequency synchronization, LS channel estimation and dynamic-static channel separation. The dynamic channel responses are then sent to MATLAB by local loopback network connection, where the beamforming and gesture recognition are accomplished. The whole system is implemented in real-time, and we can observe the CSI waveforms, beamforming patterns, and recognized gestures immediately.

To verify the performance, we select two indoor scenarios, a computer lab (scenario 1) and an apartment (scenario 2) to conduct experiments. In scenario 1, there is LOS path between the transmitter (Tx) and the receiver (Rx); in scenario 2, the Tx and Rx are placed in two rooms and thus it is an NLOS scenario.

We design six gestures to recognize, which are shown in Fig. 10. These gestures are the combinations of two basic gestures, approaching or departing the transmitter or receiver. To verify the robustness of our recognition algorithm, we invite seven volunteers to do the experiments. Volunteers may perform gestures with their own habits, e.g., with different timings, velocities and hand shapes.

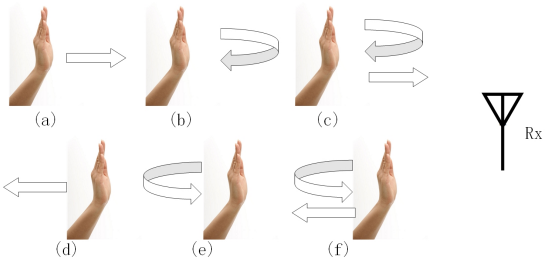


Fig. 10. Six gestures to be recognized.

B. Performance

1) Scenario 1: In scenario 1, except for accuracy rate, we will study the influence of AoA, reflection path distance and antenna number to single user and multiuser recognition. As shown in Fig. 11, the LOS path length of Tx and Rx is set to 3 m, and we select eight positions to perform gestures as marked. Positions 1-4 and 5-7 are distributed in two ellipses with two focal points located at the Tx and Rx. Thus, these two groups of positions have nearly the same dynamic path length and signal path loss. In addition, the LOS path between Tx and Rx has AoA 0°, which has orthogonal spatial channel response with position 1 and position 2 theoretically. Positions

TABLE II
POSITION PARAMETERS

Position	AoA (°)	Dynamic path length (m)
Pos1	60	5
Pos2	90	5
Pos3	135	5
Pos4	180	5
Pos5	60	7.5
Pos6	75	7.5
Pos7	90	7.5
Pos8	90	10

2, 7, 8 are distributed in the same column with AoA 90°. The distances and AoAs for all positions are listed in Table II.

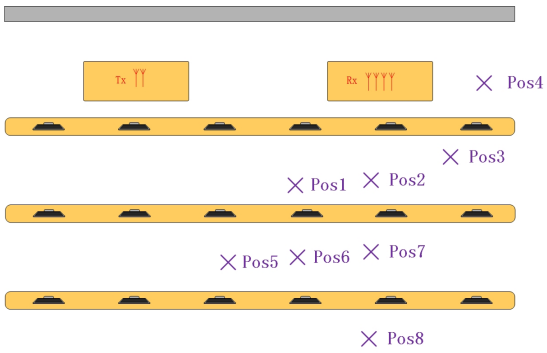


Fig. 11. The layout of a computer lab and user positions.

We first evaluate the performance of single user case, where the user stands in positions 1-4 and 7, 8 in Fig. 11. In each position, we perform six groups of gestures, each group consists of 30 candidate gestures, which is generated randomly. In the beginning of each group, we perform a preamble gesture to estimate the spatial channel and indicate the start of recognition.

The performance metrics are dynamic path SNR and gesture recognition accuracy. We use dynamic path SNR because it is a comprehensive result of transmitting power, path loss, receiver front-end noise, beamforming and phase noise suppression. To measure the SNR, we should separate motion time period T_{motion} and static time period T_{static} to calculate signal and noise power respectively. The existence of a motion is determined by Eq.(16). Outside the period of T_{motion}, we set 200 ms as the transition period, then the rest time is considered as T_{static}. The extraction of dynamic channel set $\mathbf{H}_{\text{d}}$ has been introduced in Eq.(17), and the corresponding channel set $\mathbf{H}_{\text{d,static}}$ is acquired by substituting T_{motion} to T_{static} in Eq.(17). Assume the sampling number in $\mathbf{H}_{\text{d}}$ and $\mathbf{H}_{\text{d,static}}$ are L_{d} and L_{s}, respectively. Then the SNR is calculated as

$$\text{SNR} = \frac{\|\mathbf{w}_{\text{d}}^H \mathbf{H}_{\text{d}}\|_2^2 / L_{\text{d}}}{\|\mathbf{w}_{\text{d}}^H \mathbf{H}_{\text{d,static}}\|_2^2 / L_{\text{s}}}. \tag{42}$$

The average SNR results are provided in Table III. However, in experiments we have observed that subtle changes like hand shape, speed and height will have influence on the SNR. To better present the results, we plot the cumulative distribution function (CDF) curves in Fig. 12. For each position, we get one CDF curve by the statistic of 180 times of hand gestures.

TABLE III
AVERAGE SNR

Position	SNR (dB)
Pos1	19.6
Pos2	23.8
Pos3	21.5
Pos4	17.0
Pos7	19.0
Pos8	17.6

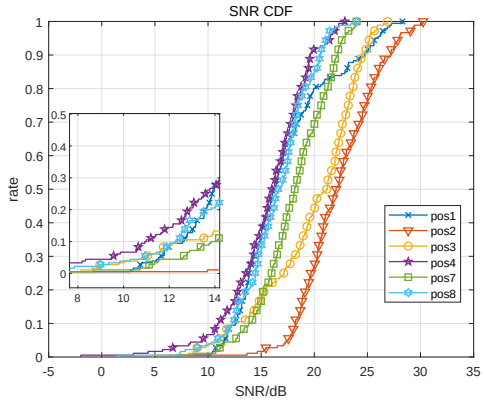


Fig. 12. CDF of the SNR for users in different positions.

From these results we can see that, SNR is good in all positions. This should attribute to IFFT and beamforming processing. The SNR is affected by both distance and AoA. For positions 2, 7 and 8, there is no doubt that the SNR reduces when the path length increases. While comparing the results of positions 1, 2, 3, 4, position 4 has the worst SNR. Position 4 is at the line of Tx-Rx LOS path, theoretically position 4 is at the blind area. However, from the experiment result, we can still conduct gesture recognition in this area and obtain a relatively high average SNR. We believe that there are two reasons to explain this phenomenon. The first is that, in indoor multipath environments, static path is not equivalent to LOS path, there are many other static reflections from different angles. The second is that, there is hand movement direction bias each time, it is impossible for the hand to have strictly consistent AoA with the LOS path. From the CDF curves of SNR, we find that position 4 has nearly 6% gestures with SNR lower than 10 dB. This proves that in blind area, SNR fluctuates with minor change of hand movement, and it has more chances to get low SNRs.

Recognition accuracy is the most important indicator for this system. In our test, both the wrong recognition results and false alarms are counted as errors. The accuracy rates are shown in Table IV. In all positions, our accuracy is higher than 96%. Although position 2 has the highest average SNR, there is no evident accuracy advantage compared with positions 1, 3 and 7, since most error detection cases happen when SNR

TABLE IV
SINGLE-USER RECOGNITION ACCURACY

Position	Accuracy (%)
Pos1	99.4
Pos2	99.4
Pos3	99.4
Pos4	96.1
Pos7	99.4
Pos8	98.3

is extremely low. Position 4 only has minor disadvantage in average SNR than position 8. But with more lower SNR cases, it has 2.2% gap in accuracy.

Next, let us evaluate the performance in multiuser scenarios. We will illustrate the beam patterns, SINRs and recognition accuracies under different kinds of gesture and user position combinations.

Beam pattern is a straightforward way to observe the effect of beamforming. To observe the beam pattern, we use the AoA of LOS path to calibrate the antenna array and remove the random phase offset on each element. After obtaining the beamforming vectors in Eq.(40), we can calculate the beamforming gain for user i at each direction θ, i.e.,

$$g_i(\theta) = |\mathbf{w}_{\mathrm{zf},i}^H \mathbf{a}_{\mathrm{R}}(\theta)|^2. \tag{43}$$

The beam pattern can thus be drawn as in Fig. 13. Since there are two users, the ZF algorithm forms a main lobe at the target user's direction to maximize signal power, and generates a null at the interference user's direction to suppress interference. For the beam patterns of both users, there are also nulls at the direction of Tx-Rx LOS path, which is at 0° in Fig. 13, to mitigate the impact of static path power leakage.

To measure the SINR of the multiuser beamforming, we need to first extract the dynamic channel set $\mathbf{H}_{\mathrm{d},1} \in \mathbb{C}^{N_r N_t \times L_{\mathrm{d},1}}$ and $\mathbf{H}_{\mathrm{d},2} \in \mathbb{C}^{N_r N_t \times L_{\mathrm{d},2}}$ separately, when the two users perform gestures one by one. Otherwise, the dynamic channel responses are mixed up, we cannot get the accurate values of signal and interference power. However, as long as other conditions are fixed, the obtained SINR estimation is actually the SINR when the two users perform gestures simultaneously. Taking user 1 as an example, the SINR is calculated by

$$\mathrm{SINR}_1 = \frac{\|\mathbf{w}_{\mathrm{zf},1}^H \mathbf{H}_{\mathrm{d},1}\|_2^2 / L_{\mathrm{d},1}}{\|\mathbf{w}_{\mathrm{zf},1}^H \mathbf{H}_{\mathrm{d},2}\|_2^2 / L_{\mathrm{d},2}}, \tag{44}$$

where the numerator represents the mean power of the target user, and the denominator represents the residual interference after ZF beamforming. In Table V, we present the average SINR under different user position combinations. In each combination, the SINRs of both users are given, and the correlation coefficient ρ_{a} of two array responses is also listed. In Fig.14, we draw the CDF curves of SINR corresponding to three groups of position combinations.

From the results we can see that, for all combinations the average SINR can achieve at least 13 dB. Compared with the SNR results in single user case, there is 4 to 9 dB degradation.

TABLE V
SINR RESULT

SINR (dB) \ Position / Combination	Pos1	Pos2	Pos3	$\lvert\rho_{\mathrm{a}}\rvert$
1 (Pos1, Pos2)	15.5	18.1	-	0.306
2 (Pos2, Pos3)	-	16.4	13.0	0.576
3 (Pos1, Pos3)	13.2	-	17.5	0.365

TABLE VI
MULTIUSER RECOGNITION ACCURACY IN SCENARIO 1

Accuracy (%) \ Method / Combination	Proposed	AoA
Pos1, Pos2	97.5	97.5
Pos2, Pos3	97.5	96.4
Pos1, Pos3	96.1	93.6

The SINR degradation is the worst in the second combination, where users in position 2 and 3 have the highest spatial channel correlation. To suppress the inter-user interference, ZF beamformer may deviate its main lobe from the signal direction, and thus reduce the signal power. Basically, the signal power decreases along with the increasing of the spatial channel correlation. Besides, since there is small change in each time of hand movement, the channel responses used in beamforming matrix design and subsequent detections are not totally invariant. This kind of mismatching will also enlarge the inter-user interference. Our measured SINR is the result of these comprehensive factors.

Let us see the recognition accuracy rate in multiuser cases. In our experiments, we provide several position combinations. In each combination, every user still perform six groups of gestures with 30 randomly generated gestures in each group. After preamble gestures, users can either perform gesture simultaneously or separately. To show the performance of our algorithm, we select an AoA based multiuser separation method as baseline [4]. It is worth noting that due to the difference in hardware and test scenario, we cannot reproduce the overall system of [4]. To compare the performance of different methods, we substitute the estimated spatial channel by an AoA based beamsteering vector, as obtained in Eq.(36), to implement the beamforming. The rest parts are the same. The recognition accuracy rates are shown in Table VI. We can see that the accuracy rates in multiuser cases are slightly lower than in single user cases, but are still above 96%. In this scenario, the proposed method is slightly better than AoA based multiuser separation. But the performance gap is small because in this scenario the spatial channel mainly consists of LOS path between the hand and receiver.

In the end, we study the influence of antenna number. We use one and two Tx antennas to compare the accuracy. In this experiment, we use position 5, 6 and 7, where position 6 are quite close to the other two, their AoA difference is only 15° and their distance is about 1 m. For fair comparison, we run

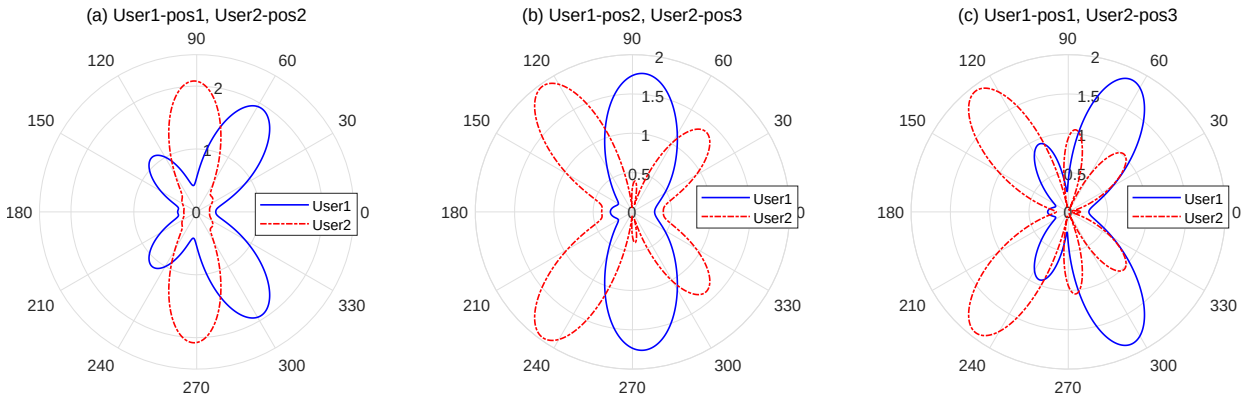


Fig. 13. Beam patterns in multiuser scenarios, where user 1 and user 2 are at different positions and have different AoAs to the receiver. The radial distance in the polar coordinates stands for the beamforming gain in linear scale.

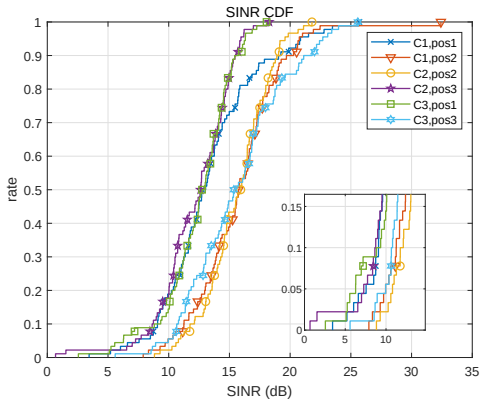


Fig. 14. CDF of the SINR for users in different combinations (C1: pos1 and pos2, C2: pos2 and pos3, C3: pos1 and pos3).

TABLE VII
MULTIUSER RECOGNITION ACCURACY WITH DIFFERENT ANTENNA NUMBER

Accuracy (%) Antenna / Combination	$N_t = 1$	$N_t = 2$
Pos5, Pos7	97.5	98.6
Pos5, Pos6	91.4	98.1
Pos6, Pos7	84.7	97.5

our real-time test and count accuracies with $N_t = 2$, and the 4×2 MIMO channels are saved so that we can reuse the same channel to calculate accuracies for $N_t = 1$. The overall accuracy results are shown in Table VII. When $N_t = 1$, the recognition accuracies obviously decrease when users are quite close, but with $N_t = 2$, all three position combinations show significant performance improvement. It also turns out that by increasing antenna number, we can overcome the restriction of spatial distance of users.

2) Scenario 2: As shown in Fig. 15, in the apartment scenario, the straight-line distance of Tx and Rx is 2.5 meters. The users sit on sofa or bed, in the same room or in different rooms. The NLOS configuration is more realistic for home applications. Besides, the transmission signal and hand reflection signal should go through a concrete wall, which leads to a severe penetration loss. Before our gesture experiments, we first measure the received signal strength and compare it with that in scenario 1. We use the frequency scan function in our hardware to measure the spectrum in these two environments. We find that in NLOS scenario, the received power drops about 20 dB.

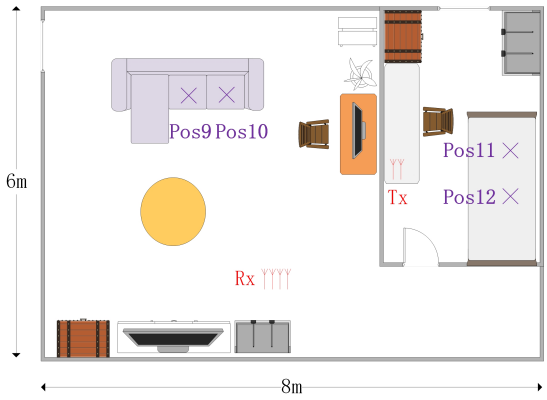


Fig. 15. The layout of an apartment and user positions.

Thus, we use 4×2 MIMO configurations to improve the detection performance. In this scenario, we only test the accuracy rate. As there are antenna arrays equipped in both the receiver and transmitter, we estimate the dominant AoA and AoD in the baseline method and recover the MIMO array response. We try three typical position combinations, corresponding to three common situations in daily lives:

1) users sit closely on the sofa;
2) users sit closely in the bedroom;
3) one user sits on the sofa, the another one sits in the bedroom.

The recognition accuracies are shown in Table VIII. From these results we can see that, in NLOS scenario the proposed method can still reach high accuracy. For all three position combinations the recognition accuracies exceed 94%. By comparision, AoA/AoD based method has performance degradation in this scenario. With NLOS, there is no dominant reflection path, which makes it hard to estimate the AoA/AoD.

TABLE VIII
MULTIUSER RECOGNITION ACCURACY IN SCENARIO 2

Accuracy (%) Method / Combination	Proposed	AoA/AoD
Pos9, Pos10	94.7	91.9
Pos11, Pos12	95.0	83.1
Pos10, Pos12	95.6	86.4

Even that we can find the AoA/AoD in the strongest path, there must be some biases between the complete spatial channel and the beamsteering vector.

Compared with LOS scenario, for the proposed method the performance drop is small despite the received signal power reduces dramatically. We believe that there are two reasons: 1) the power of hand reflection path is enhanced by beamforming with 4×2 antennas; 2) as we have proved in Eq.(5)-(8), the biggest non-ideal factor is multiplicative phase noise instead of additive white noise, and the influence of phase noise is not related to SNR.

C. Complexity Analysis

As our gesture recognition algorithm runs on a real-time system, computational complexity of each step is critical. We analyze the complexity issues from three parts. The first part is the simplified LTE raw signal processing. Since we only use one symbol in each subframe (1ms), the main complexity consists of a $N_{\text{FFT}} = 2048$ points FFT, frequency offset compensation and LS channel estimation. The second part includes preprocessing, spatial domain eigenvalue decomposition and beamforming. These steps are also required in every milliseconds. The last part is Doppler shift calculation made by DFT in N_K frequency points. When no gesture performed, this operation is skipped. Table IX shows the complexity of each module averaged in every 1 ms.

From operation time perspective, with 4×2 MIMO configuration, in C/C++ program we spent 2.1 ms in average to process a frame of LTE signal (10 ms), in which the FFT and LS channel estimation cost 0.6 ms in total, frequency offset compensation costs 1.2 ms, and the other processes cost 0.3 ms. The running time of frequency offset compensation is much longer than we analyzed, because we used sine and cosine function to calculate the compensation values, which is time-consuming for CPU implementation. If we use look-up tables, the complexity can be greatly reduced. In MATLAB, for each time window (100 ms), the gesture recognition algorithm including eigenvalue decomposition, beamforming and Doppler shift calculation costs less than 1.5 ms in total.

With this analysis, we find that the biggest computational burden comes from LTE signal processing. The gesture recognition algorithm has much less complexity. To apply gesture recognition in a mobile phone, this part does not introduce much extra complexities, because there is baseband chip demodulating LTE signal and acquiring CSI in every subframe.

TABLE IX
COMPLEXITY EVALUATION

Module	Complexity
Raw data FFT	$\mathcal{O}(N_r N_{\text{FFT}} \log_2(N_{\text{FFT}}))$
Frequency synchronization	$\mathcal{O}(N_r N_{\text{FFT}})$
LS channel estimation	$\mathcal{O}(N_r N_t N_{\text{CRS}})$
Channel IFFT	$\mathcal{O}(N_r N_t N_{\text{IFFT}} \log_2(N_{\text{IFFT}}))$
Long and short smoothing	$\mathcal{O}((L_{\text{long}} + L_{\text{short}}) N_r N_t)$
Eigenvalue decomposition	$\mathcal{O}((N_r N_t)^3)$
Beamforming	$\mathcal{O}(N_r N_t)$
Doppler shift DFT	$\mathcal{O}(N_K)$

VII. Conclusion

This paper proposed a beamforming based multiuser wireless gesture recognition method, and built a prototype system using LTE signals to verify the performance. Firstly, to estimate the spatial channel of reflection path and solve the static path leakage interference caused by phase noise, a preamble gesture is introduced and a two-layer channel estimation method was proposed. The first layer suppresses the interference caused by static path leakage, and the second layer collects power of the dynamic path. Then, in multiuser scenario, the influencing mechanisms of inter-user interference were clearly analyzed. By calculating the correlations of dynamic channel responses in time and spatial domain, we derived the eigenvalue distribution of spatial covariance matrix, and found its connection with user positions and movement speeds. Then, we proposed a beamforming method that can suppress the influence of inter-user interference and phase noise at the same time. Finally, we demonstrated experiment results on beam patterns, SINRs and recognition accuracies in different configurations, and verified the good performance of the proposed method in both LOS and NLOS scenarios. As a real-time implementation, the computational complexity of each module is also analyzed.

In the future, there are many interesting works to further improve the system performance. For example, recognizing complicated gestures by introducing more receivers, suppressing interferences caused by unknown movements in background, and upgrading the experiment platform to exploit 5G new radio (NR) signals.

References

[1] Z. Zhou, C. Wu, Z. Yang, and Y. Liu, "Sensorless sensing with WiFi," *Tsinghua Science and Technology*, vol. 20, no. 1, pp. 1–6, Feb. 2015.

[2] S. Savazzi, S. Sigg, M. Nicoli, V. Rampa, S. Kianoush, and U. Spagnolini, "Device-free radio vision for assisted living: Leveraging wireless channel quality information for human sensing," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 45–58, Mar. 2016.

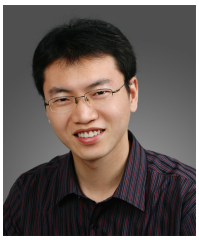
[3] L. Zhang, Q. Gao, X. Ma, J. Wang, T. Yang, and H. Wang, "DeFi: Robust training-free device-free wireless localization with WiFi," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8822–8831, Sep. 2018.

[4] J. Wang, X. Bai, Q. Gao, X. Li, X. Bi, and M. Pan, "Multi-target device-free wireless sensing based on multiplexing mechanisms," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 10 242–10 251, Sep. 2020.

[5] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using WiFi signals," in *Proceedings of ACM UbiComp 2016*, pp. 363–373.

[6] Q. Xu, B. Wang, F. Zhang, D. S. Regani, F. Wang, and K. J. R. Liu, "Wireless AI in smart car: How smart a car can be?" *IEEE Access*, vol. 8, pp. 55 091–55 112, 2020.
[7] M. Raja, V. Ghaderi, and S. Sigg, "Detecting driver's distracted behaviour from Wi-Fi," in *Proceedings of IEEE VTC 2018 Spring*, pp. 1–5.
[8] S. D. Regani, Q. Xu, B. Wang, M. Wu, and K. J. R. Liu, "Driver authentication for smart car using wireless sensing," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2235–2246, Mar. 2020.
[9] J. Wang, J. Tong, Q. Gao, Z. Wu, S. Bi, and H. Wang, "Device-free vehicle speed estimation with WiFi," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8205–8214, Sep. 2018.
[10] C. Liu, D. Fang, Z. Yang, H. Jiang, X. Chen, W. Wang, T. Xing, and L. Cai, "RSS distribution-based passive localization and its application in sensor networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2883–2895, Apr. 2016.
[11] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6258–6267, Jul. 2017.
[12] J. Wang, Q. Gao, H. Wang, Y. Yu, and M. Jin, "Time-of-flight-based radio tomography for device free localization," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2355–2365, May 2013.
[13] J. Wang, Q. Gao, M. Pan, and Y. Fang, "Device-free wireless sensing: Challenges, opportunities, and applications," *IEEE Network*, vol. 32, no. 2, pp. 132–137, Mar. 2018.
[14] H. Wang, D. Zhang, J. Ma, Y. Wang, and B. Xie, "Human respiration detection with commodity WiFi devices: Do user location and body orientation matter?" in *Proceedings of ACM UbiComp 2016*, pp. 25–36.
[15] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of ACM MobiCom 2013*, pp. 27–38.
[16] C. Wu, Z. Yang, Z. Zhou, X. Liu, Y. Liu, and J. Cao, "Non-invasive detection of moving and stationary human with WiFi," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2329–2342, Nov. 2015.
[17] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "CrossSense: Towards cross-site and large-scale WiFi sensing," in *Proceedings of ACM MobiCom 2018*, pp. 305–320.
[18] Y.-K. Cheng and R. Y. Chang, "Device-free indoor people counting using Wi-Fi channel state information for Internet of Things," in *Proceedings of IEEE GLOBECOM 2017*, pp. 1–6.
[19] X. Ma, Y. Zhao, L. Zhang, Q. Gao, M. Pan, and J. Wang, "Practical device-free gesture recognition using WiFi signals based on metalearning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 228–237, Jan. 2020.
[20] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3041–3048, Apr. 2019.
[21] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "WiFi-enabled device-free gesture recognition for smart home automation," in *Proceedings of IEEE ICCA 2018*, pp. 476–481.
[22] C. Feng, S. Arshad, S. Zhou, D. Cao, and Y. Liu, "Wi-Multi: A three-phase system for multiple human activity recognition with commercial WiFi devices," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 7293–7304, Aug. 2019.
[23] H. Kong, L. Lu, J. Yu, Y. Chen, L. Kong, and M. Li, "FingerPass: Finger gesture-based continuous user authentication for smart homes using commodity WiFi," in *Proceedings of ACM MobiHoc 2019*, pp. 201–210.
[24] I. Nirmal, A. Khamis, M. Hassan, W. Hu, and X. Zhu, "Deep learning for radio-based human sensing: Recent advances and future directions," *IEEE Communications Surveys and Tutorials*, vol. 23, no. 2, pp. 995–1019, 2nd Quart. 2021.
[25] Q. Li, R. Gravina, Y. Li, S. H. Alsamhi, and G. Fortino, "Multi-user activity recognition: Challenges and opportunities," *Information Fusion*, vol. 63, pp. 121–135, 2020.
[26] S. Yue, H. He, H. Wang, H. Rahul, and D. Katabi, "Extracting multi-person respiration from entangled RF signals," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 2, jul 2018.
[27] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person localization via RF body reflections," in *Proceedings of USENIX NSDI 2015*, pp. 279–292.
[28] R. H. Venkatnarayan, G. Page, and M. Shahzad, "Multi-user gesture recognition using WiFi," in *Proceedings of ACM MobiSys 2018*, pp. 401–413.
[29] S. Tan, L. Zhang, Z. Wang, and J. Yang, "MultiTrack: Multi-user tracking and activity recognition using commodity WiFi," in *Proceedings of ACM CHI 2019*, pp. 1–12.
[30] Y. Xie, J. Xiong, M. Li, and K. Jamieson, "MD-Track: Leveraging multi-dimensionality for passive indoor WiFi tracking," in *Proceedings of ACM MobiCom 2019*, pp. 1–16.
[31] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang, "FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, Sep 2019.
[32] W. Chen, K. Niu, D. Zhao, R. Zheng, D. Wu, W. Wang, L. Wang, and D. Zhang, "Robust dynamic hand gesture interaction using LTE terminals," in *Proceedings of ACM/IEEE IPSN 2020*, pp. 109–120.
[33] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 14)*, 3GPP TS 36.211 V14.12.0, Sep. 2019.
[34] S. Xu and Y. Tian, "Device-free motion detection via on-the-air LTE signals," *IEEE Communications Letters*, vol. 22, no. 9, pp. 1934–1937, Sep. 2018.
[35] Y. Tian, Y. He, and H. Duan, "Passive localization through channel estimation of on-the-air LTE signals," *IEEE Access*, vol. 7, pp. 160 029–160 042, 2019.
[36] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

Rui Peng (Student Member, IEEE) received his B.S. and M.S. degree in the School of Electronics and Information Engineering from Beihang University, Beijing, China, in 2015 and 2018. He worked on LTE modem development in Intel corporation, from 2018 to 2020. He is currently pursuing his Ph.D. degree in the School of Electronics and Information Engineering, Beihang University. His research interests include physical layer design of 4G/5G, MIMO precoding and wireless sensing.



Yafei Tian (Member, IEEE) received his B.S. degree in electronics engineering and Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in 2001 and 2008, respectively. He is currently an Associate Professor with the School of Electronics and Information Engineering, Beihang University. He was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2010 to 2011. His research interests lie in cellular communications, wireless sensing, and machine reasoning.

Shengqian Han (Member, IEEE) received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2004 and 2010, respectively. He is currently an Associate Professor with the School of Electronics and Information Engineering, Beihang University. From 2015 to 2016, he was a Visiting Scholar with the University of Southern California, Los Angeles, USA. His recent research interests include wireless big data and AI for communications. He served as a TPC member for numerous IEEE conferences. He is also an Associate Editor of EURASIP JWCN.