Beyond Legitimacy, Also With Identity: Your Smart Earphones Know Who You Are Quietly

Yongpan Zou, *Member, IEEE*, Haibo Lei, and Kaishun Wu, *Member, IEEE*

Abstract—User authentication and identification on smart devices has great significance in keeping data privacy and recommending personalized services. With the rising popularity of smart earphones recently, they open up a new world for users to enjoy music individually, but also bring about privacy concerns at the same time. Existing few research works propose positive sensing systems that emit and receive inaudible acoustic signals to authenticate users. However, they share shortcomings of intrusiveness to users, high power consumption, and purely focusing on authentication. Instead, in this paper, we propose a passive sensing system called EarID with low-cost customized earphones which attains user authentication and identification at once. It makes use of a embedded microphone to sense body sounds spread out through ear canals and extract 'fingerprints' as a novel biometric feature. With self-designed earphones, we design a deep learning-based real-time data processing pipeline and cope with external interference. Extensive experiments under different real-world settings show that EarID can achieve a rather low false acceptance rate of 3.4% for user authentication and a high $F1$ score of 95.5% for legitimate user identification.

Index Terms—User authentication, earable device, deep learning

1 Introduction

With the popularity of smart devices, privacy concern has become a hot spot in academia. Researchers have paid much attention to the user authentication problem on smart devices, and proposed various schemes to enhance privacy and security on such devices. Among them, biometric authentication is a mainstream scheme which makes use of fingerprint [1], [2], face [3], iris [4] or other biological features to differentiate users. Fingerprinting is currently the most widely-adopted authentication scheme in commercial smart devices, especially smartphones and tablets. However, it requires to equip additional fingerprinting sensors which increase hardware cost and are not suitable to be equipped on tiny devices such as smartwatches, smart glasses and earphones. Moreover, fingerprinting requires a user to consciously participate in the authentication process and is not appropriate for continuous authentication scenarios. In addition to fingerprinting, iris and face recognition have also been used in commercial devices. Nevertheless, they have the similar aforementioned problems of fingerprinting. Furthermore, as face and iris recognition rely on image processing algorithms, the performance of such kind of systems is highly dependent on light conditions, postures, or finger clearness. Within the scope of biometric authentication, researchers have also proposed some novel methods based on other biological features or physiological activities differing from the above, such as breathing [5], dental occlusion [6], heartbeat [7] and ear canal shape [8]. However, the works [5], [6], [7] require users to consciously put devices close to noses, mouths or chests, which degrades the user experience especially in continuous authentication cases. Several previous work [8], [9], [10] are the most similar to ours which makes use of ear canal structure as biometric feature. But they follow a active sensing approach which utilizes a microphone-speaker pair equipped in earphones to transmit and receive pre-designed acoustic signals.

In this work, we pay attention to the user authentication and identification problem with smart earphones, considering their increasing popularity in our daily life. Different from the traditional problem, we not only care about whether a user is legitimate or not, but also intend to know exactly who he/she is if identified as legitimate. That is, we propose a system that conducts user authentication first when a person tries it, and further recognizes his/her identity if he/she is authenticated as a registered one. The motivation of our consideration is due to the characteristics of smart earphones. On one hand, since earphones are a kind of private items, people are not willing to share them with others except for closest friends or family members. Thus, checking the legitimacy of a user is necessary for guaranteeing private property. On the other hand, people sometimes share their earphones with closest friends or families. By identifying who the present user is, smart earphones can switch to corresponding personal settings and provide personalized services. Considering the shortcomings of existing approaches as mentioned above, we propose a truly passive sensing system by making use of low-cost microphone sensors embedded in earphones to collect acoustics caused by putting on earphones (*i.e.*, *behavioural*

• *The authors are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China. E-mail: {yongpan, wu}@szu.edu.cn, leihaibo2019@email.szu.edu.cn.*

Manuscript received 16 June 2021; revised 2 Oct. 2021; accepted 30 Nov. 2021. Date of publication 13 Dec. 2021; date of current version 5 May 2023.
This work was supported in part by China NSFC under Grants 61802264, 62172286, U2001207, and 61872248, Guangdong NSF under Grant 2017A030312008, Shenzhen Science and Technology Foundation under Grants JCYJ20180305124807337, ZDSYS20190902092853047, and R2020A045, the Project of DEGP under Grants 2019KCXTD005 and 2021ZDZX1068, and Guangdong "Pearl River Talent Recruitment Program"under Grant 2019ZT08X603.
(Corresponding author: Kaishun Wu.)
Digital Object Identifier no. 10.1109/TMC.2021.3134654

acoustics) and body sounds spread out through ear canal (*i.e.*, *psychological acoustics*). The underlying rationale of our method are two-fold. For one thing, acoustics during putting on earphones reflect behavioural patterns of different persons, which can be explored to do user authentication similar to other related work [5], [6]. For another thing, the collected body sounds through ear canal are mostly produced by heart beating which is a unique feature for different people [5], [7]. They propagate through body structures including ear canal which act as a series of signal transformation systems. Due to the differences of these structures, their transfer functions are different which exert different impacts on body sound signals.

However, there exist two key challenges in realizing this idea. First, as the power of body sounds is minute, useful signals are likely to be submerged by ambient noises, especially when users are in motion states. It is fairly challenging to extract target signals accurately under different circumstances with conventional segmenting methods. Second, human body sounds can be affected by many factors such as emotions and moving states. It is necessary to build a robust model that can learn intrinsic features from raw signals. As for the first problem, we abandon conventional segmenting and adopt random framing to avoid segmenting failures. As for the second problem, we design a hybrid learning network. In the implementation, since low-layer signals are not accessible on commercial smart earphones, we first design such a hardware system with a low-cost microphone module and a EPS 32 MCU. The sensed data are transferred to a smartphone for further processing via Bluetooth. Note that the key contribution of our work is not proposing novel earphones hardware. Instead, we propose a novel application based on smart earphones that can be implanted on commercial products similar to previous work [8]. But as commercial earphones do not provide access to sensor data, we have to design a hardware prototype to validate the idea of user authentication with earphones. In other words, our proposed system can be easily embedded on present earphones if sensor data are available. We conduct extensive experiments to evaluate the performance of EarID under different environments. The results show that it can achieve satisfactory authentication performance with false acceptance rate and false rejection rate under 3% and 5%, respectively.

The remaining of this paper is organized as follows. In Section 2, we introduce the related work. Section 3 presents the details of EarID design. In Section 4, we introduce the implementation and experiments. Section 5 gives the evaluation of EarID performance. At last, we conclude this paper in Section 6.

2 Related Work

Our work is most related with mobile authentication methods and in-ear sensing applications. In the following, we shall give introduction to these two research areas and demonstrate the correlation between our work and existing works.

2.1 Authentication on Mobile Devices

2.1.1 Biometric Methods

Biometric authentication relies on intrinsic human physiological features to authenticate users [11]. Existing widely-used biometric features on mobile devices mainly include fingerprinting [12], face [13] and iris [4], [14]. Existing fingerprinting systems on mobile devices are mostly designed with capacitive sensors considering the trade-off of size, cost and efficiency. But this method usually fails to work when there exist dirty wastes, water, cuts or bruises on fingers and thus requires users to try over and over again [15]. Moreover, limited by sensor manufacturing, existing fingerprinting systems require additional space for sensor placement such as Home key on smartphones. Face and iris recognition are also emerging authentication methods on mobile devices which rely on image processing techniques to extract biometric features [13]. However, their performance are easily affected by light conditions and user's postures [16]. More importantly, most tiny smart devices such as smartwatches and smart earphones are not equipped with image sensors due to energy consumption and sensor size concern. Moreover, researchers also propose body structure-based authentication scheme such as the works [8], [9], [10], [17]. The works [8], [9], [10] share similar ideas which take advantage of the uniqueness ear canal structure. But they follow a active sensing approach which programs a speaker to emit sounds. By analyzing echoes bounced back by ear canal, those works accomplish user authentication. Compared with them, our work follows a passive sensing approach which only needs a microphone. This not only consumes less energy but also provides a more quiet user experience. Moreover, this work deals with a more difficult problem. That is, it achieves user authentication and identification of legitimate users at the same time.

2.1.2 Behavioral Methods

Behavioral authentication refers to authenticating users by their certain unique behaviors. Existing behavioral methods on mobile devices mainly make use of touching and/or moving gestures on screens [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. Some researchers also come up with gait-based authentication by means of inertial sensors in mobile devices [11], [28], [29], [30]. However, inertial sensor-based gait authentication is easily affected by body movements. Some other methods have also been proposed for user authentication on mobile devices, such as eye movement-based [31], [32], [33] and breathing-based [5], finger tapping/touching [34], [35], walking[36], dental occlusion [6] and heartbeat [7]. But they either require specialized hardware [31], [32], [33], communication signals [37], [38], need a user to participate in consciously [6], [7], [34], [35]. In contrast, our method only uses ubiquitous microphone sensor and works quietly even without a user's notice, since natural behavioural and psychological activities are utilized.

2.2 In-Ear Sensing Applications

With the prosperity of ear-mounted smart devices, researchers have shown increasing interest in developing in-ear sensing applications. Apart from user authentication as mentioned above, these applications can be categorized into vital sign monitoring [39], [40], [41], [42], emotion recognition [43], activity recognition including sleeping [44], [45], eating [46], [47] and movements of other body parts [48], [49], [50]. These applications make use of specialized sensors such as EEG/EMG electrodes, conductive rubber electrodes, microoptic

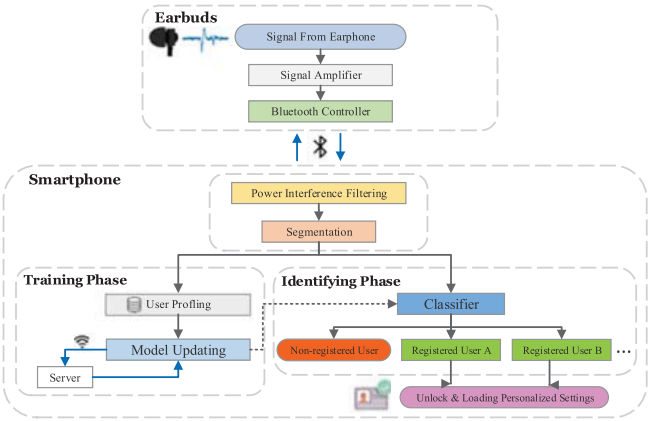

Fig. 1. The system overview of EarID.

reflective sensor and *etc*.. In contrast, our system is built with low-cost and ubiquitous microphone sensor which can be easily deployed in commercial earphones.

3 EarID DESIGN

3.1 System Overview

Fig. 1 shows the overview of EarID's data processing pipeline. EarID system consists of two parts, namely, earphones hardware and a self-designed mobile application on a smartphone. The earphones hardware consists a microphone, speaker, audio amplifier, Bluetooth module and micro controller. As the raw sensor signals are not accessible on commercial earphones, we implement the hardware by ourselves with low-cost sensors and modules which will be detailedly introduced in Section 4.1. When a user starts putting on earphones, the embedded microphone immediately activates and senses data until completes an authentication trial. This data collection process actually contains two stages, one is during putting on earphones, and the other is when earphones are worn in ears. In the former stage, the collected signals by a microphone are mainly caused by user's putting on activity, while signals in the following stage mainly depict activities of viscera such as heart, lung and *etc*.. As a result, we actually make use of both behavioral characteristics and physiological features for user authentication in EarID.

Before transmission, the collected acoustic signals are first sent an amplifier in order to boost SNR. After that, signals are transmitted to the smartphone continuously via an embedded Bluetooth unit. At the smartphone end, the application is responsible for data communication and the whole processing pipeline. Specifically, when the application receives data, it first performs power interference filtering to remove powerline interference and its harmonics. After that, we perform segmentation technique on the cleaned signal sequence to extract signals corresponding to two different stages. This is because we make use of different deep learning models to extract features on two parts of signals which shall be demonstrated in Section 3.3. Following that, we design a hybrid learning network which extracts features and accomplish authentication and identification at one time. In the following, we shall give detailed introduction to each part.

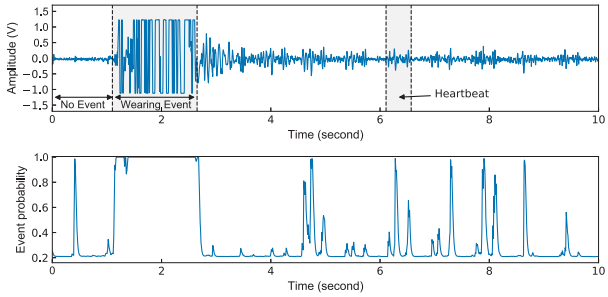

Fig. 2. The deep learning network architecture designed for EarID.

3.2 Signal Preprocessing

3.2.1 Signal Denoising

Due to the interference of power line, the collected signals contain 50 Hz components and corresponding odd harmonics. To reduce them, we first perform FFT on the obtained signal sequences and set the coefficients corresponding to those components to be zeros, and then transform the FFT coefficients back to time-domain signal sequences. The reason why we do not make use of notching filters to is that they consume more computing resources and time, which makes them less appropriate on mobile devices.

3.2.2 Wearing Event Detection

As mentioned above, a whole signal sequence is composed of two parts, namely, signals corresponding to putting on earphones and activities of body viscera respectively. As shown in Fig. 2, signals within [1,2] second indicate the wearing event and the following part are body sounds. As we utilize different models to extract features from these two parts, we first need to segment these two parts. Here we make use of a likelihood ratio test (LRT) and Hidden Markov Model (HMM)-based event detection module [51] to achieve this. This method is widely used in audio processing and has proven to be effective. It consists of the following main steps. Given two hypotheses as follows:

$$H_0 : \text{wearing event absence} : \boldsymbol{X} = \boldsymbol{N}$$
$$H_1 : \text{wearing event presence} : \boldsymbol{X} = \boldsymbol{S} + \boldsymbol{N}$$

where $\boldsymbol{S}$, $\boldsymbol{N}$ and $\boldsymbol{X}$ represent L dimensional discrete Fourier transform (DFT) coefficient vectors of acoustic signals, noises and noisy signals with their kth elements S_k, N_k and X_k respectively. Then the probability density functions conditioned on the above two hypotheses are given by

$$p(\boldsymbol{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} exp\left\{ -\frac{|X_k|^2}{\lambda_N(k)} \right\} \tag{1}$$

$$p(\boldsymbol{X}|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} exp\left\{ -\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)} \right\}, \tag{2}$$

where $\lambda_N(k)$ and $\lambda_S(k)$ represent the variances of N_k and S_k, respectively. Thus, the likelihood ratio for the kth frequency band is

$$\Lambda_k \overset{def}{=} \frac{p(\boldsymbol{X}|H_0)}{p(\boldsymbol{X}|H_1)} = \frac{1}{1+\xi_k} exp\left\{ \frac{\gamma_k \xi_k}{1+\xi_k} \right\}, \tag{3}$$

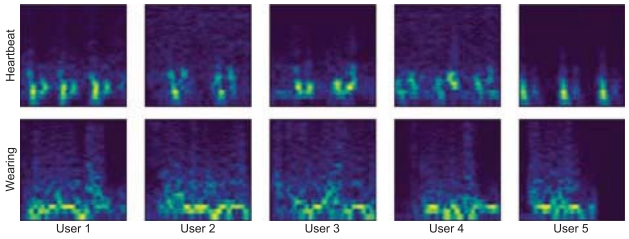

Fig. 3. The signal transformation of five randomly selected participants.

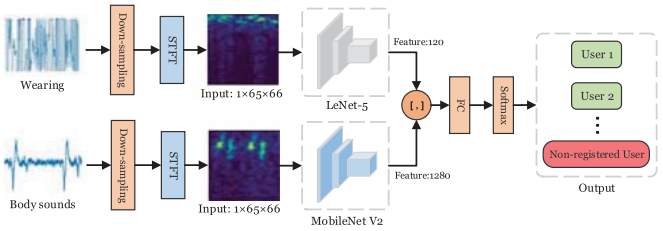

Fig. 4. The deep learning network EarNet architecture designed for EarID.

where $\xi_k \overset{def}{=} \frac{\lambda_S(k)}{\lambda_N(k)}$ and $\gamma_k \overset{def}{=} \frac{|X_k|^2}{\lambda_N(k)}$ representing the priori and posteriori signal-to-noise ratios. Consequently, the decision rule is established from the geometric mean of the likelihood ratios of different individual frequency bands, which is given by

$$log\Lambda = \frac{1}{L}\sum_{k=0}^{L-1} log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{4}$$

where η is an empirically set threshold constant which equals to 0.80 in our implementation of EarID. Applying this method on our signal sequence, we can obtain the corresponding event probability of each frame as shown in the below subfigure of Fig. 2. We can see that except for wearing earphones, there exist some other periods with high event probability which indicate occurrence of other events such as heartbeats. But the durations of those events are much shorter. Thus, we add a constraint of event duration (more than 1.5 second) to filter out other events and extract wearing events precisely.

The following signal frames mainly contain body acoustics which show high uniqueness among different people. Different from conventional operation, we do not perform segmentation technique on this part of signals based on twofold considerations. For one thing, as we can see from Fig. 2, due to the much lower SNR, it is rather difficult to accurately detect each inner-body event, which probably results in many misses in practice especially when the signals are more noisy. This will degrade the system performance or adding a user's overhead by forcing him to try more times. For another thing, although the main components of body acoustics sensed by our earphones are caused by heartbeat, other viscera also contribute useful information which is hidden in lower-power signals. As a result, for this part of signals, we adopt a simple but effective strategy, that is, randomly sampling the signals with a sliding window.

3.2.3 Signal Transformation

After obtaining signals corresponding to user behavioral and psychological features, the following step is to find proper representation of them. A straightforward idea is to extract proper features manually to train a machine learning model as done in previous work [8]. But this is not appropriate for our problem. Different from that work in which the source signals are pre-designed, source signals that we used are originated from a user's natural behavior and psychological activities. This makes it much more different to model how signals are produced and propagated. As a result, it lacks enough domain knowledge to extract proper features. Inspired by image processing, we perform the short-time Fourier transform (STFT) on two parts of signals and obtain features both in time and frequency domains. Specifically, we perform Hanning windowing and sliding STFT on signal sequences with the window size and overlap being 128 and 72 samples points, taking the trade-off between accuracy and latency into account. Fig. 3 shows the results of transforming five randomly selected participants' denoised behavioural (*i.e.*, putting on earphones) and psychological (*i.e.*, inner-organ activities) signals into spectrograms by STFT in the upper and lower rows, respectively. We can clearly see that both kinds of signals, especially the psychological signals, show unique patterns in spectrograms for different users. This indicates that it is feasible to distinguish different persons with acoustic signals collected by earphones.

3.3 Network Design for Classification

After transforming signals to two dimensional spectrograms, we design a deep learning network called EarNet as shown in Fig. 4 for user authentication and identification, which is reduced to a multi-label classification problem. The network consists of two channels, one is for behavioural signals and the other is for psychological signals. As we can see, spectrograms in two channels are fed to a LeNet-5 and MobilNet V2, both of whose fully connected (FC) layers are removed, in order to extract features. Following this, we concatenate the obtained features from two channels,forming a single feature vector, and feed it to a FC layer to finally achieve multi-label classification. The considerations of designing such a hybrid learning network are two fold. First, according to our analysis, the signals caused by putting on earphones are not powerful enough to characterize a use compared with body sounds. Thus, we choose to use a more lightweight network to extract features, for the sake of reducing computation overhead and assigning lighter weight on this channel by outputting a shorter feature vector (*i.e.*, 120 dimensions versus 1280 dimensions). Second, in each sampling session (*i.e.*, putting on earphones→sensing body sounds), sample points collected in the first stage are much less than those in the second stage, and randomly framing is performed merely on the second part of signals. This results in a much smaller training dataset in the first channel. Thus, a more lightweight network is more appropriate. In Section 5.2, we shall evaluate and discuss the performance of different network implementations. However, it is to be noted that in practical usage cases, each time a user puts on earphones, the authentication can be conducted continuously, with randomly selected frames on psychological acoustics combining with the same behavioural acoustic segment.

Fig. 5. The hardware and software of EarID system.

4 Implementation and Experiments

4.1 System Implementation

The EarID system consists of two parts, one is the earphones hardware and the other is a mobile application on a smartphone, tablet or any other smart devices. In the following, we shall give details about these two parts.

4.1.1 Hardware

As present commercial earphones do not have in-ear microphones and output raw data, we design a pair of smart earphones with low-cost electret microphones which can be bought by about 10 CNY online. We connect the microphone with an acoustic signal amplifier MAX9814 which is a low-cost, high-quality microphone amplifier with automatic gain control (AGC) and low-noise microphone bias. It features a low-noise preamplifier, variable gain amplifier (VGA), output amplifier, microphone-bias-voltage generator and AGC control circuitry. The micro-controller unit (MCU) we used is an ESP32-S2, a low-cost and low-power system on a chip with integrated Wi-Fi and dual-mode Bluetooth (BLE and BT). The ESP32 series employ a Tensilica Xtensa LX6 microprocessor in both dual-core and single-core variations and includes built-in antenna switches, RF balun, power amplifier, low-noise receive amplifier, filters, and power-management modules. We connect MAX9814 with ESP32-S2 and program it to collect acoustic signals with a sampling rate of 2000 Hz. The core hardware circuit of EarID is displayed in Fig. 7 which mainly exhibits the microphone sensor and ESP32 module. The ambient light sensor is used as an indicator of hardware states. Meanwhile, the MCU transmits collected data to a self-developed mobile application via onboard Bluetooth in real time. Although there is low energy Bluetooth mode ESP32 board, we do not use this mode due to the unstable data transmission to our smartphone. As it is more a engineering problem, we leave it as one of our system optimization in the future. To implement the prototype design, we also print a plastic earbud to integrate the microphone and speaker sensor as shown in Fig. 5.

4.1.2 Software

We develop a mobile application on Android platform which is responsible for communicating data with earphones via Bluetooth and executing the data processing pipeline. Although we select a relatively lightweight deep learning model, the training process still incurs excessive computing burden for a smartphone. As a result, we shift this part of work to a server since model training does not frequently occur when the system is put into use. Specifically, the training data are uploaded to a cloud server with specifications of 64 GB RAM, NVIDIA TITAN V GPU and Intel(R) Xeon(R) E5-2650 CPU. When the training process is finished, model parameters are transmitted back to the smartphone application for completing model deployment. It is noted that a remote server is needed only for model training which is not supported by Android platform at present and requires overwhelming computational resources. Nevertheless, after this stage, EarID application can be deployed on smartphones and directly put into use without relying on the server, since the forward predication of EarNet does not consume many resources. Even though retraining is sometimes needed in order to update the model, it only involves adjusting model parameters in the last layer and can be accomplished with a server when Internet is available. The present version of EarID application also contains other functions such as user enrollment and user guiding. In our experiments, we mainly make use of a Huawei Mate 9 smartphone with a Hisilicon Kirin 960 CPU, 6 GB RAM, 128 GB ROM and Android 9 operating system.

4.2 Data Collection

We conduct comprehensive experiments with different settings to evaluate EarID's performance. We first recruit a total number of 50 participants with 19 females and 31 males (denoted by $P_1 \sim P_{50}$) from our university aged from 18 to 35 years old, including students, staffs and faculties. Each participant is paid by 60 CNY per hour after experiments. Before starting up, we tell participants about the details of experiments to make sure that they clearly know what they should do during experiments. And we also instruct them to use the EarID system like charging the hardware when batteries run out, installing applications on smartphones and using them for data collection, and *etc*.. On the whole, our experimental settings are determined according to four impact factors, including noise level, user's movement, wearing angle of earphones and emotional state of a user. These influence factors cover different aspects of potential interference in real-world usage scenarios. In the following, we shall demonstrate the details of each experimental setting.

4.2.1 Noise Level

To evaluate the impact of noise, we collect data in common daily usage scenarios including home, workplaces, shopping malls and public transportations. To clarify the impact of different scenarios more clearly, we divide them into five categories according to their noise levels, including a silent workplace with noise less than 40 dB (*i.e.*, N_0), the same workplace with people talking and walking (40 ± 5 dB, N_1), a cafe environment (50 ± 5 dB, N_2), in a subway (70 ± 5 dB, N_3) and on a busy street (80 ± 5 dB, N_4). The noise level is measured by a SMART SENSOR Decibelmeter AS804. In each of the above settings, participants collect data for 2 hours in total at different time intervals. *We set N_0 as the baseline noise level setting*.

4.2.2 Moving State

Further, since a user's moving state affects the contact between earphones and ear canal, it is meaningful to evaluate

Fig. 6. Application scenarios of EarID.

what effect this factor indeed has on the system performance. To do this, we request each participant to collect data at seven daily moving states including *sitting static*, *typing keyboard*, *shaking head*, *jaw movements*, *speaking*, *walking* and *running*, with other impact factors being set to be baseline values at the same time. *The baseline case of this impact factor is set to be sitting static in our work.*

4.2.3 Wearing Angle

Following the above, we also evaluate how the wearing angle of earphones may affect the performance of EarID, considering that a user may change the orientation of earphones in practice. To accomplish this, we ask participants to wear smart earphones at three different angles, *i.e.*, 0°, 30° and 60°. Fig. 6 gives the schematic description of wearing orientations of the earphones. *We set* 0° *as the default wearing angle of smart earphones.*

4.2.4 Emotional State

At last, we evaluate the impact of a user's emotion. The underlying reason is that EarID relies on sensing of body sounds originated from activities of organs which are related with emotions. To clarify the impact, we utilize four public emotion stimulation datasets, namely StimFilm [52], EMDB [53], IADS [54] and SEED [55], which contain videos, audios and pictures widely used for inducing certain kinds of emotions in psychological experiments. we display these emotion stimulating materials to each participant when they are collecting data with EarID. According to the types of stimulated emotions, these materials can be categorized into four kinds including *happiness*, *neutrality*, *sadness*, and *mixed*[1]. To ensure that the materials indeed stimulate corresponding emotions, we also collect the self-report results and filter the sensing data by combing with videos' labels and self reports. *The baseline setting in terms of this factor is when a person is in neutrality.*

4.2.5 Baseline Setting

Since all the above factors affect the performance of EarID, it is impractical and unacceptable to re-train the model (*i.e.*, EarNet) with data collected under each usage scenario. Instead, we train a universal model (*i.e.*, called *baseline model*) with data collected in only one basic setting (*i.e.*, called *baseline setting*) with a specific combination of impact factors, and then evaluate it under various settings. In other words, we only train EarNet once with data collected in the *baseline setting*, and then directly put the system into use in any settings without re-training the model. Our design is based on two-fold considerations. On one hand, practical usage scenarios contain an explosive number of combinations of the above impact factors such as noise level, moving state, and the like. Consider our experimental settings, there are five noise levels, five moving states, four emotional states and three wearing angles. Consequently, there exist a total number of 300 (*i.e.*, $5 \times 5 \times 4 \times 3$) different settings, which indicates that it is rather cumbersome for users to feedback training samples and retrain the model. On the other hand, it is fairer to evaluate the system performance with a universal model instead of a case-by-case model.

As a result, it is necessary to define a *baseline setting* in which the training dataset is collected to train EarNet. In our work, the *baseline setting* is defined based on the consideration of the aforementioned four impact factors, namely, noise level, moving state, wearing angle, and emotional state, according to earphones' most frequent usage conditions in real-world scenarios. Specifically, in terms of the above factors, the *baseline setting* is with *noise level below 40 dB*, *users staying static*, *earphones worn at* 0°, and *keeping a neutral emotional state*. After training EarNet with data collected in such a specific setting, we can test its cross-setting performance in any other settings (including the *baseline setting*).

4.2.6 Specifications of Data

Since we train a *baseline model* and test its performance in all different settings, we request each participant to collect data in the *baseline setting* for a total number of 120 times (*i.e.*, called epochs). Each complete data collection epoch starts from putting earphones into ears and then lasts for a certain period of time, with a total duration of 30 seconds. As a result, in this stage each participant collects a total period of 60 minutes' data at a sampling rate of 2000 Hz in the baseline case. To construct the training dataset, for each signal sequence collected in an epoch, we first extract the segment corresponding to a putting event (see Section 3.2.2) and then randomly frame the following part into 100 pieces with each lasting for 2 seconds. We shall also evaluate the impact of these two parameters in Section 5.5. Different from the training stage, when EarID is put into practical use, it detects putting-on events in real-time and samples physiological signals with a two-second window.

In our problem, suppose m participants are treated as legitimate (*i.e.*, *registered users*) and the remaining ones are impostors (*i.e.*, *non-registered users*), we partition the dataset as follows to train the *baseline model*. For each legitimate user, 80% of his/her data are used for model training and the remaining are for model evaluation. For impostors, we utilize data from n (30 in our final model) of them in the training process, and the data of remaining p impostors for attack tests. The reason for such a data partition operation is to avoid data leakage. In Section 5, we shall give detailed analysis of the impact of m and n. There are several points to be noted about the model training process. First, since we attain authentication and identification at once, each legitimate user represents a individual class while all the impostors are treated as one class. Second, to avoid bias of the classification model, we balance the number of samples

1. The *mixed* emotion means an emotional state that can not be classified into any one of the above.

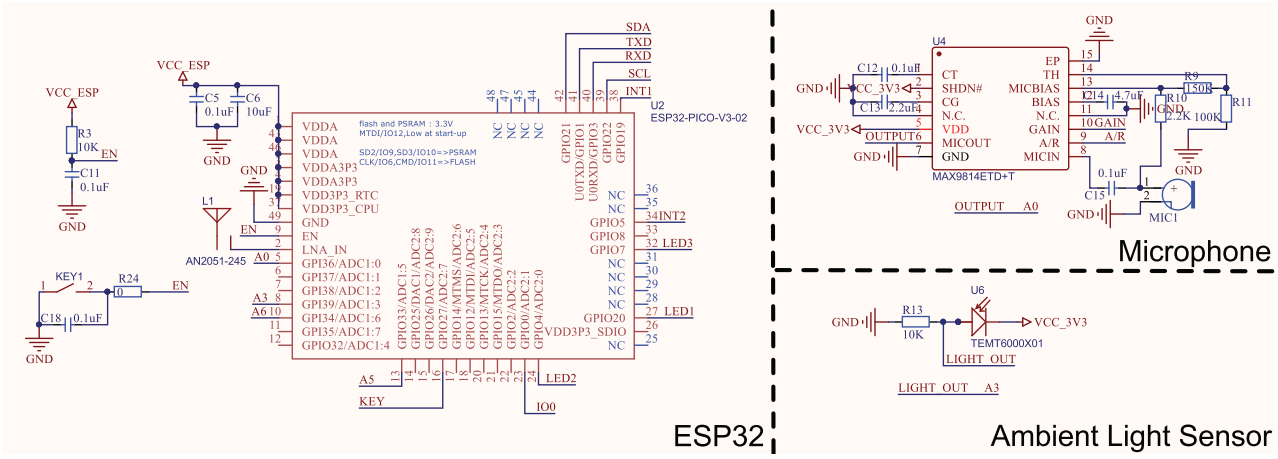

Fig. 7. The core hardware circuit of EarID system.

TABLE 1
The Performance of EarID With Different Network Implementations

Model \ Perf.	Params (Mb)	Training time (s)	FAR ± std (%)	FRR ± std (%)	Precision (%)	Recall (%)	F1 score ± std (%)
MLP	0.47	94.0	43.9 ± 6.8	18.6 ± 2.9	62.7 ± 4.4	67.5 ± 4.2	64.2 ± 4.4
LSTM	0.30	130	10.4 ± 5.6	6.4 ± 1.3	91.1 ± 3.6	92.2 ± 1.8	91.5 ± 2.7
MobileNet	2.23	302	4.5 ± 2.0	7.4 ± 1.3	94.6 ± 1.5	93.1 ± 1.1	93.8 ± 1.2
MobileNet (wearing)	2.23	302	15.7 ± 4.2	27.5 ± 5.2	78.3 ± 5.6	73.1 ± 5.3	75.2 ± 5.4
LSTM (2ch)	0.6	250	12.8 ± 7.5	10.3 ± 4.9	89.4 ± 4.7	89.1 ± 4.3	88.9 ± 4.5
MobileNet (2ch)	2.23	255	6.3 ± 2.4	13.2 ± 2.5	91.9 ± 2.6	88.3 ± 2.6	89.9 ± 2.5
EarNet	**2.98**	**380**	**3.4 ± 1.2**	**4.9 ± 2.1**	**96.0 ± 1.3**	**95.1 ± 1.7**	**95.5 ± 1.5**

from legitimate and invalid users in the training stage. After obtaining the *baseline model*, we test our system under different settings by varying each impact factor as mentioned above including the baseline case. In this stage, we request each legitimate user and impostor to try EarID for 20 times by varying each factor individually in order to evaluate its impact.

5 EarID Performance

In this part, we present the performance evaluation of EarID system. Before displaying the results, we make a brief introduction to the evaluation metrics.

5.1 Evaluation Metrics

As EarID accomplishes user authentication and identification at once, it is necessary to evaluate both aspects with properly and clearly formulated metrics considering our problem. Throughout the evaluation, we mainly adopt five widely used metrics including *false acceptance rate (FAR)*, *false rejection rate (FRR)*, *precision*, *recall* and *F1 score*. As for FAR and FRR, we treat multi-class legitimate users as one group when we calculating them. This is reasonable since these two metrics focus on evaluating the security of a system and do not care about the identity of each legitimate user. On the other hand, when we evaluate how accurately EarID can recognize legitimate users, we regard this task as a multi-label classification problem and utilize commonly used metrics namely precision, recall and F1 score.

5.2 Network Selection

In order to select an appropriate network model, we implement EarID with different models including multilayer perceptron (MLP) [56], LSTM [57], MobileNet [58] and our hybrid network architecture as shown in Fig. 4. In addition, we also evaluate system performance with different model implementation by utilizing signals caused by putting on earphones (denoted by *wearing* for short), or body sounds signals (without additional notation by default), or both of them (denoted by *2ch* for short). In LSTM-based implementations, we set key network parameters including time step, input size, size of hidden layers and number of layers to be 18, 200, 128 and 2, respectively. When both parts of signals are used (*i.e.*, the *2ch* case), we first feed them into the LSTM network individually to extract feature vectors, then concatenate them together, and connect with a FC layer for classification at last. As for MobileNet-based implementations, we need to transform one dimensional time series into two dimensional spectrograms by short-time Fourier transform (STFT) before sending them into the networks.

We evaluate these networks from different aspects besides authentication and recognition performance. Table 1 shows the evaluation results. Comparing with different network models, we can see that our hybrid network (*i.e.*, LeNet and MobileNet) achieves the best performance in terms of FAR, FRR, precision and recall which are up to 3.4%, 4.9%, 96.0% and 95.1%. But it also has the largest parameters' size (*i.e.*, 2.98 Mb) and longest training time (*i.e.*, 380 seconds). Compared with other networks, the model

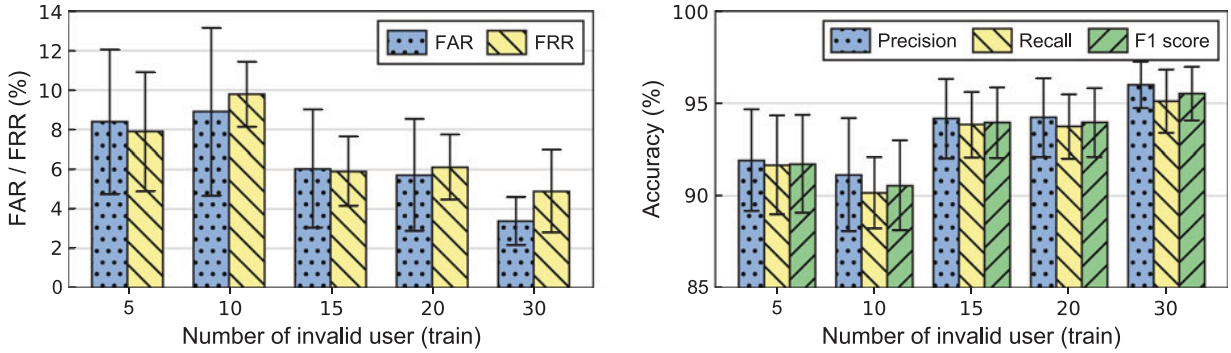

Fig. 8. The performances of EarID with data collected from different number of non-enrolled users in the training stage.

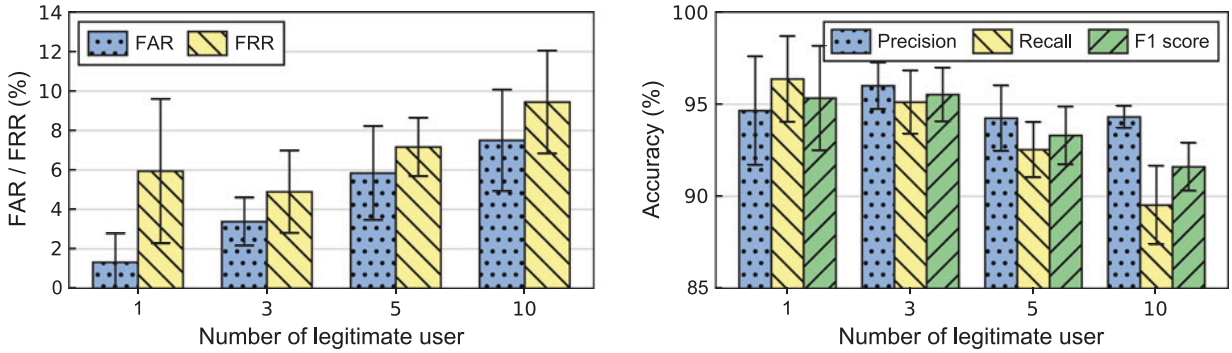

Fig. 9. The performances of EarID with different number of legitimate users.

size and training time are still affordable for present smartphones. On the other hand, from the perspective of utilized data, we can obviously observe that although the first part of signals (*i.e.*, *wearing* signals) indeed provide useful information for user authentication (FAR = 15.7%, FRR = 27.5%), they are not equivalently efficient to the second part of signals (*i.e.*, body sounds) which achieves much lower FAR (4.5%) and FRR (7.4%) with the same network implementation (*i.e.*, MobileNet). This indicates that both parts of signals can contribute to our purpose but with different weights. And this may be a possible explanation to the result that our hybrid network performs better than MobileNet and LSTM, since we use a shallower network to extract features and limit the length of feature vector smaller.

5.3 Number of Non-Registered Users

We consider two cases of the impact of the number of non-enrolled users. In the first case, we vary the number of non-enrolled users whose data are used for model training. This evaluation is helpful for determining how many illegal users are needed in constructing the training dataset. In evaluation, we feed the model with data from 5 to 30 illegal users and train it respectively. In the evaluation phase, we feed the trained models with data from the remaining participants and get the results as shown in Fig. 8. As we can see, the FAR decreases notably from 19.8% to 6.9% with the number of non-enrolled users increasing from 5 to 30. At the same time, the FRR remains almost the same. This is because with data of more illegal users are used, the model can more effectively learn the differences between biometric features of legal and illegal users, which contributes to decreasing the FAR. But since the number of legitimate users is fixed (*i.e.*, to be 3 in evaluation), increasing the number of illegal users can not enhance its ability of recognizing legitimate users. Similarly, the recognition performance also improves with the number of illegal users. Consequently, we set the default number of illegal users to be 30 in our system implementation considering the total number of participants is 50.

5.4 Number of Legitimate Users

We also consider that how the performance of EarID varies when smart earphones are set sharable to different number of close friends or families, *i.e.*, legitimate users. As a result, when we train the models, we utilize data from 1 to 10 participants as positive samples, and data from another 30 ones as negative samples. As we can from Fig. 9, with more legitimate users are considered, the FAR and FRR increase from 5.8% to 16.6%, 2.8% to 5.9% respectively. This is because increasing the number of legitimate users reduces the difference between positive and negative samples, and thus magnifies the difficulty of distinguishing them correctly. At the same time, when more legitimate users are considered, it is more difficult to separate them from each other which causes the precision and recall decrease simultaneously as shown in Fig. 9b. Considering a typical family in China has three members, we set the number of legitimate users when training the model to be 3, under which circumstance the FAR, FRR and F1 score are 7.0%, 3.9% and 94.4%, respectively.

5.5 Impact of Training Data Parameters

We also evaluate the impact of parameters of training data set, including the number of repetitions, number of segments and size of sliding window.

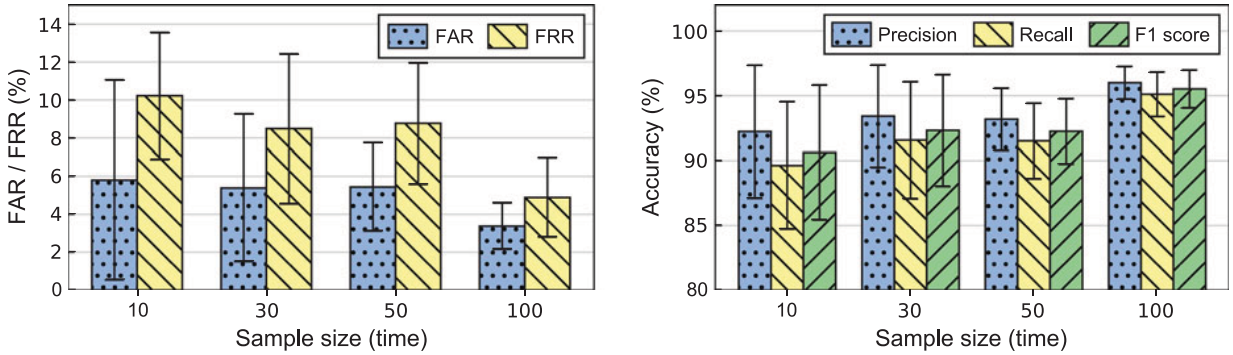

Fig. 10. The performances of EarID with different number of samples.

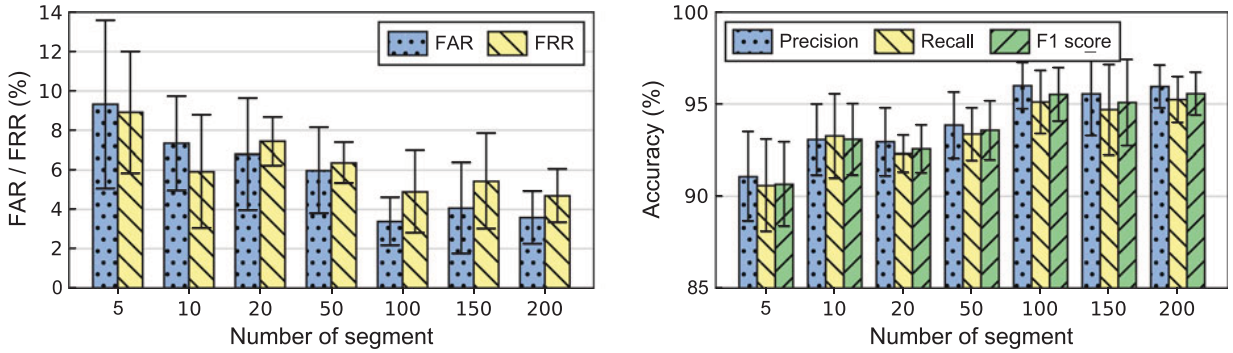

Fig. 11. The performances of EarID with different number of segments.

5.5.1 Impact of Sample Size

In our work, the sample size is defined as the number of repetitions of putting on earphones. During a repetition, each participant puts on smart earphones and collects data for certain time. We evaluate its impact by varying its value from 5 to 100 as shown in Fig. 10. It is clear that the overall performance of EarID improves when the sample size increases, in terms of all evaluation metrics. The reason is more related to the characteristics of deep learning models. That is, with more training data are used, the model can be tunned more optimally. However, there exists a trade-off between user experience and system performance, since collecting too many training samples brings about heavy burden for a user. In our work, we set this parameter to be 100 which can be further improved during the usage of EarID.

5.5.2 Impact of Number of Segments

For each signal episode, we randomly segment it into a certain number of segments. To assure a proper number, we test the performance of EarID by varying this value from 5 to 100 as shown in Fig. 11. With the increasing of number of segments, the system performance improves with FAR and FRR decreasing from 9.3% and 8.9% to 3.4% and 4.9%, respectively. At the same time, the identification performance improves in terms of precision and recall from 91.0% and 90.6% to 96.0% and 95.1%. The reason is straightforward. For a given signal sequence, extracting more segments means providing more training data samples for the deep learning model, which is helpful for tunning it more optimally and offering more powerful generalization capability. But we can also notice that with the number of segments reaches 100, the promotion of system performance becomes much slower. This is because for a fixed-length signal sequence, randomly sampling it with a fixed window size for excessive times produces many overlaps, which does not further contribute data diversity to the model learning and thus can not boost the system notably any more. As a result, considering the model training overhead, we set this value to be 100 in our implementation of EarID.

5.5.3 Impact of Window Size

We also evaluate the impact of window size (*i.e.*, frame length) when sampling the signal sequence. As shown in Fig. 12, when the window size reaches 2 seconds, the system achieves satisfactory performance with FAR, FRR, precision and recall being 3.4%, 4.9%, 95.5% and 95.1%, respectively. Compared with the case of window size being 1 second, the performance has notably improvement in all the evaluation metrics. But incrementing the window size over 2 seconds does not further improve the performance. The reason is very subtle. As mentioned previously, the second part of signals (*i.e.*, body acoustics) mainly caused by heartbeats with about 1 second per cycle. Since we do not perform segmentation but randomly framing it with a window, it probably

5.6 Impact of Noises

Fig. 13 shows the quantitative impact of five different noise levels. As we can see, although the overall system performance decreases with the increasing noise level, when it is below 70 dB, the FAR and FRR stay relatively stable around 8% and 5% respectively. Meanwhile, the user identification performance in terms of precision and recall remain about

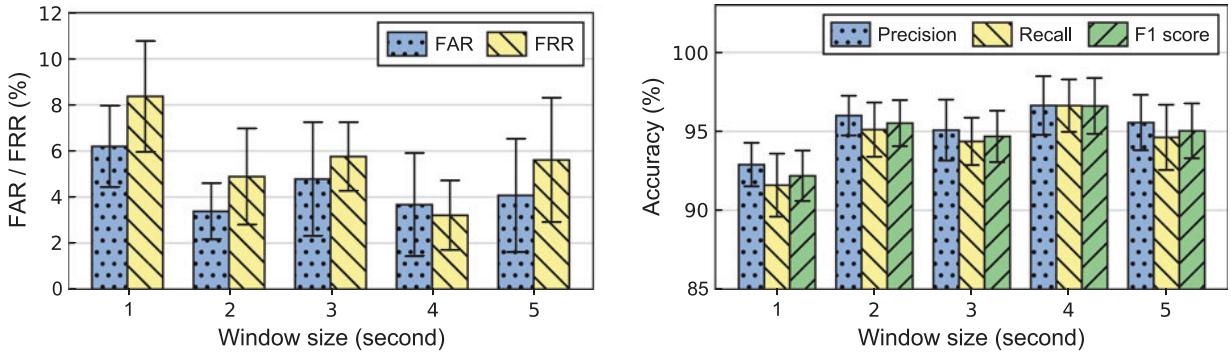

Fig. 12. The performances of EarID with different window sizes.

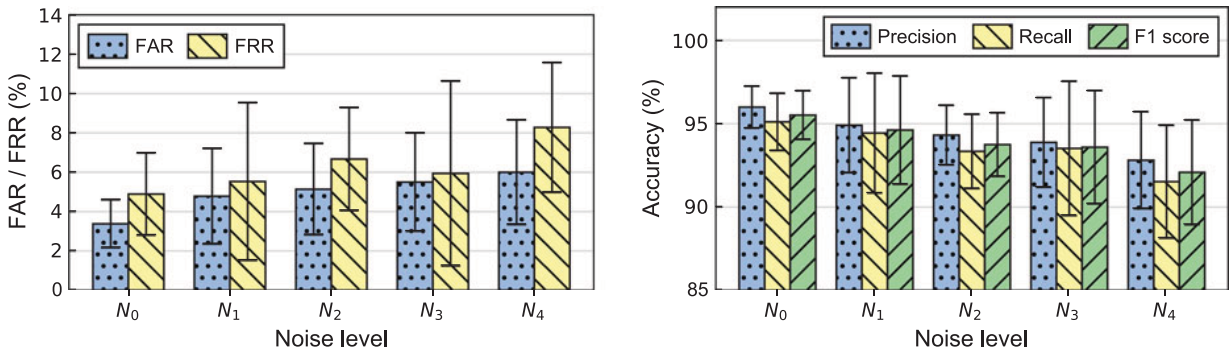

Fig. 13. The performances of EarID under different noise levels.

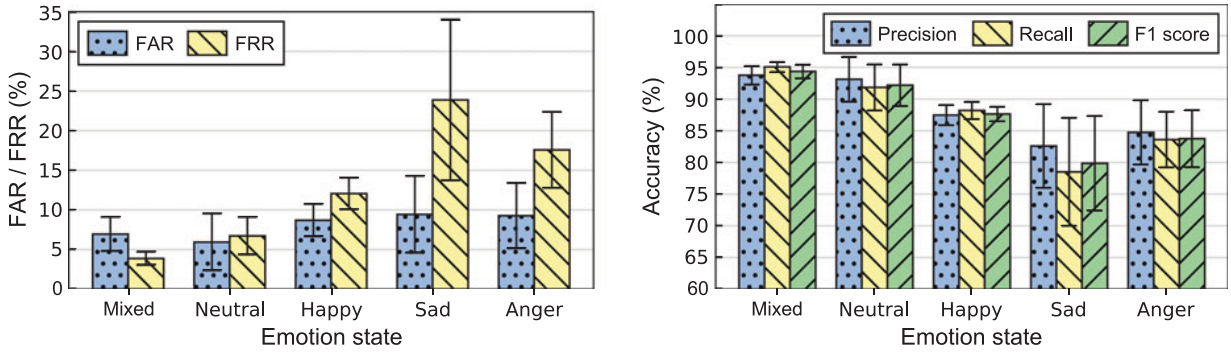

Fig. 14. The performances of EarID with a user in different emotional states.

92.7% and 94.1% respectively. This indicates that EarID can perform well in most daily life scenarios which usually contain noises below 70 dB. We can also notice that when the noise level goes up to 80 dB,the system performance degrades with FAR and FRR increasing to 10.3% and 9.4% respectively.

5.7 Impact of Emotions

Fig. 14 shows the evaluation results of EarID when users are under five different emotion states. From the results, we can obtain several key points as follows. First, emotions show more obvious impact on FRR than FAR especially in sadness and anger, since FAR keeps about 8% in different sates while FRR varies between 3.9% and 17.6%. This is because for different persons, emotions can still induce unique body sounds to distinguish them. But for the same person, emotions bring about changes of body sounds and make it more difficult to identify him/her. Second, except for sadness and anger, the other three kinds of emotions, especially neutrality and mixed show more negligible effect on the system performance. We speculate that this is because the intensities of these three emotions are less than those of sadness and anger. Third, the identification performance shows similar trends as in Fig. 14b.

5.8 Impact of Wearing Angles

Fig. 15 shows the influence of different wearing angles including 0°, 30° and 60° by rotating earphones along Y axis. As we can see, the average FAR and FRR at three cases are 3.4%, 5.4%, 4.0% and 4.9%, 6.7%, 6.2% respectively. The largest gap of FAR and FRR are 2.0% and 1.8% respectively. This gap is just comparable with variances of results. As for identification performance, the precision and recall drop by about 2% at most, from 96.0% to 94.1%, and 95.1% to 93.3% respectively. This is a notable advantage of EarID compared with EarEcho [8] whose recall drops by about 14% when rotating the earphones by 50° along Y axis. The underlying reason is that EarEcho makes use of the spatial structure between earphones and ear canals which is closely related with the wearing angles of earphones. In contrast, our system

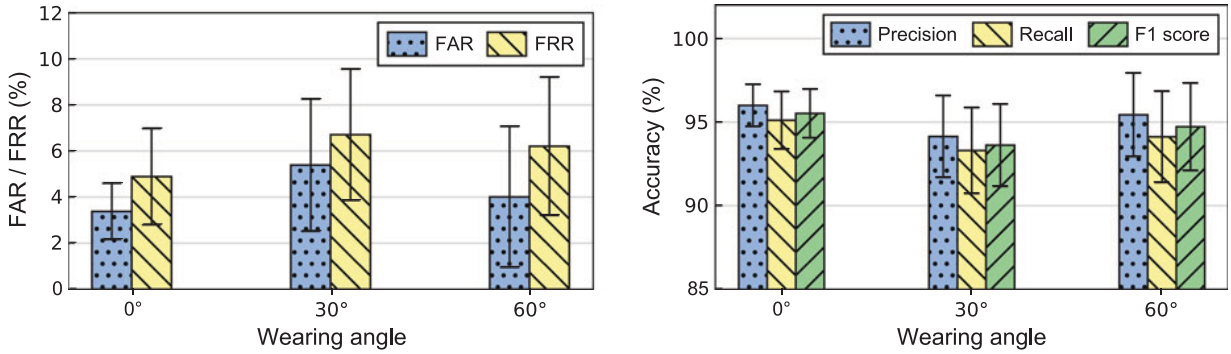

Fig. 15. The performances of EarID with earphones are worn at different angles.

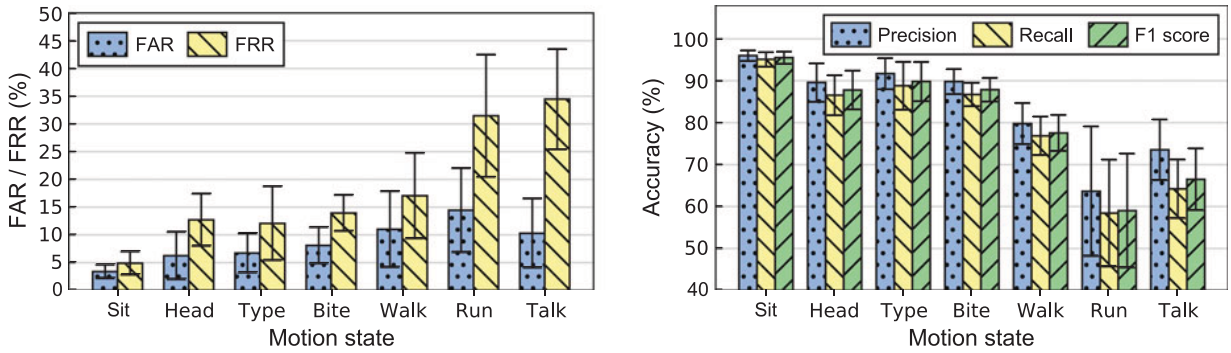

Fig. 16. The performances of EarID with a user in different motion states.

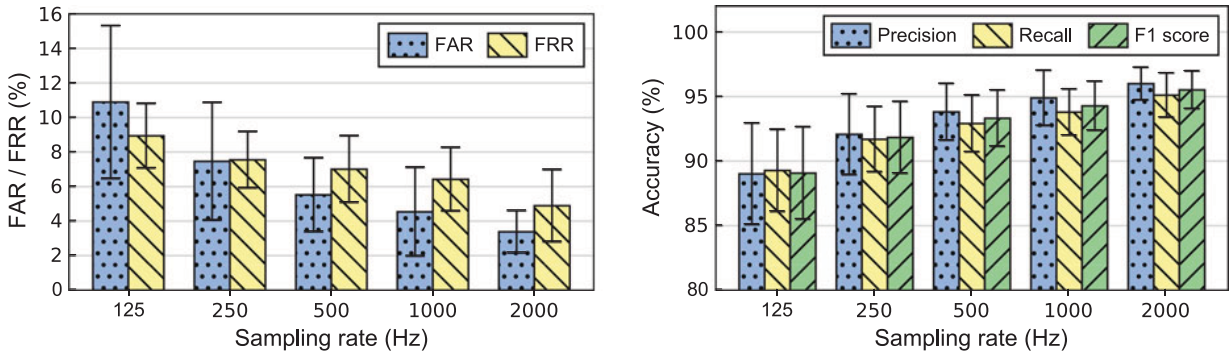

Fig. 17. The performance of EarID at different sampling rates.

relies on a user's behavioral and psychological features which are more intrinsic and robust.

5.9 Impact of Motion States

Fig. 16 shows the system performance when a user is in seven motion states. As we can see, different motions show variant impacts of which running and talking bring about more obvious performance deterioration. Compared with the baseline setting, the FAR and FRR in these two cases increase by about 9.0% and 28.1% in average, respectively. Correspondingly, the precision and recall decrease by about 27.4% and 33.8%, respectively. This indicates that EarID is more sensitive to running and talking activities. This is because these two activities have significant impact on the contact between earphones and ear canals, which makes it unstable for microphones to collect data. In contrast, the other five activities are more moderate and have slight impact on evaluation metrics with less than 8% and 11% decreases in FAR and FRR. More importantly, motion states show more slight impact on FAR than FRR, which means they mainly degrade the user experience of EarID instead of its security.

5.10 Impact of Sampling Rate

To evaluate the impact of different sampling rates, we down-sampling the collected data from 2000 Hz to 1000 Hz, 500 Hz, 250 Hz and 125 Hz, and test EarID's performance. As we can see from Fig. 17, increasing the sampling rate is helpful for boosting system performance. This is because although the biometric information contained in body sounds spans a wide frequency range. Increasing the sampling rate can effectively add the information. However, we can also see that when the sampling rate exceeds 1000 Hz, the performance increase is not obvious, which indicates most useful information distributes below 1000 Hz. Further increasing sampling rate induces more energy consumption. Thus, in our system implementation, we set the sampling rate to be 2000 Hz. Compared with active sensing scheme [8], our sampling rate

TABLE 2
The Real-Time Running Performance of EarID

Smartphone \ Metric	Release	CPU model	Battery (mAh)	Response (ms)	Energy/trial (mAh)	Memory (MB)	CPU cost (%)
XiaoMi 4C	2015	Snapdragon 808	3080	2124 ± 69	0.40	93.8 ± 0.6	71.7 ± 4.2
XiaoMi 8	2018	Snapdragon 845	3400	2031 ± 24	0.39	129.4 ± 3.5	56.0 ± 5.5
HUAWEI Mate 9	2016	Kirin 960	4000	2153 ± 31	0.38	96.2 ± 6.0	42.4 ± 3.8
HUAWEI P20	2018	Kirin 970	3400	2159 ± 33	0.35	104.5 ± 2.3	41.1 ± 2.7

is much lower (*i.e.*, 2000 Hz versus 44100 Hz) which is beneficial for reducing the computing overhead and energy consumption on resource-constrained mobile devices.

5.11 System Running Performance

In this part, we evaluate the running performance of EarID application on mobile phones in terms of CPU and memory occupation, battery consumption, and response time. As it is related with phones' hardware specifications, we conduct evaluation experiments on four kinds of smartphones including Xiaomi 4C, Xiaomi 8, Huawei Mate9, and Huawei P20. Apart from the evaluation results, we also list some basic information of them such as the releasing year on the market, CPU model, and total battery capacity as shown in Table 2.

5.11.1 Computing Resources Occupation

We also evaluate the system running performance of EarID which includes the response time, memory occupation and power consumption on the smartphone end. The response time represents how much time EarID takes to finish an authentication process. To evaluate it, we insert a piece of codes in the application to measure the time duration of an complete authentication process. Meanwhile, we make use of Android Studio Profiler to record the memory and CPU occupation when running EarID. We turn off any other third-party applications such as location services and networking connection, and set the screen lightness to be the lowest level. We conduct measurements with ten randomly selected participants, each of whom tries EarID for twenty times on each smartphone. The average results of response time, memory and CPU occupation are shown in Table 2. We can see that EarID's average response time on different smartphones is very close, with an average value about 2.1 s with a maximum standard deviation of 69 ms on the most out-of-date device (*i.e.*, Xiaomi 4C). It is noted that this response time consists of both the sampling duration (*i.e.*, window size of 2 seconds as discussed in Section 5.5.3) and the forward inference time (*i.e.*, about $31 \sim 159$ milliseconds). It indicates that the predication overhead of EarNet is totally affordable for present commercial smartphones. We can also notice that although increasing the window size is beneficial for accuracy, it makes the authentication process slower. Besides the response time, we also display the memory and CPU occupation in Table 2. We can see that the average memory and CPU cost when running EarID application on Huawei P20 is about 104.5 MB and 41.1%, respectively. With the increasing of CPU capacity, the occupation percentage of CPU cost decreases from 71.7% to 41.1% accordingly.

5.11.2 Energy Consumption

What is more, we also test the power consumption of EarID application on the four smartphones. Except for EarID APP and Bluetooth connection, we turn off all the other running applications and kill unnecessary background programs such as WiFi and cellular networking, location services and *etc*.. The screen lightness is also set to be the lowest level during experiments. We run EarID application for user authentication and identification on each smartphone continuously with an experimenter wears earphones, and record its battery level every five minutes with Android API until it decreases to 50%. The recorded battery levels of different phones during testing are shown in Fig. 18. We can see that the battery level decreases almost linearly with EarID application running on a smartphone. It takes about 2.3 hours, 2.5 hours, 3.3 hours, and 2.8 hours for the four smartphones to decrease by about 50% of battery level. During this process, a total number of 3850, 4360, 5263, and 4857 authentication testing repetitions have been accomplished on Mi 4C, Mi 8, Mate 9, and P20, respectively. As a result, we can further calculate the average energy cost of each testing repetition (*i.e.*, energy per trial) as shown in the sixth column Table 2. However, in real-world usage cases, authentication is performed once in a while instead of continuously. For example, when a user has passed authentication test and begins to listen to music, there is no need to conduct authentication incessantly. Actually, the frequency of authentication trials can be greatly less, which in turn results in much less energy consumption over the same duration.

To gain clearer knowledge of the power consumption level, we also conduct a comparative experiment as follows. We turn off all the applications including EarID and only let Bluetooth connected with hardware but without data transmission in experiments. The recored battery consumption with time is shown by the purple line in Fig. 18. We can see that in this setting the battery level decreases by 9% in three

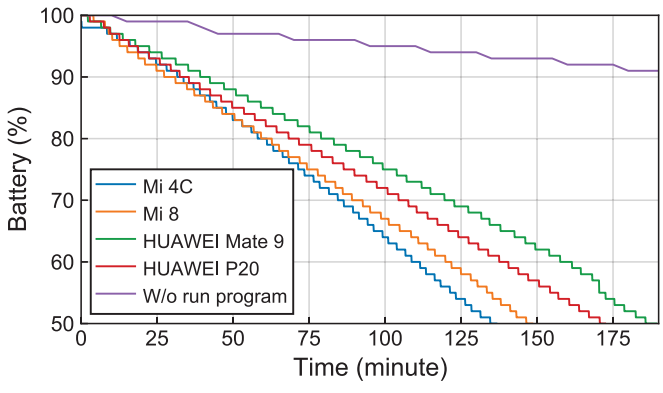

Fig. 18. The energy consumption of EarID application on smartphones.

hours which is 24% smaller than the case of running EarID application. The energy performance of EarID system can be further optimized in future work. For example, we can make use of the latest Bluetooth 5.0 which possesses much higher energy efficiency and fast enough data transmission rate compared with Bluetooth 2.0 that we used in this work.

6 Conclusion

Considering the increasing popularity of smart earphones, the privacy concern has been considered by researchers recently. In this work, we consider a different problem of accomplishing user authentication and identification simultaneously, taking both high privacy and limited shareability of earphones into account. Motivated by this, we propose a passive sensing-based system, namely, EarID that makes use of ubiquitous microphone sensor to collect behavioural and psychological acoustics, and design a hybrid learning network to extract intrinsic features. We design customized wireless smart earphone system with low-cost hardware and develop a mobile application on Android platform to showcase its usability in practice. With comprehensive experiments, we evaluate the performance of our system under different settings, and show that EarID can achieve high performance of authenticating users and identifying legitimate users.

References

[1] A. Kumar and C. Kwong, "Towards contactless, low-cost and accurate 3D fingerprint identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3438–3443.

[2] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2228–2244, Apr. 2011.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[4] J. Daugman, "How iris recognition works," in *The Essential Guide to Image Processing*. New York, NY, USA: Elsevier, pp. 715–739, 2009.

[5] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proc. ACM Mobisys*, 2017, pp. 278–291.

[6] Y. Zou, M. Zhao, Z. Zhou, J. Lin, M. Li, and K. Wu, "Bilock: User authentication via dental occlusion biometrics," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, 2018 pp. 1–20.

[7] L. Wang *et al.*, "Unlock with your heart: Heartbeat-based authentication on commercial mobile phones," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, 2018, pp. 1–22.

[8] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "Earecho: Using ear canal echo for wearable authentication," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 3, 2019, pp. 1–24.

[9] T. Arakawa, T. Koshinaka, S. Yano, H. Irisawa, R. Miyahara, and H. Imaoka, "Fast and accurate personal authentication using ear acoustics," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–4.

[10] S. Mahto, T. Arakawa, and T. Koshinak, "Ear acoustic biometrics using inaudible signals and its application to continuous user authentication," in *Proc. 26th Eur. Signal Process. Conf.*, 2018, pp. 1407–1411.

[11] A. K. Jain and A. Kumar, "Biometric recognition: An overview," in *Second Generation Biometrics: The Ethical, Legal and Social Context*. Berlin, Germany: Springer, pp. 49–79, 2012.

[12] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. Berlin, Germany: Springer, 2009.

[13] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[14] R. P. Wildes, "Iris recognition: An emerging biometric technology," *Proc. IEEE*, vol. 85, no. 9, pp. 1348–1363, Sep. 1997.

[15] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004.

[16] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Microsoft Corporation, Redmond, WA, USA, Tech. Rep. MSR-TR-2010–66, 2010.

[17] Y. Yang, Y. Wang, Y. Chen, and C. Wang, "EchoLock: Towards low-effort mobile user identification leveraging structure-borne echos," in *Proc. 15th ACM ASIA Conf. Comput. Commun. Secur.*, 2020, pp. 772–783.

[18] M. Shahzad, A. X. Liu, and A. Samuel, "Secure unlocking of mobile touch screen devices by simple gestures: You can see it but you can not do it," in *Proc. ACM Mobicom*, 2013, pp. 39–50.

[19] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann, "Touch me once and I know it's you!: Implicit authentication based on touch screen patterns," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 987–996.

[20] C. Bo, L. Zhang, T. Jung, J. Han, X.-Y. Li, and Y. Wang, "Continuous user identification via touch and movement behavioral biometrics," in *Proc. IEEE 33rd Int. Perform. Comput. Commun. Conf.*, 2014, pp. 1–8.

[21] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are how you touch: User verification on smartphones via tapping behaviors," in *Proc. IEEE Int. Conf. Netw. Protocols*, 2014, pp. 221–232.

[22] L. Yang *et al.*, "Unlocking smart phone through handwaving biometrics," *IEEE Trans. Mobile Comput.*, vol. 14, no. 5, pp. 1044–1055, May 2015.

[23] C. Liu, G. D. Clark, and J. Lindqvist, "Guessing attacks on user-generated gesture passwords," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 1, pp. 1–24, 2017.

[24] J. Liu, C. Wang, Y. Chen, and N. Saxena, "VibWrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 73–87.

[25] Y. Yang and J. Sun, "Energy-efficient w-layer for behavior-based implicit authentication on mobile devices," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.

[26] C. Liu, G. D. Clark, and J. Lindqvist, "Where usability and security go hand-in-hand: Robust gesture-based authentication for mobile systems," in *Proc. ACM Conf. Hum. Factors Comput. Syst.*, 2017, pp. 374–386.

[27] Y. Chen, J. Sun, R. Zhang, and Y. Zhang, "Your song your way: Rhythm-based two-factor authentication for multi-touch mobile devices," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 2686–2694.

[28] D. Gafurov, E. Snekkenes, and P. Bours, "Gait authentication and identification using wearable accelerometer sensor," in *Proc. IEEE Workshop Autom. Identification Adv. Technol.*, 2007, pp. 220–225.

[29] M. O. Derawi, "Accelerometer-based gait analysis, a survey," in *Proc. Norwegian Inf. Secur. Conf.*, 2010, pp. 1–12.

[30] A. Primo, V. V. Phoha, R. Kumar, and A. Serwadda, "Context-aware active authentication using smartphone accelerometer measurements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 98–105.

[31] P. Kasprowski and J. Ober, "Eye movements in biometrics," in *Proc. Int. Workshop Biometric Authentication*, 2004, pp. 248–258.

[32] R. Bednarik, T. Kinnunen, A. Mihaila, and P. Fränti, "Eye-movements as a biometric," in *Proc. Conf. Image Anal.*, 2005, pp. 780–789.

[33] C. Song, A. Wang, K. Ren, and W. Xu, "EyeVeri: A secure and usable approach for smartphone user authentication," in *Proc. IEEE 35th Annu. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[34] W. Chen *et al.*, "Taprint: Secure text input for commodity smart wristbands," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–16.

[35] X. Xu *et al.*, "TouchPass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, 2020, pp. 1–13.

[36] Z. Luo, W. Wang, Q. Huang, T. Jiang, and Q. Zhang, "Securing IoT devices by exploiting backscatter propagation signatures," *IEEE Trans. Mobile Comput.*, early access, Jun. 02, 2021, doi: 10.1109/TMC.2021.3084754.

[37] W. Wang, S. He, L. Sun, T. Jiang, and Q. Zhang, "Cross-technology communications for heterogeneous IoT devices through artificial doppler shifts," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 796–806, Feb. 2018.

[38] X. Xiao, W. Wang, T. Chen, Y. Cao, T. Jiang, and Q. Zhang, "Sensor-augmented neural adaptive bitrate video streaming on uavs," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1567–1576, Jun. 2019.

[39] A. Martin and J. Voix, "In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1256–1263, Jun. 2017.

[40] S. Vogel, M. Hülsbusch, T. Hennig, V. Blazek, and S. Leonhardt, "In-ear vital signs monitoring using a novel microoptic reflective sensor," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 882–889, Nov. 2009.
[41] N. Bui *et al.*, "eBP: A wearable system for frequent and comfortable blood pressure monitoring from user's ear," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–17.
[42] J.-H. Park, D.-G. Jang, J. W. Park, and S.-K. Youm, "Wearable sensing of in-ear pressure for heart rate monitoring with a piezoelectric sensor," *Sensors*, vol. 15, no. 9, pp. 23 402–23 417, 2015.
[43] C. Athavipach, S. Pan-Ngum, and P. Israsena, "A wearable in-ear eeg device for emotion monitoring," *Sensors*, vol. 19, no. 18, 2019, Art. no. 4014.
[44] D. Looney, V. Goverdovsky, I. Rosenzweig, M. J. Morrell, and D. P. Mandic, "Wearable in-ear encephalography sensor for monitoring sleep. Preliminary observations from nap studies," *Ann. Amer. Thoracic Soc.*, vol. 13, no. 12, pp. 2229–2233, 2016.
[45] A. Nguyen, R. Alqurashi, Z. Raghebi, F. Banaei-Kashani, A. C. Halbower, and T. Vu, "A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring," in *Proc. 14th ACM Conf. Embedded Netw. Sensor Syst. CD-ROM*, 2016, pp. 230–244.
[46] C. Min, A. Mathur, and F. Kawsar, "Audio-kinetic model for automatic dietary monitoring with earable devices," in *Proc. 16th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2018, pp. 517–517.
[47] S. Bi *et al.*, "Auracle: Detecting eating episodes with an ear-mounted sensor," in *Proc. ACM Interactive Mobile, Wearable Ubiquitous Technol.*, 2018, pp. 1–27.
[48] H. Manabe, M. Fukumoto, and T. Yagi, "Conductive rubber electrodes for earphone-based eye gesture input interface," *Pers. Ubiquitous Comput.*, vol. 19, no. 1, pp. 143–154, 2015.
[49] A. Bedri, D. Byrd, P. Presti, H. Sahni, Z. Gue, and T. Starner, "Stick it in your ear: Building an in-ear jaw movement sensor," in *Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. 2015 ACM Int. Symp. Wearable Comput.*, 2015, pp. 1333–1338.
[50] T. Amesaka, H. Watanabe, and M. Sugimoto, "Facial expression recognition using ear canal transfer function," in *Proc. 23rd Int. Symp. Wearable Comput.*, 2019, pp. 1–9.
[51] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[52] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cogn. Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.
[53] S. Carvalho, J. Leite, S. Galdo- Álvarez, and O. F. Gonçalves, "The emotional movie database (emdb): A self-report and psychophysiological study," *Appl. Psychophysiology Biofeedback*, vol. 37, no. 4, pp. 279–294, 2012.
[54] M. M. Bradley and P. J. Lang, "The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual," Univ. of Florida, Gainesville, FL, USA, Tech. Rep. B-3, 2007.
[55] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng.*, 2013, pp. 81–84.
[56] C. M. Bishop *et al.*, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.
[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[58] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

Yongpan Zou (Member, IEEE) received the BEng degree in chemical machinery from Xi'an Jiaotong University, China, and the PhD degree from the CSE Department, Hong Kong University of Science and Technology, Hong Kong, in 2013 and 2017 respectively. He is currently an assistant professor with the College of Computer Science and Software Engineering, Shenzhen University, China, since September 2017. His research interests include wearable/mobile/ubiquitous computing and HCI.

Haibo Lei is currently working toward the postgraduate degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interest include ubiquitous sensing, mobile computing, and Internet of Things (IoT).

Kaishun Wu (Member, IEEE) received the BEng degree from Sun Yat-sen University, China, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2007 and 2011, respectively. He is currently a distinguished professor with the College of Computer Science and Software Engineering, Shenzhen University, China. His research interests include wireless communications and mobile computing. He won several best paper awards of international conferences such as IEEE Globecom 2012, IEEE MASS 2014.