Earmonitor: In-ear Motion-resilient Acoustic Sensing using Commodity Earphones

XUE SUN, Northwest University, China
JIE XIONG, University of Massachusetts Amherst, USA
CHAO FENG*, Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, Northwest University, China
WENWEN DENG, Northwest University , China
XUDONG WEI, Northwest University, China
DINGYI FANG, Internet of Things Research Center, Northwest University, China
XIAOJIANG CHEN, Internet of Things Research Center, Northwest University, China

Earphones are emerging as the most popular wearable devices and there has been a growing trend in bringing intelligence to earphones. Previous efforts include adding extra sensors (e.g., accelerometer and gyroscope) or peripheral hardware to make earphones smart. These methods are usually complex in design and also incur additional cost. In this paper, we present Earmonitor, a low-cost system that uses the in-ear earphones to achieve sensing purposes. The basic idea behind Earmonitor is that each person's ear canal varies in size and shape. We therefore can extract the unique features from the ear canal-reflected signals to depict the personalized differences in ear canal geometry. Furthermore, we discover that the signal variations are also affected by the fine-grained physiological activities. We can therefore further detect the subtle heartbeat from the ear canal reflections. Experiments show that Earmonitor can achieve up to 96.4% Balanced Accuracy (BAC) and low False Acceptance Rate (FAR) for user identification on a large-scale data of 120 subjects. For heartbeat monitoring, without any training, we propose signal processing schemes to achieve high sensing accuracy even in the most challenging scenarios when the target is walking or running.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

ACM Reference Format:
Xue Sun, Jie Xiong, Chao Feng, Wenwen Deng, Xudong Wei, Dingyi Fang, and Xiaojiang Chen. 2022. Earmonitor: In-ear Motion-resilient Acoustic Sensing using Commodity Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 182 (December 2022), 22 pages. https://doi.org/10.1145/3569472

*Corresponding author.

Authors' addresses: Xue Sun, sunxue@stumail.nwu.edu.cn, Northwest University , xi'an, China; Jie Xiong, jxiong@cs.umass.edu., University of Massachusetts Amherst, USA; Chao Feng, chaofeng@nwu.edu.cn., Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, and Northwest University, China; Wenwen Deng, Northwest University , China; Xudong Wei, Northwest University, China; Dingyi Fang, Internet of Things Research Center, and Northwest University, China; Xiaojiang Chen, xjchen@nwu.edu.cn, Internet of Things Research Center, and Northwest University, China.

2474-9567/2022/12-ART182 $15.00
https://doi.org/10.1145/3569472

1 INTRODUCTION

Earphones are one of the most popular wearables in our everyday life. To improve user experience and support new features such as touch control, there has been a growing trend in bringing intelligence to earphones. For example, Jabra Elite Sport [17] equips the earphones with multiple sensors to support motion tracking. Huawei Freebuds [19] embed bone vibration sensors to enhance the quality of voice recording. Apple's Airpods pro [1] uses an extra inward-facing microphone to analyze noise coming from outsides so that anti-noise waveforms can be generated to cancel the noise for better audio quality. This technology is called active noise cancellation (ANC) and it is expected that the ANC-enabled earphones with an extra in-ear microphone embedded will account for 75% of the earphone market by 2027 [31].

In addition to the basic music play and voice communication functions, new functions can now be realized on earphones. HeadFi [9] designed a peripheral hardware device to be connected to cheap headphones to enable touch-based gesture control, user identification, and heartbeat monitoring on headphones. Although the peripheral hardware design works well, connecting an extra piece of hardware is still troublesome in some real-life scenarios. OESense [21] utilized earphones to capture the bone conduction sounds to realize three applications, i.e., step counting, daily activity sensing and hand-to-face gesture recognition. In comparison to previous works, we introduce Earmonitor, a low-cost earphone with a minimal hardware modification that can achieve user identification and heartbeat monitoring. Specifically, Earmonitor utilizes acoustic signals to characterize the fine-grained ear canal structure. Based on the fact that the size and geometry of each individual's ear canal are unique [23, 38], accurate user identification can be achieved. We further discovered that while the signal reflection is able to characterize the ear canal structure, the fine-grained signal variation can be utilized to characterize subtle physiological activities such as heartbeat. This is because heartbeats cause the blood vessel under the skin of the ear canal to vary and the subtle change can be detected through the signal variation.

In this paper, we propose to add a very cheap microphone component (less than 10 cents) to cheap low-end earphones ($5) to equip it with the capability of fine-grained sensing. The basic idea is that we can extract the unique ear canal features and ear canal dynamics from acoustic signals reflected back from the ear canal and received at the in-ear microphone. Though promising, to realize fine-grained sensing using the ear canal reflected signals, several challenges need to be addressed:

- While the added microphone can capture the reflected signal for sensing, the direct path signal between the speaker and the added microphone is much stronger which can severely interfere with sensing. Traditional method uses insulation material to block the direct path signal. This method works well for communication but it has been shown in a recent work that the leakage from the insulation material is still strong enough to interfere with sensing [7]. This is because higher frequency causes a larger signal leakage. As ultrasound is exploited for sensing, the leakage is more severe during the sensing process while the human voice for communication is usually in the low-frequency range.
- Another challenge is that fine-grained sensing only works in very constrained conditions, i.e., the target needs to be stationary. Once the human target is moving such as walking or running, fine-grained sensing fails because the fine-grained information (e.g., heartbeat) is buried in the large-scale human movements. We observe that even speaking can significantly degrade the sensing performance of heartbeat sensing.
- The third challenge is that the captured sensing information can be very unstable. It can be affected by factors such as the earphone wearing position (depth and angle) and the target's status (e.g., head motion). Small changes may not be a big problem for tracking but can be a severe issue for applications such as user identification in which we would like to extract unique and stable features.

To address the strong direct-path interference (first challenge), we propose to cancel this direct path interference out. This is because although the direct-path interference is strong, the interference does not change as long as we keep the signal strength unchanged because the relative positions of speaker and microphone are fixed. Therefore,

we can measure this direct path interference beforehand when there is no signal reflection and later subtract it away from the received signal containing both reflection and interference. To fully cancel this interference out, we need to precisely align the pre-recorded interference and the received signal in time domain. We observed that traditional correlation-based method does not work well here because it results in sample-level offset and the residue owing to this the sample-level offset is still strong enough to severely interfere with fine-grained sensing. We therefore adopt a phase-based method to align the two signals at a sub-sample level granularity to fully cancel out the strong interference to make fine-grained sensing possible.

To address the severe movement interference induced by walking and running, we employ the Variational Mode Decomposition (VMD) method to decompose the phase information into multiple frequency components. Traditional methods assume the frequencies of targeted activity (e.g., heartbeat) and interfering activity (e.g., walking) are different and employ filters to remove the interference out. However, the walking frequency and heartbeat frequency of human targets vary in a large range. Therefore, traditional methods may fail in some real-world scenarios. In this work, we propose a novel solution to identify the interference component based on one key observation: due to large signal variations, there are harmonics associated with the component of walking/running. We can therefore leverage this property to identify the frequency component related to walking/running without requiring any pre-knowledge of these activities.

To tackle the third challenge to push the proposed system one step towards real-life adoption, we propose to utilize the Mel-frequency cepstral coefficient (MFCC) feature for sensing based on one key observation: while other features such as transfer function can be severely affected by human motions and device wearing positions, MFCC is more stable in the presence of strong interference. This is because transfer function feature is related to the ear canal whose shape and size can be easily affected by human motions such as head moving or speaking. On the other hand, the MFCC feature characterizes the vibration of eardrum, which is less affected by body motions or earphone wearing positions. However, we want to point out that although MFCC feature is more stable, it is not as unique as the transfer function feature. Therefore, in this work, we employ both transfer function and MFCC features for sensing. The main contributions of this work are summarized as follows.

- We devise Earmonitor, a low-cost design to equip cheap commodity earphone with the power of sensing. The design has a small form factor which can be integrated into existing earphones.
- Earmonitor can perform fine-grained sensing based on the subtle changes of the ear canal such as user identification and heartbeat sensing even when the target is speaking or walking. We believe working in the presence of interference is one important step towards real-life adoption of wireless sensing.
- Besides conducting experiments in the lab, we also conducted a clinical study with patients with atrial fibrillation and premature beat to show the effectiveness of the proposed system. The results of the experiment show that Earmonitor works well in real-world settings.

The rest of this paper is organized as follows. Sec. 2 introduces the literature about acoustic sensing including user identification and physiological state sensing. Sec. 3 presents the background knowledge. Sec. 4 introduces the system overview and signal processing. Sec. 5 and Sec. 6 present the detailed system design. Sec. 7 presents the implementation and evaluation of our system followed by a limitation discussion section and conclusions.

2 RELATED WORK

In this section, we first discuss the literature on acoustic sensing. Then we present the related work on user identification and physiological state sensing.

Acoustic Sensing. Acoustic signals have been applied to enable fine-grained sensing such as respiration monitoring [40, 45] and heartbeat sensing [39, 46]. BreathListener [45] utilized the Energy Spectrum Density (ESD) of acoustic signals to capture breathing pattern in driving environments and eliminated motion interference by applying the Ensemble Empirical Mode Decomposition (EEMD) scheme. Zhang *et al.* [46] utilized the smart speak

to monitor the respiration and heartbeat of static human targets. As the signal utilized for sensing is reflected from human chest, the performance is affected by body motions and even the clothes the target is wearing. In contrast, Earmonitor obtains the heartbeat information from signal reflected from the ear canal, which is more robust against external interference.

User Identification. Biometrics based on physiological and behavioral characteristics, such as fingerprint [32], face [33], voice [10], touch [20] and gait [44] have been widely used for user identification. The ear canal-based identification provides a hidden and secure user identification scheme on earphones [11–13, 42]. EarEcho [13] utilized an in-ear microphone to extract the transfer function (frequency response) of signal reflected from ear canal to depict ear canal geometry. EarDynamic [42] extracted the channel response of signal reflected from ear canal that corresponded to the ear canal deformation to achieve user identification when user was speaking. HeadFi [9] realized user identification in both static and dynamic scenes through designing a peripheral hardware device connected to headphones. EarGate [11] utilized the uniqueness of human gait for user identification. In contrast, Earmonitor combines the feature MFCC with double-ear transfer function to combat against interference induced by speaking and body movements. Meanwhile, we transmit signals with a large bandwidth to increase the diversity of ear canal features in the frequency domain for user identification. As identification is a short-period process, we found that employing a large bandwidth signal does not degrade user experience even though the sound is audible.

In-Ear Physiological Sensing. There is a growing trend in embedding sensors in earphones to sense physiological information (e.g., heartbeat, blood pressure and brain wave) of human targets [5, 9, 15, 22, 27]. For example, Goverdovsky *et al.* [15] leveraged the in-ear EEG sensor to sense the brain wave. EBP [5] utilized a custom hardware to measure the blood pressure from user's ears. Pressler *et al.* [29] proposed an ear-canal based respiration sensing system. They utilized the in-ear microphone to pick up the respiration sound. Park *et al.* [27] utilized in-ear piezoelectric sensor to infer the heart rate. Martin *et al.* [22] demonstrated the feasibility of measuring heartbeat and breathing utilizing in-ear microphone when an individual was stationary. HEARt [6] utilized deep learning techniques to enable the heart rate monitoring in the presence of body motions (e.g., walking, running and speaking). In contrast, Earmonitor is based on lightweight signal processing to remove movement interference to achieve accurate heartbeat monitoring. Earmonitor can work in the presence of strong interference including speaking, head movement and body movement (i.e., walking and running), moving one step towards more practical acoustic in-ear sensing.

3 BACKGROUND AND PRELIMINARY

In this section, we present the background knowledge about our system. We introduce the basic structure of human ear and the anatomical view of the vessels around the ear, as well as the background of heartbeat.

3.1 The Structure of Human Ear

Ear is the most important part of human auditory system that receives sound waves and converts them into neural signals that are transmitted to the brain. The ear pinna and ear canal are located in the outer ear, as shown in Fig. 1(a). Due to the geometry complexity of ear canal [42], each person's ear canal structure is unique. The uniqueness lies in many aspects including length, size, and shape of the ear canal. The length of our ear canal ranges from 2 *cm* to 3 *cm* while the volume of ear canal ranges from 0.7 *ml* to 1.5 *ml* [30]. The curvature and cross-section of the ear canal are also different for different people. Therefore, it is possible to utilize acoustic signals to depict the ear canal information to achieve user identification.

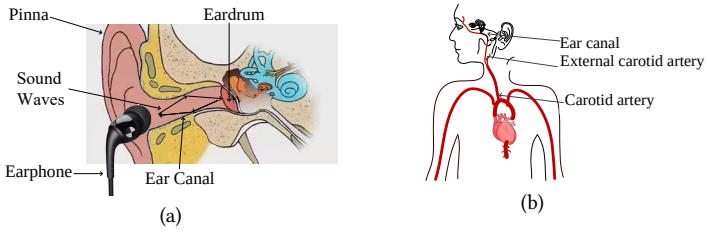


Fig. 1. (a) The structure of the ear; (b) Anatomical view of the vessels around the ear.

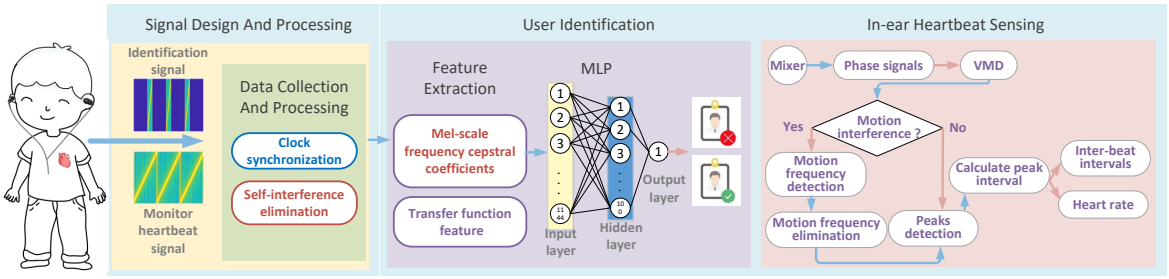


Fig. 2. System architecture of Earmonitor.

3.2 Background of Heartbeat

Heartbeat-induced blood pulse is transmitted from the *aorta* in the heart's left ventricle to various tissue organs throughout the body. There are many *aorta* branches, and among them the branch that flows to the head and neck is called the *common carotid artery*, which is further divided into the *intra-cervical artery* and the *external carotid artery*. The *external carotid artery* is sent along the way to the *maxillary artery* which is located near the ear canal wall, as shown in Fig. 1(b). When the heart beats, the *aorta* movement drives the *external carotid artery* to move and causes a pressure on the ear canal surface. It causes the deformation (e.g., expansion or compression) of the ear canal, which caused the *Ear Canal Dynamic Motion*. The variance of the ear canal surface can then be used to obtain the heartbeat information. In our work, the reflection signals bounced off the ear canal are affected by *ear canal deformation* caused by heartbeats. We thus utilize the tiny variation in the reflected signals to capture heartbeat information.

4 SYSTEM DESIGN

In this section, we first present the overview of the proposed Earmonitor system. We then introduce the signal design of Earmonitor. Lastly, we present the method to deal with interference.

4.1 System Overview

The system framework and processing flow are shown in Fig. 2, which consists of three major components: signal design and processing, user identification, and in-ear heartbeat sensing.

Signal Design and Processing. In the signal design and processing phase, the in-ear speaker transmits acoustic signals to probe the ear canal. The built-in microphone captures the signal reflected from the user's ear canal. Then we synchronize the received signal which contains both reflection and direct path signals with

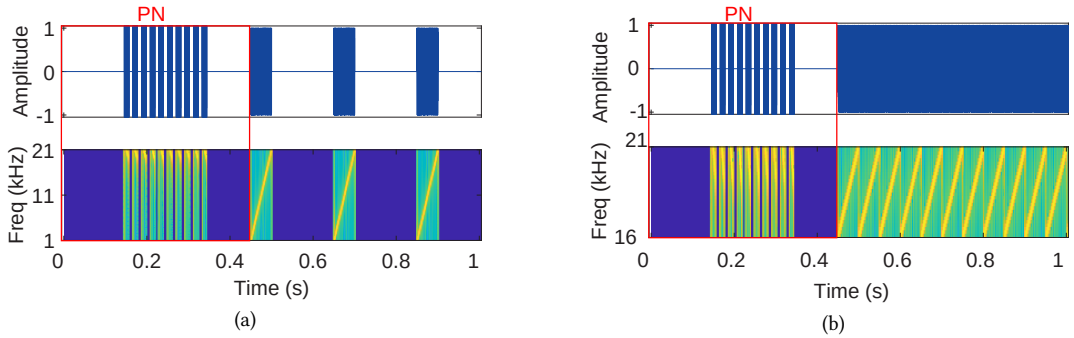


Fig. 3. An illustration of transmitted signals in time domain and frequency domain: (a) Signal design for user identification; (b) Signal design for heartbeat monitoring.

the pre-measured signal which only contains direct path signal by applying the proposed sub-sample level phase-based synchronization method. We put the earphone in an open space without any reflector in front of it to capture the direct path (speaker to microphone) signal as the pre-measured signal. After synchronization is achieved, we can remove the direct-path signal (self-interference) to obtain clean ear canal reflection signal for sensing.

User Identification. In the user identification phase, we extract both the transfer function and MFCC features from the canal-reflected signals. We then feed the extracted features into the trained classifier for user identification. The MFCC feature is less affected by body motions and therefore the proposed system can work even when the human target is in a non-static state.

Heartbeat Sensing. In order to obtain the in-ear heartbeat information, we extract the fine-grained phase variation induced by heartbeat. To deal with strong interference caused by large body movements such as walking, we leverage the Variational Mode Decomposition (VMD) scheme to decompose the signal into components of different frequencies. The key observation enabling us to differentiate the heartbeat component from other components such as walking is that large movement such as walking induces more than one frequency components. Besides the base frequency component, there are harmonics whose frequencies are integer times of the base frequency. We can then apply this property to identify the frequency component corresponding to heartbeat.

4.2 Signal Design and Processing

In this section, we introduce how to design the transmitted signals for sensing and the rationale behind the design.

4.2.1 Signal Design for User Identification. For user identification, the transmitted signal is designed as a FMCW chirp with frequency sweeping from 1 kHz to 21 kHz. The duration of the chirp is 0.05 s and there is a 0.15 s gap between adjacent chirps. A pseudo-noise (PN) preamble is added at the beginning of the transmitted signals for synchronization. Fig. 3(a) illustrates the transmitted signal in time domain and frequency domain.

The reason for such a design is twofold: on one hand, a larger bandwidth can provide richer frequency-domain characteristics for user identification. On the other hand, a longer chirp can achieve a higher SNR. However, a long chirp may cause the reflected signal to collide with the transmission of next chirp as the length of ear canal is only 2-3 cm. To ensure a higher SNR and avoid the collision issue, we set the chirp period as 50 ms and include a large gap (i.e., 150 ms) between adjacent chirps.

4.2.2 Signal Design for Sensing Heartbeat. Different from user identification based on the ear canal structure, the heartbeat information is contained in the subtle ear canal changes and therefore it is more difficult to be

Fig. 4. The components of the earphone.

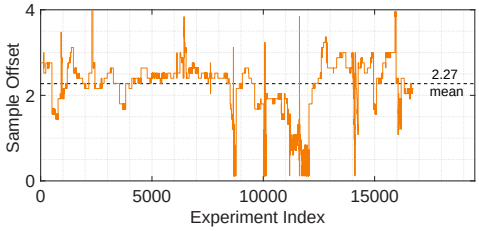


Fig. 5. System delay of received signals.

extracted. Moreover, the heartbeat information is pulse-like with intervals in between. If we transmit chirp signal with gaps in between as Fig. 3(a), the heartbeat information may not always be detected. So, we design the signal for heartbeat sensing as continuous FMCW chirps with the frequency sweeping from 16 kHz to 21 kHz. The chirp period is still set as 50 ms. Fig. 3(b) illustrates the transmitted signal in time domain and frequency domain, respectively. The reason for such a design is as following: the reflected signals is mixed with the transmitted signals and then passed through a low-pass filter. The mixed signal can be represented as:

$$m(t) = \sum_{i=1}^{n} \alpha_i cos(2\pi(\frac{B}{T}\tau_i t + f_0\tau_i - \frac{B}{2T}\tau_i^2)) \tag{1}$$

In order to extract the tiny heartbeat information, we focus on the phase of the mixed signal, $\varphi = 2\pi(f_0\tau_i - \frac{B}{2T}\tau_i^2)$ in Equation 1. As the secondary term $\frac{B}{2T}\tau_i^2$ is much smaller than $f_0\tau_i$, we omit the secondary term and obtain the phase value as $\varphi = 2\pi f_0\tau_i$. Considering the comfort of long-term heartbeat sensing, we choose a frequency band that is inaudible to human ears. The heartbeat-induced ear canal variation is on the scale of sub-millimeter level [22]. We set the starting frequency $f_0 = 16k$ Hz and the phase variation can be calculated as:

$$\varphi = 2\pi f_0\tau_i = \frac{2\pi * 16000 * 2 * 0.0001}{343} = 0.02\pi \tag{2}$$

4.2.3 Self-interference Elimination. As shown in Fig. 4, as one simple microphone is added in our design, there is direct path interference from the speaker to microphone. This interference signal is much stronger than the ear canal reflected signal. The good news is that the relative locations of microphone and speaker are fixed and therefore we can measure the interference beforehand and subtract away the direct path interference from the received signal. We record the direct path signals by putting the earphone in a open space without any reflector in front to measure the direct path signal. In practice, we found that we still encounter the challenge of non-negligible residual interference after simple subtraction. This issue is mainly caused by the random system delay before the acoustic signal is transmitted out. This makes the pre-recorded signal not aligned exactly with the randomly-delayed version transmitted for sensing. The traditional correlation-based alignment method does not work well because it still results in sample-level offset. We applied correlation-based method to align the pre-recorded signal and the measured signal. As shown in Fig. 5, for more than 16000 trials, there exists an offset of $0 - 4$ samples. One sample offset corresponds around 7 mm distance error which is too large for fine-grained heartbeat sensing requiring an accuracy on the scale of sub-millimeter.

To address this problem, we discover that the offsets at all signal paths are the same. As the distance between the microphone and speaker is fixed, the propagation time of the direct path signal is a constant and therefore we can adopt the direct path signal as a reference to calculate the random time offset for each signal reception. We

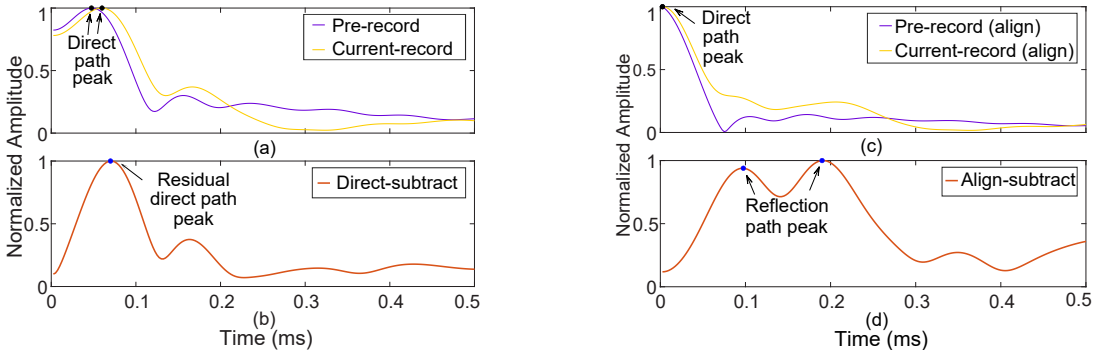


Fig. 6. Characteristics of time offset, the direct path and the reflections: (a) There is an offset between the pre-recorded and current-recorded signals; (b) The direct path peak still exists after subtraction; (c) Align the two signals at sub-sample level granularity; (d) The direct path interference is eliminated after subtraction.

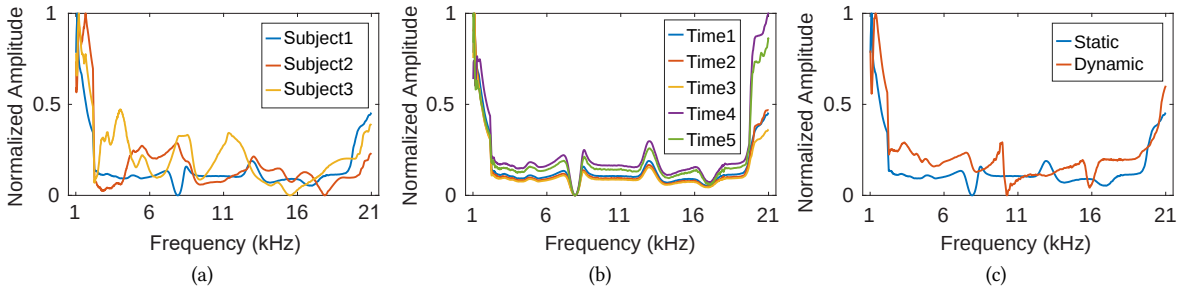


Fig. 7. Transfer function feature: (a) Different subjects; (b) Different time; (c) Static and dynamic TF feature.

first employ the FFT operation to identify the direct path peak which is much stronger than other reflection paths. The location of the identified peak presents us a coarse-grained sample-level time offset (t_i). We further employ the phase information of the signal to obtain the sub-sample level time offset (t_d). By adding t_i and t_d together, we obtain the accurate time offset by taking the direct path as the reference for each signal reception. After removing the time offset, the signal received can be precisely aligned with the pre-recorded direct path signal for subtraction to obtain clean reflection signal for sensing. Fig. 6 presents the direct path interference before and after the phase-based alignment. Fig. 6(a) plots the pre-recorded signals (with offset) and current-recorded signals (with offset). Without a tight alignment, we can see that after the subtraction, the direct path peak is still large in Fig. 6(b). By applying the phase-based alignment scheme, the direct path peak of the pre-recorded and current-recorded signals are moved to the timestamp "0" (without offset) as shown in Fig. 6(c). After the signals are precisely aligned, Fig. 6(d) shows that the direct path interference is successfully eliminated and the reflected signals are well-preserved.

5 USER IDENTIFICATION

In this section, our goal is to extract unique and stable ear canal features which include transfer function and MFCC features for user identification.

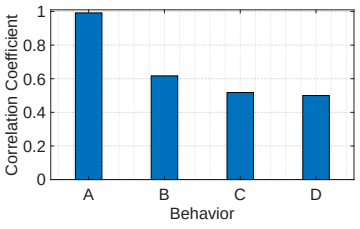
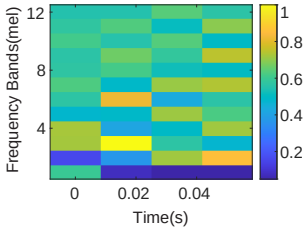
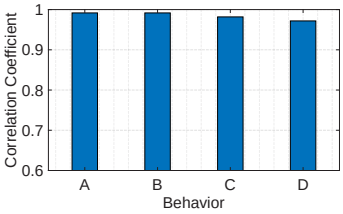


Fig. 8. TF feature similarity.



(a) MFCC



(b) Dynamic similarity

Fig. 9. MFCC features.

5.1 Transfer Function Feature Extraction

The ear canal's unique physical structure can be used to identify users. The frequency response of reflected signal can depict the geometric structure of the ear canal. We define the transfer function (TF) of the ear canal's frequency response as:

$$H(f) = \frac{P_{yx}(f)}{P_{xx}(f)} \tag{3}$$

where P_{yx} is the cross power spectral density between the recorded and transmitted signal. P_{xx} is the self power spectral density of the transmitted signal. We observe that when the information from both ears are exploited, a high identification accuracy can be obtained compared to only one ear's information is utilized. We therefore define the double-ear transfer function as:

$$H_d(f) = \frac{H_r(f)}{H_l(f)} \tag{4}$$

with $H_r(f)$, $H_l(f)$ representing the transfer function of right ear canal and left ear canal, respectively.

We estimate $H(f)$ using Welch's averaged, modified periodogram [43]. We extract the TF feature with 2400 Discrete Fourier Transform (DFT) points and a 20 ms hamming sliding window (50% overlap). The result is shown in Fig. 7(a). We can see that different subjects show obvious different TF features. In Fig. 7(b), we can also see that the TF features for the same person are stable over time. These results demonstrate the feasibility of employing the TF feature to depict the ear canal's unique structure for user identification.

5.1.1 Dynamic Deformation of Ear Canal. We further discover that when a user is shaking the head, speaking or chewing, these activities cause the ear canal to vary. The TF features of the same person when the person is static and dynamic are inconsistent, as shown in Fig. 7(c). When people are speaking or chewing, the jaw movement drives the *temporomandibular* joints to move. The joints connecting the jaw with skull are located near the ear canal and they impact the ear canal structure. Specifically, when we open mouth or chew, the ear canal's diameter can be altered by up to 2.5 *mm* [28]. Therefore, the TF feature is easily affected by body motions. We conducted a benchmark experiment to verify this. We calculate the correlation coefficient to quantity the similarity between two TF features. We can see in Fig. 8 that when the target is static (Case A), the TF feature is stable with a similarity coefficient close to 1. However, when the target is moving his head (Case B), speaking (Case C) and chewing (Case D), the features extracted are quite different from that collected when the target is in static state. The similarity coefficient decreases to 0.45 when the target is chewing. These results demonstrate that TF feature is unique for user identification. However, the feature changes when target is moving and this can cause false negative in user identification.

5.2 MFCC Feature Extraction

While the TF feature is unique, it is not stable when the target is in dynamic state. To enhance the robustness of user identification, we then extract another feature, i.e., the Mel-frequency cepstral coefficient (MFCC) from

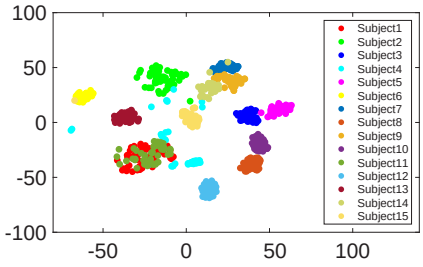


Fig. 10. MFCC feature of different subjects.

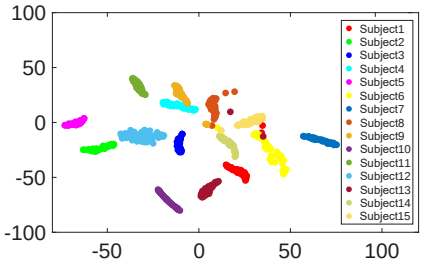


Fig. 11. Combined MFCC and TF features.

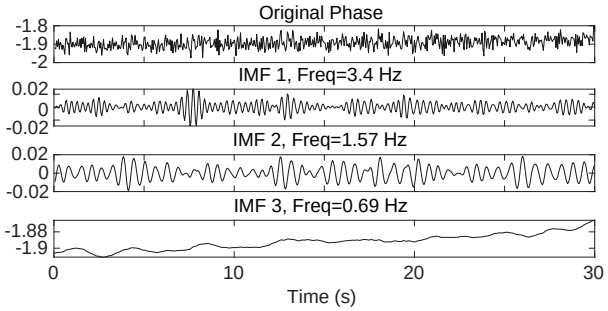


Fig. 12. The decomposed phase signals. The IMF 2 corresponds to heartbeat information.

the ear canal reflected signal. When acoustic signal is transmitted from the earphone to the ear canal, it causes the eardrum to vibrate. The MFCC feature depicts the characteristics of the eardrum vibration, which is less affected by body movements [35]. Fig. 9(a) shows the MFCC feature extracted from the signals reflected from one subject's ear canal. We calculate the MFCCs of the received signals in each sliding window with a size of 40 ms and a overlap of 20 ms between adjacent windows. The number of filterbank channels is set to 40 and the 12 order cepstral coefficients are computed in each 40 ms window. As shown in Fig. 9(b), we can see that the MFCC features are more robust against body motions (Case A is static, Case B is head moving, Case C is speaking and Case D is chewing). The similarity coefficients are close to one even the subject is performing activities.

Although the MFCC features are more stable, we discover that MFCC features do not show significantly different patterns among different subjects. As shown in Fig. 10, we first extract the MFCC from the 15 subjects, and we use the t-SNE scheme [37] to project MFCC into a two-dimensional space. We can see that MFCC features do not show clear different patterns among subjects. MFCC feature is not unique enough for user identification with a large number of users and therefore we combine it with the TF feature to achieve high accuracy and high robustness at the same time. We perform the feature fusion operation on the two features using the concat method [36]. We then visualize the combined feature for different subjects in Fig. 11. We can see that the combined feature can clearly separate subjects.

6 HEARTBEAT SENSING

In this section, we present how to extract information of the tiny heartbeat and estimate the inter-beat intervals under both stationary and dynamic scenarios (i.e., speaking, walking, and running).

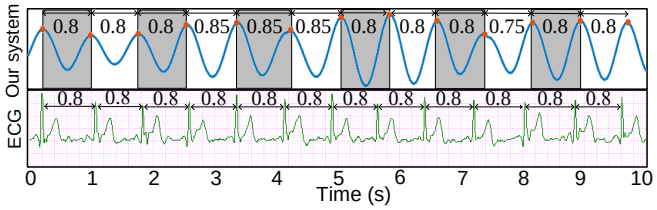
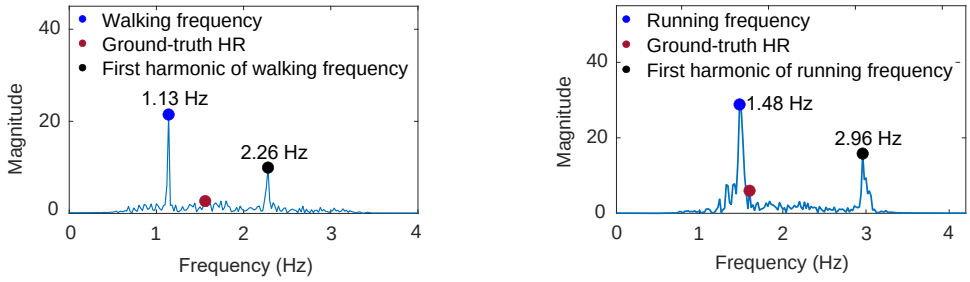


Fig. 13. Heartbeat information result compared with ECG.



(a) Walking (b) Running

Fig. 14. The FFT of motion signals.

6.1 Heartbeat Information Extraction in Static Scenarios

According to the analysis in the Sec 4.2.2, we know that both heartbeat and body movements can induce phase variation of signals. When the subject is stationary, the variation in phase is mainly caused by heartbeat and heartbeat induces periodic signal variations at a specific frequency.

To obtain the heartbeat frequency and fine-grained inter-beat interval information, we need to separate the signal components of different frequencies. Empirical model decomposition (EMD) [18] is a signal analysis method, which decomposes a signal into components of different frequencies. However, EMD method may result in overlapping of components. Therefore, we employ the Variational Mode Decomposition (VMD) [8] method to decompose the superimposed signals. This method can effectively overcome the frequency component overlapping issue through restricting the bandwidth size of each frequency component.

There is one important parameter, i.e., the number of decomposition layer e in the process of VMD scheme. The value of e is decided based on the component frequency. VMD obtains the components from high frequency to lower frequency. When the frequency of current component layer is higher than that of the previous component layer, the decomposition process terminates and the number of layers is determined. We repeat the process 100 times and obtain the average value as the information, i.e., intrinsic mode function (IMF) of each signal component layer.

Based on the above VMD decomposition method, Fig. 12 demonstrates the decomposition result of the phase information extracted from the ear canal reflected signal. We plot the three IMFs in descending order of the frequency. It can be observed that IMF 1 is the high-frequency noise while IMF 2 is the heartbeat component. The corresponding average heart rate is 1.57 Hz. IMF 3 is the low-frequency noise.

Note that not only the average heart rate matters, the heartbeat interval is also an important metric. We employ an interpolation method [34] to smooth the extracted signal waveform. We compare the extracted inter-beat intervals with the ground-truth ECG waveform collected using a dedicated ECG sensor. As shown in Fig. 13, the inter-beat intervals are consistent with the ground-truth data collected from a medical ECG device.

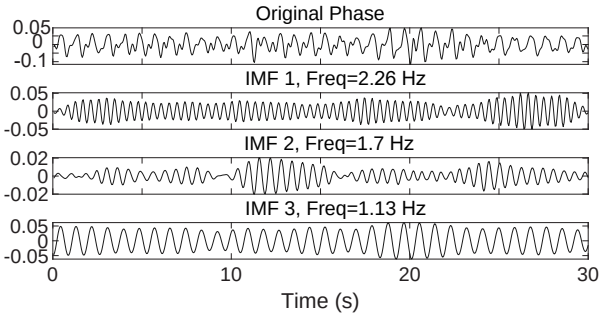


Fig. 15. The decomposed walking signals.

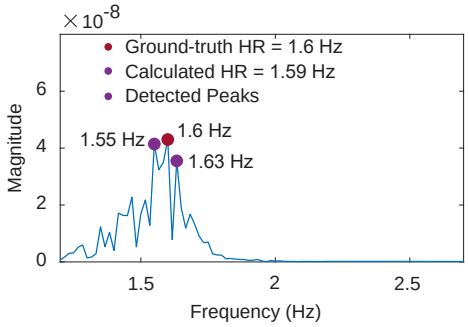


Fig. 16. The FFT of remaining signals.

6.2 Heartbeat Information Extraction when Target is Moving

When target is not stationary (e.g., walking and running), it is much more challenging to obtain the target's heartbeat information. Frequency domain analysis does not work well because the target motion frequency and heartbeat frequency can be quite similar. Moreover, as the magnitude of heartbeat signal is much weaker than that of target motions, heartbeat signals can be buried in motion signals without being detected.

6.2.1 Motion Frequency Estimation. As shown in Fig. 14, we perform FFT operation on the motion signals. We observe that there is a dominant peak at 1.13 Hz for walking as shown in Fig. 14(a). Meanwhile, there is also a dominant peak at 1.48 Hz for running as shown in Fig. 14(b). The motion magnitude is much larger than that of heartbeat, which can cause the heartbeat information not being captured. Fortunately, based on the VMD decomposition method, we discover a key observation that large motions such as walking/running induce more than one frequency components. Besides the base frequency component, there are harmonics whose frequencies are integer multiple of the base frequency. We can therefore employ this property to identify those frequency components corresponding to the large motions. As shown in Fig. 15, we plot the three IMFs in descending order of the frequency. We can see that the frequency of IMF 1 is twice that of IMF 3. We can therefore identify these two frequency components as walking-related and exclude them from being considered as heartbeat signal.

6.2.2 Heartbeat Information Extraction. After identifying the motion-related frequency components, we subtract these components from the original phase data. We then apply a band-pass filter to remove those noise clearly out of the range of heartbeat frequency (0.8 Hz-2.5 Hz) and then apply the FFT operation to the remaining signal. The result is shown in Fig. 16 and we can see multiple peaks. This is expected because the human heart rate is not a constant but varies slightly during a period of time such as one minute. We therefore detect all the peaks with amplitude larger than 0.9 of the maximum peak (i.e., the peak at 1.6 Hz). We then calculate the mean of the peaks as the average heart rate. We also observe that the frequency components (1.55 Hz, 1.6 Hz and 1.63 Hz) are slightly lower than the VMD component (1.7 Hz). This is because the high frequency noise is removed. In this experiment, the ground-truth heart rate is 1.6 Hz. The result shows that we can successfully extract heartbeat information even when the target is moving.

7 EVALUATION

In this section, we conduct a series of experiments to evaluate the performance of Earmonitor. We also conduct a clinical study with hospitalized cardiac patients to show the effectiveness of our heartbeat sensing system. Finally, we evaluate the effect of sensing signal on user's comfortableness and music play.

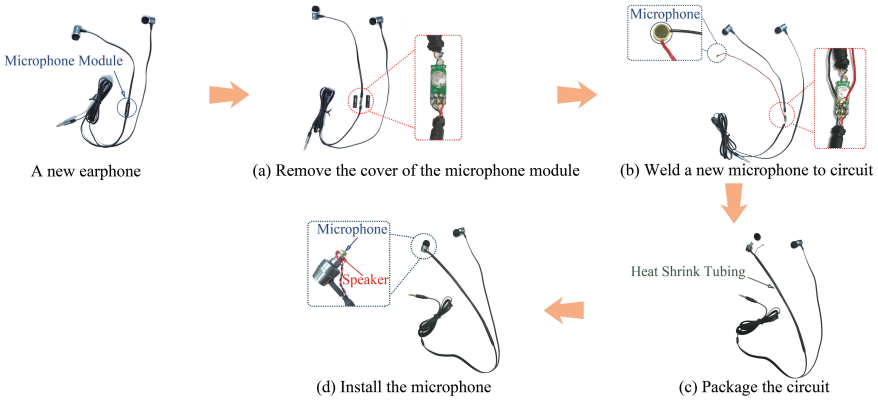


Fig. 17. The process of making the earphone.

Table 1. Demographic summary of the participants.

Age group	Number
10-20	46 (34 females, 12 males)
21-40	67 (31 females, 36 males)
41-60	5 (2 females, 3 males)
61-90	2 (2 males)

Table 2. The information of heart rate sensing about the different activities.

Environment	Subject	Activity	Length (second)	Trial
Living room	120	Sleeping	60	20
Office	120	Sitting, Speaking, Head Moving	60	20
Activity room (treadmill)	120	Walking, Running	60	20

7.1 Experiment Setup

We built our system utilizing a cheap low-end earphone ($5). The microphone chip (less than 10 cents) is attached to the front side of the earphone's speaker. The detail steps are illustrated in Fig. 17:

(a) We take a new earphone and remove the cover of the microphone module as shown in Fig. 17(a);
(b) We weld the two wires of the new microphone to the circuit board of the microphone module (the red wire is connected to the positive pole and the black one is connected to the negative pole) as shown in Fig. 17(b);
(c) We utilize a heat shrink tube to package the soldered microphone wire and the original earphone wire together as shown in Fig. 17(c);
(d) We take off the silicone earbuds that cover the earphone speaker and put the microphone on the speaker. Then we put the silicone earbuds back to cover the speaker and the microphone as shown in Fig. 17(d).

We connect the earphone to two devices which are a MacBook Pro laptop and a Samsung Galaxy S6 smartphone to control the signal transmissions and recordings. We connect the device and earphone via a 3.5 mm audio interface.

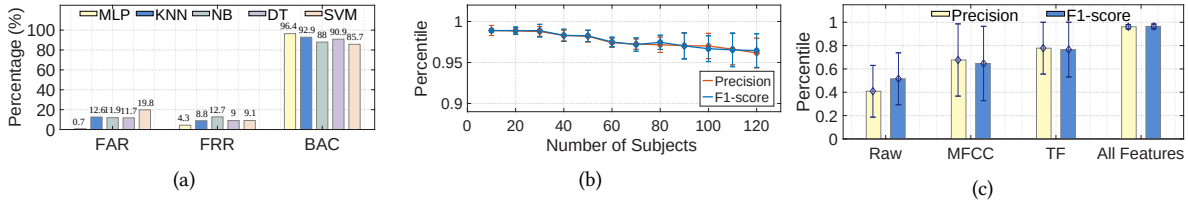


Fig. 18. Overall performance of user identification: (a) Different classification algorithms; (b) Different features and (c) Different individuals.

7.2 Data Collection, Ground-truth and Metrics

Data Collection. We recruit a total of 120 volunteers (53 males and 67 females) aged between 10 and 83 for experiments. The demographic summary of the participants is shown in Tab. 1. The volunteers wear our earphones in a way they feel comfortable. They are also asked to perform different motions to see the effect of motion on system performance. The experiments were IRB-approved by the host university.

For heart rate monitoring, we collect data for a total of six activities. These activities are sleeping, sitting, speaking, head moving, walking and running. For each activity, we repeat the data collection process 20 times. The length of each collection lasts for 60 seconds. We collect the activity data in three environments, i.e., a living room, an office, and an activity room. We summarize the detailed information about the six activities in Tab. 2.

Specifically, for sleep activity, we collected data in a living room. During the process of data collection, the participants lay on the bed with a comfortable posture and breathed naturally. For sitting, speaking, and head moving activities, we collected data in an office environment. For sitting, the participants sat in a chair and breathed naturally. For speaking, the participants read aloud a book without other body movements. For head moving, the participants randomly shook their heads. Each data collection process lasts for 60 s. For walking and running activities, we collected data in an activity room with treadmills. For walking activity, the participants walked on the treadmill at a speed of 2 km/h for 60 s. For running activity, the participants run on the treadmill at a speed of 4 km/h for 60 s.

Ground-truth. In our system, we utilize two devices, i.e., BIOPAC MP160 [24] and Polar H10 [16] to measure the ground-truths. We adopt MP160 for ground-truth measurements in static scenarios and use Polar H10 for ground-truths in dynamic scenarios.

Metrics. Besides the common metrics (i.e., Precision, F1-score, FAR and FRR), we also adopt the following metrics to evaluate the performance of our system.

- Balanced Accuracy (BAC): $BAC = \frac{TPR+TNR}{2}$, in which TPR and TNR are true positive rate and true negative rate, respectively.
- Mean Absolute Error (MAE): the mean absolute difference between the estimated heart rate H^E and the actual heart rate H^A, i.e., $MAE = \frac{1}{N}\sum_{i=1}^{N}|H^E - H^A|$.
- Mean Percentage Error (MPE): the mean percentage difference between H^E and H^A, i.e., $MPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|H^E-H^A|}{H^A}$.

7.3 User Identification Performance

7.3.1 Overall Performance. We first evaluate the overall performance of user identification. We consider different machine learning algorithms for classification, including Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), Naive Bayesian (NB), Decision Tree (DT), and Support Vector Machine (SVM). We employ 10-fold cross-validation to split collected user data. The result is shown in Fig. 18(a). Among the 5 algorithms, MLP achieves the best overall performance (0.7% FAR, 4.3% FRR, and 96.4% BAC). We therefore adopt MLP for the rest of our evaluation. We then evaluate our system performance with increasing the number of subjects. The

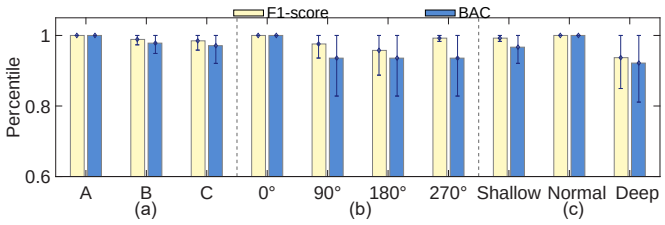
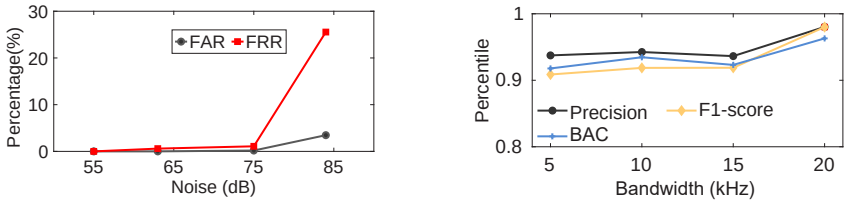


Fig. 19. System robustness: (a) Body motions (A: head moving, B: speaking, C: chewing); (b) Earphone wearing angles; (c) Earphone wearing depths.

Time	BAC	F1-score	Precision
One hour	0.984	0.984	0.983
One day	0.974	0.972	0.975
One week	0.954	0.952	0.954
One month	0.952	0.951	0.952

Table 3. Long-term user identification.



(a) Environmental noise

(b) Bandwidth

Fig. 20. Robustness quantification.

result is shown in Fig. 18(b), as the number of subjects increased from the 10 to 120, the precision only decreases slightly from 99.4% to 96.4%.

7.3.2 Contribution of Each Feature. We now quantify the performance gain obtained from each feature. As shown in Fig. 18(c), the raw data without feature extraction achieves a poor performance with a precision of 42% and a F1-score of 50%. When only MFCC feature or only TF feature is employed, the achieved precision is about 70% and 80% respectively. With the proposed scheme to combine the two features, our system can achieve a much higher precision of 96%.

7.3.3 System Robustness. We now evaluate the system robustness in various challenging scenarios.

- **Body Motion.** We evaluate the system performance in the presence of different body motions: head moving (Case A), speaking (Case B) and chewing (Case C). In Fig. 19(a), we can see that the average F1-score and BAC values are about 98% in the presence of head moving and speaking motions. Chewing motion has the most significant effect on the performance and the BAC value is still above 90%. We believe this is because chewing alters the ear canal's geometry structure [4], which severely impacts ear canal's frequency response.
- **Earphone Wearing Position.** To illustrate the impact of earphone wearing position on system performance, we conduct experiment at four earphone wearing angles θ (0°, 90°, 180° and 270°) and three in-ear depth d (Shallow, Normal and Deep). The results are shown in Fig. 19(b) and Fig. 19(c). We can see that different wearable behaviors do have a slight effect on system performance. However, the average F1-score and BAC are always above 90%.
- **Resistance to Environmental Noise.** Different environmental noise may influence the sensing signals and thus will impact the performance of our system. We choose four environments with different sound noise levels that are 55 dB, 63 dB, 75 dB and 84 dB, respectively. As shown in Fig. 20(a), when the environmental noise is below 80 dB, FAR and FRR are below 3%. FAR and FRR increase dramatically when the noise sound level is 84 dB. It is worth noting that noise above 80 dB is rare in our daily life.

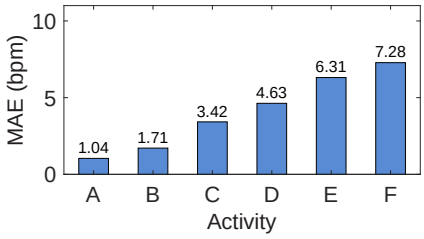


Fig. 21. The MAE of different activities (A: sleeping, B: sitting, C: speaking, D: head moving, E: walking, F: running).

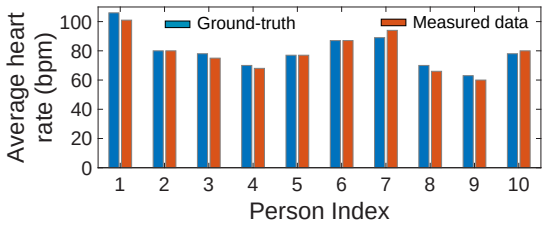


Fig. 22. The heartbeat rates of different subjects.

- **Impact of Chirp Bandwidth.** Fig. 20(b) shows the user identification performance under different chirp bandwidths. We can see that when the bandwidth is relatively small (5 kHz), the precision is around 93%. As we increase the bandwidth to 20 kHz, the precision has a noticeable increase to about 98%.
- **Long-term User Identification.** To evaluate the long-term stability of the system performance, we track three volunteers over one month and conduct the data collection once a day for one month. We record the user identification performance over time. The result is shown in Tab. 3. We can see that the performance only slightly degrades. The BAC value decreases from 98.4% (one hour) to 95.2% over one month.

7.4 Heartbeat Monitoring Performance

7.4.1 Different Activities. We evaluate the performance of heartbeat monitoring with 120 subjects in the presence of different activities. These activities are sleeping (Case A), sitting (Case B), speaking (Case C), head moving (Case D), walking (Case E) and running (Case F). The results are shown in Fig. 21. We can see that the MAEs (mean absolute errors) for sleeping, sitting, speaking, head moving, walking and running are 1.04 *bpm*, 1.71 *bpm*, 3.42 *bpm*, 4.63 *bpm*, 6.31 *bpm* and 7.28 *bpm*, respectively. In summary, a much smaller MAE can be achieved for static scenarios (Case A and B). The average heartbeat rate is within the range of 40 to 180 *bpm*. We want to point out that although the MAE for running is higher, the error rate is still quite low (around 6%) because the heartbeat rate is much higher (i.e., a mean heart rate of 120 *bpm*) during the running process. For heartbeat rate measurement, an error lower than 10% is usually considered good and this is also the requirement adopted by a lot of commodity devices [3].

7.4.2 Different Subjects. We now zoom in to evaluate the performance of heartbeat monitoring for each subject. We pick 10 subjects and show the measured heartbeat rates and the ground-truths in Fig. 22. We can see that the ground-truths vary across persons but the errors are always small.

7.4.3 Long-term Tracking. In this experiment, we monitor the heart rate of one volunteer continuously for 20 minutes under different activities (i.e., sleeping, sitting, speaking, head moving, walking and running). We set 5 s as the window size to calculate the average heart rate under each activity and show the results in Fig. 23. We can see that the measured heart rates are very close to the ground-truths when the target is sleeping, sitting, speaking and moving head. When the target is walking or running, there are larger deviations but the overall trend still matches the ground-truth well, demonstrating the effectiveness of the proposed system in the most challenging scenarios.

7.4.4 Impact of Music Play. Listening to music is the key function of earphones. We first evaluate whether our design affects the quality of music play through the subjects' perceptions. We ask each subject to listen to music

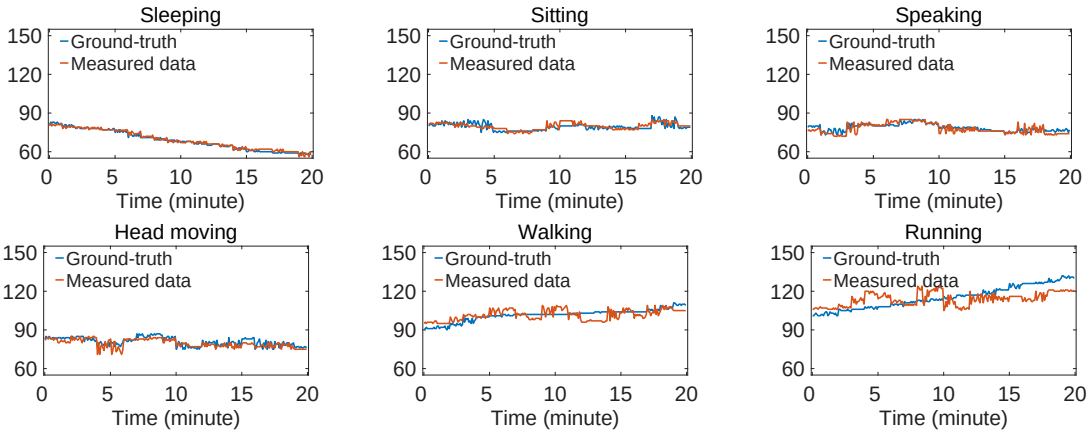
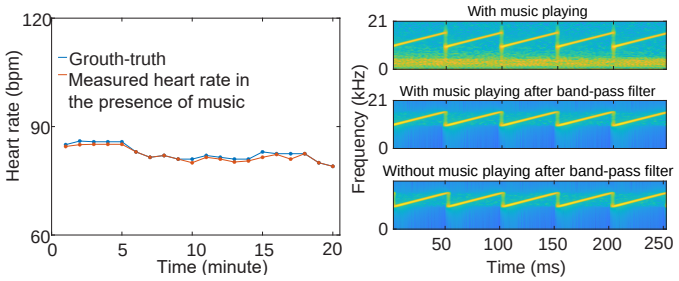


Fig. 23. Long-term tracking performance under different activities.



(a) (b)

Fig. 24. The impact of music play.

(a) (b)

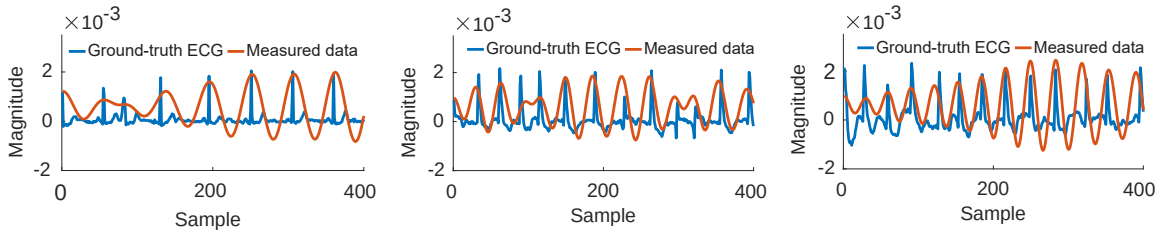
Fig. 25. Clinical study with cardiac patients.

with unmodified and modified earphones respectively and collect their perception about the audio quality. All subjects report that no difference can be perceived between the two earphones.

We then evaluate if the proposed system can monitor the target's heart rate and be used to listen to music simultaneously. We track one volunteer's heart rate for 20 minutes and at the same time, the volunteer is listening to music using the same earphone. As shown in Fig. 24(a), we can see that the measured heart rates are still very close to the ground-truths even in the presence of music playing. This is because the frequency of music is usually below 4 kHz which is far below the frequency of sensing signal (16 kHz - 21 kHz). Therefore, a band-pass filter can easily remove the music signal for sensing as shown in Fig. 24(b). The result demonstrates the feasibility of using Earmonitor to measure heart rate in the presence of music play.

7.5 Clinical Study with Cardiac Patients

In this section, we attempt to assess the performance of our heartbeat monitoring system in real-world hospital settings. We conducted a clinical study with hospitalized cardiac patients with diverse cardiac abnormalities including coronary disease, atrial fibrillation, premature beat and other symptoms. The experiments conducted were IRB-approved. Table. 4 shows the demographic of six cardiac patients.



(a) Premature beats (b) Atrial fibrillation (c) Coronary disease

Fig. 26. Clinical study shows the heartbeat information for cardiac patients, blue and red lines represent ECG device and our system.

Table 4. Demographic information.

ID	Age	Gender	Diagnosis	ID	Age	Gender	Diagnosis
1	81	male	coronary disease	4	59	male	atrial fibrillation
2	68	male	atrial fibrillation	5	58	male	atrial fibrillation
3	65	female	atrial fibrillation	6	46	female	premature beat

Fig. 25 shows the experiment scene with the proposed system deployed in a hospital setting. The patients were sitting or lying on the hospital bed and our earphones were worn in the patient's ears. Fig. 26(a), Fig. 26(b) and Fig. 26(c) show the heartbeat information of the premature beats, atrial fibrillation and coronary disease, respectively. The mean absolute error in the R-R intervals between measured and the ground-truth ECG is 0.05 s, 0.06 s and 0.09 s, respectively. In addition, within the context of clinical practice, the normal R-R interval is 0.6 s to 1.2 s. An R-R interval below 0.4 s or above 1.5 s is considered an abnormal R-R interval. For example, Fig. 26(a) has a slow heartbeat with an R-R interval of 1.5 s. An error of 0.05 s corresponds to an error percentage of 3.3%.

7.6 User Experience Survey

In this section, we evaluate the effect of sensing signal on user's comfortableness, as well as the effect of sensing signal on music.

7.6.1 The Effect of Sensing Signal on User's Comfortableness. We adopted a questionnaire-based method, which is commonly used in social and health sciences [14]. We invited 120 volunteers to wear the earphone and participate in our experiment. We asked the volunteer to rate their feeling on a scale of 1 to 5 with 1 indicating "Absolutely not uncomfortable" and 5 indicating "Extremely uncomfortable". The detailed scales are shown in Tab. 5.

For user identification, we send chirp signal with a bandwidth of 20 kHz (1 kHz to 21 kHz). The signal transmission lasts for one second. The chirp length is fixed as 50 ms and 5 chirps are transmitted with 150 ms gaps in between. For heart rate sensing, the bandwidth is 5 kHz (16 kHz to 21 kHz). Signal transmissions last for one minute and the 50 ms chirp is transmitted continuously. The signal volume is set as 6% of the maximum volume. The result is shown in Tab. 5. We can see that although the transmitted signals are in the audible range, for user identification, almost 100% of volunteers report that the signals do not cause any discomfort. This is because we use a small volume and it is a short process which lasts only for one second. For heartbeat sensing, we can see that 9% of volunteers report absolutely not feeling uncomfortable and around 74% of volunteers report they are basically not feeling uncomfortable. However, 13% of volunteers report they feel slightly uncomfortable and 4% of volunteers report that they feel uncomfortable. One interesting observation is that the 4% volunteers are in the age range of 10-12 years old. We believe this is because young people have a higher chance to hear the near-ultrasound signals. To address this issue, we embed the sensing signal into music [25] and transmit the combined signal for sensing. The new result is also presented in Tab. 5. We can see that when we embed the

Table 5. Does continuous or repeated action of playing sound cause you uncomfortable?

Score (Explanation)	User identification	Heartbeat sensing	Signal embedded with music
1 (Absolutely not)	99%	9%	100%
2 (Basically not)	1%	74%	0%
3 (Slightly)	0%	13%	0%
4 (Very much)	0%	4%	0%
5 (Extremely)	0%	0%	0%

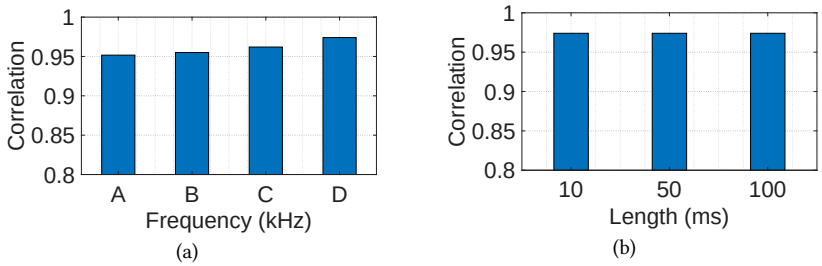


Fig. 27. Objective evaluations about the effect of chirps on music: (a) different frequency chirps (A:1 kHz - 21 kHz, B: 6 kHz - 21 kHz, C: 10 kHz - 21 kHz, D: 16 kHz - 21 kHz); (b) different length chirps with the frequency from 16 kHz to 21 kHz.

Table 6. The survey about the effect of sensing signal on music quality.

Score (Explanation)	User identification	Heartbeat sensing
1 (Clear difference)	0%	0%
2 (Some difference)	0%	0%
3 (Slight difference)	0%	0%
4 (Basically no difference)	34%	0%
5 (No difference at all)	66%	100%

small sensing signal into music, no volunteers report uncomfortable. We believe embedding sensing signals into music is an efficient method to improve user experience with acoustic sensing.

7.6.2 The Effect of Sensing Signal on Music. We then evaluate the effect of sensing signal on music from both objectively and subjectively perspectives. Objective audio quality metrics such as Perceptual Evaluation of Audio Quality (PEAQ) [26] and Perceptual Objective Listening Quality Assessment (POLQA) [2] are widely used in evaluating the audio quality. However, these two metrics are not suitable here since they are specifically designed for the quality evaluation of phone calls.

Based on the principles of these two metrics, we conduct an objective evaluation of the audio quality by computing the correlation coefficient between the music with chirp signals and the original music. We vary the chirp parameters such as the chirp frequency and chirp length, and embed the sensing chirps into music. The results are shown in Fig. 27. We can see that correlation coefficients are always above 95%. This means the music quality is not affected by the chirp signals.

To further understand the effect of chirps on music, we conduct a subjective evaluation and adopt the mean opinion score (MOS) [41] as the metric to evaluate the users' opinions on music quality change. We invited 120 volunteers to wear the earphone and participate in our experiment. We asked the volunteers to listen to a music

clip of 30 s without chirp and with embedded chirp separately. We asked the volunteers to rate their feeling on the second clip with chirp signal embedded by taking the first clip without chirp signal embedded as the reference. The rating is on a scale of 1 to 5 with 1 indicating "Clear difference" and 5 indicating "No difference at all". The detailed scales are shown in Tab. 6.

To evaluate the effect of user identification signals on music, we embed chirps in a 30 s music clip. Note that for user identification, we only send 1 s of chirp signal. For heart rate sensing, we continuously send chirps for 30 s and the chirp length is 50 ms. The results are also shown in Tab. 6. We can see that for user identification, the average MOS score is 4.7 with 66% of volunteers reporting "No difference at all". This is because although the transmitted chirps are in the audible range, the chirps only last for a small period. For heartbeat rate sensing, although the chirps last longer, as the signal is in the frequency range above 16 kHz which is much higher than the music frequency, it does not affect the music quality and the MOS score is 5.0.

8 LIMITATION AND DISCUSSION

We briefly discuss the limitations of our system.

- **Implementation on wireless earphones.** The proposed system is currently only implemented on wired earphones. This is because wireless earphones support the use of speaker and microphone at the same time only in the hands-free mode. However, in this mode, due to the transmission capability of Bluetooth, the audio sampling rate is limited to 16 kHz and therefore can not support ultrasound signal transmissions.
- **Finer-grained sensing.** We demonstrated the feasibility of employing earphones for heart rate sensing in this work. However, it is still challenging for the proposed system to obtain fine-grained ECG information. We believe a potential method to obtain more fine-grained information is to effectively combine machine learning models with signal processing techniques.
- **Other vital signs.** While tracking heart rate and breath rate have been demonstrated to be feasible, sensing another important vital sign, e.g., blood pressure is still challenging.

9 CONCLUSION

In this paper, we propose Earmonitor hosted on low-cost earphones by simply adding a cheap in-ear microphone to achieve fine-grained sensing including user identification and heartbeat monitoring. We also conducted a clinical study with hospitalized cardiac patients to show the effectiveness of the proposed system in real-world settings.

ACKNOWLEDGEMENTS

This work is supported by NSFC A3 Foresight Program Grant 62061146001. This work is also partially supported by the National Natural Science Foundation of China under Grant Nos. 62272388, 62172332, and the International Cooperation of Shaanxi (2019KWZ-05, 2020KWZ-013). Chao Feng is the corresponding author.

REFERENCES

[1] Apple. 2020. https://www.apple.com/airpods-pro/. (2020).
[2] John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. 2013. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society* 61, 6 (2013), 366–384.
[3] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. 2020. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine* 3, 1 (2020), 1–9.
[4] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. 2018. Auracle: Detecting eating episodes with an ear-mounted sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–27.

[5] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[6] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2021. Motion-resilient Heart Rate Monitoring with In-ear Microphones. *CoRR* abs/2108.09393 (2021). arXiv:2108.09393 https://arxiv.org/abs/2108.09393

[7] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.

[8] Konstantin Dragomiretskiy and Dominique Zosso. 2013. Variational mode decomposition. *IEEE transactions on signal processing* 62, 3 (2013), 531–544.

[9] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard E Howard. 2021. HeadFi: bringing intelligence to all headphones.. In *MobiCom*. 147–159.

[10] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[11] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: Gait-based User Identification with In-ear Microphones. *arXiv preprint arXiv:2108.12305* (2021).

[12] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice In Ear: Spoofing-Resistant and Passphrase-Independent Body Sound Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[13] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[14] Bill Gillham. 2008. *Developing a questionnaire*. A&C Black.

[15] Valentin Goverdovsky, David Looney, Preben Kidmose, and Danilo P Mandic. 2015. In-ear EEG from viscoelastic generic earpieces: Robust and unobtrusive 24/7 monitoring. *IEEE Sensors Journal* 16, 1 (2015), 271–277.

[16] Polar H10. 2019. https://www.polar.com/us-en/products/accessories/h10_heart_rate_sensor/. (2019).

[17] Wireless Bluetooth Sports Headphones. 2020. https://www.jabra.cn/sports-headphones/. (2020).

[18] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences* 454, 1971 (1998), 903–995.

[19] Huawei. 2020. https://consumer.huawei.com/en/headphones/freebuds-pro/. (2020).

[20] Jian Liu, Chen Wang, Yingying Chen, and Nitesh Saxena. 2017. VibWrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 73–87.

[21] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: employing occlusion effect for in-ear human sensing. *arXiv preprint arXiv:2106.08607* (2021).

[22] Alexis Martin and Jérémie Voix. 2017. In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring. *IEEE Transactions on Biomedical Engineering* 65, 6 (2017), 1256–1263.

[23] Henrik Møller, Dorte Hammershøi, Clemen Boje Jensen, and Michael Friis Sørensen. 1995. Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society* 43, 4 (1995), 203–217.

[24] BIOPAC MP160. 2016. https://www.biopac.com/product/mp150-system-221/. (2016).

[25] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[26] Bruno Paillard, Philippe Mabilleau, Sarto Morissette, and Joël Soumagne. 1992. PERCEVAL: Perceptual evaluation of the quality of audio signals. *Journal of the Audio Engineering Society* 40, 1/2 (1992), 21–31.

[27] Jang-Ho Park, Dae-Geun Jang, Jung Wook Park, and Se-Kyoung Youm. 2015. Wearable sensing of in-ear pressure for heart rate monitoring with a piezoelectric sensor. *Sensors* 15, 9 (2015), 23402–23417.

[28] Chester Pirzanski and Brenda Berge. 2005. Ear canal dynamics: Facts versus perception. *The Hearing Journal* 58, 10 (2005), 50–52.

[29] G.A. Pressler, J.P. Mansfield, H. Pasterkamp, and G.R. Wodicka. 2004. Detection of respiratory sounds at the external ear. *IEEE Transactions on Biomedical Engineering* 51, 12 (2004), 2089–2096. https://doi.org/10.1109/TBME.2004.836525

[30] Daniel M Rasetshwane and Stephen T Neely. 2011. Inverse solution of ear-canal area function from reflectance. *The Journal of the Acoustical Society of America* 130, 6 (2011), 3873–3881.

[31] Grand View Research. 2020. https://www.grandviewresearch.com/press-release/global-earphones-headphones-market/. (2020).

[32] Aditi Roy, Nasir Memon, and Arun Ross. 2017. Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems. *IEEE Transactions on Information Forensics and Security* 12, 9 (2017), 2013–2025.

[33] Ferdinando S Samaria and Andy C Harter. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE workshop on applications of computer vision*. IEEE, 138–142.

[34] Ronald W Schafer and Lawrence R Rabiner. 1973. A digital signal processing approach to interpolation. *Proc. IEEE* 61, 6 (1973), 692–702.

[35] Yang Shao, Zhaozhang Jin, DeLiang Wang, and Soundararajan Srinivasan. 2009. An auditory-based feature for robust speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4625–4628.
[36] Quan-Sen Sun, Sheng-Gen Zeng, Yan Liu, Pheng-Ann Heng, and De-Shen Xia. 2005. A new method of feature fusion and its application in image recognition. *Pattern Recognition* 38, 12 (2005), 2437–2448.
[37] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[38] Susan E Voss and Jont B Allen. 1994. Measurement of acoustic impedance and reflectance in the human ear canal. *The Journal of the Acoustical Society of America* 95, 1 (1994), 372–384.
[39] Lei Wang, Kang Huang, Ke Sun, Wei Wang, Chen Tian, Lei Xie, and Qing Gu. 2018. Unlock with your heart: Heartbeat-based authentication on commercial mobile phones. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 3 (2018), 1–22.
[40] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.
[41] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).
[42] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.
[43] Peter Welch. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics* 15, 2 (1967), 70–73.
[44] Wei Xu, ZhiWen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. Acousticid: gait-based human identification using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.
[45] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.
[46] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can" Hear" Your Heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.