IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 24, 2019, accepted July 15, 2019, date of publication July 25, 2019, date of current version August 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931088

A Survey on Human Behavior Recognition Using Smartphone-Based Ultrasonic Signal

ZHENGJIE WANG, YUSHAN HOU, KANGKANG JIANG, CHENGMING ZHANG, WENWEN DOU, ZEHUA HUANG, AND YINJING GUO

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding authors: Zhengjie Wang (cieewangzj@163.com) and Yinjing Guo (gyjlwh@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471224, in part by the Qingdao Postdoctoral Applied Research Project under Grant 2015180, and in part by the Shandong Province Key Research and Development Plan (Public Welfare Special) Project under Grant 2018GHY115022.

ABSTRACT With the rapid progress of the Internet of Things (IoT) technology, human behavior recognition has become an important research topic in the field of ubiquitous computing and has obtained quite a number of research achievements. Accurate human behavior recognition can enhance the quality of human–computer interaction and facilitate the development of various sensing applications. With the popularity of smartphones and the improved performance of sensors such as speakers and microphones built in smartphones, the behavior recognition technique based on ultrasound signal of the smartphone has gained more attention and achieved several research results. In this paper, we first review the common behavior recognition techniques including light, video, sound, and frequency radio and outline their main characteristics. Then, we introduce the fundamental principle of human behavior recognition based on ultrasound signals. Specifically, these systems treat speakers and microphones embedded in smartphones as the transceiver and leverage the received signal changes caused by human movement including phase differences, frequency shift, and time of flight (ToF) to recognize human behavior. Next, we investigate the state-of-the-art studies and applications and analyze the signal processing techniques such as data collection, signal preprocessing, feature description, and behavior recognition approach. Afterward, according to the purpose of these applications, we classify them into five groups and compare them in detail including hand gesture recognition, activity recognition, hand trajectory tracking, vital sign monitoring, and lip reading. Finally, we conclude by discussing the limitations, challenges, and open issues involved in behavior recognition based on ultrasound signal of smartphone.

INDEX TERMS Doppler effect, human behavior recognition, smartphone, ultrasonic signal.

I. INTRODUCTION

With the significant advances in computer technology, human behavior recognition has become an important research topic and has attracted a variety of research efforts. The purpose of human behavior recognition is to develop effective techniques to model and understand human behavior from sensor data. Although there are various human behaviors, we concentrate on some typical movements. These behaviors not only include simple daily actions, such as waving a hand, walking, and driving, but also cover health monitoring, such as heartbeat and respiration monitoring. Accurate behavior recognition can enhance the quality of human-computer interaction and facilitate various applications, such as health monitoring, home entertainment, daily activity recognition, etc. [1]. Recently, several effective physical models and recognition algorithms have been proposed and related experiments under realistic settings have been performed. Many studies apply device-based (users need to wear device on the body) approaches and they can measure accurately more data and make precise control to sensing procedures [2]. These studies show that many systems can achieve satisfactory recognition accuracy and may be employed in many scenarios. In addition, we find that device-free (users do not wear any sensor) behavior recognition approaches have been widely studied due to their non-intrusive manner [3]. Specifically, these systems can decrease disturbance to daily life and enable to monitor targets for long periods.

The associate editor coordinating the review of this manuscript and approving it for publication was Haiwen Liu.

TABLE 1. Comparison of the four types of recognition methods.

Method	System/Work	Signal	Behavior Recognized
Vision-based	U. Lee *et al.* [8], J.beh *et al.* [9], X. Li *et al.* [10], Paulo *et al.* [11], A. Bux *et al.* [12]	Images	hand gesture recognition, activity recognition, individual authentication
Light-based	LiSence [13], GestureLite [14], LiGest [15]	Light	hand gesture recognition, activity recognition
RF-based	GRfid [16],	RFID	activity recognition, hand gesture recognition, vital sign detection, individual authentication
	WiFinger [17],	CSI	
	J. Wang *et al.* [18],	RSS	
	W. Li *et al.* [19],	Radar signal	
	WiSee [20]	OFDM	
Audio-based	J. Liu *et al.* [21], M. Chen *et al.* [22], SoundWrite II [23], WordRecorder [24], BodyScope [25], SoundTrak [26], WritePad [27]	Audible sound	hand gesture recognition, activity recognition, hand trajectory tracking, vital sign monitoring, lip reading
	Y. Huang *et al.* [28], T. Chiang *et al.* [29] K. Kalgaonkar *et al.* [30], A. Ghosh *et al.* [31], T. Wang *et al.* [32] Multiwave [33], DopGest [34]	Ultrasonic signal	
	Swadloon [35] LLAP [36], D. Graham *et al.* [37]		

Human behavior usually changes the signal propagation path and generates signal variation or shadow effect. The relationship between human behavior and signal variation is unique and can be exploited to recognize the behavior. According to the types of signals received at the receiving devices, we classify them into four groups: vision [4], light [5], radio frequency (RF) [6], and acoustic signal [7], as shown in Table 1. This table lists some typical applications and compares recognized human behavior using these four modalities. From the literature, we discover that these four modalities can be utilized to recognize common human gestures, such as hand gesture and daily activity. Besides, no more recognized behaviors are mentioned in light-based method. As for the other three modalities, they can be leveraged to develop many other applications, such as identity authentication, vital signal detection, lip reading recognition. From the types of recognized behaviors, they can implement similar functions. Therefore, which modality is selected depends on the environmental condition. If we have a good light condition and line-of-sight (LOS) scenario, we can select the video-based method. If we want to monitor target indoor or through the wall scenario, we can select RF signal because we can easily access the devices and RF can propagate across walls without disturbing subjects. If we can use the audio device, we can select the audio-based method.

A. VISION-BASED METHOD

Quite an amount of research on behavior recognition adopts the vision-based approach due to the ubiquitous availability of digital camera. The idea of these approaches is that the color features are effectively extracted by exploring image processing techniques and human behaviors are identified by leveraging recognition algorithms. For example, U. Lee *et al.* [8] first extracts finger features using the related technique from data with depth information to detect the presence of fingers and then recognizes the finger gestures based on the detection results. Based on Hidden Markov Model (HMM), Beh *et al.* [9] achieved the accurate recognition of hand motions. Li *et al.* [10] proposed an activity recognition method. They first use the activity mask generated by a conditional generative adversarial network (cGAN) to locate the activities and then estimate activities by a Visual Geometry Group-Long Short-Term Memory (VGG-LSTM) network. Borges *et al.* [11] presented a survey about human behavior recognition based on the video. First, they divide the human activities based on visual recognition into four classes which are interactions, human gestures, actions, and behaviors. Then they categorize recognition approaches into three groups, including hybrid approaches, appearance-based approaches, and motion-based approaches. Moreover, they introduce some techniques used in human action recognition methods. Bux *et al.* [12] presented a survey of human activity recognition. They first introduce the segmentation technique used for activity recognition and divide them into two groups, including background construction-based and foreground extraction-based. Then they study the methods of feature extraction and classification. Although computer vision-based recognition approaches are widely deployed and achieve satisfactory recognition accuracy under several scenarios, many environmental factors would affect the recognition performance, such as the line of sight path, users' skin color, distance between camera and user, light condition, and so on, making it challenging to design a widely applicable system.

B. LIGHT-BASED METHOD

There is also an increasing research interest to treat light as a sensing signal to realize human behavior recognition in recent years. The principle of these systems is that human posture may block light propagation and generate shadow. Apparently, the unique relationship between shadow and human posture can be built and explored to develop the sensing

system. For example, LiSence [13] can reconstruct a 3D user skeleton posture using the shadow generated from the blocked visible light by the human body. GestureLite [14] can classify 10 pre-defined gestures using the shadow caused by hand. Different from the LiSense, it utilizes the ambient light instead of lights mounted on the ceiling. And it uses machine learning method to realize the classification of gestures performed by the user. Although GestureLite achieves a high accuracy of recognition, it usually recognizes the participant postures and is not robust for strange subjects. Similarly, LiGest [15] also uses the ambient light to recognize hand gestures, and it is robust for strange subject, the position and orientation of the user, and the lighting condition. LiGest first utilizes the training samples to learn the unique shadow patterns of hand movements and then matches the unknown hand postures with the learned patterns to realize the recognition of gestures. Although the light-based human behavior recognition technology achieves high accuracy, it usually needs users to deploy some photodiodes and customized LEDs.

C. RF-BASED METHOD

Recently, behavior recognition approaches based on wireless signals have brought a lot of attention due to the low deployment cost and widely applicable scenarios. The basic idea of these approaches is that the wireless signal propagates in a multi-path manner and the wireless channel keeps stable when there are no people in the environment. However, if a person moves within the range of wireless signal coverage, the location and gesture of the user will affect the signal propagation path, which results in channel disturbance due to the signal changes caused by reflection, scatter, refraction, and so on. Therefore, we can recognize different human behaviors based on signal change characteristics and channel distortion patterns. Specifically, we first extract the signal disturbance caused by the user's movement and detect the feature changes, and then identify human behaviors by exploring recognition algorithms or classification models to facilitate various applications.

Based on the deployed equipment of RF signal, we divide them into two types, such as commercial off-the-shelf devices (COTS) and bespoken devices. The former includes many devices that can use radio frequency identification (RFID) [16], received signal strength (RSS) [18], and channel state information (CSI) [17]. The latter includes some software defined radio (SDR) platforms [19], [20].

We first explored several representative applications based on COTS devices. For example, GRfid [16] is a hand gesture recognition system that identifies different hand gestures by leveraging the phase changes of the captured RFID signal. Similarly, WiFinger [17] is also a hand motion recognition system that analyzes the relationships between human behaviors and the received signals. Specifically, WiFinger system is based on the observation that the user's finger leads to a unique pattern in CSI data while performing a certain gesture. Wang *et al.* [18] presented an approach which can be utilized to recognize activities and gestures, simultaneously realize the localization. They extract features from RSS data by their designed sparse autoencoder network and then leverage the SoftMax regression classifier to obtain an activity label.

We then investigate these applications with bespoken devices. W. Li *et al.* [19] proposed a novel method that uses unsupervised classification with HMM based on the micro-Doppler radar to realize human activity recognition. It builds a passive radar system based on an SDR platform to obtain the Doppler information. WiSee [20] is a novel gesture recognition system based on the WiFi signal. It extracts Doppler shifts of orthogonal frequency division multiplexing (OFDM) signal emitted from USRP-N210 device. Although most of the recognition approaches based on wireless signals do not require LOS path and can be easily deployed, they usually are sensitive to the influence of environment and user changes due to the multi-path effects, which results in the lack of stable performance of these approaches under different settings.

D. AUDIO-BASED METHOD

Acoustic signal has been widely studied in speech synthesis, music information retrieval, and natural language processing in decades. At the same time, human behavior based on acoustic signal has drawn more attention among researchers. We consider two types of acoustic signals: audible sound and ultrasound. The former can be used to sense environment sound, extract sound features, and identify human actions. The latter can be utilized to measure signal variation, extract propagation path information, and recognize person activities.

1) AUDIBLE SIGNAL

a: DEVICE-FREE PATTERN

Specifically, there are a large number of device-free applications about sensing audible sound for human behavior recognition. For instance, Liu *et al.* [21] proposed a keystroke recognition system by exploring the mm-level acoustic ranging using a smartphone. It exploits two microphones on the smartphone to measure mm-level distance difference and identify keystroke based on the distance and position. Similarly, SoundWrite II [23] is an audible sound-based text and stroke recognition system. It includes stroke input detection, stroke recognition, and text recognition using stroke combination. It captures the sound signals reflected by moving a finger on the table surface, extracts time and frequency features, and identifies stroke using pattern classification approaches. WordRecorder [24] is a passive sensing system for handwriting recognition. It captures sound signals generated by pens and paper, and then sends the collected data to a smartphone for text recognition. Chen *et al.* also proposed a device-free hand gesture and handwriting recognition system, called Ipanel [22], based on acoustic signals from finger friction on a surface. When a user's finger slides on a nearby surface, the acoustic signal can be captured and analyzed to obtain the unique features from spatio–temporal and

frequency domain. These features are converted into images and are fed into CNN to identify finger motion.

b: DEVICE-BASED PATTERN

Besides these device-free applications, some recognition systems which require the participant to wear some acoustic sensors have been developed. For example, BodyScope [25] is an activity recognition system based on a wearable acoustic sensor. It measures the sound produced in a person's mouth and throat and utilizes support vector machine (SVM) to classify 12 user activities including eating, drinking, speaking, laughing, and coughing, etc. Zhang *et al.* [26] presented SoundTrack, a position tracking system that identifies the finger position in 3D space using a finger ring with a miniature speaker and a smartwatch linking 4 microphones. Specifically, when the speaker wore on a user's finger transmits the sound signal at a specific frequency, the microphone arrays on the target device capture the sound signal affected by the finger motion. Then users calculate the finger's position in 3D space by exploiting phase information extracted from the received acoustic signals. WritePad [27], a kind of passive sound sensing system, was proposed to recognize the numbers written on the hand back. It leverages the smartwatch to collect the sound signals created by writing numbers on the user's hand back and then establishes a hybrid convolutional neural network (CNN) to realize the number recognition.

2) ULTRASONIC SIGNAL

Besides audible sound signal, the ultrasonic signal has been deeply studied and widely used for human behavior recognition. Here, we divide the human behavior recognition based on ultrasonic signal into two groups, such as custom device-based and COTS-based. Custom device-based pattern usually needs users to wear devices on the body. Therefore, this pattern provides some different options for human behavior recognition compared with COTS-based pattern.

a: CUSTOM DEVICE-BASED PATTERN

Some systems recognize human behavior by using the custom device. For example, Huang *et al.* [28] designed a hand gesture recognition system based on the ultrasonic signal. This system requires users to wear eight sEMG sensors and one ultrasound probe on the arm and then utilizes the ultrasound imaging technique to depict the detected hand gestures. Moreover, this system not only can detect discrete finger movements but also predict continuous finger angles. Chiang *et al.* [29] presented a system of pedestrian dead reckoning. This system requires users to wear the shoes with sensors, one microphone on the one shoe and two buzzers on the other shoe. In this system, authors argued that the relative movement of people's feet would cause different Doppler shift and the Doppler shift can be used to infer the step length, step event, and the step orientation. Furthermore, they can utilize this inferred information to localize the user's location.

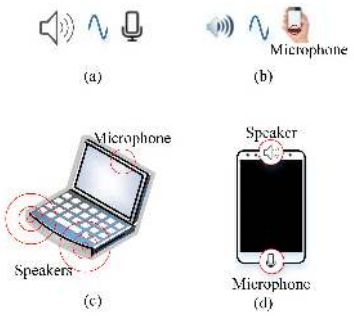

FIGURE 1. Four types of ultrasonic signal-based recognition. (a) Deploying sound sensors in the environment. (b) Utilizing the only microphone built in smartphone. (c) Using a laptop or desktop. (d) Using speakers and microphones built in smartphone.

b: COTS-BASED PATTERN

Here we investigate these applications that apply COTS speakers and microphones as the transceivers. As shown in Fig. 1, we consider four types of ultrasound-based recognition studies according to the deployed equipment: deploying sound sensors in environment [30]–[32], utilizing only microphone built in smartphone [35], using laptop or desktop [33], [34], and using speakers and microphones built in smartphone [36], [37]. Deploying sound sensors in an environment means that we can install independent speaker and microphone, as shown in Fig. 1(a). Utilizing only microphones built in smartphone refers to these systems which use the microphones built in smartphone to collect ultrasonic signals and deploy extra speakers emitting sound signals instead, as shown in Fig. 1(b). Using laptop or desktop refers to utilizing the speakers and microphones embedded in computer to recognize behaviors, as shown in Fig. 1(c). While using speakers and microphones built in smartphone indicates that speakers transmit ultrasonic signals and microphones built in the same mobile receive echo signals, as shown in Fig. 1(d). We introduce some representative applications based on the above four methods as follows.

Firstly, Kalgaonkar and Raj [30] proposed a dynamic gesture recognition system based on ultrasound, which consists of one transmitter and three receivers. This system extracts frequency data of the echoes and compares them with the original signal to recognize the behavior. Ghosh *et al.* [31] proposed a non-intrusive and device-free sensing system to recognize multiple types of user group activities using ultrasonic sensors. It utilizes the customized ultrasonic sensors deployed in the monitoring area to sense human activities. Wang *et al.* [32] proposed a contactless and real-time respiration monitoring system by sensing the chest translation. It leverages the Doppler effect to build a mathematical model depicting the link between Doppler shift and the direction of airflow. Secondly, Swadloon [35], a direction finding and indoor localization approach, was proposed based on smartphone. Ultrasonic signals first are sent from anchor speakers, then the user shakes the smartphone or walks with the smartphone. During the smartphone motion, the ultrasonic signal is received and the disturbance of the signal is analyzed

to determine the motion direction displacement relative to anchors using Doppler effect.

Thirdly, Pittman *et al.* proposed Multiwave [33], a complex gesture (e.g., triangle, arrow) recognition system using speaker and microphone embedded in the laptop based on Doppler effect. This system validates its performance with 14 complicated gestures and achieves 94% accuracy with two speakers. Liu *et al.* proposed DopGest [34], which is a hand posture recognition system using speakers and microphone built in laptop. Two speakers transmit the sound signal in different frequency respectively and the DopGest recognizes hand motions by combing k-nearest neighbor (KNN) classifier with dynamic time warping (DTW) algorithm after extracting Doppler shift feature. Fourthly, LLAP [36] was proposed to measure the distance between hand and device and track the hand trajectory. It utilizes the speaker built in smartphone and microphone as the transceiver and transforms the phase changes caused by hand motions into the distance of hand movement to track the hand. Differently, Graham *et al.* [37] proposed a system to measure the distance between mobile and object. This system first turns the mobile phone into an active sonar system and then measures the time of flight to realize the distance measurement.

Among these COTS-based applications, as shown in Fig. 1, Fig. 1(a) and Fig. 1(b) need to deploy sound sensors in the environment, which might result in extra deployment cost. And Fig. 1(c) utilities the laptop to recognize human behavior, which is not convenient to deploy. Fig. 1(d) leverages smartphone with built-in speakers and microphones to implement behavior recognition. The systems based on Fig. 1(d) provide many advantages, such as zero device cost, convenient deployment, excellent recognition accuracy, wide application scenarios, and long-term monitoring without disturbing participants. Therefore, we concentrate on these applications that leverage smartphone as the hardware devices to transmit and receive the ultrasonic signal.

Currently, there has been encouraging progress in human behavior recognition using ultrasonic signal based on built-in speakers and microphones of smartphone, such as distance measurement [36], [37], encounter profiling [38], hand gesture recognition [39]–[46], activity recognition [47], lip reading [48]–[50], respiration detection [51], [52], Parkinson's diagnosis [53], hand trajectory tracking [36], [54]–[59], multi-device interaction [60]–[62], direction finding and localization [63]–[65], context sensing [66], [67], indoor mapping [68], [69], acoustic imaging [70], grip sensing [71], touch force sensing [72]–[74].

Compared with the conventional approaches (vision, RF, light), the smartphone-based ultrasonic approach has some merits. We illustrate the merits from the following aspects. For the vision method, the ultrasound of smartphone approach does not require strict environmental conditions. In addition, it can achieve real-time recognition results using lower computation cost, making it cost-effective for most behavior recognition. For the light-based method, the smartphone-based approach does not require extra device except smartphone and can identify more human behaviors, such as heartbeat rate and respiration rate. Besides, the light-based method is unsuited to sleeping conditions because we usually turn off the light. Differently, the ultrasound methods are not affected by the light condition and can conduct the monitoring for a whole day. As for the RF-based method, the smartphone-based approach has many advantages such as low-cost deployment (compared with bespoke devices), no need of modification of the device drivers (compared with CSI), more accurate recognition (compared with RSS), and less number of devices (compared with RFID). In all, although ultrasonic-based recognition technology has many strengths, more efforts are still needed to fully understand its limits and enhance its robustness and recognition accuracy.

There are two types of human behavior recognition using an ultrasonic signal from smartphone. The first one is a device-based pattern. Specifically, users conduct some actions when wearing or holding a smartphone. In this scenario, the smartphone usually moves with the users' movement and the sound signal from moving smartphone is used to recognize human behaviors [57], [75]. The second one is a device-free pattern. Specifically, a smartphone is placed at a place, such as tables or desks near the participant. In the monitoring procedures, the smartphone usually keeps stationary and the changed signal from the still smartphone is utilized to identify human behavior. It is noteworthy that operating on the touch screen is excluded from the device-free pattern because the user needs to touch the devices.

This paper investigates the state-of-the-art human behavior recognition applications based on an ultrasonic signal from the smartphone in the device-free pattern. Specifically, these studies utilize the ubiquitous smartphone with built-in speakers and microphones without any hardware modification. The speakers emit an inaudible sound signal and the microphones receive the changed signal caused by human behavior. Then the changed ultrasonic signal will be analyzed for human behavior recognition. These applications usually include whole-body activity [47], [76], hand waving gesture [40]–[43], hand trajectory tracking [36], [54]–[56], [58], [59], lip reading [48]–[50], and vital sign monitoring [51]–[53], [77]. Notably, we primarily focus on the applications that solely leverage the speakers and microphones of the smartphone. The behavior recognition systems leveraging the other built-in sensors of smartphone (e.g., gyroscope, accelerometer) are beyond the range of our research.

The contributions of this paper can be summarized as follow. Firstly, we present a comprehensive review of recent progress in human behavior recognition based on the ultrasonic signal of a smartphone. To the best of our knowledge, this paper is the first survey on ultrasound behavior recognition with speakers and microphones of the smartphone. Secondly, we analyze the fundamental principle of human behavior recognition based on ultrasonic signal and present a typical framework that exploits speakers and microphones built in smartphone as sensing devices. Meanwhile, we

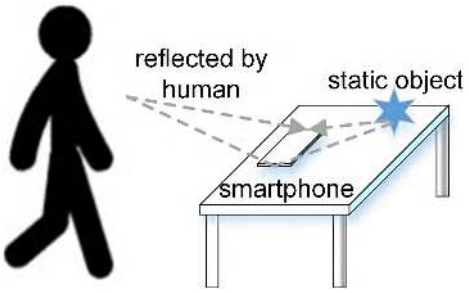

FIGURE 2. The fundamental principle of the ultrasonic-based human behavior recognition using smartphone.

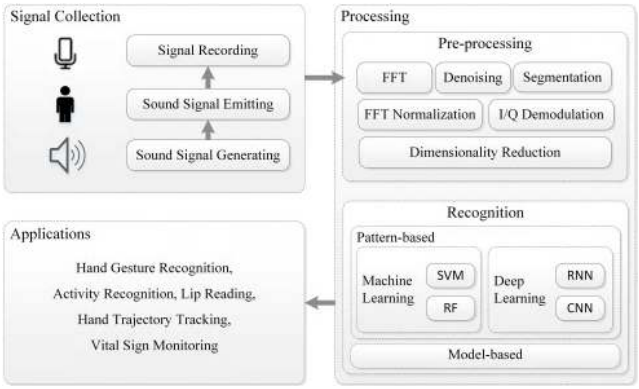

FIGURE 3. General processing framework of human behavior recognition based on the ultrasonic signal of a smartphone.

summarize the general signal preprocessing, feature description, and action recognition algorithms. Finally, we investigate the state-of-the-art applications and make an in-depth analysis and comparison, including hand gesture recognition, activity recognition, lip reading, hand trajectory tracking, and vital sign monitoring.

The rest of the paper is organized as follows. In this survey, we first introduce the basic principle of human behavior recognition based on the ultrasonic signal of a smartphone and provide a typical framework of human behavior recognition based on the ultrasonic signal of smartphone in Section II. Afterward, we introduce the signal generation, signal properties, signal preprocessing, feature description, and behavior recognition in Section III. Then we summarize the related behavior recognition applications of ultrasound signal based on smartphone in Section IV. And we discuss the theoretical limitations of these approaches and some open research topics in Section V. Finally, we conclude by summarizing our research work in Section VI.

II. BASIC PRINCIPLE OF BEHAVIOR RECOGNITION AND SYSTEM FRAMEWORK

In this section, we will briefly introduce the fundamental principle of human behavior recognition and system framework based on ultrasound signals from the smartphone.

A. FUNDAMENTAL PRINCIPLE

As shown in Fig. 2, a typical hardware device for ultrasonic-based behavior recognition system solely needs a smartphone with speakers and microphones. We call the system as active sonar sensing because the speaker emits the ultrasonic signal. Therefore, we can actively transmit sound signal and make precise control to the sound modulation by elaborating signal waveform and parameter. The modulated sound signal is emitted by the speakers and the changed signal is captured by the microphone in the same phone. These changes stem from environmental variation, such as parts of bodies and nearby moving persons. When a user moves in the coverage of the speakers and microphones, the sound signal is affected by the movement. Therefore, the captured signal comprises of ambient noise, reflection from participant movement and furniture, and interference from another nearby person. The captured signal is transformed and filtered to eliminate various noise. Then, the signal is analyzed and useful information is extracted to represent the movement. In other words, a user's movement near the speakers and microphones would affect the signal propagation paths and change the phase and frequency of received signal during ultrasonic signal propagation, which can be used to correlate the signal changes with the corresponding behavior patterns. In addition, we can utilize the time of flight (ToF) of the sound signal to localize the position of the target to track the continuous movement of the hand. Therefore, ToF is generally used to track a target or recognize the trajectory of an object.

B. SYSTEM FRAMEWORK

As shown in Fig. 3, a typical system framework using the ultrasonic signal of smartphone comprises many important components such as signal collection, signal preprocessing, and behavior recognition. We introduce the function of these components to present a clear description of the framework. The first component is the signal collection which includes signal generation, signal transmission, and signal reception. Based on the requirements of the system, we can employ different types of modulation techniques to generate a different signal. The signal can be emitted by real-time calculation or by playing the recorded sound file. Then, when we need to recognize some behaviors, we receive the changed signal caused by human behavior and analyze the variation to get the movement pattern. Because we receive the raw sound signal from the microphone, the captured signal contains much noise from hardware, ambient factors, other parts of body, and another nearby person movement. Thereby, we must exploit effective algorithms to eliminate noise and identify useful information for the next step. Afterward, we build feature vectors from the extracted signal to feed them into a classifier and recognize human behavior. We also utilize the geometric model to track target position. The detailed description of the framework is presented in Section III.

III. PROCESSING OF BEHAVIOR RECOGNITION

In this section, we present a detailed description of the processing framework for human behavior recognition based on

TABLE 2. Comparison of five waveforms.

Signal	Characteristics	Advantages and disadvantages
CW	constant amplitude and frequency	SNR will be increased; however, it will result in poor spatial resolution.
Chirp	good autocorrelation properties	It can improve resolution and sensitivity.
FMCW	low cross-correlation, high autocorrelation	great robustness
OFDM	orthogonal subcarriers divided by the bandwidth independently transmit data	processing complexity could be reduced and the sync of transmitter and receiver could be achieved effectively.
ZC	constant amplitude, orthogonal to its delayed versions	perfect for synchronization

the ultrasonic signal of a smartphone, as shown in Fig. 3. First, we introduce the signal collection from two aspects, including the types of the emitted signals and the basic properties of signal extracted from the echoes including Doppler shift, phase, and ToF. Then we analyze some preprocessing methods and the feature extraction. Finally, we review the common recognition approaches for behavior recognition.

A. SIGNAL GENERATING

Most applications discussed here apply the following patterns: the speakers first emit ultrasound signal, and then the microphone receives the signal and the system makes further processing. Since we send the sound signal, we can make more control over the sound signal. Waveform generation plays a crucial role in the acoustic sensing system since it determines the characteristics of the signal. We can benefit from the good waveform design when conducting denoising and feature extraction. According to the characteristics of sound waveforms, many types of sound signals can be employed, such as continuous wave (CW) signal [36], [52], chirp signal [40], [68], frequency modulated continuous wave signal (FMCW) [51], [77], OFDM signal [54], Zadoff-Chu sequence (ZC) [45], as shown in Table 2. They have their advantages and limitations. Long CW signal is generally used to improve signal-noise ratio (SNR); however, the spatial resolution is not satisfactory. FMCW is another signal that is commonly used in some applications. One of the benefits is its strong anti-interference ability. In addition, the OFDM signal is adopted due to its low processing complexity. ZC sequence [45] has constant amplitude and is perfect for synchronization due to its orthogonality with its delayed versions [7]. We compare these five commonly used waveforms in Table 2. This table lists some main waves, their characteristics, disadvantages, and advantages.

Besides signal waveform, there are many factors that need to be considered when using the ultrasonic signal. As shown in Table 3, these factors include signal frequency, experimental devices, and the number of sensors. This table exhibits and compares the frequency of the ultrasound signal used in behavior recognition. From the table, we can directly acquire the related hardware information about each system. As the most popular smartphone operation system, Android is applied by almost all brands of smartphones, such as Samsung and Huawei.

Generally, ultrasonic signal refers to the sound signal with a frequency above the average people's hearing range. Usually, the hearing range of the average user is within 16 kHz while a young user may hear the sound with a frequency above 20 kHz [78]. Therefore, the range of frequency of ultrasonic sound in these applications covers from 16 kHz to 23 kHz. This frequency range usually meets our monitoring requirements. Therefore, to satisfy most application scenarios, 20 kHz sound frequency is a better choice. Modern smartphones are usually equipped with many microphones and speakers to enhance sound signal quality and suppress the noise from environment and hardware devices. For simple gesture, one speaker and one microphone are enough to recognize it. For position tracking, one speaker and two microphones or two speakers and one microphone is a good choice because multiple sensors provide more information about human movement space.

B. BASIC PROPERTIES OF SIGNAL

After receiving the signal from the microphone, we usually extract three types of properties to realize human behavior recognition, including Doppler effect, phase, and ToF. In this section, we will concentrate on these three basic properties and interpret the reason why they can be utilized for behavior recognition.

1) DOPPLER EFFECT

The Doppler Effect is proposed by the Austrian physicist Christian Doppler in 1842 to describe the frequency change phenomenon that the wavelength of the object's radiation changes due to the relative motion of the source and the observer. According to this phenomenon, we can recognize different human behaviors. Based on Doppler effect of the ultrasonic signal, different motions will lead to distinct frequency changes. Therefore, we can utilize this attribute to realize hand posture recognition, lip reading, human activity identification, etc. For example, the speaker emits the ultrasonic signal at a given frequency, and the microphone captures the sound signal. During the signal propagating in air, the user's hand moves away from the microphone, which results in frequency shift, as shown in Fig. 4. Specifically, we can observe that the microphone will receive a lower frequency sound signal due to the hand moving away from the microphone. On the contrary, the microphone will capture the sound signal with the higher frequency due to the hand moving towards the microphone.

Furthermore, the frequency shift can be computed by the following formulas [34],

$$f' = \left(\frac{v + v_0}{v - v_0}\right) f \tag{1}$$

$$\Delta f = f' - f \tag{2}$$

TABLE 3. Comparison of systems in adopted signal, signal frequency, experimental devices, and sensors used.

System	Adopted signal	Signal frequency	Experimental devices	Sensors used
Dolphin [43]	continuous tone	21 kHz	MI One, Samsung S3	one speaker, one microphone
AudioGest [40]	sine acoustic wave	19kHz	Samsung Galaxy S4	one speaker, one microphone
SonicOperator [41]	sine acoustic wave	21 kHz	MI Note, Vivo X7, MI 5, HTC One	one speaker, one microphone
H. Watanabe *et al.* [42]	---	20 kHz	Huawei P9 Lite	one speaker, one microphone
B. Fu *et al.* [47]	continuous periodical signal	20 kHz	Google Nexus 5X	one speaker, one microphone
B. Fu *et al.* [76]	continuous periodical signal	20 kHz	Google Nexus 5X	one speaker, one microphone
LLAP [36]	CW sound signal	17 - 23 kHz	Samsung Galaxy S5, iPhone 6s	one speaker, two microphones
FingerIO [54]	OFDM signal	18 - 20 kHz	Samsung Galaxy S4	one speaker, two microphones
Strata [55]	BPSK signal	18 - 22 kHz	Samsung Galaxy S4	one speaker, two microphones
EchoTrack [56]	chirp signal	16 - 23 kHz	Nexus 6P	two speakers, one microphone
SteerTrack [58]	sinusoidal signal	20 kHz	Google Pixel, LG G4, HTC U Ultra, Samsung Galaxy S6, Huawei Mate8	one speaker, two microphones
DMT [59]	---	left speaker: 17 kHz right speaker: 19kHz	Huawei GRA-TL00	two speakers, one microphone
ApneaApp [51]	FMCW signal	18 - 20 kHz	Samsung Galaxy S4, Samsung Galaxy S5, HTC One, Galaxy Nexus	one speaker, one microphone
SonarBeat [52]	CW sound signal	18 - 22 kHz	Samsung Galaxy S6, Samsung Galaxy S7 Edge	one speaker, one microphone
W. Wang *et al.* [53]	---	16.8kHz - 21.7kHz	iOS platform, such as iPhone 6/6s/7	one speaker, one microphone
ACG [77]	FMCW signal	17 - 19 kHz	Google Nexus 6P	one speaker, two microphones
SilentTalk [48]	---	19 kHz	Samsung Galaxy Note III	one speaker, one microphone
SilentKey [49]	CW signal	17.5 kHz	Samsung Galaxy S6, Samsung Galaxy S7	one speaker, one microphone
LipPass [50]	pilot tone	20 kHz	Nexus 6P, Galaxy S6, Galaxy Note 5, Huawei Honor 8	speaker, microphones

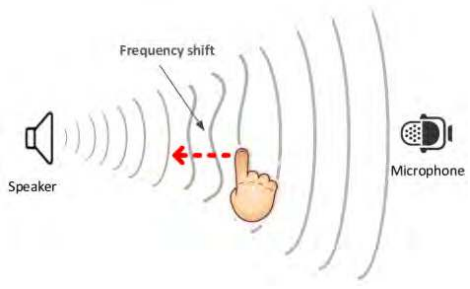

FIGURE 4. Doppler shift [59].

where f' and f are the frequency of the received sound signal from the microphone and the frequency of the original signal from the speaker, respectively; v and v_0 refer to the speed of sound in air and the hand or object relative to the microphone, respectively.

2) PHASE

The phase information of the sound signal is an important feature to depict the position of a time point on the waveform cycle and can be affected by the objects located in the propagation paths. It is usually utilized to measure the distance and track trajectory. The basic idea of using phase information for human behavior recognition is that the ultrasound signal can be reflected by the objects located in the propagation paths, which results in phase changes of the received signal. Specifically, a moving object would affect ultrasound signal propagation and change the signal phase during the sound signal propagating between the receiver and transmitter. Therefore, phase changes can be utilized to recognize human behaviors. For example, W. Wang *et al.* proposed a trajectory tracking scheme, called LLAP [36]. This system first extracts the phase change information of the ultrasound signal received by the microphone and then

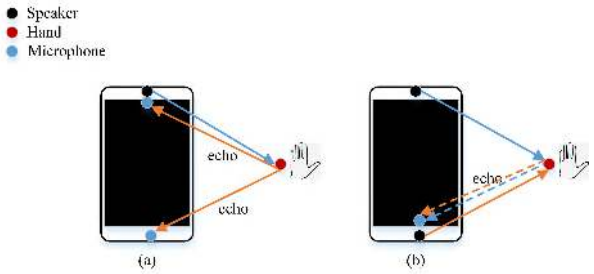

FIGURE 5. Two models of ToF. (a) Using one speaker and two microphones. (b) Using two speakers and one microphone.

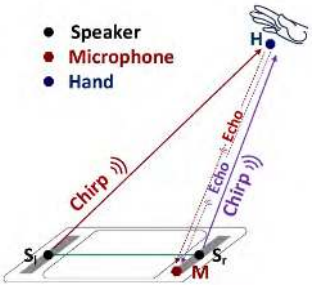

FIGURE 6. Localization scenario [56].

converts the phase changes into distance information to achieve the tracking of the hand movement.

The phase information can be expressed by (3) according to [36], which can be used to recognize behaviors. And we will analyze how this phase information obtained in Section III.C.

$$\varphi_p(t) = -(\frac{2\pi f d_p(t)}{c} + \theta_p) \tag{3}$$

where $\varphi_p(t)$ means the phase information of path p; $d_p(t)$ is the time-varying path length; c and f are the speed of sound in air and the frequency of the original signal, respectively; θ_p is the initial phase lag caused by hardware delay.

3) TOF

In addition to the commonly used phase information and Doppler shift, the time of flight can be used to track a target. It means that the speakers transmit sound signal, the signal will be reflected by the user's hand, and the microphone captures the reflected signal. The time difference between transmitting signal and capturing the reflected signal can be used to calculate the distance from the speaker through hand to the microphone. Then the distance will be utilized to localize the hand position to realize hand tracking. According to the studies on hand tracking based on the ultrasonic signal of a smartphone, there are two kinds of models usually used to calculate time difference (see in Fig. 5). Fig. 5(a) depicts that one speaker transmits the ultrasonic signal and two microphones capture the echo signal to localize the hand [58]. Fig. 5(b) shows that two speakers emit a signal and one microphone receives the echo signal to localize the hand [56]. Since signal emitted from speaker is reflected by hand, a constant propagation distance can draw an ellipse on a plane. If we want to determine the position of the hand, we may exploit the smartphone that owns two microphones and one speaker or two speakers and one microphone. Because the smartphone can generate two elliptical propagation paths and these two paths may intersect at one point. This point is the position of the hand. Because modern smartphones are usually equipped with more than one speaker and microphone, the experiments can be conducted without extra cost.

For example, as shown in Fig. 6, two speakers (S_r, S_l) built in smartphone emit a sound signal and the microphone (M) captures the reflected signal. The distance of signal propagating in the air can be expressed by (4) and (5) [56].

$$d_{S_lHM} = (t_4 - t_1) \times c + d_{S_lM} \tag{4}$$

$$d_{S_rHM} = (t_5 - t_3) \times c + d_{S_rM} \tag{5}$$

where d_{S_lHM} and d_{S_rHM} are the distance from S_l to M across H and from S_r to M across H, respectively. c is the speed of sound in the air; d_{S_lM} and d_{S_rM} refer to the distance from S_l to M and from S_r to M, respectively; t_1 and t_3 refer to the time that the microphone receives a sound signal transmitted by S_l and S_r, respectively; t_4 and t_5 are the time that the microphone receives reflected sound signal transmitted by S_l and S_r, respectively.

C. PREPROCESSING

The original signal captured by a microphone cannot be used to extract features directly due to the much noise. In addition, the signal drift of original signal stemming from the time elapses and device diversity also decreases the effectiveness of data. To improve recognition accuracy, we need to preprocess the raw signal before extracting effective features. We can take many effective processing methods such as fast Fourier transform (FFT), denoising, FFT Normalization, audio signal segmentation, and I&Q demodulation, etc. In this section, we provide an analysis of these processing approaches. Since the processing methods are different in the measurement of Doppler shift and phase information, we will analyze these methods by dividing them into two groups: Doppler shift and phase information.

1) DOPPLER SHIFT

a: FFT

The captured sound signal is a time sequence, and it represents the time-domain sound signal data. However, we could not obtain sufficient information from the time-domain data. Since the time-frequency diagram contains more useful information, we want to transform time-domain space to frequency-domain space. Fortunately, we can conduct FFT to convert the time-domain data into frequency-domain data.

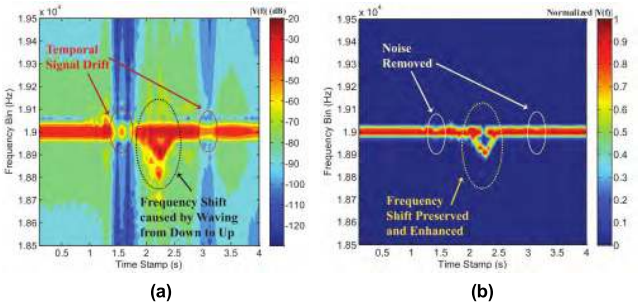

FIGURE 7. The spectrograms of the original signal and the signal after processing [40]. (a) The spectrogram of the original signal. (b) The spectrogram of the signal after the FFT-based normalization.

After getting the frequency domain data, we analyze the signal characteristics and eliminate noise. For example, we can perform some signal processing approaches such as bandpass filter or FFT normalization to obtain clean data.

b: DENOISING

Denoising can be used at two aspects including noise removal and computation complexity reduction. For example, environmental noise may occur at high frequency or low frequency which can be removed using denoising methods, such as filters and threshold vector. At the same time, filters can be used to improve computation efficiency when sampling. Specifically, the ambient noise (e.g., human activities, etc.) in the received signals can change signal waveform and decrease the quality of data. With this interference, we cannot accurately extract Doppler shift features. Therefore, we first identify these noises and then eliminate them by using various methods. For example, we can remove environment noise with threshold vector [43], keep carrier frequency and suppress noise with bandpass filter [76].

c: FFT NORMALIZATION

In addition to the ambient noise, another challenge that needs to be tackled is how to remove signal drift caused by time-elapse and hardware diversity. We can observe from Fig. 7(a) that the amplitude of different frequency bins has unpredictable signal drift. Although the magnitudes in 19 kHz bin vary from -83dB to -24dB, their relative amplitude is robust and keep in a stable range. As a result, audio spectrogram has a similar shape. Because we utilize Doppler shift to recognize behavior, we pay more attention to sharp frequency change such as peak and trough values. According to the characteristics of this noise, we can utilize FFT-based normalization (see formula (6)) to normalize the amplitude of frequency bins for each time stamps to eliminate signal drift.

$$|Y(f)| = \frac{Y(f) - Y_{min}(T)}{Y_{max}(T) - Y_{min}(T)} \tag{6}$$

where $Y_{max}(T)$ and $Y_{min}(T)$ are the maximum and minimum amplitude of frequency bins at time T, respectively; $Y(f)$ refers to the amplitude of frequency f at time T; $|Y(f)|$ means the normalized value of $Y(f)$.

For example, at the time stamp 0.5 s, we can obtain from Fig. 7(a) that $Y_{max}(T)$ and $Y_{min}(T)$ are approximately -30 dB and -95 dB respectively; $Y(19kHz)$ is about -35 dB; then $|Y(19\,kHz)|$ (see in Fig. 7(b)) is approximately 0.92 according to (6). As Fig. 7(b) shows, the signal drift is eliminated, the signal spectrum becomes smoother, and stable frequency information is kept after FFT-based normalization.

d: SEGMENTATION

If users move or perform the same motion continuously for a period of time, we should divide recorded signals into single motion elements for recognition. SilentTalk [48] recognizes soundless lip movements using the ultrasonic signal. This system identifies lip movements to recognize the content of talking without sound. Therefore, the lip movement is similar to the normal talking in spite of soundless. As a result, we can leverage the language rules to improve recognition accuracy because the content of talking is meaningful words or sentences. In other words, although we recognize lip motions without sound, we can exploit the pronunciation rules to enhance recognition precision. SilentTalk takes two steps to segment the received signal. Specifically, the sentence is first divided into words with a short silent interval based on English speech and then the word is split into syllables by inter-syllable segmentation. This system utilizes a sliding time window and short-time Fourier transform (STFT) to implement the above procedures. Different from SilentTalk, AudioGest [40] utilizes the Doppler effect to recognize hand gestures. Therefore, it focuses on frequency changes of the received signal and depicts them with an audio spectrogram. It converts the signal frequency shift into a color image and utilizes the image to depict frequency feature. Specifically, it first subtracts the normalized spectrum values and then squares the amplitude of frequency bins. After processing the whole image with Gaussian smoothing, binarization is conducted to segment the Doppler shift zone comprising peak pixels. The frequency shift is identified using image processing techniques.

e: DIMENSIONALITY REDUCTION

Dimensionality reduction refers to the reduction of the number of stochastic variables by obtaining some main variables which can represent the whole information. It discards some redundant information of the data and retains the useful information. Its purpose is to reduce computation complexity and improve system robustness. It can usually be used in feature selection and feature extraction using machine learning or deep learning algorithms. Currently, there are some common methods used in human action recognition with the audio signal on the smartphone. For example, B. Fu *et al.* [47] reduce the dimensions of the data by Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Random Projection. They validate the effect of dimension reduction with classification score using random forest classifier. From the comparison results, as shown in Fig. 8,

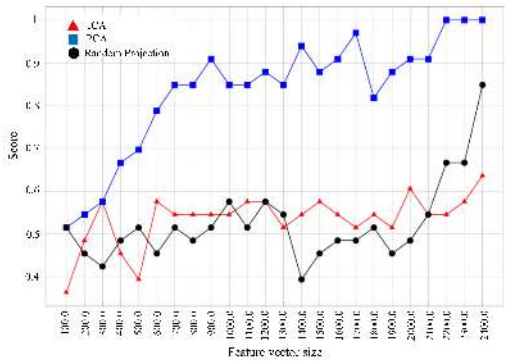

FIGURE 8. Comparison results of three types of dimensionality reduction techniques [47].

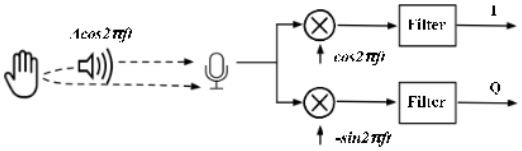

FIGURE 9. The demodulation structure.

we can observe that the three techniques have a gradual increase in the classification score as the size of the feature vector becomes larger. The results show that PCA holds the best accuracy of dimension reduction than the other methods and its classification score keeps stable when the size of the feature vector reaches a certain value.

2) PHASE INFORMATION

Different from the processing procedures of Doppler shift extraction, the received signal should be demodulated in order to obtain the phase information. We analyze the processing procedures of phase information as follows.

a: I/Q DEMODULATION

Doppler shift and phase can be used to recognize human behaviors. To obtain phase information from the captured signals, the captured signals first need to be demodulated into In-phase (I) component and Quadrature (Q) component. The I component refers to the component whose direction is the same as the received signal, while the Q component is orthogonal to the received signal. The demodulation structure is shown in Fig. 9.

As shown in Fig. 9, we obtain the two components [36]. Assuming that the emitted signal is expressed as A cos $(2\pi ft)$, the signal propagates in the air in the multi-path pattern, and the signal received from path p is shown in (7):

$$R_p(t) = 2A_p^{'} cos(2\pi ft - \frac{2\pi f d_p(t)}{c} - \theta_p) \tag{7}$$

where $2A_p^{'}$ means the amplitude of the received sound signal; $d_p(t)$ is the time-varying path length; c refers to the speed of sound in air; f means the frequency of original signal; θ_p is the initial phase lag caused by hardware delay.

Then this received signal is multiplied with $\cos(2\pi ft)$ and $-\sin(2\pi ft)$ respectively. After that, filters are used to remove the high-frequency components to get the following two components:
In-phase (I) component:

$$I_p(t) = A_p^{'} cos(\frac{-2\pi f d_p(t)}{c} - \theta_p) \tag{8}$$

Quadrature (Q) component:

$$Q_p(t) = A_p^{'} sin(\frac{-2\pi f d_p(t)}{c} - \theta_p) \tag{9}$$

We then get the phase for path p by combining the above two components as the real and imaginary parts into a complex signal, as shown in (3).

b: FILTERING

From Fig. 9 we can find that filters are used before we obtain the I and Q components. According to the characteristics of applications, different filters will be used for a different purpose. For example, we can eliminate high-frequency components with a low pass filter [52] and enhance computational efficiency with suitable sampling rate by using decimation and interpolation steps, such as cascaded integrator comb (CIC) filter [36]. After getting these two components, there still is the impact of other noise, which can be removed for calibration with the median filter [52].

D. FEATURE DESCRIPTION

Features refer to that the most effective characteristics of the original information, such as statistical information, velocity, direction, size of the target, and texture, etc. Generally, the original signal contains an amount of redundant information. Thus, we should concentrate on meaningful data. According to the state-of-the-art applications, there are two commonly used information for feature description in human behavior recognition using smartphone-based inaudible acoustic sensing: Doppler shift and phase. We will introduce how these two types of information are used for human behavior recognition in this section. And Table 4 shows the comparison results of the two information.

1) DOPPLER SHIFT

As we know, human movements will impact the frequency of acoustic signals, which results in a Doppler shift. We can utilize the frequency changes to recognize human behaviors. To obtain the frequency change information, we can analyze the time-frequency diagram of the echo signal, as shown in Fig. 10. From the diagram we can see that the diagram includes time, frequency, and amplitude, which depicts the relationship between time and signal strength on all frequencies. The diagram can also be utilized to estimate the hand moving speed, duration in air, and hand waving range by specific algorithms [40]. Specifically, the time-frequency diagram is different for different human behaviors. Fig. 10(a) represents the diagram of waving hand from right to left, while Fig. 10(b) depicts the diagram of waving hand from up to down. From Fig. 10, we observe that the time-frequency

TABLE 4. Comparison of extracted features.

Extracted feature	System	Characters	Advantages and disadvantages
Doppler shift	AudioGest [40], Dolphin [43], SonicOperator [41], H. Watanabe *et al.* [42], B. Fu *et al.* [47], B. Fu et al. [76], SilentTalk [48], DMT [59], LipPass [50], ApneaApp [51],	obvious changes in the frequency domain	sensitive to the multipath effect and complex noisy environments
Phase	FingerIO [54], LLAP [36], Strata [55], ACG [77], W. Wang *et al.* [53], SonarBeat [52]	be measured easily in the time domain	low latency and complexity

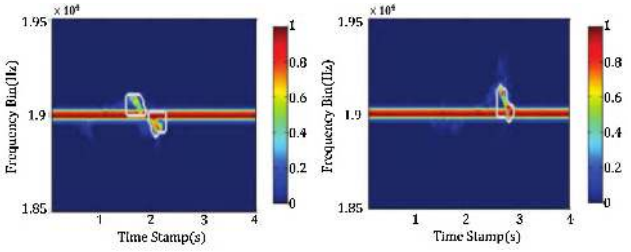

FIGURE 10. The time-frequency diagram of different movements [40]. (a) Waving hand from right to left. (b) Waving hand from up to down.

diagram describes the unique relationship between human behavior and frequency variations. Thus, we can recognize human behaviors according to the unique relationship between behaviors and frequency variations. For example, AudioGest [40] analyzes the distinct diagram of different hand postures to realize the hand movement recognition, and its recognition accuracy is up to 96%.

2) PHASE

The recognition of human behaviors can also be achieved by extracting phase changes of the received signal as the feature. And the phase change information is easier to be measured in the time domain. Specifically, the time domain complex signal obtained by (3) describes the relationship between phase changes and the distance of hand movement. Therefore, the phase change information can be used to track human hand trajectory and measure the distance. Obviously, when the hand moves towards the microphone, the phase of the dynamic vector (caused by the hand movement) will increase, as shown in Fig. 11. We can observe that the hand motion leads to the complex signal changes in a specific pattern. Since the phase of the sound signal will increase by 2π when the sound signal propagation path decreases by c/f, we can obtain the hand movement distance by calculating the phase changes. The static vector in Fig. 11 represents the reflection

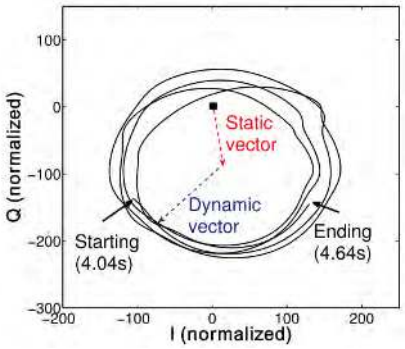

FIGURE 11. The complex signal changes due to hand motion [36].

of static objects (e.g., wall, desk, etc.) and can be removed to obtain the dynamic vector.

E. BEHAVIOR RECOGNITION

In this part, we analyze the behavior recognition techniques. We divide this part into two categories including pattern-based and model-based. The former considers the problems as a classification problem. We can train the classifier by collecting a lot of data to obtain ideal classifier parameters. The latter argues that the problems can be addressed by using mathematical models. We can get recognition label by some simple calculation without need of a large amount of data.

1) PATTERN-BASED

The aim of specific human behavior recognition (e.g., hand gesture recognition, activity recognition, and individual authentication) is to label the unknown behavior with known dataset. Therefore, this kind of human behavior recognition can be converted into a classification problem. Generally, classification problems can be effectively addressed using pattern-based methods and need more data compared with model-based methods. These methods solely utilize some common algorithms and calculate the parameters of algorithms to realize pattern classification. In this section, we will introduce some common classification methods used in smartphone-based human behavior recognition. These methods usually include some machine learning methods, such as SVM, random forest (RF), etc. Furthermore, some teams expect that the features can be extracted automatically to improve the classification accuracy. Therefore, they use deep learning methods to train neural networks to recognize and classify behaviors. Several commonly used classification methods are shown in Table 5. Although these methods can be used to classify human behaviors, all of them have limitations and advantages.

a: MACHINE LEARNING

SVM. SVM is a common method of discrimination and a supervised learning model in the field of machine learning. To make indivisible samples of low-dimensional space linearly separable, SVM transforms the samples into

TABLE 5. Comparison of classification methods.

Method		System/work	Advantages	Disadvantages
Machine learning	SVM	Dolphin [43], B. Fu *et al.* [47], SonicOperator [41], LipPass [50], SilentKey [49], B. Fu *et al.* [47]	faster classification and better accuracy when the sample size is small.	difficult to implement when the sample size is large.
	RF	B. Fu *et al.* [47], H. Watanabe *et al.* [42]	high accuracy, strong anti-over-fitting ability.	long training time, easy to fall into over-fitting on some noisy sample sets.
Deep learning	RNN	SonicOperator [41]	good performance in processing sequence data based on time and sequence with different lengths.	gradient disappearance occurs as the time interval increases.
	CNN	B. Fu *et al.* [76], UltraGesture [46]	local perception, weight sharing, multi-convolution kernel.	difficult to process sequence with different lengths.

multi-dimensional feature space by the nonlinear mapping algorithm. In other words, it constructs an optimal hyperplane to segment data to achieve the classification. In behavior recognition applications based on the ultrasonic signal of a smartphone, SVM is widely used for classification and recognition due to its good classification performance, such as [47], Dolphin [43], SonicOperator [41]. However, SVM is difficult to implement when the dimension of the sample is high because it is difficult to find the ideal hyper-plane.

RF. Random forest is a kind of classifier that contains multiple decision trees. It utilizes these decision trees to train the classification model and identify different behaviors. The random forest can generate a high accurate classifier, process multi-dimensional data without feature selection, and maintain high accuracy even a large part of features are missing. Due to its good performance, some systems utilize random forest classifier to recognize human behaviors based on the ultrasonic signal of a smartphone, such as [42], [47]. However, it will be easy to fall into over-fitting on some noisy sample set classification or regression problems.

b: DEEP LEARNING

In addition to the methods mentioned above, some teams begin to identify human behaviors with deep learning. Deep learning is a method of machine learning and learns features from dataset automatically without the need for feature extraction [79], [80]. It combines lower-level features to form more abstract high-level features. And the behavior recognition technique based on deep learning trains neural network models to perform human behavior recognition. For example, convolutional neural network (CNN) [46], [76] and recurrent neural network (RNN) [41] are introduced into human behavior recognition. Compared with the SVM method in [47], the recognition accuracy of using CNN [76] has been greatly improved.

2) MODEL-BASED

In this part, we present the model-based method used in behavior recognition applications. The model-based method usually applies mathematical model to describe the problems. Therefore, it usually needs small data compared with pattern-based method. The difficulty of model-based method is how to develop an appropriate model to illustrate the specific problem. Therefore, we usually design a specific model for a problem. For the behavior recognition using ultrasonic signal of smartphone, the applications involve user localization, hand trajectory tracking, and vital sign monitoring. To localize the user's hand and track the hand movement, the authors usually establish a time-based model to calculate distance using ToF or extract phase information to measure the change of phase caused by the target movement. As aforementioned in Section III.B, there are two time models based on time of flight utilized to realize hand localization. One is leveraging two speakers to transmit sound signals and one microphone to capture the echoes [56]. The other is employing one speaker to emit the signal and two microphones to capture the reflected signal [58]. Although the distance from one speaker to one microphone through the hand can be calculated with the time of flight, it is not sufficient to localize the hand because the hand could be any position of the ellipse. Therefore, two microphones or two speakers could be employed to yield two ellipses whose intersection point is the position of the hand.

Besides time model, the phase information of the reflected signal can be used for hand tracking. For 1D hand moving distance measurement, the phase change can be used to convert into the distance change to realize hand movement distance estimation [36]. For 2D hand tracking, the initial position of the hand and the phase change, which are obtained from the captured signal, are combined to achieve hand movement tracking [55]. Additionally, the phase change information caused by human motions can be utilized to monitor vital signs and for health assistance [52], [77].

IV. APPLICATIONS

In this section, we review the applications of human behavior recognition using smartphone-based ultrasonic signal and analyze their main characteristics in detail. Specifically,

TABLE 6. Gesture recognition.

Work	Pre-processing	Feature extracted	Experimental scenarios	Recognized behaviors	Accuracy
Dolphin [43]	Noise reduction, Normalization	Doppler shift	Quiet environment, Outdoor, Noisy environment	24 gestures	94%
AudioGest [40]	FFT normalization, Audio signal segmentation	Doppler shift	Living Room, Bus, Cafe, HDR Office	6 gestures	95.1%
SonicOperator[41]	Noise elimination, Data normalization	Doppler shift	Subway station, Restaurant, Indoor	24 gestures	95%
H. Watanabe *et al.* [42]	FFT normalization, Gaussian smoothing	Doppler shift	---	6 gestures	Increased by 15.3%

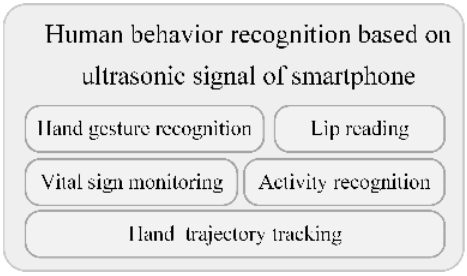

FIGURE 12. The categories of applications.

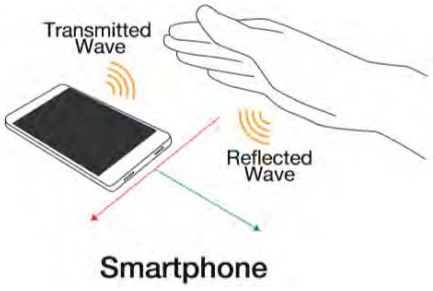

FIGURE 13. Recognize hand gestures by using the ultrasonic signal based on a smartphone [42].

we divide these applications into the following categories: hand gesture recognition, activity recognition, hand trajectory tracking, vital sign monitoring, and lip reading, as shown in Fig. 12. We present many comprehensive tables (Table 6 - Table 10) to illustrate the crucial components of different systems and explain the characteristics of these state-of-the-art applications.

A. HAND GESTURE RECOGNITION

Human gesture recognition plays a vital role in human-computer interaction (HCI) research because it can greatly enhance communication quality and help to develop various applications. Here, we focus on the hand gesture recognition using the ultrasonic signal from smartphone. With convenient hand motion identification, we can easily operate smart devices and enrich control commands. The structure of hand gesture recognition using the ultrasonic signal based on the smartphone is shown in Fig. 13. The speakers emit an ultrasound signal and the microphone captures the echo signal reflected by the user's hand. In this section, we review the hand gesture recognition systems based on the ultrasonic signal of smartphone and compare these systems in pre-processing techniques, experimental scenarios, recognized behaviors, and accuracy in Table 6.

In 2014, Q. Yang *et al.* proposed an in-air gesture recognition system based on the continuous inaudible sound signal, called Dolphin [43]. Dolphin emits a 21 kHz ultrasonic signal using the loudspeaker and captures gesture-reflecting signals using the microphone. It extracts the Doppler shift from the received signal and recognizes gestures using continuous Doppler shift sequences. Different from other systems, Dolphin employs a two-step recognition method. It first defines 10 groups of gestures called predefined groups and some groups include some finer granular gestures. Therefore, for an unknown gesture, Dolphin first classifies the gesture into a predefined group with a manual method. If this predefined group has no finer granular gestures, Dolphin obtains its label as the classification result of the unknown gesture. Otherwise, Dolphin further classifies the gesture into a finer granular gesture with machine learning methods.

Specifically, Dolphin collects data and generates a frequency vector with 60 points by using FFT. After noise elimination and normalization, it obtains a concise shift sequence M that depicts the basic change of movement direction and comprises -1 and 1. Based on the M, Dolphin determines the predefined group of the unknown gesture. The process of determining gesture's predefined group is called manual method because it solely compares two digital sequences comprised of -1 and 1. Because some gestures have a similar movement pattern, we can categorize them into a group using the compressed vector M. If we cannot identify the label of this gesture, we need to make a further judgment in the group. Based on the group, we can further use common machine learning algorithms to train classifiers to recognize them. When needing further to determine the finer granular gesture, a larger vector with 1800 points is constructed and fed into machine learning algorithm such as Native Bayes, K-nearest neighbor classifier, Bayes Net, Random Tree, Large Linear

classifier (Liblinear), and SVM, to obtain the finer granular gesture's label as the classification result of unknown gesture. The reason to employ two steps recognition process is that it is difficult to recognize a set of 24 gestures using a simple machine learning classifier. Three different android devices are used to validate the system performance with 7 classifiers in 3 environments. This system recognizes 24 pre-defined gestures with an average accuracy up to 94%, such as simultaneous waving of both hands and one-handed in different directions.

In 2016, W. Ruan *et al.* proposed a training-free hand gesture recognition system using inaudible audio signal, AudioGest [40]. This device-free system first analyzes weak echo signal mixed with ambient noise and then conducts FFT to identify the Doppler shift caused by hand movement. The authors apply FFT normalization to eliminate the audible noise and signal drift and obtain audio signal segmentation using Gaussian smooth filter. Next, the system interprets the relation between audio spectrograms and hand actions. It builds a model to estimate hand speed using the frequency shift. Based on the interpretation, the authors obtain hand moving speed, direction, duration, and distance. The authors conduct comprehensive experiments to validate the system performance with three electronic devices and 3900 hand gestures from 5 users in 5 test environments. Results demonstrate that AudioGest achieves an accuracy up to 95.1% with 6 hand gestures (e.g., up-down, down-up, etc.). Furthermore, it incorporates three aspects including moving speed, in-air duration, and waving range into hand gesture recognition. Besides, it considers six directions of motion and three levels of speed, duration, and waving range (e.g., slow speed, medium speed, fast speed, etc.), as shown in Fig. 14. Theoretically, AudioGest can recognize up to $6\times3\times3\times3=162$ gestures by combining these waving attributes. Practically, it can accurately identify 6 hand gestures due to the measurement accuracy of one microphone.

In 2017, X. Li *et al.* presented an in-air hand recognition system based on ultrasound signal and Doppler effect, called SonicOperator [41]. This system utilizes the Doppler shift caused by different hand gestures to realize the classification for gestures. SonicOperator first performs FFT to transform time domain data into frequency domain data, then eliminates ambient noise and normalizes the sample data. Afterward, it utilizes the transfer learning method to train RNN and alters the training objective function to transfer the knowledge of feedforward neural network into RNN. The system validates its performance with a set of 24 gestures and 36000 samples in 3 environments. The experimental results show that SonicOperator can recognize 24 pre-defined gestures with an average accuracy of 95%.

In 2018, H. Watanabe *et al.* [42] proposed a method using a smartphone with a cover shielding microphone to improve gesture recognition accuracy. They find that the frequency changes usually are similar when the user performs some hand gestures, which makes it difficult to classify them. They think that a cover on a microphone can modify the

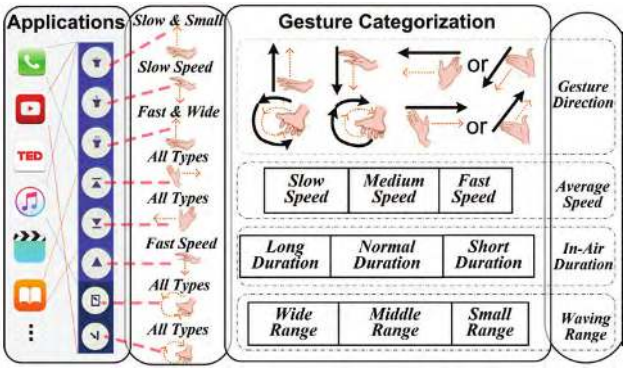

FIGURE 14. Hand gestures by combining waving attributes [40].

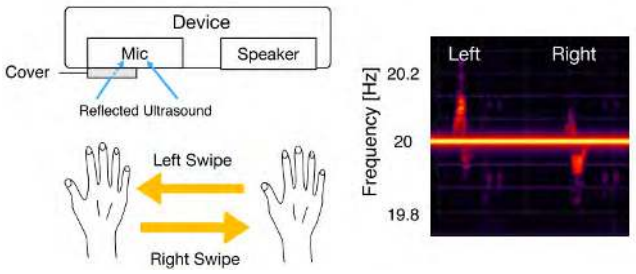

FIGURE 15. Spectrograms of different gestures when the microphone is covered [42].

propagation path and change the characteristics of the received signal. Therefore, the Doppler effect will be altered and more features can be obtained, leading to the improvement of recognition accuracy. The authors investigate the time series and extract 10 features. Meanwhile, they examine spectrogram and compute 9 features. The authors utilize 19 features to validate the system performance with 8 participants performing 6 hand gestures using 4 different types of devices covered four shield forms (e.g., half shield, hole, sponge, and directivity). As shown in Fig. 15, there is obvious difference between the right swipe and left swipe when the microphone is covered. The experimental results show that the average recognition accuracy is improved 15.3% by covering the microphone.

B. ACTIVITY RECOGNITION

Accurate recognition of human activities is the research direction of future intelligent life. By recognizing basic activities in daily life (e.g., bending, kicking, stretching arms, etc.), the interaction quality between people and the environment or smart devices can be improved, and the scope of intelligent applications can also be expanded. In this section, we review the activity recognition systems based on the ultrasonic signal of smartphone and compare these systems in preprocessing techniques, experimental scenarios, recognized behaviors, and recognition accuracy, as shown in Table 7.

In 2017, B. Fu *et al.* [47] proposed a whole-body activity recognition system based on the inaudible sound signal. It utilizes a speaker and a microphone on the same smartphone to

TABLE 7. Activity recognition.

Work	Pre-processing	Feature extracted	Recognized behaviors	Accuracy
B. Fu *et al.* [47]	FFT normalization, Dimensionality reduction	Doppler shift, Peak-to-peak distance	Bicycle, Squat, Toe-touch	73%~92%
B. Fu *et al.* [76]	Bandpass filtering	Doppler shift	Bicycle, Squat, Toe-touch	Bicycle: 88%, Squat: 91%, Toe-touch: 97%

FIGURE 16. Motions and the position of smartphone [76].

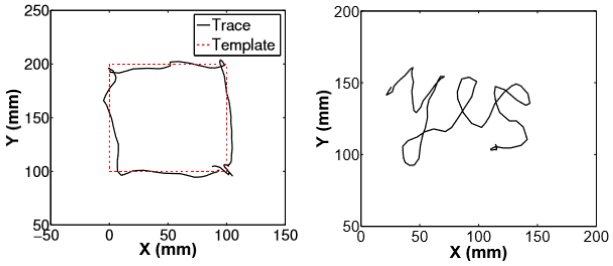

FIGURE 17. Sensing results of drawing in air. (a) Drawing square. (b) Drawing word [36].

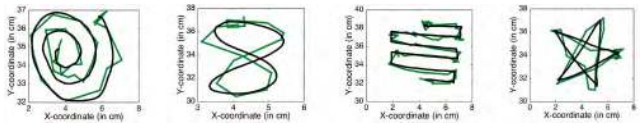

FIGURE 18. Tracking accuracy. The black lines refer to the ground truth trace while the green lines refer to the trace that FingerIO tracks [54].

emit and capture ultrasound signal and analyzes the sound spectrogram based on Doppler shift. The distance between amplitude peaks can be extracted to facilitate improving classification accuracy. Because the frequency changes caused by person actions can be efficiently identified from the spectrogram, authors recognize three motions (bicycles, squats, and toe touches, see in Fig. 16) by using machine learning classification methods. This system achieves an accuracy between 73% and 92% for different motions and classification methods. In the next year, they designed the CNN model [76] to classify the three motions with an accuracy of 88% for bicycles, 91% for squats, and 97% for toe touches.

C. HAND TRAJECTORY TRACKING

The handwriting input is an important requirement for many smart devices and has been widely studied because it can achieve fast information input and enrich human-computer interaction procedures. However, those early handwriting recognitions usually need to touch the screen, which may be inconvenient in some scenarios. To further improve the quality of HCI, researchers have extended the early touch-based recognition to in-air trajectory tracking. With this HCI method, we can easily input information on small screens or use screen simultaneously. Currently, there are some studies about hand trajectory tracking using smartphone-based inaudible acoustic signals, which greatly promotes the progress of smart sensing techniques and HCI means. In this section, we review the hand tracking systems based on the ultrasonic signal of smartphone and compare these systems in preprocessing techniques, experimental scenarios, recognized behaviors, and recognition accuracy, as shown in Table 8.

In 2016, W. Wang *et al.* proposed a hand gesture trajectory tracking scheme called LLAP [36]. It leverages the speakers and microphones on same smartphone and builds a device-free hand gesture tracking system. The idea of this system is that ultrasound signal phase changes caused by hand or finger can be effectively measured and converted into movement direction and movement distance. This system achieves mm-level accuracy for hand movement distance by using sound phase features and implements two-dimensional hand gesture recognition by using multiplexing continuous waves. The authors use I/Q demodulation to obtain the complex signal and then separate it into static vector and dynamic vector. The former stems from LOS path or static objects and the latter comes from the hand moving. Therefore, the authors obtain the moving distance by calculating the phase changes according to the dynamic vector. This system achieves a hand gesture tracking with higher accuracy, lower latency, and much higher speed solution. For 2-D tracking, the LLAP obtains a tracking error of 4.57mm and achieves accuracy of 92.3% for 26 Latin letters and 91.2% for some words such as ''yes,'' ''can'', and ''bye'' (see in Fig. 17). At the same time, it implements a low latency of less than 15ms on the smartphone.

The same year, R. Nandakumar *et al.* proposed a fine-grained finger trajectory tracking solution, called FingerIO [54], which can track any pattern that user draws and achieve millimeter-level tracking accuracy (see in Fig. 18). FingerIO utilizes the speakers and microphones on the same smartphone. Specifically, the speaker emits an inaudible

TABLE 8. Hand trajectory tracking.

Work	Pre-processing	Feature extracted	Experimental scenarios	Recognized behaviors	Accuracy
LLAP [36]	I/Q demodulation, CIC filtering	Phase	Static, Music and Speech environments	26 letters, 11 words	Letters:92.3% Words:91.2%
FingerIO [54]	Set distance threshold, Fine-tune distance estimate	Phase, Time difference	Office	All kinds of shapes	Median error: 8 mm with smartphone
Strata [55]	Bandpass filtering, Frame detection	Phase	Student office	Diamond, Triangle, Circle	0.6 cm drawing error
EchoTrack [56]	Bandpass filtering, Multipath elimination	ToF	Laboratory	Straight line, Triangles	76% within 3 cm error, 48% within 2 cm error
SteerTrack [58]	Eliminating symbol time offset	Time of arrival (ToA)	Real driving environments: local road and highway	Track the rotation angle of the steering wheel	4.61 degrees error
DMT [59]	FFT normalization	Doppler shift	Laboratory, Hall, Outdoor	All kinds of finger gestures	More than 95%

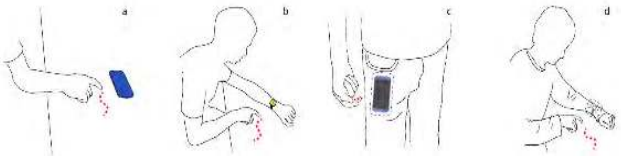

FIGURE 19. Applications. (a) Any surface could be a writing surface. (b) A FingerIO device with smartwatch form. (c) Track movements when the smartphone is in the pocket. (d) Track with an occluded smartwatch [54].

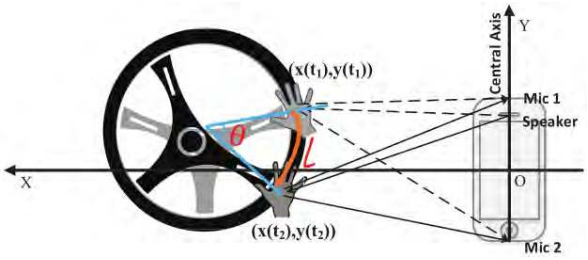

FIGURE 20. Leveraging the hand movement trajectory to track the rotation angle of the steering wheel [58].

sound signal in the form of OFDM with cyclic suffix and the microphones capture the echoes affected by finger moving. With OFDM signal, it corrects the sample errors and enhances finger recognition accuracy. It validates the system performance using a smartphone and a smartwatch with the built-in speakers and microphones. FingerIO achieves an average tracking accuracy of 8 mm with a smartphone and 1.2 cm with a smartwatch in 2D finger tracking. It can work well in the interaction space of 0.5 m $\times$ 0.5 m region and even the phone is in the pocket which occludes from the finger, as shown in Fig. 19.

The tracking accuracy of the above two schemes is still limited to the effect of multipath propagation and other motions. To reduce the impact of these two factors, S. Yun *et al.* proposed Strata [55] in 2017. Strata first estimates the relative distance change of the finger by extracting phase changes from the channel tap which corresponds to the finger movement. Moreover, the absolute distance of the finger is estimated by Strata using the changes in the channel impulse response. Then, Strata combines the relative distance

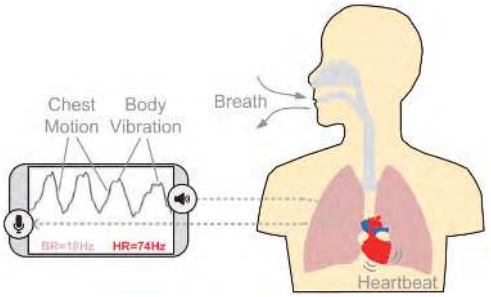

FIGURE 21. Vital sign - heartbeat monitoring using the ultrasonic signal based on a smartphone [77].

and the absolute distance to achieve high precision trajectory tracking. It can estimate the distance change according to the phase and also can calculate the absolute distance from the channel variation. And the comprehensive experiments validate the system performance. The median tracking errors of Strata, LLAP, and improved FingerIO are 0.3 cm, 0.7 cm, and 1.5 cm in 1D environment and 1.01 cm, 1.9 cm, and 3.47 cm in 2D scenario. Strata can get the average error within 0.6 cm when drawing a simple shape in a 2D space. It also achieves low latency with the position update every 12.5ms.

In 2018, X. Xu *et al.* presented SteerTrack [58], a device-free steering tracing system tracks the rotation angle of steering wheel based on the ultrasonic signal from a smartphone, as shown in Fig. 20. It utilizes the speaker built in smartphone to emit ultrasonic signal, captures sound echoes by the microphone, and then leverages the relative correlation coefficient (RCC) and reference frame to analyze the hand movement trajectory. Then, it maps the steering wheel in 3D to 2D ellipse. It designs a method based on geometrical transformation to estimate the rotation angle of the steering wheel according to the hand movement trajectory. Extensive experiments with 5 participants and 5 different smartphones for 6 weeks are conducted to evaluate the system performance. It can recognize three different steering maneuvers with an accuracy of 97.73%. SteerTrack also can estimate the rotation angle of the steering wheel with an average accuracy of 4.61 degree error.

TABLE 9. Vital sign monitoring.

Work	Pre-processing	Feature extracted	Experimental scenarios	Recognized behaviors	Accuracy
ApneaApp [51]	---	Doppler shift	Harborview sleep laboratory	Sleep apnea	99.2%
SonarBeat [52]	I/Q demodulation, Low-pass filtering, Adaptive median filtering	Phase	Office, Bedroom, Movie theater	Breathing rate	Media error: 0.2 bpm
W. Wang *et al.* [53]	I/Q demodulation, High-pass filter	Phase	Various use cases	Tremor detection	---
ACG [77]	Dual microphone cancellation, Baseband signal phase processing, Adaptive filter	Phase	Standard office environments	Heart rate, Heartbeat interval	Estimation error: heart rate: 0.6bpm, heartbeat interval: 19ms

In 2019, W. Liu *et al.* proposed DMT [59], a device-free finger motion tracking system based on the ultrasonic signal from a smartphone. It uses a smartphone with multiple speakers and microphones to transmit and receive an inaudible sound signal. DMT leverages a Fourier fitting algorithm to detect minute Doppler shift. With the frequency changes, authors develop a geometric method to track the hand trajectory and recognize hand gestures by an image matching algorithm. It effectively eliminates the temporal signal drift and diverse-device drift based on a threshold method. Afterward, using tracking model, it obtains the hand trajectory by linking the finger position. It leverages the particle filter to determine the initial position and exploits an image matching algorithm to match the gesture. Extensive experiments validate system performance. DMT achieves Doppler shift recognition accuracy of more than 90% and finger motion identification accuracy of more than 95%.

D. VITAL SIGN MONITORING

Human vital signs (e.g., breathing, heartbeat, hand trembling, etc.) reflect a person's physical health condition. Accurate measurement and long-term monitoring of these vital signs enable us to obtain the participant's physical condition timely and take precautions in advance. Therefore, vital sign monitoring has drawn wide attention due to its significance to one's health. Instead of wearing-based monitoring techniques that require subjects to wear or attach some sensors on the body and are infeasible in some scenarios, the device-free pattern provides more advantages which are suitable for long-term monitoring. The vital sign monitoring using smartphone-based ultrasonic signal is increasingly drawing more attention, such as heartbeat monitoring (see in Fig. 21), diagnosis of Parkinson's disease, breathing monitoring, etc. In this section, we review the vital sign monitoring systems based on the ultrasonic signal of smartphone and compare these systems in preprocessing techniques, experimental scenarios, recognized behaviors, and accuracy, as shown in Table 9.

In 2015, R. Nandakumar *et al.* designed a contactless system, called ApneaApp [51], to identify apnea events during sleep with a smartphone. It measures the chest and abdomen movements from respiration and uses FMCW signal as the transmitted sound signal. Authors leverage this signal to obtain time difference calculated from frequency shift and extract the amplitude changes of chest movements due to respiration. Meanwhile, multiple breathing signals can be tracked simultaneously because they experience different propagation time. Other non-breathing body movements can be determined due to different reflection features. And they proposed some algorithms including central apnea algorithm, obstructive apnea algorithm, and hypopnea algorithm and analyzed the echoes for apnea-hypopnea index estimating. Therefore, many apnea events including central apnea, obstructive apnea, and hypopnea, can be effectively detected. Extensive experiments with 37 participants and four different sleeping positions demonstrate that this system achieves concurrent breathing movement tracking from multiple users. ApneaApp achieves a respiration frequency accuracy of 99.2% when the smartphone is placed within 1 m away from the user. This system also computes the mean error of Apnea-Hypopnea Index (AHI) which is the average rate of apnea events during the sleep duration. The mean error of AHI is 1.9 event/hr.

In 2017, Wang *et al.* proposed SonarBeat [52], a device-free vital monitoring system to monitor the breathing rate using the phase changes of the received sound signal. It detects the periodic rise and fall of signal caused by the chest movement. Specifically, the speaker emits an ultrasonic signal as a CW radar. The microphone of the same phone captures the reflected echoes. This system first obtains I/Q values after signal preprocessing and then extracts a useful phase to estimate the breathing rate after phase unwrapping. Comprehensive experiments with 5 participants in three different scenarios for three months are conducted to validate system performance. It achieves a mean estimation error of 0.2 bpm for breathing rate. It also confirms its robustness to a different direction, different ranges, and different breathing rates of different users.

In addition to monitoring the breathing, Wang *et al.* [53] proposed a tremor detection application using smartphone-based inaudible acoustic sensing for the early diagnosis of Parkinson's disease in 2017. It detects the hand movements by extracting the phase changes of the reflected signal and then determines whether the hand is static, moving or trembling.

Although users do not need to hold the smartphone, their hand needs to move within 30 cm around the mobile phone. Additionally, the tremor detection application can further measure some parameters such as trembling frequency and trembling magnitude to determine the intensity of the tremor.

In 2018, Qian *et al.* proposed a passive heartbeat monitoring system using an ultrasonic sound signal, called Acoustic-cardiogram (ACG) [77]. It leverages the speaker embedded in a smartphone to emit the inaudible FMCW sound signal and utilizes the two microphones to receive the echoes containing heart rate and heartbeat rhythm. Two microphones can effectively eliminate the pseudo self-interference coming from direct power leakage by comparing the difference of the signals of them. This system first obtains the heartbeat signal from the mixed signal including ambient noise and respiration, then calculates the phase changes of the sound signal. Next, authors mitigate the echo from speaker to receiver and utilize the spectrogram of the baseband signal and PCA analysis to obtain chest motion from breathing and body vibration from the heartbeat. Then it recognizes heart rate using the peak value of frequency domain, extracts the heartbeat data from environment noise with IIR comb notch filter, and adopts EM algorithm to obtain individual heartbeats. The results of the experiments with 10 subjects show that it can monitor heart rate with a median error of 0.6 beats per minute and estimate heartbeat rhythm with a median average error of 19 ms.

E. LIP READING

The traditional speech recognition systems have some shortcomings: it can easily degrade the performance due to the ambient noise; it cannot be applied to special places (e.g., study room, sickroom, etc.) that require being quiet; it is not suitable for the speech impairment. Nowadays, the smartphone-based lip language recognition system can provide a promising solution to these questions and has attracted more attention of researchers due to the rapid development of smartphones. The basic principle of these systems is that users make soundless lip movements and the lip motions will reflect the ultrasonic signal emitted by the speaker (see in Fig. 22). Then the echo signal received at the microphone will be analyzed to realize lip language recognition and individual authentication. In this section, we review the lip reading systems based on the ultrasonic signal of smartphone and compare these systems in preprocessing techniques, experimental scenarios, recognized behaviors, and recognition accuracy, as shown in Table 10.

In 2017, J. Tan *et al.* proposed SilentTalk [48], a device-free lip language recognition system based on the Doppler effect of ultrasonic signal. The speaker transmits ultrasonic signals and the microphone from the same smartphone receives the echoes affected by mouth movements. Then, a band-pass Butterworth filter and an adaptive filter are used to remove interference and time difference of arrival (TDoA) are utilized to suppress multi-path noise. The system first analyzes the frequency shift and then qualifies the relationship between the

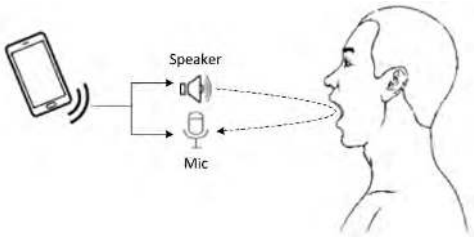

FIGURE 22. Soundless lip movement recognition using the ultrasonic signal based on a smartphone [48].

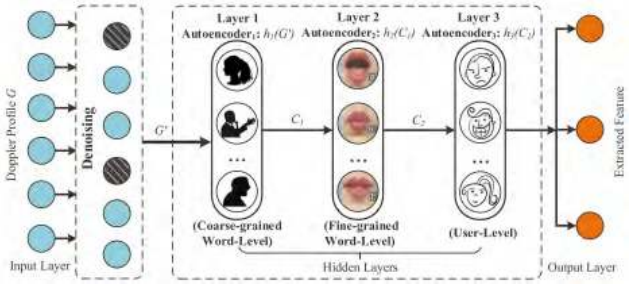

FIGURE 23. Structure of autoencoder-based DNN model [50].

frequency change and mouth movement using a Frequency Shift Detection Model (FSDM). Then, it employs a Continuous Lip Reading Model (CLRM) on FSDM to recognize continuous lip reading. The experiment with 10 volunteers and 12 basic mouth motions is conducted to validate the system performance. SilentTalk can recognize 12 different English syllables (e.g., b, f, d, etc.) with an accuracy up to 95.4%. It also achieves the average accuracy of identifying short sentences whose length is less than 6 words up to 74.8%.

Different from the SilentTalk which recognizes the human language, LipPass [50] is a user identity authentication system using the speaker and microphone embedded in a smartphone. In 2018, L. Lu *et al.* proposed the device-free recognition system by leveraging distinct behavioral schemes of user's lip movements. LipPass analyzes the Doppler shift caused by peoples' lip movements and finds the unique lip movement patterns of users. It utilizes a 3-layer deep neural network (see in Fig. 23) to extract useful features from Doppler profiles to describe the lip movement rule. Next, authors apply SVM and Support Vector Domain Description (SVDD) to build a binary classifier and a spoofer identifier. For multiple words, authors develop a weighted voting method to enhance authentication accuracy. The experiment with 48 participants in 4 real scenarios is conducted to validate the system performance. LipPass achieves average 90.21% identification accuracy and 93.1% detection accuracy of a spoofer.

Similarly, Tan *et al.* proposed another authentication system in 2018, called SilentKey [49]. This system leverages a speaker to transmit 17.5 kHz continuous wave signal and a receiver to capture the ultrasonic signal to identify lip reading. A sequence of specific mouth actions of a user can be used to build a unique feature and determine a user's identification. Specifically, it analyzes the signal changes

TABLE 10. Lip reading.

Work	Pre-processing	Feature extracted	Experimental scenarios	Recognized behaviors	Accuracy
SilentTalk [48]	Bandpass filtering, LMS filtering	Doppler shift	Library, Office, Seminar room	12 basic mouth movements, Short sentences up to 6 words	Basic mouth movements:95.4% short sentences:74.8%
SilentKey [49]	Bandpass filtering	Signal amplitude	Noise and quiet environments	Individual authentication	TPR: 70% - 83.1% TNR:86.7% - 90.7%
LipPass [50]	Passphrase Segmentation	Doppler shift	laboratory, pub, train station, dark laboratory,	Individual authentication	90.21%

caused by the minute mouth movements and extracts unique feature description reflecting unique user identification. It utilizes a Hilbert transform to analyze signal's envelopes to build mouth movement profiles. It extracts three kinds of features including rhythm, duration, and envelope differences of reflections to construct feature vector. It leverages DTW to quantify the envelope differences, utilizes an improved DTW to measure the distortion degree of signal and speaking rhythm, and applies a peak detection method to determine the start and end position of the echo signals and the duration time. The authors recognize user's identity using SVM with above 3D feature vectors. The experimental result with 50 participants shows that SilentKey achieves a rate of true positive (TPR) around 70% to 83.1% and a rate of true negative (TNR) around 86.7% to 90.7% for individual authentication. SilentKey can face some attacks because the input cannot be duplicated. It also studies the impact of basic mouth movements, such as letters, syllables, and numbers on echo signals and designs effective mouth movement sets for ultrasonic identification recognition.

V. LIMITATIONS AND FUTURE DIRECTIONS

Over the last decades, researchers have proposed many promising human behavior recognition methods based on the ultrasonic signal. With the popularity of the smartphone, behavior recognition studies based on the inaudible acoustic signal are constantly emerging because they leverage the powerful sensing capability of built-in speakers and microphones, which extends the smartphone to an active sonar device. However, speakers and microphones embedded in smartphones are designed for common conversation, which usually does not meet our requirements when treating them as acoustic sensors. Therefore, we will face many challenges when employing a smartphone as a signal transceiver. Moreover, different types of mobile phones own their distinct hardware features, which increases more difficulties when developing behavior recognition algorithm. Although many researchers have presented many effective methods and developed some prototype systems to solve these problems, we have to consider many limitations when developing related applications. Meanwhile, we also try our best to design potential applications in this field. In this section, we discuss the limitations as well as possible future research directions of behavior recognition technology based on ultrasonic signal of the smartphone. Many research methods based on audible sound can be employed in ultrasonic applications.

A. LIMITATIONS

1) HETEROGENEITY

Currently, there are various smartphone brands and they are comprised of several electronic devices, leading to diverse hardware characteristics. These significant differences bring about a considerable impact on algorithm design when measuring physical signal and processing data using these phones. Device diversity hinders the scalability of the algorithms proposed because many parameters of the algorithm may depend on the specific devices. Meanwhile, hardware signal drift and noise interference are also crucial factors in signal measurement. Because accurate data collection is the premise of behavior recognition, we have to consider these features when developing and evaluating the algorithm. A possible solution to this problem is to apply transfer learning algorithm.

2) ROBUSTNESS

Most applications of behavior recognition leverage phase [36], [54], Doppler shift [40], [42], and ToF [56], [58] to measure signal reflection, propagation distance, and moving velocity of hand actions. These features are sensitive to the test environments and can easily be affected by the test scenarios. Complicated indoor environment and multi-path effect can decrease recognition accuracy. Therefore, many evaluations of the system usually are conducted in quiet scenarios where there is seldom environment interference. For example, the behavior recognition accuracy may decrease when identifying a person's movement if there are other persons passing the test area. How to design a more robust algorithm to mitigate environmental noises is still a challenging problem.

3) MOVEMENT TYPES

We can divide these applications into two groups, such as vital sign monitoring and action recognition. The former identifies the periodic and rhythm features to estimate respiration rate and heart rate. It just calculates a single frequency value. The latter measures the acoustic signal, segments it into discrete sequences, and feeds them into a classification algorithm. Although theoretically, we can classify a number of action types, we just identify some limited actions due to

weak measured data or insufficient data collection. The typical applications are AudioGest [40] and Ipanel [22] that can recognize more actions compared with other implementations. To recognize more movements, we require more data processing algorithms to differentiate more types of human movements and enhance the resolution of signal measurements.

4) STANDARDIZED DATASET

Most researchers evaluate their system using some specific actions conducted by some participants, which makes the test data work well under the specific environment. However, it is difficult to compare and evaluate system performance because there are not a common standardized dataset and estimation criteria. Therefore, standardized human gesture data are required to evaluate the current various recognition algorithms. As for development of standard dataset, we can obtain some valuable experience from many successful datasets, such as CIFAR10, IMDB, Fashion-MNIST, ImageNet, MNIST, etc. When we want to build a successful dataset, we must consider various factors, such as types of actions, age of users, noise level of experimental scenarios, interference nearby, distance between user and smartphone, orientation of the smartphone, volume of the speaker, number of the participants, etc. Besides these common factors, we hope that the test covers more daily activities (e.g., walking, running) and more smartphone states (e.g., in the pocket, NLOS condition). As for the vital signal monitoring, it is recommended to measure the data using medical devices approved by medical management institution. We also hope the dataset can be used at model-based, pattern-based or deep learning-based human behavior recognition. Therefore, the size of the dataset should meet the need of large amount data.

B. FUTURE DIRECTIONS

1) MULTI-MODAL DATA FUSION

Many hand gesture applications based on inaudible acoustic signal have achieved a good recognition accuracy using unimodal measurement data. However, multi-modal data will improve identification performance because we can take many data fusion algorithms, which enables us to suppress noise and increase sufficient data. For example, we can use multiple speakers and microphones because the modern smartphone contains many sound sensors to enhance communication quality [77]. Besides, we can incorporate signal phase, Doppler frequency shift, and ToF into data measure and algorithm design, which can effectively improve the accuracy of data collection and increase recognition precision [21].

2) APPLICATION EXTENSION

Nowadays, behavior recognition based on ultrasonic signal has been applied in many scenarios. Besides common gesture recognition, its application area has been extended to disease diagnosing, information input, and vital sign detection, etc. Although current research has achieved some essential progress, deeper studies and more advanced algorithms are still required to enlarge the acoustic sensing research ranges. For example, we hope to diagnose more diseases, input more data, and present more helpful health-care assistance. Besides, many novel applications are applying acoustic-based methods such as liveness detection [81], [82].

3) SECURITY ISSUES

Smartphone-based human behavior recognition using ultrasonic signals is a novel method for HCI. With zero-cost deployment and convenient interaction scheme, this method has drawn more research interest. However, it brings some issues about security. As an example, the human cannot hear the ultrasonic signals, thus it might be utilized to recognize human motion [83], identify keystroke sequences [84], guess Android unlock pattern [85], and infer user's input text [86], which will steal people's information and violate people's privacy. Seriously, this method can be utilized to conduct ultrasound attacks and perform control commands [87].

4) DEEP LEARNING

Nowadays, deep learning has gained striking attention due to its overwhelming performance advantages in image processing, speech recognition, natural language processing, and recommendation systems [79]. Naturally, we want to utilize its excellent capability of feature representation and feature extraction to improve the recognition accuracy of human behavior identification based on the ultrasound signal. Recently, there are many studies based on audible sound signal employing the deep learning method. For example, Ipanel [22] applies CNN to extract features to recognize common user gestures (e.g., click, flip, scroll, zoom, etc.) and handwriting (10 numbers and 26 alphabets). Ipanel achieves encouraging recognition accuracy and provides compelling evidence that the application of deep learning can significantly improve identification precision. And the performance of the system outperforms that of the other studies. These results indicate that deep learning can largely enhance system performance and increase recognition accuracy. Therefore, we can apply deep learning methods to human behavior recognition based on ultrasonic signal of the smartphone.

VI. CONCLUSION

Recently, IoT has made encouraging progress and many novel applications have been increasingly emerging. It enables us to link and access more devices and perform more human-machine interactions. Therefore, we need more convenient interaction methods to communicate with devices effectively. Among the existing interaction means, the method based on ultrasonic signal holds many striking advantages and its applications are constantly surging. It can provide fine-grained action identification and does not interfere with a person's normal life. Apart from interaction, sensing based on ultrasound signal has many potential applications, such as vital signal monitoring, identification authentication, information stealing, and data input, etc. With the popularity of smartphone, human behavior recognition leveraging the embedded-in speakers and microphones is becoming a hot

research topic because it presents cost-effective deployment and convenient interaction pattern. These studies allow us to implement ubiquitous data acquisition and information processing, which significantly enriches the IoT applications.

In this paper, we first present conventional human behavior recognition and review the state-of-the-art applications applying the following techniques, such as video, light, RF, and audio. Then we introduce the fundamental principle of the ultrasonic-based recognition system and analyze the main signal properties from the echo signal, including phase, frequency shift, and ToF. Next, we present the architecture of the behavior recognition system based on ultrasonic signal and summarize the contents from basic signal selection to behavior recognition. Specifically, the speaker first transmits ultrasonic signal, the microphone captures the echoes changed by human behavior from microphone, then the system conducts signal preprocessing procedures to eliminate noises and the main algorithm to extract useful features and classify human actions. Afterward, we investigate in detail the state-of-the-art applications of the behavior recognition using smartphone-based ultrasonic sensing in five areas, including hand gesture recognition, activity recognition, hand trajectory tracking, vital sign monitoring, and lip reading. We present many tables to exhibit and compare the crucial components of different systems and interpret the characteristics of these state-of-art applications. Finally, based on current study trends, we discuss the limitations and open issues involved in human behavior recognition based on ultrasonic signal of the smartphone.

REFERENCES

[1] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A survey on ontologies for human behavior recognition," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 1–33, Apr. 2014.
[2] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2013.
[3] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li, "Device-free WiFi human sensing: From pattern-based to model-based approaches," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 91–97, Oct. 2017.
[4] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2015.
[5] T. Li, Q. Liu, and X. Zhou, "Practical human sensing in the light," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services*, Singapore, 2016, pp. 71–84.
[6] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 90–102.
[7] C. Cai, R. Zheng, and M. Hu, "A survey on acoustic sensing," Jan. 2019, *arXiv:1901.03450v1*. [Online]. Available: https://arxiv.org/abs/1901.03450v1
[8] U. Lee and J. Tanaka, "Finger identification and hand gesture recognition techniques for natural user interface," in *Proc. 11th Asia–Pacific Conf. Comput. Hum. Interaction*, Bengaluru, India, 2013, pp. 274–279.
[9] J. Beh, D. Han, and H. Ko, "Rule-based trajectory segmentation for modeling hand motion trajectory," *Pattern Recognit.*, vol. 47, no. 4, pp. 1586–1601, Apr. 2014.
[10] X. Li, Y. Zhang, J. Zhang, Y. Chen, H. Li, I. Marsic, and R. S. Burd, "Region-based activity recognition using conditional GAN," in *Proc. ACM Multimedia Conf.*, Mountain View, CA, USA, 2017, pp. 1059–1067.
[11] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1993–2008, Nov. 2013.
[12] A. Bux, P. Angelov, and Z. Habib, "Vision based human activity recognition: A review," in *Advances in Computational Intelligence Systems*. Cham, Switzerland: Springer, 2017, pp. 341–371.
[13] T. Li, C. An, Z. Tian, A. T. Campbell, and X. Zhou, "Human sensing using visible light communication," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 331–344.
[14] M. Kaholokula, "Reusing ambient light to recognize hand gestures," Dartmouth Comput. Sci., Dartmouth College, Hanover, NH, USA, Tech. Rep. TR2016-797, May 2016.
[15] R. H. Venkatnarayan and M. Shahzad, "Gesture recognition using ambient light," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–28, Mar. 2018.
[16] Y. Zou, J. Xiao, J. Han, K. Wu, Y. Li, and L. M. Ni, "GRfid: A device-free RFID-based gesture recognition system," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 381–393, Feb. 2017.
[17] S. Tan and J. Yang, "WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition," in *Proc. 17th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Paderborn, Germany, 2016, pp. 201–210.
[18] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6258–6267, Jul. 2017.
[19] W. Li, Y. Xu, B. Tan, and R. J. Piechocki, "Passive wireless sensing for unsupervised human activity recognition in healthcare," in *Proc. 13th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Valencia, Spain, 2017, pp. 1528–1533.
[20] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, Miami, FL, USA, 2013, pp. 27–38.
[21] J. Liu, Y. Wang, G. Kar, Y. Chen, J. Yang, and M. Gruteser, "Snooping keystrokes with mm-level audio ranging on a single phone," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 142–154.
[22] M. Chen, P. Yang, J. Xiong, M. Zhang, Y. Lee, C. Xiang, and C. Tian, "Your table can be an input panel: Acoustic-based device-free interaction recognition," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Mar. 2019, vol. 3, no. 1, pp. 1–21.
[23] G. Luo, M. Chen, P. Li, M. Zhang, and P. Yang, "SoundWrite II: Ambient acoustic sensing for noise tolerant device-free gesture recognition," in *Proc. IEEE 23rd Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Shenzhen, China, Dec. 2017, pp. 121–126.
[24] H. Du, P. Li, H. Zhou, W. Gong, G. Luo, and P. Yang, "WordRecorder: Accurate acoustic-based handwriting recognition using deep learning," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1448–1456.
[25] K. Yatani and K. N. Truong, "BodyScope: A wearable acoustic sensor for activity recognition," in *Proc. ACM Conf. Ubiquitous Comput.*, Pittsburgh, PA, USA, 2012, pp. 341–350.
[26] C. Zhang, Q. Xue, A. Waghmare, S. Jain, Y. Pu, S. Hersek, K. Lyons, K. A. Cunefare, O. T. Inan, and G. D. Abowd, "SoundTrak: Continuous 3D tracking of a finger using active acoustics," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Jun. 2017, vol. 1, no. 2, pp. 1–25.
[27] M. Chen, P. Yang, S. Cao, M. Zhang, and P. Li, "WritePad: Consecutive number writing on your hand with smart acoustic sensing," *IEEE Access*, vol. 6, pp. 77240–77249, 2018.
[28] Y. Huang, X. Yang, Y. Li, D. Zhou, K. He, and H. Liu, "Ultrasound-based sensing models for finger motion classification," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1395–1405, Sep. 2018.
[29] T.-H. Chiang, K.-Y. Ou, J.-W. Qiu, and Y.-C. Tseng, "Pedestrian tracking by acoustic Doppler effects," *IEEE Sensors J.*, vol. 19, no. 10, pp. 3893–3901, May 2019.
[30] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic Doppler sonar," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 1889–1892.
[31] A. Ghosh, D. Chakraborty, D. Prasad, M. Saha, and S. Saha, "Can we recognize multiple human group activities using ultrasonic sensors?" in *Proc. 10th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Bengaluru, India, 2018, pp. 557–560.
[32] T. Wang, D. Zhang, L. Wang, Y. Zheng, T. Gu, B. Dorizzi, and X. Zhou, "Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2959–2973, Apr. 2019.
[33] C. R. Pittman and J. J. LaViola, Jr., "Multiwave: Complex hand gesture recognition using the Doppler effect," in *Proc. 43rd Graph. Interface Conf.*, Edmonton, AB, Canada, 2017, pp. 97–106.

[34] Q. Liu, W. Yang, Y. Xu, Y. Hu, Q. He, and L. Huang, "DopGest: Dual-frequency based ultrasonic gesture recognition," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, China, Oct. 2018, pp. 293–300.

[35] W. Huang, Y. Xiong, X. Y. Li, H. Lin, X. Mao, P. Yang, and Y. Liu, "Shake and walk: Acoustic direction finding and fine-grained indoor localization using smartphones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Toronto, ON, Canada, Apr./May 2014, pp. 370–378.

[36] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2016, pp. 82–94.

[37] D. Graham, G. Simmons, D. T. Nguyen, and G. Zhou, "A software-based sonar ranging sensor for smart phones," *IEEE Internet Things J.*, vol. 2, no. 6, pp. 479–489, Dec. 2015.

[38] H. Zhang, W. Du, P. Zhou, M. Li, and P. Mohapatra, "DopEnc: Acoustic-based encounter profiling using smartphones," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2016, pp. 294–307.

[39] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: Using the Doppler effect to sense gestures," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Austin, TX, USA, 2012, pp. 1911–1914.

[40] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Heidelberg, Germany, 2016, pp. 474–485.

[41] X. Li, H. Dai, L. Cui, and Y. Wang, "SonicOperator: Ultrasonic gesture recognition with deep neural network on mobiles," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, USA, 2017, pp. 1–7.

[42] H. Watanabe and T. Terada, "Improving ultrasound-based gesture recognition using a partially shielded single microphone," in *Proc. ACM Int. Symp. Wearable Comput.*, Singapore, 2018, pp. 9–16.

[43] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, "Dolphin: Ultrasonic-based gesture recognition on smartphone platform," in *Proc. IEEE 17th Int. Conf. Comput. Sci. Eng.*, Chengdu, China, Dec. 2014, pp. 1461–1468.

[44] Z. Xu, K. Wu, and P. Hu, "Ultrasonic waves based gesture recognition method for smartphone platform," *Comput. Eng. Appl.*, vol. 54, no. 2, pp. 239–245, 2018. 10.3778/j.issn.1002-8331.1608-0081.

[45] K. Sun, T. Zhao, W. Wang, and L. Xie, "VSkin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, New Delhi, India, 2018, pp. 591–605.

[46] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "UltraGesture: Fine-grained gesture sensing and recognition," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Hong Kong, Jun. 2018, pp. 1–9.

[47] B. Fu, D. V. Gangatharan, A. Kuijper, F. Kirchbuchner, and A. Braun, "Exercise monitoring on consumer smart phones using ultrasonic sensing," in *Proc. 4th Int. Workshop Sensor-Based Activity Recognit. Interact.*, Rostock, Germany, 2017, pp. 1–6.

[48] J. Tan, C.-T. Nguyen, and X. Wang, "SilentTalk: Lip reading through ultrasonic sensing on mobile phones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.

[49] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "SilentKey: A new authentication framework through ultrasonic-based lip reading," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Mar. 2018, vol. 2, no. 1, pp. 1–18.

[50] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1466–1474.

[51] R. Nandakumar, S. Gollakota, and N. Watson, "Contactless sleep apnea detection on smartphones," in *Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Services*, Florence, Italy, 2015, pp. 45–57.

[52] X. Wang, R. Huang, and S. Mao, "SonarBeat: Sonar phase for breathing beat monitoring with smartphones," in *Proc. 26th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Vancouver, BC, Canada, 2017, pp. 1–8.

[53] W. Wang, L. Xie, and X. Wang, "Tremor detection using smartphone-based acoustic sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., ACM Int. Symp. Wearable Comput.*, Maui, HI, USA, 2017, pp. 309–312.

[54] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Santa Clara, CA, USA, 2016, pp. 1515–1525.

[55] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Niagara Falls, NY, USA, 2017, pp. 15–28.

[56] H. Chen, F. Li, and Y. Wang, "EchoTrack: Acoustic device-free hand tracking on smart phones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.

[57] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatTracker: High precision infrastructure-free mobile device tracking in indoor environments," in *Proc. 15th ACM Conf. Embedded Netw. Sensor Syst.*, Delft, The Netherlands, 2017, pp. 1–14.

[58] X. Xu, J. Yu, Y. Chen, Y. Zhu, and M. Li, "SteerTrack: Acoustic-based device-free steering tracking leveraging smartphones," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Hong Kong, Jun. 2018, pp. 1–9.

[59] W. Liu, W. Shen, B. Li, and L. Wang, "Toward device-free micro-gesture tracking via accurate acoustic Doppler-shift detection," *IEEE Access*, vol. 7, pp. 1084–1094, 2018.

[60] K.-Y. Chen, D. Ashbrook, M. Goel, S.-H. Lee, and S. Patel, "AirLink: Sharing files between multiple devices using in-air gestures," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Seattle, WA, USA, 2014, pp. 565–569.

[61] M. T. I. Aumi, S. Gupta, M. Goel, E. Larson, and S. Patel, "DopLink: Using the Doppler effect for multi-device interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Zürich, Switzerland, 2013, pp. 583–586.

[62] Z. Sun, A. Purohit, R. Bose, and P. Zhang, "Spartacus: Spatially-aware interaction for mobile devices through energy-efficient audio sensing," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Services*, Taipei, Taiwan, 2013, pp. 263–276.

[63] W. Huang, Y. Xiong, X. Y. Li, H. Lin, X. Mao, P. Yang, Y. Liu, and X. Wang, "Swadloon: Direction finding and indoor localization using acoustic signal by shaking smartphones," *IEEE Trans. Mobile Comput.*, vol. 14, no. 10, pp. 2145–2157, Oct. 2015.

[64] H. Chen, F. Li, and Y. Wang, "EchoLoc: Accurate device-free hand localization using COTS devices," in *Proc. 45th Int. Conf. Parallel Process. (ICPP)*, Philadelphia, PA, USA, 2016, pp. 334–339.

[65] Y.-C. Tung and K. G. Shin, "EchoTag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 525–536.

[66] F. Li, H. Chen, X. Song, Q. Zhang, Y. Li, and Y. Wang, "CondioSense: High-quality context-aware service for audio sensing system via active sonar," *Pers. Ubiquitous Comput.*, vol. 21, no. 1, pp. 17–29, 2017.

[67] I. Bisio, A. Delfino, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Ultrasounds-based context sensing method and applications over the Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3876–3890, Oct. 2018.

[68] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatMapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Niagara Falls, NY, USA, 2017, pp. 42–55.

[69] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang, "Smartphone-based acoustic indoor space mapping," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Jun. 2018, vol. 2, no. 2, pp. 1–26.

[70] W. Mao, M. Wang, and L. Qiu, "AIM: Acoustic imaging on a mobile," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, Munich, Germany, 2018, pp. 468–481.

[71] N. Kim and J. Lee, "Towards grip sensing for commodity smartphones through acoustic signature," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Int. Symp. Wearable Comput.*, Maui, HI, USA, 2017, pp. 105–108.

[72] Y.-C. Tung and K. G. Shin, "ForcePhone: Software lets smartphones sense touch force," *IEEE Pervasive Comput.*, vol. 15, no. 4, pp. 20–25, Oct./Dec. 2016.

[73] Y.-C. Tung, "Acoustic sensing: Mobile applications and frameworks," Ph.D. dissertation, Dept. Comput. Sci. Eng., Univ. Michigan, Ann Arbor, MI, USA, 2018.

[74] Y.-C. Tung and K. G. Shin, "Expansion of human-phone interface by sensing structure-borne sound propagation," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services*, Singapore, 2016, pp. 277–289.

[75] Y. Zhang, J. Wang, W. Wang, Z. Wang, and Y. Liu, "Vernier: Accurate and fast acoustic motion tracking using mobile devices," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1709–1717.
[76] B. Fu, F. Kirchbuchner, A. Kuijper, A. Braun, and D. V. Gangatharan, "Fitness activity recognition on smartphones using Doppler measurements," *Informatics*, vol. 5, no. 2, p. 24, May 2018.
[77] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu, "Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1574–1582.
[78] D. Ensminger, and L. J. Bond, *Ultrasonics: Fundamentals, Technologies, and Applications*, 3rd ed. New York, NY, USA: CRC Press, 2011, pp. 1–3.
[79] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[80] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proc. 2015 ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Osaka, Japan, 2015, pp. 283–294.
[81] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Vienna, Austria, 2016, pp. 1080–1091.
[82] C. Huang, H. Chen, L. Yang, and Q. Zhang, "BreathLive: Liveness detection for heart sound authentication with deep breathing," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Mar. 2018, vol. 2, no. 1, pp. 1–25.
[83] R. Nandakumar, A. Takakuwa, T. Kohno, and S. Gollakota, "CovertBand: Activity information leakage using music," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, Sep. 2017, vol. 1, no. 3, pp. 1–24.
[84] T. Zhu, Q. Ma, S. Zhang, and Y. Liu, "Context-free attacks using keyboard acoustic emanations," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Scottsdale, AZ, USA, 2014, pp. 453–464.
[85] P. Cheng, I. E. Bagci, U. Roedig, and J. Yan, "SonarSnoop: Active acoustic side-channel attacks," Aug. 2018, *arXiv:1808.10250*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180810250C
[86] I. Shumailov, L. Simon, J. Yan, and R. Anderson, "Hearing your touch: A new acoustic side channel on smartphones," Mar. 2019, *arXiv:1903.11137*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190311137S
[87] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *Proc. 15th USENIX Conf. Netw. Syst. Design Implement.*, Renton, WA, USA, 2018, pp. 547–560.

ZHENGJIE WANG was born in Liaoyang, China, in 1972. He received the B.S. degree in power engineering from the North China University of Water Resources and Electric Power, Henan, China, in 1995, the M.S. degree in computer software and theory from Northeast University, China, in 2003, and the Ph.D. degree in computer application technology from the China University of Mining and Technology (Beijing), Beijing, China, in 2013.

Since 2003, he has been a Lecturer with the College of Electronic and Information Engineering, Shandong University of Science and Technology. He is the author of two books and more than ten articles. His research interests include human behavior recognition, people activity inference, person counting, identity authentication, and people tracking using Wi-Fi devices and smartphones.

YUSHAN HOU was born in Jinan, Shandong, China, in 1995. She received the B.S. degree from the Shandong University of Science and Technology, in 2017, where she is currently pursuing the M.S. degree in communication and information system. Her research interests include deep learning, machine learning, and signal processing.

KANGKANG JIANG was born in Jining, China, in 1993. She received the B.S. degree from Jining University, in 2017. She is currently pursuing the M.S. degree in communication and information system with the Shandong University of Science and Technology. Her research interests include image processing, machine learning, and deep learning.

CHENGMING ZHANG was born in Zaozhuang, China, in 1995. He received the B.S. degree from the Qilu University of Technology, in 2018. He is currently pursuing the M.S. degree in electronic and communication engineering with the Shandong University of Science and Technology. His research interests include machine learning, deep learning, and signal processing.

WENWEN DOU was born in Tai'an, Shandong, China, in 1996. She received the B.S. degree from Weifang University, in 2018. She is currently pursuing the M.S. degree in communication and information system with the Shandong University of Science and Technology. Her research interests include deep learning, machine learning, and signal processing.

ZEHUA HUANG was born in Binzhou, China, in 1996. He received the B.S. degree from the Shandong University of Science and Technology, in 2018, where he is currently pursuing the M.S. degree in communication and information systems. His research interests include machine learning, deep learning, and signal processing.

YINJING GUO was born in Jining, China, in 1966. He received the B.S. degree in radar engineering from Ordnance Engineering College, in 1989, and the M.S. degree in communication and electronic systems and the Ph.D. degree in weapon systems and application engineering from the Beijing Institute of Technology, in 1992 and 2004, respectively.

Since 1996, he has been a Professor with the College of Electronic and Information Engineering, Shandong University of Science and Technology. He has published over 90 papers, among which more than 40 articles have been retrieved by EI and SCI. His research interests include wireless communications, electromagnetic compatibility theory and applications, special radars, and unmanned aerial vehicles.

Dr. Guo is a member of the Qingdao Senior Experts Association. He has received the National Science and Technology Award. He has served as the Local Chair for the first, second, and third International Conference on Intelligent Information Technology Applications and the third IEEE International Conference on Communication and Mobile Computing. He is a Reviewer of the National Natural Science Foundation of China. He is also a Reviewer of many international journals.