HearFit: Fitness Monitoring on Smart Speakers via Active Acoustic Sensing

Yadong Xie* Fan Li* Yue Wu* Yu Wang†
* School of Computer Science, Beijing Institute of Technology, Beijing, China.
† Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA.
Email: {ydxie, fli, ywu}@bit.edu.cn, wangyu@temple.edu

***Abstract*—Fitness can help to strengthen muscles, increase resistance to diseases and improve body shape. Nowadays, more and more people tend to exercise at home/office, since they lack time to go to the dedicated gym. However, it is difficult for most of them to get good fitness effect due to the lack of professional guidance. Motivated by this, we propose HearFit, the first non-invasive fitness monitoring system based on commercial smart speakers for home/office environments. To achieve this, we turn smart speakers into active sonars. We design a fitness detection method based on Doppler shift and adopt the short time energy to segment fitness actions. We design a high-accuracy LSTM network to determine the type of fitness. Combined with incremental learning, users can easily add new actions. Finally, we evaluate the local (i.e., intensity and duration) and global (i.e., continuity and smoothness) fitness quality of users to help to improve fitness effect and prevent injury. Through extensive experiments including over 7,000 actions of 10 types of fitness with and without dumbbells from 12 participants, HearFit can detect fitness actions with an average accuracy of 96.13%, and give accurate statistics in various environments.**

I. INTRODUCTION

Nowadays, more and more people pay attention to their health status and body shape, causing a rapidly growing popularity of fitness. Effective fitness can bring numerous benefits, such as increasing muscle strength, improving body shape and decreasing risks of cardiovascular diseases [1], [2]. However, due to the fast pace of life, it is inconvenient for many people to go to dedicated gyms. Besides, it usually cost much to become a gym's membership and consult personal trainers. Therefore, plenty of people begin to do fitness at home/office, which is more conducive to save time and money. But due to lack of effective supervision and professional effect evaluation, they usually get unsatisfactory fitness effect.

The usual way to get fitness instruction is to hire a personal trainer, which is not suitable at home/office. There are several fitness APPs on the market [3]–[5] that can provide fitness guidance. However, these APPs can neither monitor fitness processes in real time, nor provide fine-grained fitness statistics. In order to facilitate people to exercise at home/office, researchers propose many solutions. Some fitness monitors are realized by analyzing videos [6]–[8], but they depend on good lighting conditions and may involve privacy issues. There is a new trend of using wearable devices for fitness monitoring. RecoFit [9] uses inertial sensors worn on forearms to track strength training. But these devices need to be worn all the time during fitness, which add extra burden to users. Thus, some researchers use RF signal for non-invasive fitness monitoring. TTBA [10] monitors free-weight exercise by attaching RFID tags on dumbbells and using RSSI to recognize the motion types, but it can only monitor exercise that require dumbbells. FitAssist [11] can recognize exercise based on Wi-Fi devices. However, the Wi-Fi based approaches usually need a pair of antennas, and users need to exercise between them. Besides, Wi-Fi signal is easy to be interfered by environmental factors like the movements of other people. Therefore, it is an urgent need to develop a non-invasive, easy-to-deploy and anti-interference fitness monitoring system to provide fine-grained fitness statistics and help to improve fitness effect.

Toward this end, we take one step forward to study the feasibility of using acoustic sensing, which achieves great success in human activity recognition [12]–[16], for fitness monitoring. In addition, smart speakers like Amazon Echo and Google Home are becoming popular. As of 2019, there are 157 million smart speakers in U.S. homes and an average 2.6 smart speakers on each household with smart speaker [17], which bring great potential in fitness monitoring without attaching any devices to users. In this paper, we design and implement HearFit, the first fitness monitoring system based on commercial smart speakers for home/office environments, HearFit aims to provide users with fine-grained fitness statistics to improve fitness effect in a more comprehensive method.

The key idea of HearFit is that the smart speaker emits inaudible ultrasonic signal, the signal is reflected by the user and then received by the microphone array, which is then processed to extract the user's fitness statistics. However, there are multiple challenges to realize HearFit. Firstly, people perform non-fitness activities most of the day, so HearFit needs to distinguish fitness from daily activities automatically. Secondly, at home/office, the reflected signal of the exercising user is easily interfered by the activities of surrounding people (e.g., walking or typing), which requires HearFit to be anti-interference. Finally, HearFit needs to classify and evaluate the fitness actions accurately, and allows users to add new actions.

To address these challenges, we first analyze the fitness

Fan Li is the corresponding author. The work of Fan Li is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No.62072040, 61772077 and Beijing Natural Science Foundation under Grant No. 4192051.

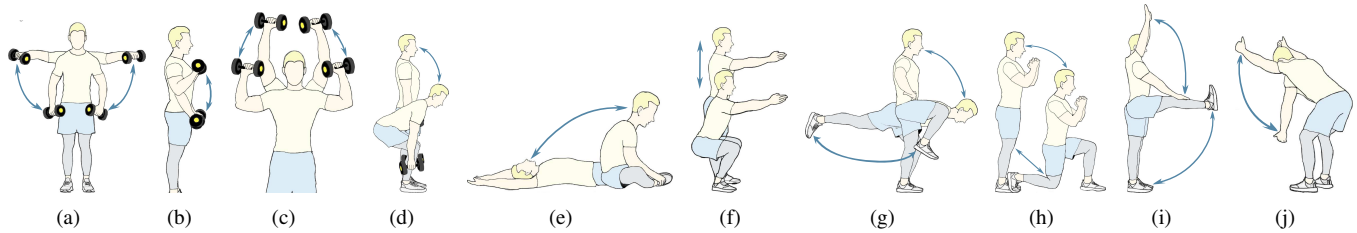


Fig. 1: Illustration of 10 types of fitness actions: (a) Lateral raise; (b) Dumbbell curl; (c) Shoulder press; (d) Dumbbell deadlift; (e) Sit-up; (f) Squat; (g) One leg deadlift; (h) Lunge; (i) Front kick; (j) Y stretch.

activities and find that they are typically more repetitive than non-fitness activities. Thus, we use the autocorrelation of acoustic signal to determine the activity mode of the user. In order to minimize the interference of surrounding people, we use the microphone array to locate the user. But the close distance between each microphone (i.e., about $5cm$) makes it impracticable to calculate an accurate distance between the user and the smart speaker through triangulation. So we calculate the accurate direction of the user by Generalized Cross-Correlation with phase transform (GCC-PHAT) [18] and use beamforming to amplify the reflected signal of the user. Finally, we design a method to segment each fitness action based on Short Time Energy (STE) [19], and design a Long Short-Term Memory (LSTM) network to identify the type of each action. Besides the actions that trained in advance, users can add new actions. Combined with incremental learning [20], HearFit can update the network with a small training set of new actions. To evaluate the fitness effect, we study the Frequency, Intensity, Time and Type (FITT) principle [21] commonly used in fitness evaluation, and define 4 metrics to evaluate the fitness process from local view and global view.

The prototype of HearFit is built using a Raspberry Pi 4B, a circular 6-microphone array and a omni-directional speaker, which is the same as most of smart speakers in the market. To evaluate the performance of HearFit, we recruit 12 volunteers (9 males and 3 females) and record their fitness process in 4 different rooms. The experiments involve 10 typical types of fitness, as shown in Fig. 1. Finally, we collect over $7,000$ actions of all types of fitness. Results demonstrate that HearFit can accurately identify fitness type in various indoor environments, and provide fine-grained fitness statistics.

Our contributions are summarized as follows:

- We propose a fitness monitoring system, HearFit, using smart speakers. It can provide users with fine-grained fitness statistics and help users to improve fitness effect in a comprehensive method. To the best of our knowledge, we are the first to use smart speakers for fitness monitoring.
- We turn smart speakers into active sonars and present several methods to extract features from reflected signal. We design a novel method to segment actions and design an LSTM network to identify the type of actions. We define 4 metrics from local and global views to help to improve fitness effect.
- Using the features of the microphone array, HearFit has good anti-interference ability. Combined with incremental learning, users can easily add new actions to the system.
- We evaluate HearFit using a hardware prototype and conduct extensive experiments with 12 volunteers in different environments. The results show that HearFit can identify fitness actions with an average accuracy of 96.13%, and provide accurate fitness effect evaluation.

II. RELATED WORK

Recent works show that fitness monitoring can be broadly categorized as:

Camera based methods. Many researches leverage cameras for fitness monitoring and activity recognition. Fitness Mate [6] uses Microsoft Kinect to monitor and guide physical exercise. MotionMA [22] extracts models of movement demonstrated by users and provides real-time feedback on how to improve their performance by using Kinect. A Hidden Semi-Markov Model based approach [23] monitors and evaluates body motions during a rehabilitation training by extracting features from skeleton joint trajectories acquired by the RGB-D camera. These methods, however, highly depend on good lighting conditions. And people usually refuse to deploy cameras at home/office due to privacy consideration [24].

Wearable device based methods. The development of wearable devices provides a new way for fitness monitoring. iCoach [25] is a smart fitness glove with inertial measurement units, which can recognize exercise and assess workout quality. FitCoach [26] leverages smartphones or smartwatches mounted on upper arms to achieve fine-grained fitness interpretation and smart exercise guidance. GymSoles [27] is an insole prototype that provides feedback on the centre of pressure at the feet to assist users with maintaining the correct posture while performing squats and dead-lifts. However, these methods require users to wear additional devices and can only recognize fitness of the body part where the device is attached.

Wireless device based methods. There are also some works using wireless devices to recognize fitness. FEMO [29] monitors fitness by attaching passive RFID tags on the dumbbells and adopts Doppler shift for action recognition and assessment. But, it can only monitor exercise that require dumbbells and users need to prepare dedicated RFID readers. Motion-Fi [30] can count repetitive motions and enable multi-users to perform motions leveraging wireless backscattering. SEARE [31] is a fitness monitor which provides users with health management during exercise using CSI-waveform shape

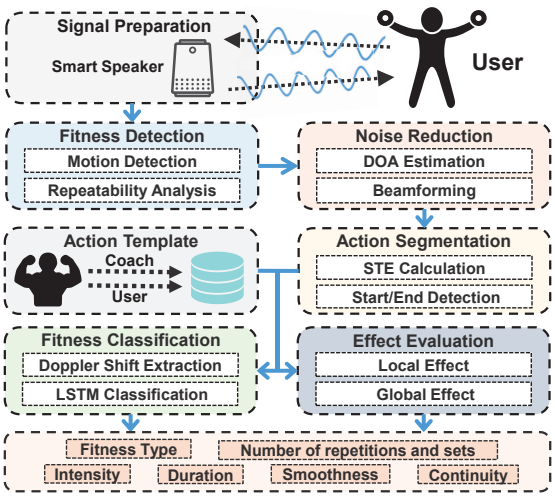


Fig. 2: System architecture of HearFit.

as features. But the Wi-Fi based approaches usually need a pair of antennas, and users need to exercise between the antennas. Besides, Wi-Fi signal is easy to be interfered by environmental factors like movements of other people.

We are the first to achieve non-invasive fitness monitoring using acoustic sensing [12]–[16] based on smart speakers. Unlike the above works, HearFit automatically monitors fitness and evaluates fitness effect of users without wearing extra devices. Through taking advantages of the omni-directional speaker and microphone array, HearFit can sense the fitness in any direction and has good anti-interference ability.

III. SYSTEM DESIGN

This section details the core techniques behind the segmentation, classification and evaluation of fitness.

A. System Overview

Fig. 2 shows the architecture of HearFit. It mainly includes 6 parts, which are *Signal Preparation*, *Fitness Detection*, *Noise Reduction*, *Action Segmentation*, *Fitness Classification* and *Effect Evaluation*. In order to classify fitness actions and evaluate effect, we recruit professional fitness coaches to collect standard fitness data as *Action Template*. After observing the coaches, we divide a fitness process into repetitions and sets. A repetition is one complete cycle of a fitness action, while a set usually contains certain number of repetitions.

In *Signal Preparation*, we design an acoustic signal considering various factors. The speaker emits the signal, which is then reflected by surrounding objects and received by the microphone array. Note that here we get 6 signals from 6 microphones. Then, HearFit performs *Fitness Detection* to detect whether there are fitness activities around the smart speaker. Specifically, we adopt a band-pass filter on the reflected signal and search for Doppler shifts that may be caused by motions. If the Doppler shifts are repetitive, the corresponding motions are considered to be fitness actions. Once the fitness is detected, HearFit performs *Noise Reduction*. The Doppler shifts from all the 6 microphones are processed to get the Direction of Arrival (DoA) of signals. To increase the signal strength of the fitness and suppress noise, we use the beamforming to synthesize 6 signals. In *Action Segmentation*, we calculate the STE to capture the energy pattern of the signal, and analyze the slope of the STE to detect the start and end time of each repetition. By analyzing the time interval between each repetition, we can calculate the numbers of repetitions and sets.

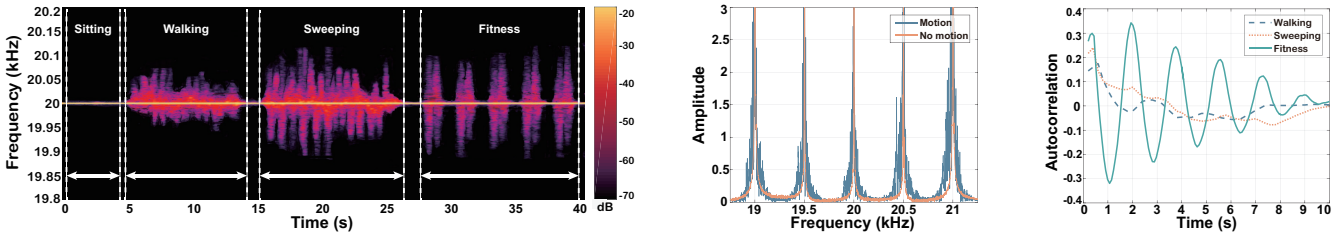
HearFit determines the type of each repetition in *Fitness Classification*. In detail, we perform Fast Fourier Transform (FFT) on the signal to extract effective features of Doppler shifts. The features are sent to an LSTM network, which is trained by standard fitness data, to determine the fitness type. After fitness, users usually care about fitness effect to ensure that muscles are strengthened and injury risks are low. So *Effect Evaluation* aims to provide feedback to users. The feedback includes two aspect: local effect and global effect. Local effect denotes the quality of each repetition by evaluating intensity and duration. The global effect denotes the continuity and smoothness of each set. Finally, HearFit provides users with the statistics to help them improve fitness effect. Through incremental learning, users can add new actions to *Action Template* with a cost of a small amount of training.

B. Signal Preparation

To monitor a user's fitness process, we turn the smart speaker to an active sonar and observe Doppler shifts to detect the user's motion. Doppler shift is the change in frequency of a signal in relation to an observer who is moving relative to the signal source. Formally, the Doppler shift Δf is determined by $\Delta f = (2v/c) \cdot f$, where f is the emitted frequency, c is the speed of sound and v is the speed of relative movement. In HearFit, the smart speaker is not only the signal source but also the observer. The user who reflects the signal can be seen as a virtual signal source. To accurately detect the user's motion, the emitting signal needs to be carefully designed.

At home/office, the acoustic signal can be reflected by multiple objects, such as roof, wall and furniture. Under this condition, wide-band signal can work well [32]. Thus, we design a wide-band emitting signal through considering several factors. First, the signal should be inaudible to people. According to [33], the frequency above $18kHz$ is already inaudible to kids who are sensitive to sounds. Second, higher frequency of a sound results in a more obvious Doppler shift. However, due to the upper frequency limit of most speakers (i.e., $22kHz$) [34] and the sampling rate supported by most of the microphones (i.e., $44.1kHz$), the highest frequency has to be lower than $22kHz$. Third, frequency selective fading may occur with single frequency due to the multipath effect of the signal, which greatly decreases the system performance. It can be reduced by emitting signal with multiple frequencies. Therefore, comprehensively considering the factors, we design a signal with 5 frequency components (i.e., $19kHz$, $19.5kHz$, $20kHz$, $20.5kHz$ and $21kHz$). The signal is defined as:

$$s(t) = \sum_{n=1}^{5} A_n \sin\left(2\pi f_n t + \phi_n\right), \tag{1}$$



(a) Doppler profile of several activities.

(b) Doppler shifts of motion and no motion.

(c) Autocorrelation of 3 types activities.

Fig. 3: An example of the process of fitness detection.

where f_n denotes the n-th frequency component. A_n and ϕ_n are the amplitude and initial phase, respectively. We select the 5 frequency components for two reasons. First, when users exercise, the speed of their movements usually do not exceed $1.5m/s$, which can cause a Doppler shift of $185.2Hz$ under a sound of $21kHz$. Since $185.2Hz$ is smaller than the half of intervals between frequency components, their Doppler shifts do not overlap in the frequency domain. Second, we find that more than 5 frequency components would make the signal audible even at a low volume due to sub-harmonics.

At the receiving end, the 6-microphone array collects reflected signals at a sampling rate of $44.1kHz$. Through the analysis of the collected data, we find that each type of fitness action has a unique pattern on its Doppler shift, which suggests that Doppler shift can be used in fitness monitoring.

C. Fitness Detection

HearFit first detects motions around the smart speaker. After that, it needs to distinguish between fitness actions and non-fitness activities (e.g., sitting, walking or sweeping) contained in those motions. We first conduct a preliminary experiment on recordings of several human activities. For example, a user first sits, then walks to a location to sweep the floor, and finally performs 5 fitness actions. Fig. 3(a) shows the corresponding Doppler profile of $20kHz$ signal collected by one microphone. We can observe that fitness actions are more repetitive than non-fitness activities. This is because in order to improve fitness effect, users usually divide a fitness process into multiple sets, each set contains multiple identical fitness actions with similar intervals between each two actions. Based on this observation, we design a two-step detection method to determine the start time of the fitness.

Motion detection. First, we check whether there are motions around the smart speaker through analyzing the amplitudes of Doppler shifts. By observing Fig. 3(a), we can find that when the user is sitting, there is the reflected signal by static objects without Doppler shifts. When the user has motions, the reflected signal contains Doppler shifts. So, to detect Doppler shifts, HearFit emits one-second signal every $10s$, then FFT is performed on the signal received by each microphone. Fig. 3(b) shows the FFT results under conditions with motions and without motions. We can see that when the user has motions, there are several peaks with larger amplitude around the center of each frequency component. On the contrary, when there is no motion, the amplitude of each frequency component sharply raises and drops to a negligible value. Based on such observation, we extract two closest peaks on each side of each frequency component. Finally, we obtain 120 peaks (i.e., 2 peaks $\times$ 2 sides $\times$ 5 frequency components $\times$ 6 microphones) for each one-second signal and calculate the average value of these peaks. If the value exceeds 1, we consider that there exists motions.

Repeatability analysis. HearFit starts to emit consecutive signal once it detects motions. We add a window with length of $13s$ that slides $1s$ each time on the reflected signal, and adopt a band-pass filter on each window to remove out-band interferences. Spectral subtraction is also used to remove frequency components, and only the parts of Doppler shifts are remained. HearFit further determines whether the motions are fitness activities by calculating autocorrelation [35] of the filtered signal. Autocorrelation is the correlation of a signal and its delayed copy. It can indicate whether a signal is periodic. We denote a signal as X_t and its copy that delay for a period of k as X_{t+k}, then its autocorrelation is:

$$R(k) = \frac{E\left[(X_t - \mu)(X_{t+k} - \mu)\right]}{\sigma^2}, \tag{2}$$

where μ and σ are the expectation and standard deviation of X_t, respectively. Fig. 3(c) shows the autocorrelation of the 3 types of activities. We can see that fitness actions have obvious repetitive patterns, while non-fitness activities hardly have long-term repetitive pattern. Then, we count the peaks that have autocorrelation values larger than a threshold (empirically set as 0.1). If each of two consecutive windows both has more than 4 peaks, we consider there exists fitness actions.

HearFit can only adopt repeatability analysis without motion detection to detect fitness. It needs HearFit to emit consecutive signal and analyze data all the time. However, by observing usage scenarios, we find that there are usually no human activities around the smart speaker in most of the time. Thus, we design motion detection that HearFit just emits and collects 1-second signal every $10s$ to make it more energy-efficient.

D. Noise Reduction

At home/office, the reflected signal from the fitness user is easily interfered by the activities of surrounding people (e.g., walking or typing on the computer). In order to minimize these interferences after HearFit detects fitness, we take advantages

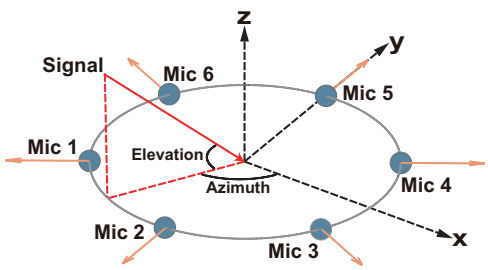


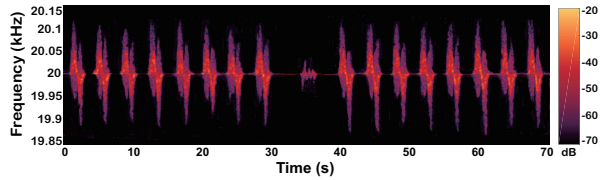
Fig. 4: Coordinate system of the microphone array.

of microphone array to directionally enhance the reflected signal of the fitness user. For convenience, we establish a suitable coordinate system, as shown in Fig. 4. For an array with M microphones, the azimuth angle of the m-th microphones is given by $\varphi_m = (-(M-1)/2+m-1)\cdot 360/M$. The azimuth angle is defined as the angle from the x-axis toward the y-axis in the xy-plane. The elevation angle is defined as the angle from the xy-plane toward the z-axis.

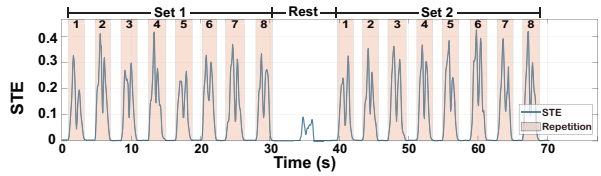
DOA estimation. First, we need to get the user's direction relative to the smart speaker. Since the reflected signal arrives at each microphone at different time, we use Time Difference of Arrival (TDOA) based method to estimate the user's direction. The GCC-PHAT algorithm [18], [36] is a commonly used method based on time delay estimation. It assumes that signal source is located in the array far field, so the DOA are the same for all microphones. We use the signals of the 6 microphones processed in repeatability analysis and estimate the correlations between each signal pairs by GCC-PHAT and find the largest peak in each correlation. The peak identifies the delay between the two signals that arrive at microphones in a pair. Finally, a least-squares estimation is used to derive the azimuth and the elevation angles of the user. Since most of users do not change their locations during fitness, the DOA estimation only needs to run once at the beginning of fitness.

Beamforming. Now we have an initial estimation of the azimuth and the elevation angles of the user. Then we use a time-delay beamforming algorithm to suppress interference from other people and increase the SNR of the reflections from the user. The time-delay beamforming compensates a reflected signal coming from a specific direction for the arrival time differences across the microphones. It includes two steps: time alignment and synthesis. Time alignment is achieved by transforming signals into the frequency domain and applying linear phase shifts corresponding to a time delay. The individual signals are then synthesized and converted back to the time domain. Finally, we get one copy of synthesized Doppler shifts from the signals of 6 microphones.

Most of smart speakers have circular LED lights used for various functions, such as indicating the direction of the voice command. Such LED lights can also be used for indicating the direction of fitness. In order to improve the usability of HearFit, we also allow users to use voice commands to start fitness monitoring. Then, we determine the direction of the user by detecting the direction of the user's voice commands.



(a) Doppler profile of 16 lunges.



(b) Segmentation result of 16 lunges.

Fig. 5: The process of action segmentation.

E. Action Segmentation

In order to accurately recognize fitness types and provide fitness effect evaluation, we segment each repetition. In other words, we need to detect the start time and end time of each repetition. We find that when the user moves, the energy of the Doppler shifts also changes, so we calculate the STE [19] of the signal and segment each repetition based on the STE.

STE calculation. STE is widely used in speech recognition to separate voiced and unvoiced speech segments. We first normalize the signal using the Min-Max Normalization. Then, we apply a sliding window with length of $0.5s$ that slides $0.1s$ each time on the signal. We define the STE of time t as:

$$E_t = \sum_{\delta=t-l+1}^{t} \left[s(\delta)w(t-\delta)\right]^2, \tag{3}$$

where $s(\delta)$ is the amplitude of the signal at time δ, w is the hanning window and l is the length of the window. Fig. 5(a) shows the Doppler profile of two sets of lunges and Fig. 5(b) shows the corresponding STE. We further find a unique pattern of the STE: 1) at the beginning of a repetition, STE increases rapidly; 2) STE drops slightly when the repetition stops in the middle; 3) STE increases again when the user returns to the initial pose; 4) STE decreases sharply when the user finishes a repetition. Based on this pattern, we design a method to detect the start and end time of each repetition.

Start/End detection. Detecting the start and end time of each repetition not only helps HearFit to classify fitness actions, but also provides basis for the effect evaluation. The common segmentation methods are peak detection [37] and KL divergence algorithm [29], but they are not applicable in our system. Peak detection usually segments repetitions by detecting peaks or troughs in sensing data. However, it is only applicable when there is no rest interval between consecutive repetitions. KL divergence algorithm uses the discrete probability distribution of each window to identify repetition and rest interval. But the window size has great influence on system performance, as a small window may cause a repetition to be divided into several parts, while a large window may bring a large amount of computation.

Different from them, we segment each repetition in a simple but effective way based on the slope of STE. We set thresholds sth and eth for the slope of the start and end position, respectively. Since each repetition contains start and end time, we alternately search for the start and end time in the STE. We choose the time t with STE E_t less than 0.03, and calculate its slope by $\Theta_t = (E_{t+1} - E_t)/\Delta t$. If Θ_t is greater than sth, and $E_{t+2} > E_{t+1}$, we consider time t as a start time. Then, we search for the corresponding end time after time t. If $\Theta_{t+\psi}$ is less than eth, and $E_{t+\psi-1} > E_{t+\psi}$, we consider time $t+\psi$ as corresponding end time. We then continue to search for the start and end time of the next repetition until the user finishes fitness. Fig. 5(b) shows the result of the segmentation, we can see that all start and end time are correctly identified. Note that the result of the segmentation may include tiny movements during the rest between two sets, we filter out the tiny movements by analyzing the maximum energy of each repetition. Finally, if the time interval between two repetitions is more than $7s$, we consider that they belong to different sets. If there is no repetitive pattern within $90s$ after a repetition, we consider the fitness is finished.

F. Fitness Classification

After segmentation, HearFit aims to identify the fitness type of each repetition. Considering that repetitions in a set are usually the same, HearFit only classifies the first 3 repetitions in each set to reduce the amount of calculation. In the early stage, we recruit professional fitness coaches to collect standard fitness data as action template. These data are used to train an LSTM network as classifier. Before training the network, we first extract effective features from these data.

Doppler shift extraction. To accurately classify repetitions, it is necessary to extract reliable features from each repetition's reflected signal. The frequency-domain information stored in the signal is widely used in activity recognition [38], [39]. Our previous experiments also verify that each type of fitness action has unique Doppler profile in the frequency domain. In addition, each repetition is a consecutive action lasts for a period, if we directly perform FFT on the signal, the time domain information is lost. So we design a time-frequency feature extraction method.

For the signal of each repetition, we further divide it into 8 blocks on average. To obtain frequency feature (i.e., Doppler shifts), each block is processed by a 4096-point FFT, which means dividing the sampling rate into 4096 points evenly. Since the frequency we use is mainly between $19kHz$ and $21kHz$, the information in other frequency can be ignored. In other words, we just focus on the 1765th to 1950th points. Through experiments, we find that the amplitude of FFT results is better for training network than the phase. Thus, the amplitude of each block is used to form a 186-dimension feature vector, and each repetition gets 8 feature vectors.

LSTM classification. After getting feature vectors, we use a deep learning method to classify repetitions. Traditional classifiers usually treat each block as independent and ignore the temporal context. So, we design an LSTM network to exploit

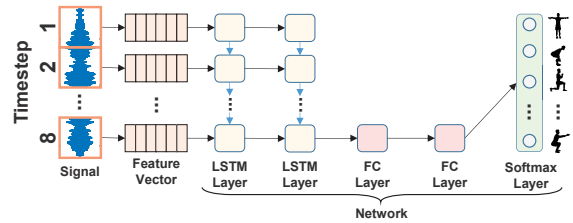


Fig. 6: Structure of the LSTM network.

the temporal dependencies within blocks [40]. The architecture of LSTM is recurrent, which means that it considers not only the current block but also previous blocks. It has 8 timesteps and each timestep takes a block as input. Finally, it generates a classified result after the last timestep.

Fig. 6 shows the structure of the designed network, which has 2 LSTM layers, 2 fully connected (FC) layers and 1 softmax layer. LSTM layers are the most essential layer, each of which can transform the input to compressed representations that can characterize fitness actions through an unsupervised manner. At t-th timestep, the LSTM layer can map the input p_t into a compressed vector h_t as follow:

$$h_t = \sigma\left(W\left[h_{t-1}, p_t\right] + b\right) \cdot tanh\left(C_t\right), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, W and b denote the weight and bias, respectively. C_t denotes the status at t-th timestep. Then, we add two FC layers after the last timestep. Finally, the softmax layer calculates a class probability vector, and the repetition is then assigned to the class with the highest probability. In addition, we add a Batch Normalization between any two layers, which can prevent overfitting.

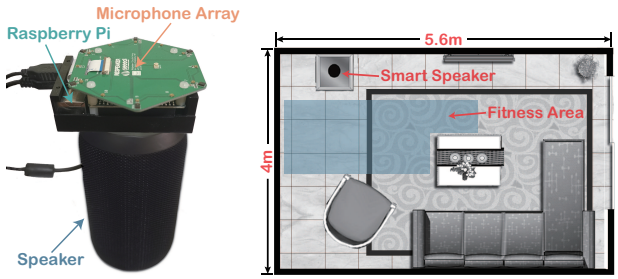
The network is trained by using standard fitness data in action template and then it can distinguish actions in the template. To enhance the usability, HearFit allows users to add new actions besides template. We borrow ideas from incremental learning [20] to make the network recognize new actions. Specifically, the structures and parameters of the first 3 layers are kept unchanged, and the last 2 layers are retrained by the new actions and all existing actions. This method can reduce the size of data set during retraining. Users only need to complete several repetitions of the new actions.

G. Effect Evaluation

The effect evaluation aims to provide fine-grained fitness statistics to users. By studying the FITT principle [21] commonly used in fitness evaluation, HearFit measures fitness qualities from two aspects: local effect and global effect.

Local effect. Local effect focuses on evaluating each repetition's intensity and duration.

The intensity reflects the energy expended in each repetition. Each repetition usually consists of two parts: extension and retraction. The balance of their intensities can enhance bidirectional muscle contraction strength and avoid disequilibrium. In HearFit, we use STE to represent intensity. By observing Fig. 5(b), we can find that the STE of each repetition includes two obvious peaks, representing the intensities of extension



(a) Prototype of HearFit. (b) Indoor environment: living room.

Fig. 7: An example of experimental device and environment.

and retraction, respectively. We define the balance of intensity as $BI = \alpha - I_e/I_r$, where I_e and I_r are the intensities of extension and retraction, α represents the standard balance obtained from the action template. Finally, we obtain the balance of intensity in each repetition.

The duration reflects the time spent in each repetition. A shorter duration may make the muscles stretch and contract too fierce, which can significantly increase the risk of injury. A longer duration may have a negative effect on muscle flexibility and reaction speed. In addition, the closer to the standard data the duration is, the more standard the repetition is. The duration of each repetition can be directly obtained after the segmentation (Section III-E). Suppose D is the duration of the repetition, D_s is the corresponding standard duration in the action template, we can calculate their difference $d = D_s - D$ as the standard of the duration.

Global effect. Global effect focuses on the overall performance of each set. We design two metrics, continuity and smoothness.

Continuity reflects the consistency of rest intervals between two repetitions within a set. Effective fitness should maintain a steady rhythm, and unstable rest intervals usually indicate that the intensity of fitness is too high or too low. In order to evaluate the continuity of each set, we adopt the kurtosis as the metric. In probability theory, kurtosis can describe the sharpness of the probability distribution in a real-valued variable. Formally, suppose $R = [r_1, r_2, \cdots, r_n]$ is the rest intervals of a set, the kurtosis can be calculated as:

$$Kurt = \frac{\sum_{i=1}^{n} (r_i - \mu)^4}{\left(\sum_{i=1}^{n} (r_i - \mu)^2\right)^2} - 3 = \frac{\mu^4}{\theta^4} - 3, \qquad (5)$$

where μ and θ are the mean and standard deviation of the R. A larger $Kurt$ indicates a better continuity of a set.

Smoothness reflects the consistency of intensities within a set. The similar intensities of repetitions in a set means that the user can control the muscles well, so as to improve the fitness effect. In order to evaluate the smoothness of each set, the average of I_e and I_r is used as the intensity of each repetition. Similar to continuity, we calculate the intensity kurtosis of a set. And a larger intensity kurtosis indicates a better smoothness of a set.

When a user completes a fitness process, he/she can connect the smartphone to the smart speaker. Then, his/her fitness statistics in this process are automatically synchronized to the smartphone. According to the statistics provided by HearFit, users can improve fitness effect next time.

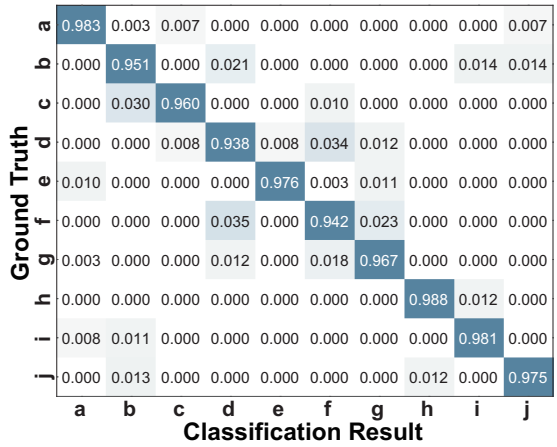


Fig. 8: Overall performance of HearFit.

IV. IMPLEMENTATION AND EVALUATION

In this section, we introduce the implementation details and provide the evaluation results.

A. Experiment Setup

We implement HearFit using a Raspberry Pi 4B, a circular 6-microphone array and a omni-directional speaker as shown in Fig. 7(a), which is the same as most of commercial smart speakers. To evaluate the performance of HearFit, we recruit 12 volunteers (9 males and 3 females, aged from 17 to 46), including fitness coaches, people who regularly exercise and people who rarely exercise. Their fitness process are recorded in 4 different indoor environments, including a bedroom, a living room, a study room and an office, the size of which are $5.7m \times 3m$, $5.6m \times 4m$, $2.8m \times 3.6m$ and $4.5m \times 4.1m$, respectively. We choose the sizes of the rooms that are most common in residential areas. Fig. 7(b) shows the floor plan of the living room, where the location of the smart speaker is convenient for users to operate. The experiments involve 10 typical types of fitness. We collect over $7,000$ repetitions of all types of fitness for training and evaluation.

B. Evaluation Methodology

We mainly evaluate HearFit from the following aspects.

Precision. The percentage of repetitions which are correctly classified into type A in all repetitions classified into A.

Recall. The percentage of repetitions which are classified into type A in the repetitions truly belong to A.

F1-score. It is the harmonic mean of precision and recall. In our system, the F1-score for specific types of fitness is defined as $F1 = 2 \cdot \frac{Precision * Recall}{Precision + Recall}$.

Confusion matrix. Each row and each column of the matrix represent the ground truth and the classification result, respectively. Each entry represents the percentage of a certain type of fitness that is classified into each class.

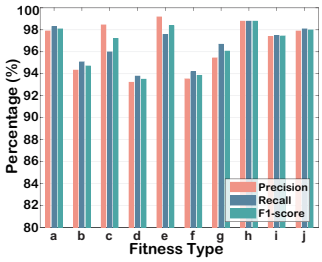


Fig. 9: Precision, recall and F1-score.

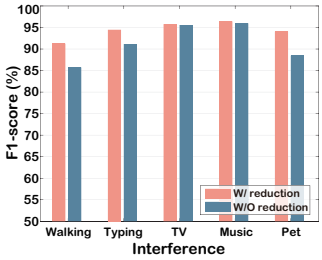


Fig. 10: F1-score under different interferences.

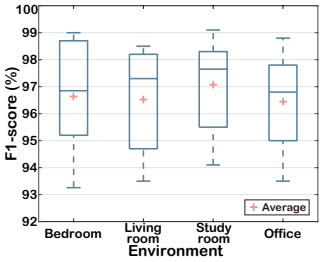


Fig. 11: F1-score under different environments.

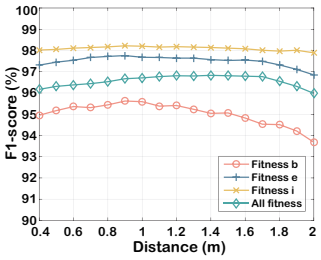


Fig. 12: F1-score under different distances.

C. Overall Performance

We first evaluate the overall performance of fitness classification based on the LSTM network. Fig. 8 shows the confusion matrix of 10 types of fitness actions performed by all volunteers in 4 different environments. HearFit achieves an average accuracy of 96.13% for classifying all types of actions. Particularly, we find the dumbbell deadlift and squat have slightly lower accuracy than other types of actions. This is because the two actions mainly contain vertical movements of the upper body, which makes some repetitions be classified to another type. But the lowest accuracy is also over 93.50%.

Fig. 9 shows the corresponding precision, recall and F1-score. The precision are no less than 93.24%, while the recall are no less than 93.80%. The average F1-score is 96.61%, which is slightly better than FitCoach (around 95%) [26]. But FitCoach uses wearable devices which require users to wear and can only recognize fitness of the body part where the device is located. HearFit also performs better than wireless signal based methods in [28] and [29], the F1-score of which are 93% and 91%, respectively.

D. Impact of Different Factors

1) Impact of interference: The reflected signal of the user would be affected by surrounding activities. We study the impacts on HearFit when there is someone walking, typing, watching TV, listening to music and a pet wandering around the user, in terms with and without noise reduction (Section III-D). From Fig. 10, we can see that F1-score is relatively low when there is a walking person, and the closer to the user the walking person is, the lower the F1-score is. But according to proxemics [41], the social distance between people is larger than $2m$. Thus, people barely get too close to the user. Moving pets and people who are typing on a keyboard also have slight impacts. Since these movements are tiny or closer to the floor, we can use noise reduction to filter them out. Watching TV and listening to music have almost no impacts on HearFit. Since the sounds of TV and music mainly have low-frequency component while HearFit emits and collects high-frequency signal. No matter in which condition, it is obvious that noise reduction significantly improve the performance of HearFit.

2) Impact of environment: We evaluate the impacts of four indoor environments that have different furniture densities and layouts. The distribution of the F1-score for each environment is shown in Fig. 11. It shows that different environments have almost no impacts on system performance as long as there is line-of-sight (LOS) signal. Moreover, when there are small objects between users and smart speakers, such as benches, boxes or dustbins, the F1-score of sit-up slightly decreases by 0.4%. This may be because when the user lies on the floor, the small objects block some reflected signal of user's body, while only block little signal when he/she takes other actions.

3) Impact of distance: Then we study the impact of the distance between the smart speaker and the user. Fig. 12 shows the F1-score of all fitness actions and 3 types of them at different distances. It can be seen that HearFit has good performance when the distance is less than $2m$, which is sufficient in most of indoor environments. In addition, the F1-score of dumbbell curl (Fitness b) decreases relatively faster than others, since only forearms are used in this action and the reflected signal of forearms is relatively weak. Note that a smaller distance does not always result in a higher F1-score, for HearFit can not capture the reflected signal of the user's whole body if the distance is too small.

4) Impact of face orientation: It is possible for users to face to different orientations in different environments. We define the degree as $0°$ when a user completely faces to the smart speaker. When the user turns left, the degree decreases, otherwise the degree increases. We measure the F1-score of 5 angles at 3 distances, and the results are shown in Fig. 13. The F1-score of all angles exceeds 93%, and it keeps relatively high when the angle is between $-45°$ and $45°$. If the angle exceeds this range, the F1-score gradually decreases. A special condition is that when the user turns back to the smart speaker, the movements of some parts of his/her body (e.g., forearm) can not be detected by HearFit. So, we suggest that users face to their smart speakers as much as possible.

5) Impact of user: 6 volunteers are involved in this experiments with height ranging from $1.6m$ to $1.86m$ and weight ranging from $46kg$ to $82kg$, including fitness coaches, people who regularly exercise and people who rarely exercise. We divide their fitness levels into 3 classes: master, normal and novice. Fig. 14 shows the distribution of F1-score for 6 volunteers at different levels. We can see that HearFit has good performance on both master and normal users. However, due to the fact that novices have almost no fitness training, they can not keep stable when doing some actions (e.g., one leg deadlift), leading to a slightly lower F1-score. But after a few training, the F1-score of novice users can reach the normal

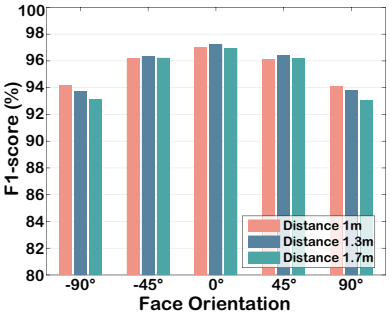


Fig. 13: F1-score under different orientations.

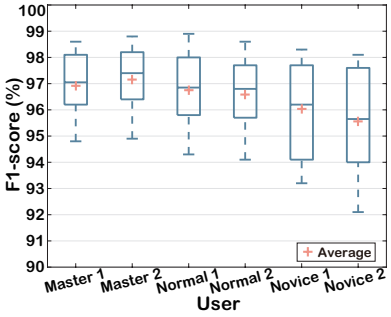


Fig. 14: F1-score under different fitness levels.

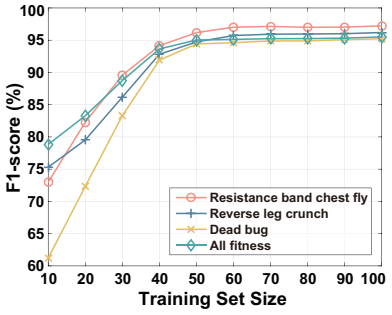
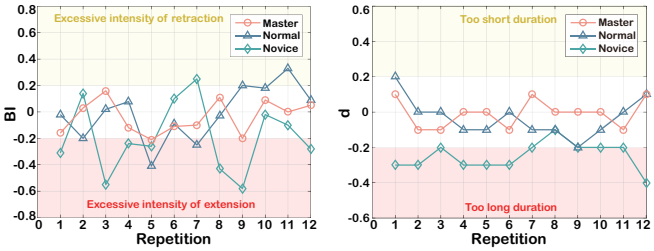
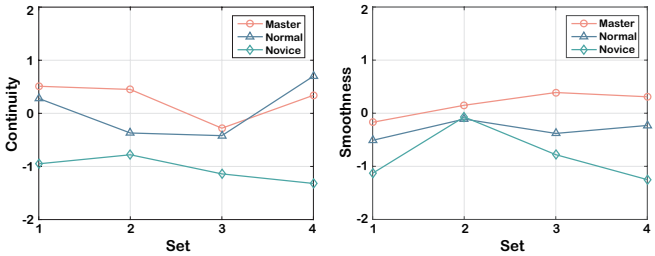


Fig. 15: F1-score for new actions.



(a) Intensity. (b) Duration.

Fig. 16: Local effect of 3 volunteers.



(a) Continuity. (b) Smoothness.

Fig. 17: Global effect of 3 volunteers.

level. On a whole, HearFit performs well under different fitness levels and different body types.

6) Impact of new action: When users add new actions, HearFit needs to retrain the last 2 layers of the LSTM network. Thus, we study how many repetitions of new and previous actions should be used to achieve a good performance. We add 3 new actions that are reverse leg crunch, dead bug and resistance band chest fly. Fig. 15 shows the F1-score under different training set sizes of each action. We can see that when the training set is greater than 50, the average F1-score of all fitness actions is greater than 95%. In other words, to add a new action, HearFit only needs to collect 50 repetitions. Compared with retraining all layers, this method based on incremental learning can reduce training data by 82% and training time by 96%.

E. Performance of Effect Evaluation

In order to test the performance of effect evaluation, we ask 3 volunteers at different levels to do 4 sets of squats with each set containing 12 repetitions. We measure the quality from aspects of local effect and global effect.

1) Local effect: Fig. 16(a) shows the intensity of each repetition in the last set. It can be seen that the master user can maintain a relatively stable and standard intensity. The normal user begins to lose stability in the last few repetitions. The novice user has the most unstable intensity due to lacking fitness training. Fig. 16(b) shows the duration of each repetition in the last set. Similar to intensity, the durations of the master user and the normal user are close to the standard duration, while most of durations of the novice user are larger than the standard duration.

2) Global effect: Finally, we evaluate the globe effect of the 4 sets. Fig. 17(a) shows the continuity of each set. We can see the master user has the best continuity, and the continuity of the normal user is only slightly lower than that of the master user. When the number of sets increases, the continuity of the novice user decreases significantly, since the physical strength of the novice user is insufficient. The same trend appears in smoothness evaluation, as shown in Fig. 17(b). When users view fitness statistics, we can provide them with standard fitness data as a reference. They can improve the fitness effect according to the statistics provided by HearFit.

V. CONCLUSION AND FUTURE WORK

In this paper, we design and implement HearFit, the first non-invasive fitness monitoring system based on smart s-peakers working at home/office, which aims to provide users with fine-grained fitness statistics to improve fitness effect. HearFit extracts features from fitness actions based on Doppler profiles, then adopts an LSTM network as classifier. Finally, it evaluates fitness effect from the local view and the global view. HearFit is anti-interference by adopting beamforming, it also allows users to add new actions conveniently. The extensive experiments show that it achieves an average classification accuracy of 96.13%, and can provide accurate fitness statistics.

HearFit is suitable to work under LOS conditions, which means that there are no obstacles between most of the user's body and the smart speaker. When multiple users use the same smart speaker, HearFit needs users to pick their own statistics from the smartphone. As our future work, we will use multiple smart speakers at different locations to overcome NLOS, and use deep learning to identify multiple users automatically.

REFERENCES

[1] J. J. Smith, N. Eather, P. J. Morgan, R. C. Plotnikoff, A. D. Faigenbaum, and D. R. Lubans, "The health benefits of muscular fitness for children and adolescents: a systematic review and meta-analysis," *Sports Medicine*, vol. 44, pp. 1209–1223, 2014.

[2] G. Schuler, V. Adams, and Y. Goto, "Role of exercise in the prevention of cardiovascular disease: results, mechanisms, and new perspectives," *European Heart Journal*, vol. 34, no. 24, pp. 1790–1799, 2013.

[3] S. Rocca, "GymBook," https://www.gymbookapp.com/, 2020.

[4] Under Armour, "myfitnesspal," https://www.myfitnesspal.com/, 2020.

[5] Nexercise, "Sworkit," https://www.sworkit.com/, 2019.

[6] P. S. Madanayake, W. A. D. K. Wickramasinghe, H. P. Liyanarachchi, H. M. D. M. Herath, A. Karunasena, and T. D. Perera, "Fitness Mate: Intelligent workout assistant using motion detection," in *IEEE ICIAfS*, 2016, pp. 1–5.

[7] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE THMS*, vol. 45, no. 1, pp. 51–61, 2015.

[8] O. Çeliktutan, C. B. Akgül, C. Wolf, and S. Bülent, "Graph-based analysis of physical exercise actions," in *ACM MIIRH*, 2013, pp. 23–32.

[9] D. Morris, T. S. Saponas, A. Guillory, and I. Kelner, "RecoFit: Using a wearable sensor to find, recognize, and count repetitive exercises," in *ACM CHI*, 2014, pp. 3225–3234.

[10] F. Ding, Q. Zhang, R. Zhao, and D. Wang, " TTBA: An RFID-based tracking system for two basic actions in free-weight exercises," in *ACM Q2SWinet*, 2018, pp. 7–14.

[11] Y. Zhu, D. Wang, R. Zhao, Q. Zhang, and A. Huang, " FitAssist: Virtual fitness assistant based on WiFi," in *ACM MobiQuitous*, 2019, pp. 328–337.

[12] H. Chen, F. Li, W. Du, S. Yang, M. Conn, and Y. Wang, "Listen to your fingers: User authentication based on geometry biometrics of touch gesture," *ACM IMWUT*, vol. 4, no. 3, pp. 75:1–75:23, 2020.

[13] K. Sun, T. Zhao, W. Wang, and L. Xie, "VSkin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *ACM MobiCom*, 2018, pp. 591-605.

[14] H. Chen, F. Li, and Y. Wang, "EchoTrack: Acoustic device-free hand tracking on smart phones," in *IEEE INFOCOM*, 2017, pp. 1–9.

[15] 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 *IEEE INFOCOM*, 2018, pp. 1466–1474.

[16] Y. Xie, F. Li, Y. Wu, S. Yang, and Y. Wang, "D^3-Guard: Acoustic-based drowsy driving detection using smartphones," in *IEEE INFOCOM*, 2019, pp. 1225–1233.

[17] National Public Radio, "The smart audio report," https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/, 2020.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE TASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[19] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *IEEE TAEECE*, 2013, pp. 208–212.

[20] E. Belouadah and A. Popescu, "DeeSIL: Deep-shallow incremental learning," in *ECCV Workshops*, 2018.

[21] A. C. of Sports Medicine, "ACSM's guidelines for exercise testing and prescription," in *Lippincott Williams & Wilkins*, 2013.

[22] E. Velloso, A. Bulling, and H. Gellersen, "MotionMA: Motion modelling and analysis by demonstration," in *ACM CHI*, 2013, pp. 1309–1318.

[23] M. Capeccia, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, A. Monteriù, L. Romeo, and F. Verdini, "A Hidden Semi-Markov Model based approach for rehabilitation exercise assessment," *Journal of Biomedical Informatics*, vol. 78, pp. 1–11, 2018.

[24] F. Kirchbuchner, T. G. Puppendahl, M. R. Hastall, M. Distler, and A. Kuijper, "Ambient intelligence from senior citizens perspectives: understanding privacy concerns, technology acceptance, and expectations," in *Springer AmI*, 2015, pp. 48–59.

[25] Y. Zou, D. Wang, S. Hong, R. Ruby, D. Zhang, and K. Wu, "A low-cost smart glove system for real-time fitness coaching," *IEEE IOT*, pp. 1–1, 2020.

[26] X. Guo, J. Liu, and Y. Chen, "FitCoach: Virtual fitness coach empowered by wearable mobile devices," in *IEEE INFOCOM*, 2017, pp. 1–9.

[27] D. S. Elvitigala, D. J. C. Matthies, L. David, C. Weerasinghe, and S. Nanayakkara, "GymSoles: Improving squats and dead-lifts by visualizing the users center of pressure," *ACM CHI*, 2019, pp. 1–12.

[28] X. Guo, J. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "Device-free personalized fitness assistant using WiFi," *ACM IMWUT*, vol. 2, no. 4, pp. 165:1–165:24, 2018.

[29] H. Ding, L. Shangguan, Z. Yang, J. Han, Z. Zhou, P. Yang, W. Xi, and J. Zhao, " FEMO: A platform for free-weight exercise monitoring with RFIDs," in *ACM SenSys*, 2015, pp. 141–154.

[30] N. Xiao, P. Yang, Y. Yan, H. Zhou, and X. Li, "Motion-Fi: Recognizing and counting repetitive motions with passive wireless backscattering," in *IEEE INFOCOM*, 2018, pp. 2024–2032.

[31] F. Xiao, J. Chen, X. Xie, L. Gui, L. Sun, and R. Wang, "SEARE: A system for exercise activity recognition and quality evaluation based on green sensing," *IEEE TETC*, pp. 1–1, 2018.

[32] V. P. Sivonen, "Directional loudness and binaural summation for wideband and reverberant sounds," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 2852–2861, 2007.

[33] H. Zhang, W. Du, P. Zhou, M. Li, and P. Mohapatra, "DopEnc: Acoustic-based encounter profiling using smartphones," in *ACM MobiCom*, 2016, pp. 294–307.

[34] H. Chen, F. Li, X. Hei, and Y. Wang, "CrowdX: Enhancing automatic construction of indoor floorplan with opportunistic Encounters," *ACM IMWUT*, vol. 2, no. 4, pp. 159:1–159:21, 2018.

[35] T. Wang, D. Zhang, Y. Zheng, T. Gu, X. Zhou, and B. Dorizzi, "C-FMCW based contactless respiration detection using acoustic signal," *ACM IMWUT*, vol. 1, no. 4, pp. 170:1–170:20, 2018.

[36] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *IEEE HSCMA*, 2008, pp. 69–72.

[37] C. Shen, B. Ho, and M. Srivastava, "MiLift: Efficient smartwatch-based workout tracking using automatic segmentation," *IEEE TMC*, vol. 17, no. 7, pp. 1609–1622, 2018.

[38] 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," *ACM IMWUT*, vol. 3, no. 1, pp. 3:1–3:21, 2019.

[39] L. Lu, J. Yu, Y. Chen, Y. Zhu, X. Xu, G. Xue, and M. Li, "KeyListener: Inferring keystrokes on QWERTY keyboard of touch screen through acoustic signals," in *IEEE INFOCOM*, 2019, pp. 775–783.

[40] N. Y. Hammerla, S. Halloran, and T. Plotz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *ACM IJCAI*, 2016, pp. 1533–1540.

[41] E. T. Hall, " The hidden dimension," *Doubleday Anchor Books*, 1966.