EarPrint: Earphone-Based Implicit User Authentication With Behavioral and Physiological Acoustics

Yongpan Zou, *Member, IEEE*, Jianhao Weng, Haibo Lei, Danyang Wang, Victor C. M. Leung, *Life Fellow, IEEE*, and Kaishun Wu, *Fellow, IEEE*

***Abstract*—With the increasing pervasiveness of smart earphones, it is appealing to propose more unobtrusive and convenient wearable authentication methods. Researchers have designed earphone-based authentication systems which utilize high-frequency audio signals to scan ear canal structure. Nevertheless, they possess shortcomings of low unobtrusiveness and robustness. In this article, we put forward an earphone-based passive authentication system which makes use of physiological and behavioral acoustic signals caused by a user's natural actions, including putting on earphones and inner organs' activities, respectively. By introducing attention mechanism into the network design, our method adaptively weighs two channel signals, and extracts stable fingerprints for different people, which relieves model retraining for unseen users and improves its scalability. We have built a real-time prototype called EarPrint by designing the earphones and a mobile application, and conducted comprehensive experiments under diverse settings. Experimental results demonstrate that EarPrint has low false acceptance rate (FAR) and equal error rate (EER) less than 1% and 5% in most cases, respectively.**

***Index Terms*—Deep metric learning (DML), smart earphones, user authentication.**

I. Introduction

With privacy protection becoming a significant issue, researchers have spared great effort to the user authentication problem on commercial smart devices. Biometric features, such as fingerprint [1], [2], face [3], and iris [4] have been widely used by commercial authentication systems. But they all require users to consciously participate in the authentication processes, which makes them not suitable for continuous authentication scenarios. Besides, some other biological features or physiological activities, such as breathing [5], heartbeat [6], dental occlusion [7], and ear canal shape [8], [9], [10] have been also utilized in previous works. Nevertheless, with systems proposed in the works [5], [6], [7], users have to place devices close to their noses, mouths or chests, which degrades their user experience vastly. The related works [8], [9], [10] make use of a pair of speaker and microphone to emit and receive modulated acoustic signals to scan ear canals, and extract their structure characteristics to differentiate different users. But this active sensing approach possesses the following shortcomings. For one thing, although high-frequency acoustic signals are used, they are not definitely inaudible for any users due to the audible sounds leakage phenomenon [11] and differences in hearing sensitivities among people with different ages and genders [12]. As a result, those active sensing systems may cause uncomfortable intrusion to users. For another thing, this approach is rather dependent on the closed cavity formed by earphones and ear canals, which means that it is highly sensitive to the contact between both of them.

In this article, we adopt a different technical route from the above works [8], [9], [10] and propose an earphone-based passive authentication system called EarPrint with only microphone sensors. By figuring out the identity of a user, smart earphones can switch to his/her personal settings and provide personalized services. Different from existing related works, EarPrint makes use of microphones embedded in earphones to sense users' apparent *behavioral* (i.e., putting on and/or adjusting earphones) and invisible *physiological* activities (i.e., heartbeats, breathing, etc.), both of which produce distinct acoustic signals among different persons. The underlying rationale of our method is two-fold. On the one hand, *behavioral acoustics* induced by putting-on events reflect one's behavioral pattern which can be potentially used for authentication purpose similar to touching screens [13], [14], [15], dental occlusion [7], etc. in previous works. On the other hand, *physiological acoustics*, produced by inner organs and propagating through body structures, uncover different characteristics of

Manuscript received 16 October 2023; revised 27 December 2023 and 9 May 2024; accepted 12 June 2024. Date of publication 20 June 2024; date of current version 25 September 2024. This work was supported in part by the China NSFC under Grant 62172286 and Grant U2001207; in part by the Guangdong NSF under Grant 2022A1515011509; in part by the Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007; in part by the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603; in part by the Guangdong "Pearl River Talent Plan" under Grant 2019JC01X235; in part by the Project of DEGP under Grant 2023KCXTD042; and in part by the Tencent "Rhinoceros Birds"—Scientific Research Fund for Young Researchers of Shenzhen University. *(Corresponding author: Kaishun Wu.)*

Yongpan Zou, Jianhao Weng, Haibo Lei, and Danyang Wang are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, Guangdong, China (e-mail: yongpan@szu.edu.cn; wengjianhao2021@email.szu.edu.cn; leihaibo2019@email.szu.edu.cn; Yolanda_swam@163.com).

Victor C. M. Leung is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518060, Guangdong, China (e-mail: vleung@ieee.org).

Kaishun Wu is with the Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511453, China (e-mail: wuks@hkust-gz.edu.cn).

Digital Object Identifier 10.1109/JIOT.2024.3417622

people's physiological activities and structures, which have the potential to be a novel biometric feature as well. We find that these two channels can supplement each other and enhance the authentication performance. Someone may doubt about the intrusiveness caused by putting on earphones especially in continuous authentication cases. In fact, when the system is used in real life, we only collect and store *behavioral acoustics* at the moment when a user puts on earphones. But *physiological acoustics* can be continuously sampled and combined with the latest behavioral segments to form complete samples. As a result, there is no need to put on and take off earphones repeatedly in continuous authentication.

However, there are three key challenges to deal with in order to design EarPrint. First, even though we know that both kinds of signals supplement each other, the effective information derived from these signals often differs due to uncontrollable external interference and irregular user movements. Therefore, quantifying the influence of each channel is a challenging task. Second, as *physiological acoustics* are similar to body ambient sounds, their intensities are rather minute without obvious inducing events. Hence, it is rather difficult to extract physiological acoustics accurately with traditional signal detection methods as most works do. Third, to guarantee the unobtrusiveness and scalability of system, it requires to build a lightweight universal authentication model that works effectively in various cases without retraining.

To deal with the above challenges, we introduce channel and spatial attention mechanism in our model design, in order to adaptively weigh two channel acoustics in a sample level. Moreover, we give up conventional events detection-based segmenting but take advantage of random framing for *physiological acoustics* to avoid detection failures. Lastly, we follow a feature matching approach that relies on extracting embeddings of users' data, instead of training a classifier. By adopting this approach, there is no requirement to retrain the model each time an unseen user registers or authenticates. As for the implementation of EarPrint, since commercial earphones do not provide access to low-level signals, we design a smart earphone system with low-cost microphone sensors and an EPS32 board as the micro-controller (MCU). The collected acoustic signals are transmitted through Bluetooth to a smartphone for further data processing. We have conducted comprehensive experiments to evaluate EarPrint's performance in various settings. The experimental results demonstrate that EarPrint achieves excellent performance with equal error rate (EER), false acceptance rate (FAR) and false rejection rate (FRR) less than 5%, 1%, and 11% in common cases, respectively.

In a nutshell, the main contributions of our work can be summarized as follows.

1) We have proved the feasibility of making use of body acoustics for passive user authentication, and designed attention-based deep metric learning (DML) network for retraining-free and multiuser authentication. Compared with existing works, our method outperforms in zero retraining cost, high scalability, and good performance.
2) We have built a real-time system called EarPrint which includes a pair of low-cost smart earphones and an Android application. We have also conducted comprehensive experiments in diverse settings to evaluate its performance. The results show that EarPrint can achieve good authentication performance in practical settings and is likely to be a plug-in application on commercial earphones.

The remaining of this article is organized as follows. In Section II, we introduce the related work. Section IV presents the details of EarPrint design. In Section V, we introduce the implementation and experiments. Section VI gives the evaluation of EarPrint performance. At last, we conclude this article in Section VIII.

II. Related Work

In this section, we shall introduce wearable authentication methods and in-ear sensing applications in a detail which are closely related with our work.

A. Wearable Authentication

Due to the constraints of hardware, such as limited size and low-cost sensors, traditional biometric authentication methods, such as fingerprinting [16], face [3], and iris [4], [17] recognition are no longer suitable for wearable devices. To handle the user authentication problem on wearables, researchers have explored different methods based on heartbeat [18], [19], gait [20], dental occlusion [7], breathing [5], [21], structure of ear canals [8], [9], [10], ear electroencephalogram [22], [23], inertial sensors [24], [25], [26], etc. Among them, the works [22], [23] make use of specialized EEG sensors which are not commonly equipped in commercial devices. The works [5], [7] require a user to intentionally put a device close to his/her mouth, which violates the implicit authentication. The work [24] requires users to consciously emit audible sounds, making it unsuitable for use in public settings. The works [25], [26] require model retraining for new users and do not support multiuser authentication.

The works [8], [9], [10], [19], [20], [21], [27] are most related to ours which make use of commercial earphones for authentication. Nevertheless, EarGait [20] depends on walking gait and thus can not be used when a user is static. In contrast, our system makes use of nonstopping human physiological and implicit behavioral activities. EarEcho [8] and other two similar works follow an active sensing approach which emits high-frequency acoustics to scan the structure of ear canals. They can cause intrusion to users and make them feel uncomfortable. HeartPrint [19] is similar to our previous work [27], as both of them utilize classification models to identify different users via in-ear acoustic sensing. But they require to collect samples from unseen users and retrain the models, which increases users' overhead. In contrast, EarPrint makes use of DML and eliminates model retraining. What is more, EarPrint supports multiuser authentication. Breathsign [21] is similar to our work which makes use of bone-conducted breathing sounds as a novel biometric characteristic. However, a breathing cycle lasts longer than a heartbeat. And Breathsign [21] takes four breathing cycles to

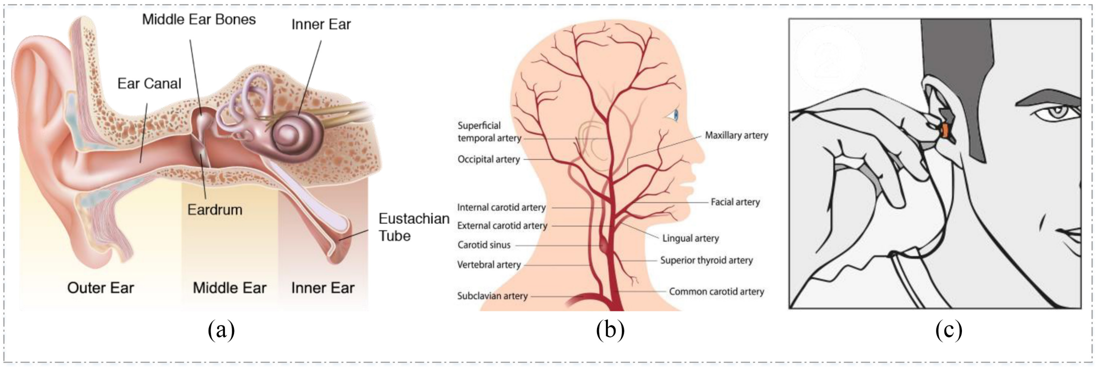


Fig. 1. Demonstration of feasibility of EarPrint. (a) Human auditory system. (b) Head blood supply. (c) Putting on earbuds activity.

achieve a high accuracy, which makes its delay of authentication process longer than our work. We shall show the detailed comparison with existing earable authentication methods in Section VI-I.

B. In-Ear Sensing Applications

In-ear sensing applications have attracted increasing research interest due to the prosperity of earable devices. In addition to user authentication, these applications cover a wide areas, including vital sign monitoring [28], [29], [30], [31], emotion recognition [32], activity recognition, such as sleeping [33], [34], eating [35], [36], and movements of other body parts [37], [38], [39]. They mostly rely on some specialized sensors, such as microoptic reflective sensor, conductive rubber electrodes, and EEG/EMG electrodes. In comparison, EarPrint is implemented with ubiquitous and low-cost microphone sensors, which makes it easily deployed in commercial earphones.

III. Feasibility Study

To lay a solid foundation of our work, we first validate the feasibility by means of theoretic analysis and data measurement in this section.

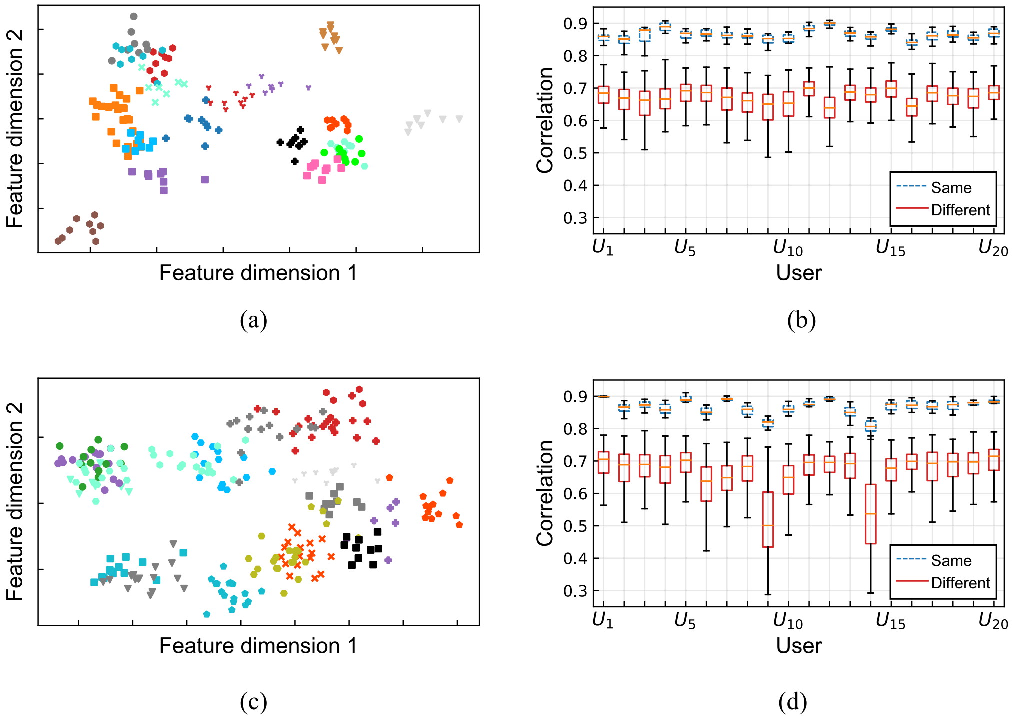
A. Theoretic Foundation

Since EarPrint depends on both physiological and behavioral features, it is necessary to figure out the causes of these two kinds of acoustic signals in order to reveal the underlying rationale of our work. On the one hand, as shown in Fig. 1(a) and (b), a person's carotid artery and jugular vein are in rather close proximity to his/her ear drum (also known as tympanic membrane) which is a great conductor of sound. As a result, when blood flows through vessels, the induced periodic vibrations can be captured by the ear drum and propagated to outer ear thorough the ear canal. What is more, when an in-ear microphone is placed in one's ear canal, the radiation impedance at the entrance of the outer ear increases, which results in an increase in the level of sounds generated in the auditory meatus by bone conduction, especially at low frequencies. As a result, the low-frequency components of bone-propagated sounds will be boosted due to the loss of outer ear sound pathways whenever the ear canal orifice is occluded. This is the so-called *occlusion effect* [40]. This finally increases the signal-to-noise ratio (SNR) of microphone measurements. As is well known, blood flows are originated from heartbeats, and controlled by periodical opening and closure of valves located between the atria and ventricles, and between the ventricles and major arteries. Since heartbeat has been proved to contain an individual's physiology and provide a unique identity for each person, we conjecture that physiological acoustic signals captured by a microphone in the ear canal also have the potential to be utilized for user authentication.

On the other hand, human behaviors are proved to exhibit certain degree of uniqueness and have been utilized to distinguish persons [41], [42]. Moreover, a previous work [43] demonstrates that the activity of picking a smartphone up shows distinctness of different people and can be used for authentication. These motivate us to explore the possibility of making use of acoustic signals caused by rubbing between an earplug and ear canal cavity during putting on earphones. Our considerations are two-fold. For one thing, combining different kinds of signals is helpful for boosting the system performance. For another thing, considering the material of most earplugs is silica gel, the rubbing acoustic signals mainly depend on how the earplugs contact with one's ear canal.

B. Data Measurement

To validate the theoretical feasibility as demonstrated above, we conduct quantitative measurements of the uniqueness of both physiological and behavioral acoustic signals. Specifically, we collect both kinds of signals of 20 participants (labelled as $U_1 \sim U_{20}$, 8 males and 12 females, aged 20–43) with the hardware and mobile application developed in Section V-A, and following the same experimental routine as described in Section V-B. Note that the data are collected only in the *baseline setting* as described in Section V-B4, namely, with noise level lower than 40 dB, participants keeping static, and earphones worn at 0°. Each participant collects data for 120 times in total, each of which lasts for 30 s starting from putting earphones into ears. After that, we perform simple signal processing operations, including denoising and transformation, and then feed the obtained samples into a VGG19 network pretrained on the ImageNet data set which



(a) (b)

(c) (d)

Fig. 2. Embeddings and cross-correlations of physiological and behavioral signals collected in the feasibility experiments. (a) and (c) Different colors and markers represent different users. (b) and (c) Same and different represent intra and inter correlation, respectively.

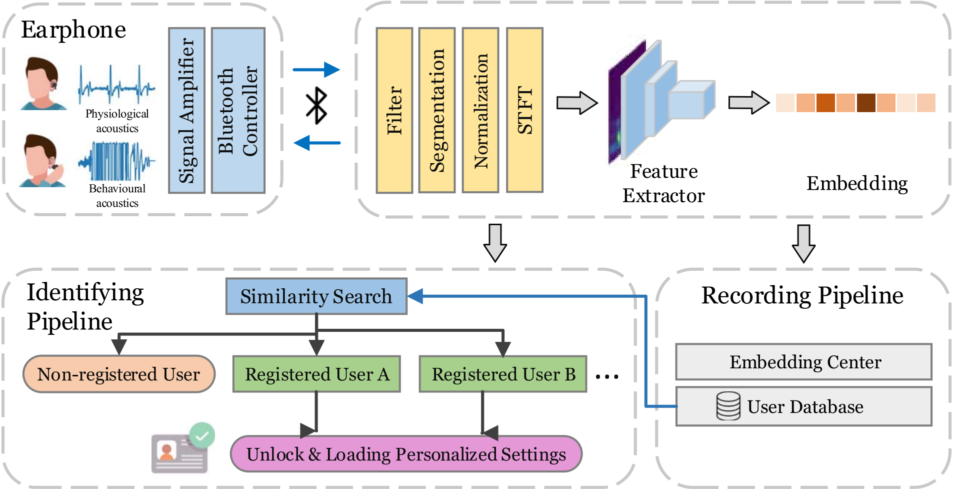


Fig. 3. System overview of EarPrint.

outputs 512-D embeddings correspondingly. To display the embeddings, we further apply linear discriminant analysis (LDA) in order to reduce the dimensions from 512 to 2. As shown in Fig. 2(a) and (c), physiological and behavioral embeddings form different clusters for different users, which indicates that both kinds of signals contain unique information of person identity. We also calculate the intra and inter correlations of the 20 participants' physiological and behavioral embeddings. Specifically, as for the intra correlations, we first calculate the average embedding of each participant, and then calculate its correlations with other embeddings. As for the inter correlations, we calculate the correlations between each participant's average embedding and all the other participants' embeddings. As we can see from Fig. 2(b) and (d), the average intra correlations of physiological and behavioral embeddings are 0.89 and 0.87, respectively. And the average inter correlations of physiological and behavioral embeddings are 0.70 and 0.69, respectively. The results imply that it is feasible to make use of both kinds of signals for user authentication.

IV. EarPrint Design

A. System Overview

Fig. 3 gives an overview of EarPrint's architecture. As we can see, the whole system is composed of two components, including earphone hardware and a self-developed mobile application. The self-designed earphone contains a microphone, a speaker, an audio amplifier, a Bluetooth module, and a micro-controller whose details will be introduced in Section V-A. When a user puts on earphones into his/her ears, the microphones activate immediately and collect data until an authentication process completes. A whole data-collection process actually has two stages of which one is during putting on earphones, and the other is when earphones are worn in ears. In the first stage, acoustic signals are mainly caused by the rubbing between an earpiece and the ear canal during putting-on activities. Acoustic signals collected in the next stage are mainly induced by activities of viscera, including heart, lung, etc. In a nutshell, we make use of both behavioral characteristics and physiological features for user authentication in EarPrint.

The collected signals are first boosted in the amplifier, and then transmitted to the mobile application continuously via Bluetooth. The application undertakes data communication and the whole data-processing pipeline. To be specific, after receiving data, the system first filters out power interference and its harmonics due to the defects of hardware. After that, it detects wearing events (WETs) and segment both two parts of signals into 2 s and normalize them. In other words, both the extracted behavioral and physiological signal segments last for 2 s. Then it converts signal segments into spectrograms, feeds them into a feature extractor to obtain 128-D embeddings. When using EarPrint for the first time, user's signals are required to be collected and then system stores user's embedding center in database. Generally, a embedding will be matched in the database to obtain the identity information. In the following, we shall give detailed introduction to each part.

B. Signal Preprocessing

1) Putting-on Events Detection: After collecting acoustic signals, we first apply a three-order Butterworth low-pass filter with a cut-off frequency of 50 Hz to remove high-frequency noises. As we utilize different feature extractors for two channels, we need to extract the two parts of signals at first. Here, we utilize the same method as in our previous work [27], i.e., a likelihood ratio test (LRT) and hidden Markov model (HMM)-based event detection module [44].[1] Consequently, we can obtain the corresponding events probabilities of each frame. We can see that except for putting on earphones, there exist some other periods with high event probabilities indicating the occurrence of other events. However, we can notice that the duration of these events is much shorter, which is helpful for filtering them out. Hence, we restrict the duration of detected events to more than 1.5 s so as to extract WETs precisely.

As for the remaining part (i.e., *physiological acoustics*), we do not perform signal detection-based segmentation based on the following consideration. Due to the low SNR, it is very difficult to detect every physiological event accurately, which probably results in server miss rate in more diverse and noisy practical usage scenarios. This shall decrease EarPrint's performance and worsen users' experience by forcing them to try more times. As a result, after extracting the WE, we adopt a simple but effective strategy-sampling the physiological signals randomly with a fixed-width sliding window. Note that a complete process includes a WE sample and multiple physiological samples which can constitute different pairs of inputs.

2) Signal Transformation: After the above, we adopt soft normalization method with the 5th and 95th quantiles of signals as done in the work [45]. Following that, we perform the short-time Fourier transform (STFT) on two channels of signals with a Hanning window of which the window size and overlap are 128 and 72 samples points, respectively. The final resulting time-frequency spectrogram is scaled to a size of 64×64.

C. Attention-Based Authentication Network

As aforementioned, the network design should fulfil the following goals. First, it can extract intrinsic features of different users, so that samples of the same person form a tight cluster in the feature space, while samples of different persons separate as far as possible. This is also a basic requirement of any authentication model. Second, the model should be able to deal with unseen users without collecting massive data for retraining the model. In other words, it is lightweight to register and authenticate new users. Consequently, it is improper to design a classification model as most previous works do. Third, there are two channels of inputs, namely, behavioral and physiological signals, it is required to adaptively weights them in order to make the system adjust to different environments. Inspired by FaceNet [46], our proposed network aims to generate an embedding that characterizes each user instead of a specific category. The whole pipeline is shown in Fig. 4. In the following, we introduce the details of each component.

1) Feature Extractor: In our system, we make use of both behavioral and physiological signals for user authentication. However, it is improper to treat them equally as they contribute differently due to the following reasons. On the one hand, we find that behavioral signals are more easily to be affected owing to that subjects sometimes adjust their earphones irregularly after wearing them. This indicates that behavioral and physiological signals differ in stability. On the other hand, both channels reflect different patterns of a user in two aspects and possess distinct characterizing abilities. The above observations motivate us to assign different weights to them. As a result, we design an attention mechanism-based learning network that extracts feature representations with assigning weights automatically. The architecture of designed feature extractor is shown in Fig. 4 which can be formulated as follows:

$$\boldsymbol{x} = \mathrm{FC}\big(\big[\mathrm{AvgPool2}d(\Theta_1(X_p));\, \boldsymbol{W} \cdot \mathrm{AvgPool2}d(\Theta_2(X_b))\big]\big) \quad (1)$$

where FC represents a couple of fully connected layers. $\boldsymbol{x}$, X_p, X_b denote the output embedding, physiological channel spectrograms, and behavioral channel spectrograms, respectively. Both Θ_1 and Θ_2 are ResNet18 [47] without the last layer in our design. Actually, they can also be other appropriate feature

[1] Limited by the page space, we omit the details of this method and suggest readers to refer to [27].

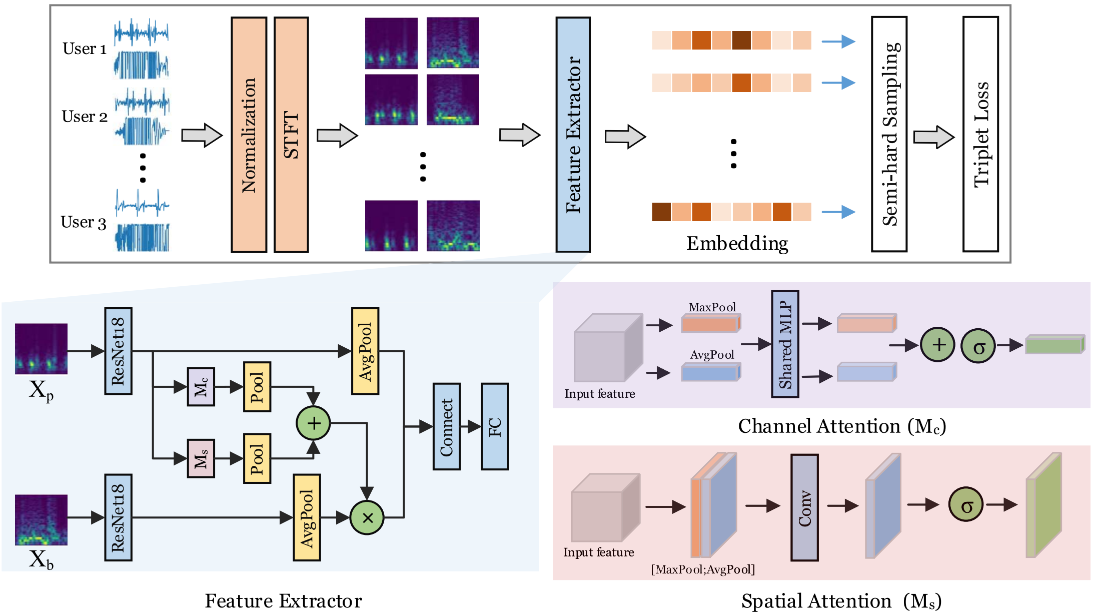


Fig. 4. Deep learning network architecture designed for EarPrint.

extractors. $\boldsymbol{W}$ indicates the weight of behavioral features. The final output of the network is a 128-D feature vector.

To obtain the weight $\boldsymbol{W}$, we borrow the idea of convolutional block attention module (CBAM) [48] to design lightweight channel attention (denoted by M_c) and spatial attention (denoted by M_s) as defined in

$$\begin{aligned}\boldsymbol{W} &= \text{AvgPool1}d(M_c(\boldsymbol{F}) + \text{AvgPool2}d(M_s(\boldsymbol{F})\\ M_c(\boldsymbol{F}) &= \sigma(\text{MLP}(\text{AvgPool}(\boldsymbol{F}) + \text{MLP}(\text{MaxPool}(\boldsymbol{F})\\ M_s(\boldsymbol{F}) &= \sigma\left(\text{Conv}^{7\times 7}\left(\left[\text{AvgPool}(\boldsymbol{F});\,\text{MaxPool}(\boldsymbol{F})\right]\right)\right)\end{aligned} \quad (2)$$

where $\boldsymbol{F}$ denotes output of $\Theta_1(X_p)$ and σ represent the sigmoid function. The key idea of CBAM is that given an intermediate feature map, it sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement. The rationale of adopting this module is that as convolution operations extract informative features by blending cross-channel and spatial information together, it emphasizes meaningful features along those two principal dimensions: 1) channel and 2) spatial axes. In our work, we make use of this mechanism to weigh behavioral and physiological channels adaptively.

2) Feature Similarity: To authenticate the current user, we need to calculate the similarity between his/her feature vector and the stored feature vectors of legitimate users. A higher similarity indicates a higher probability that the current user is a registered legitimate one. Since the network's output is a 128-D feature vector, to ensure the orthogonality of the similarity measure, we have chosen cosine similarity as the metric to quantify the similarity between feature vectors which can be defined as

$$s(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\boldsymbol{x_1} \cdot \boldsymbol{x_2}}{\|\boldsymbol{x_1}\|_2 \cdot \|\boldsymbol{x_2}\|_2} \quad (3)$$

where $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ denote two different embeddings. Here, we take advantage of cosine similarity to measure the distance between two embeddings as it is widely used in commercial authentication systems, such as face and fingerprint recognition.

3) Training Pipeline: After feature extraction, embeddings of data samples can be obtained as shown in Fig. 4. Following that, we adopt semi-hard sampling combined with triplet loss function, for the sake of faster convergence speed and better learning performance. Each time we train the model, we input an anchor a, a positive example p, and a negative sample n. Note that a positive sample belongs to the same category with an anchor, and vice versa. The loss function $\mathcal{L}$ for training our feature extractor can be written as

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N} \max\{d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0\} \quad (4)$$

where a_i, p_i, and n_i represent an anchor, a positive sample, and a negative example, respectively. *margin* is a tradeoff parameter which is set to be 0.2 empirically in our system implementation. d is the distance between any two samples which can be computed by

$$d = 1 - s(\boldsymbol{x_1}, \boldsymbol{x_2}). \quad (5)$$

The network is built using PyTorch and trained on a server equipped with an NVIDIA TITAN V. Some of the hyperparameters during the training process are listed in Table I. It is to be noted that the training of the neural network and the usage of the system are independent. In other words, new users do not need to retrain the model when utilizing the system in practice.

4) Recording Pipeline: When a valid user uses EarPrint for the first time, she needs to log a couple of samples and obtain the average embedding of them. It is noted that a complete sample contains signals of putting on earphones and inner body sounds with a duration of 30 s. Furthermore, the body sounds are randomly cut into 100 segments, each of

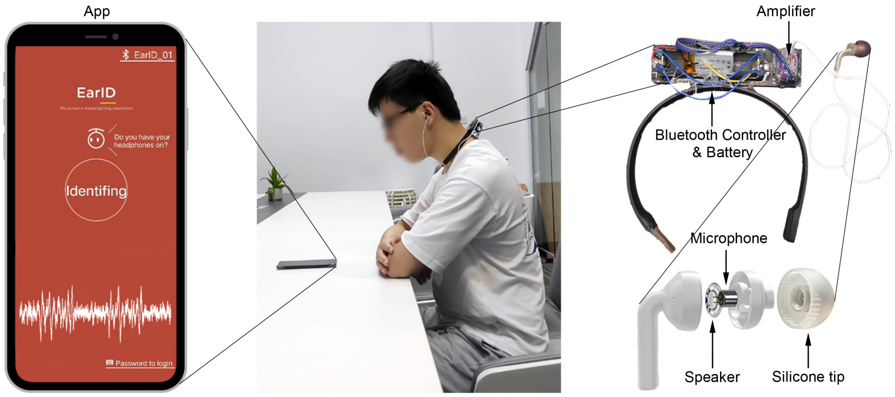


Fig. 5. Hardware and software of EarPrint system. This figure is borrowed from our previous work [27].

TABLE I
HYPERPARAMETERS USED FOR MODEL TRAINING

Hyperparameter	Value
batch size	512
epoch	20
learning rate	0.001
optimizer	Adam

which is combined with putting on signals forming an input sample. Consequently, each user can obtain a total number of 1000 embeddings (denoted by E_k). Then we can compute the center of all these embeddings by (6) as follows:

$$\text{Center}_k = \frac{1}{N}\sum_{x_i \in E_k} x_i \quad (6)$$

where Center_k denotes the center of each user's embeddings E_k, and N denotes number of each user's embeddings. For each valid user, we store his/her corresponding central embedding in the system.

5) Authentication Pipeline: In the authentication phase, when a user puts on earphones, the hardware collects acoustic signals and feeds into the data processing pipeline to finally extract the corresponding embedding. After that, the obtained embedding is matched with embeddings stored in the system and the similarity is calculated using (3). When the similarity is above a certain threshold, the user is identified. In our implementation, the similarity threshold is set to be 0.4. We also evaluate the impact of this parameter in Section VI-A.

V. IMPLEMENTATION AND EXPERIMENTS

A. System Implementation

In the following, we shall give details of the hardware and software of EarPrint.

1) Hardware: Since smart earphones do not output raw microphone signals, we design and implement smart earphones with cheap electret microphone sensors at a price of 10 CNY for each. The microphone sensors are connected with a low-cost, high-quality acoustic signal amplifier MAX9814 which integrates a preamplifier, variable gain amplifier (VGA), output amplifier, microphone-bias-voltage generator and AGC control circuitry. For more detailed specifications of MAX9814, please refer to this website [49]. The micro-controller unit (MCU) that we used in our earphones is an ESP32-S2 chip which is embedded with Wi-Fi and dual-mode Bluetooth (BLE and BT) modules. The ESP32 employs a Tensilica Xtensa LX6 microprocessor in both dual-core and single-core variations, and consists of built-in antenna switches, RF balun, power amplifier, and power-management modules. We connect the MCU with MAX9814 and control it to sample acoustic signals at 2000 Hz. At the same time, the Bluetooth module transmits collected data to the mobile application at the maximum speed. Although there is an ultralow power mode of Bluetooth module on the ESP32 board, we do not use it in our implementation as we find that the data transmission is not stable enough. Considering that this is an engineering problem, we leave it as future work to further optimize the hardware of EarPrint. To make the hardware aesthetic and comfortable, we also print a plastic earbud to pack microphone and speaker sensors, and a neckband to support the MCU board as shown in Fig. 5. Note that as the hardware in this article is the same with that used in our previous work [27], we directly cite the figure showing the appearance of hardware here.

2) Software: To perform user authentication and display results, we also develop a mobile application software on an Android smartphone, which is mainly responsible for data communication with earphones and executing data processing pipeline. We deploy the user authentication network which has been trained on a desktop or cloud server on a mobile device. This mobile application also contains some other functions, such as user enrolment and user guiding. In our experiments, we 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. As for the off-line training stage, we utilize a server with 64 GB RAM, NVIDIA TITAN V GPU, and Intel Xeon E5-2650 CPU. The network parameters will be frozen once the training process is finished, even though it will be deployed on different smartphones and have different user.

B. Data Collection

We have conducted extensive experiments under different settings considering practical impact factors to evaluate

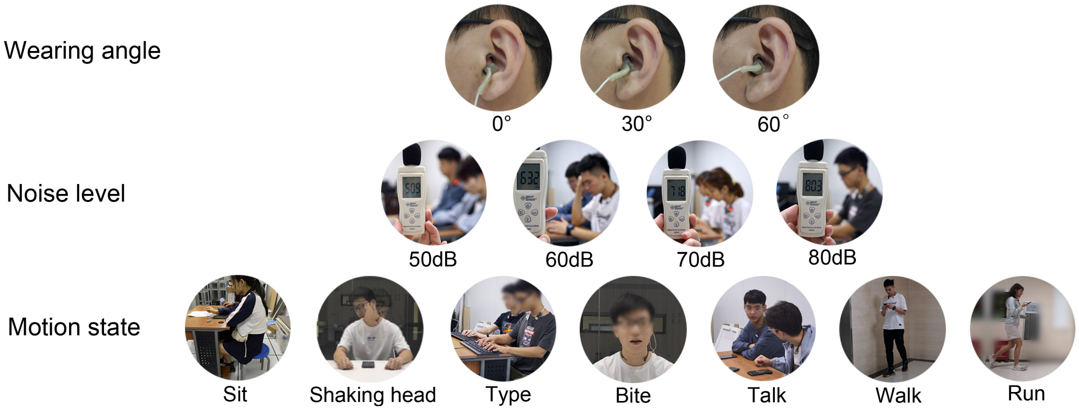


Fig. 6. Experimental settings.

EarPrint's performance. Specifically, we recruit 50 participants (denoted by $U_1 \sim U_{50}$), including students, staffs, and faculties with 31 males and 19 females from our campus aged between 18 and 35 years old. Before experiments, we explain the experimental details to all the participants in order to ensure that they clearly understand what they need to do during experiments, including how to charge the hardware, install applications on smartphones, and operate the application during data collection. Experimental settings in this work is classified according to four main factors, namely, noise level, users' motion states, wearing angle of earphones, and time gap between data collection sessions. The above impact factors cover different aspects of practical interference in real world. In the next, we describe the details of different experiment settings.

1) Noise Level: We carry out data collection in daily scenarios, such as residential houses, shopping malls, and public transportation in order to evaluate the impact of noise. To quantify this, we divide the above different settings into the following categories according to the intensity of noises which include a silent workplace with noise level lower than 40 dB (i.e., N_0), the same workplace with people talking and walking (50 ± 5 dB, N_1), a cafe environment (60 ± 5 dB, N_2), in a subway (70 ± 5 dB, N_3) and on a busy street (80 ± 5 dB, N_4). We measure the noise level with a SMART SENSOR Decibelmeter AS804. Under each setting, participants collect data for 2 h in all at different sessions each of which lasts for 10 min, in order to evaluate the robustness of our method over time. We set N_0 as the baseline setting with respect to this factor.

2) Moving State: In common sense, users' moving states will affect the contact between earphones and ear canals, and thus has impact on the quality of obtained signals. As a result, it is necessary to evaluate how this factor affects the system performance. To do this, we ask participants to collect data when they are in seven common moving states, including *sitting static*, *shaking heads*, *typing keyboard*, *jaw movements*, *running*, *speaking*, and *walking*. In each state, participants collect data for a total number of 2 h. At the same time, other factors like the noise level and wearing angle are set to be baseline values. As for this factor, we set *sitting static* as the baseline case.

3) Wearing Angle: Similarly, the wearing angle of earphones also has impact on the collected acoustic signals. And users may also wear earphones at different angles in practice. Hence, it is necessary to evaluate the impact of this factor. Specifically, we request each participant to wear earphones at three different angles, i.e., 0°, 30°, and 60°. Fig. 6 shows how a user wears earphones at these angles. At each angle, the data collection lasts for 2 h for each participant with all the other factors set to be baseline cases as well. By default, we set 0° as the *baseline* value of wearing angle in the evaluation.

4) Baseline Setting: In this work, we take four main impact factors, including noise level, moving state, wearing angle, and time gap into account when evaluating the system performance. Considering different combinations of these factors, there are tens of experimental settings. Nevertheless, it is infeasible to train an authentication model in each setting, since these exist an explosive number of settings in practice as aforementioned. For example, we consider five different noise levels, five moving states, three wearing angles, and three time gaps. Consequently, there are a total number of 225 testing settings. It results in heavy burden and terrible user experience to train a model in each setting. It is necessary to define a *baseline setting* in which data are collected to train a *baseline model* and evaluate the impact of each factor clearly. In this work, the *baseline setting* is selected according to most frequent usage scenarios of earphones, which is with noise level below 40 dB, users staying static and testing just after recording and the earphones worn at 0°.

5) Specifications of Data: As we build a *baseline model* and test it in diverse settings, we ask each participant to collect data in the *baseline setting* for 120 times in total. Each complete process of data collection lasts for 30 s in total, starting from putting on earphones and then remaining a certain period of time. As result, in this stage each participant collects a total period of 60 min data with a sampling rate of 2000 Hz in the baseline case.

For each scenario evaluation, the data set will be randomly divided into a training set (40 people) and a test set (10 people). The training set is utilized to train the feature extractor, while the test set is further divided randomly into legitimate users and attackers (individuals attempting to trick the system by using signals from unauthorized users). Under baseline conditions, the test set consists of 1 legitimate user and 9 attackers. Additionally, we need to extract a certain amount of data from legitimate users to obtain user feature vector templates. These templates are then used to calculate similarity with other data, resulting in evaluation metrics.

Due to the use of multiple random sampling methods, all evaluations for various scenarios will be repeated 10 times. The average values and standard deviations will be provided to demonstrate the reliability of the conclusions. Fig. 7 shows how the training and testing data sets are partitioned.

In the testing phase, we will use data recorded from legitimate users in the baseline setting to calculate the feature vector templates. Subsequently, we will validate these templates using data recorded by the same legitimate users in the target scenarios. For the other nonlegitimate users in the test set, we will not make scene-based distinctions. We assume that unauthorized users will attempt various methods to attack our identity authentication system.

It is worth mentioning that we have not conducted imitation experiments. This is because the physiological and behavioral signals we collect are extremely faint and only present within the ear canal. These signals cannot be perceived by the external environment, and it is difficult to obtain or mimic them. This highlights the security of our system.

VI. EarPrint Performance

In this part, we evaluate EarPrint with three main metrics, namely, EER, FAR, and FRR. FAR is the percentage of unauthorized users accepted by the system; FRR is the percentage of a valid user incorrectly rejected by the system; EER is the point at which the system's FRR and FAR are equal. In general, the superior performance of an authentication system is reflected in lower FRR and FAR. However, these two metrics are usually contradictory, meaning that when FRR decreases, FAR may increase; and vice versa. As for the EER, the ideal case is to have FAR and FRR as close to each other as possible, which indicates that an authentication system can maintain a low error rate while preserving a certain level of sensitivity and specificity. As a result, the system is not only cautious in rejecting nontargets, but also effective in identifying targets, achieving a good balance in two aspects. Therefore, a lower EER reflects a better performance of an authentication system in practical applications. In summary, these metrics are essential for evaluating the effectiveness of a system. Achieving a lower value for FAR, FRR, and EER indicates superior system performance.

A. Determination of Similarity Threshold

The impact of similarity threshold δ on FAR and FRR is a tradeoff for an authentication system. A larger threshold leads to lower FARs but higher FRRs, and vice versa It indicates that a larger threshold is beneficial to better security against imposters but also causes inconvenience to a valid user. To obtain the quantitative relationship between them, we vary δ from 0 to 1 and test FAR and FRR, respectively. Fig. 8 shows the result. We can see that when δ equals to 0.29, EarPrint achieves an average EER of 4.23% with a standard deviation of 0.63%. Hence, when we evaluate the FAR and FRR performance, we choose δ to be 0.4, with which the FAR and FRR are 0.58% and 10.9%, respectively, as we value security more than convenience in our system design. This is reasonable considering that a user does not perform any actions during an authentication process, which means multiple trials do not cause trouble to users.

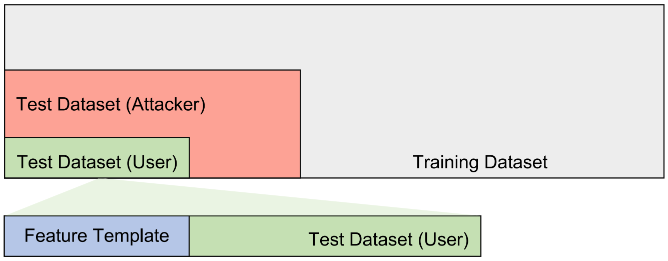


Fig. 7. Partitioning of training and testing data sets.

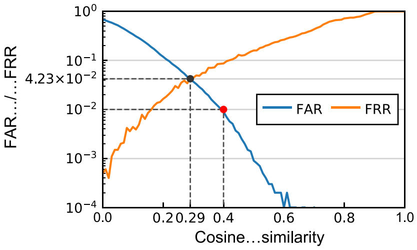


Fig. 8. FAR and FRR vary with different similarity thresholds.

B. Impact of Noises

We evaluate the impact of external noise at five different intensity levels and obtain the results as shown in Fig. 10. As we can see, although the overall performance degrades with the increasing noise level, the FAR, EER and FRR stay relatively stable less than 4.89%, 0.96%, and 11.8%, respectively, when the noise level is below 60 dB. Even when the noise level rises above 70 dB, the EER, FAR, and FRR increase by 2.11%, 0.81%, and 4.29%, respectively. The results show that EarPrint has good robustness against noises, especially in terms of defence imposters' attack. It is reasonable to claim that EarPrint still works well in a majority of daily scenarios in which the noise level is below 70 dB. Even when the noise level further increases to 80 dB or above, the EER and FAR can still remain relatively stable, but FRR goes up noticeably. It also reveals that EarPrint can extract unique features from body sounds of different users.

C. Impact of Moving States

Fig. 9 shows EarPrint's performance when participants are in different motion states as aforementioned. We can see that, different motions show diverse impact of which talking and running show the most significant impact. Except for the above two states, the FAR and EER keep under 3.5% and 18.3%, respectively. Compared with the baseline case (i.e., sit), the FAR only increase slightly by less than 2.9%. Even when a user is running, the FAR of EarPrint can stay around 5.5%, which indicates that EarPrint can defend attacks in most life

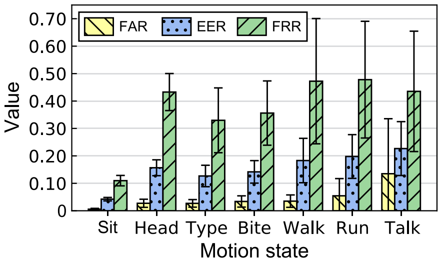


Fig. 9. System performance at different motion states with the original model.

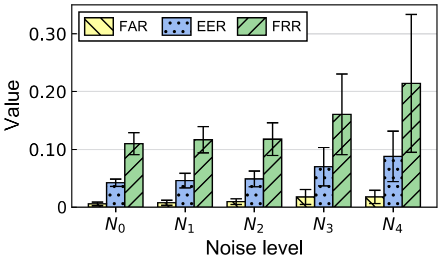


Fig. 10. System performance under different noise levels.

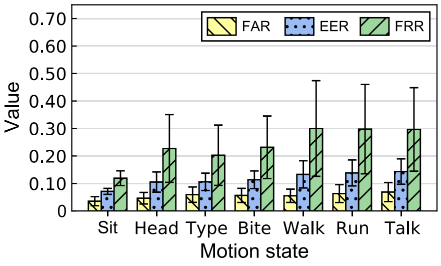


Fig. 11. Impact of motion states on EarPrint with two model training strategies.

scenarios. Running and talking continuously affect the contact between earphones and ear canals, which makes it unstable to sense acoustic signals with embedded microphones. This is the reason why the performance of EarPrint degrades more sharply in these two motion states. On the other hand, we can also notice that the FRR of EarPrint is more sensitive to motions, which means that it is more likely for a user to perform multiple authentication trials. To deal with this problem, we use one sample collected during running and talking, respectively, to calculate the average feature vector as the template. Fig. 11 shows the results with this strategy. We can see that it can effectively improve the authentication performance.

D. Impact of Wearing Angles

Fig. 12 shows the authentication performance of EarPrint when participants wear earphones at different angles, including 0°, 30°, and 60°. The maximum gaps of EER, FAR and FRR at three wearing angles are about 0.4%, 0.36%, and 1.0%, respectively. These differences are even less than the variances of results which demonstrates the stability of EarPrint against wearing angle. This is a distinct advantage over previous work based on active sensing, such as EarEcho [8] whose recall drops by about 14% after rotating earphones by 50° along Y axis. This is because EarEcho relies on the 3-D structure of the cavity between earphones and ear canals. This structure is dependent on how earphones are worn in ears. In comparison, EarPrint makes use of users' intrinsic behavioral and psychological features which are more robust.

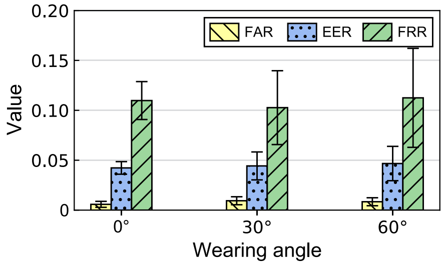


Fig. 12. System performance at different wearing angles.

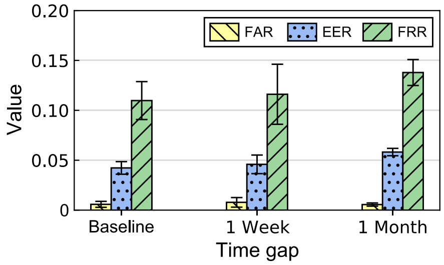


Fig. 13. System performance after different periods.

E. Performance at Different Time Gaps

In order to ensure the stability of the system for a long time, we collected test data again at intervals of one week and one month, respectively. Fig. 13 shows EarPrint's performance after one week and two weeks compared with the baseline case. Note that the baseline performance in this case is a mixture of data collected from three experiments, from which templates and test sets are randomly drawn. The evaluation method of one-week time interval and one-month time interval is to use the data extracted from the first experiment as a template, and use the data after the interval for testing. We can see that the FAR nearly does not change in the three cases, with a negligible difference of 0.2%. Meanwhile, the FRR encounters a slight increase from 10.9% in the baseline case to 13.7% after two weeks. The underlying reason may be that at different time, subjects may have different physiological states, such as getting a cold, coughing, etc. This would change the patterns of *behavioral acoustics*. Fortunately, with a nearly unchanged FAR, our model still does not mistake different subjects. To handle the problem of rising FRR, we can update the centre of embeddings [i.e., Center$_k$ in (6)] with newly collected samples in real-world usage scenarios. It should be noted that this update process does not need retrain the model, and even does not involve users' conscious participation.

F. Impact of Model Training Parameters

In this section, we mainly evaluate the impact of training parameters, including the number of repetitions, number of segments, and size of sliding window.

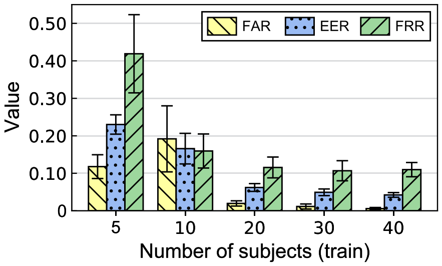


Fig. 14. System performance with different number of training subjects.

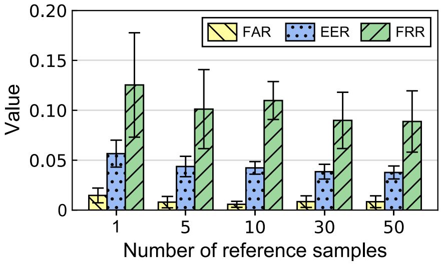


Fig. 16. System performance with different number of reference samples.

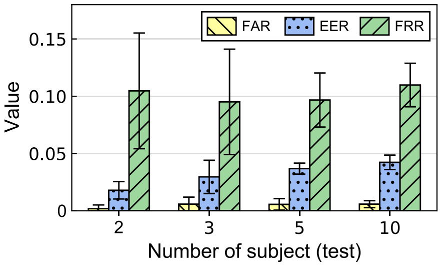


Fig. 15. System performance with different number of testing subjects.

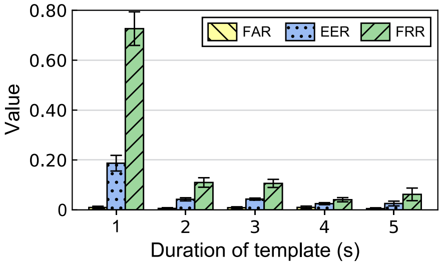


Fig. 17. System performance with different framing sizes.

1) Number of Training Subjects: We first evaluate how the number of training subjects affects the performance by utilizing data of 5–40 participants to train the network model. To do this, each time we randomly select different numbers of participants as trainers and the rest as testers for 10 times, then compute the average EER, FAR and FRR, and finally obtain the results as shown in Fig. 14. We can see that the three metrics decrease with the number of subjects taking part in the training process. In other words, the more trainers are involved, the better the performance is. The underlying reason is that data from more subjects is helpful for training the model to learn intrinsic fingerprints of different persons, which boosts the authentication performance of EarPrint. But we can also notice that when the number of training subjects exceeds 20, the EER, FAR and FRR decrease slightly by less than 1.98%, 1.38%, and 0.88%, respectively. In our implementation, we make use of data of 40 subjects for model training for the sake of optimal performance.

2) Number of Testing Users: We also evaluate the impact of number of testers by asking the remaining 10 participants to act as valid users and imposters. Specifically, we randomly select 2–10 participants of which each acts as a valid user and the rest are imposters in one time. It is noted that these persons have not participated in model training, which means that their data are not used in the training phase. As shown in Fig. 15, the maximum difference in terms of EER, FAR and FRR in four different testing cases is 2.44%, 0.4%, and 1.47%, respectively. These differences are very close to the variances in different testing cases. It demonstrates that there is nearly no difference of EarPrint's performance when different number of unseen subjects take part in testing the system.

3) Impact of Reference Samples: The number of reference samples used for extracting an unique feature vector for each user has great impact on the authentication performance. In an intuitive sense, providing more reference samples is helpful for obtaining a stable feature vector. To quantify this, we utilize 1–50 reference samples to get average feature vector, and then test the performance of EarPrint as shown in Fig. 16. The overall trend is that EER, FAR and FRR decrease with the number of reference samples, from 5.66%, 1.47% and 12.54%–3.77%, 0.84%, and 8.88%, respectively. What is more, we can notice that when the number of reference samples exceed 10, the EER and FAR keep relatively stable. Considering reducing the effort of users to logging data, we set this parameter to be 10 when we implement EarPrint.

4) Impact of Window Size: As aforementioned, we make use of a sliding window with a fixed length (i.e., window size) to sample *physiological acoustics*. To evaluate how the window size affects results, we vary its value from 1 s to 5 s and obtain corresponding results as shown in Fig. 17. As we can see, when this value is set to be 2 s, EarPrint shows good performance with EER, FAR and FRR being 4.23%, 0.58%, and 10.98%, respectively. In comparison with the case of 1 s window size, EarPrint's performance has notable improvement in terms of all the metrics. But when the window size exceeds 2 s, the performance does not improve obviously any more. The underlying reason is interesting. As mentioned above, physiological acoustic signals are mainly induced by heartbeating whose period is about 0.6 s to 1 s. As we randomly segment physiological acoustics with a sliding window, it is likely to miss the core part of a heartbeat containing useful information for user authentication when the window size is small. However, although a larger framing window is useful for improving EarPrint's performance, it also means that longer authentication time is required, which degrades user experience. Consequently, we set the window size to be 2 s in our implementation.

5) Impact of Number of Segments: In the model training stage, participants conduct each experiment session for 10 min as described in Section V-B. As a result, it produces multiple

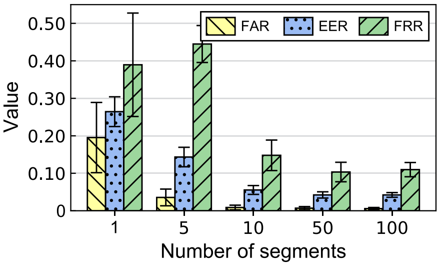


Fig. 18. System performance with different number of segments.

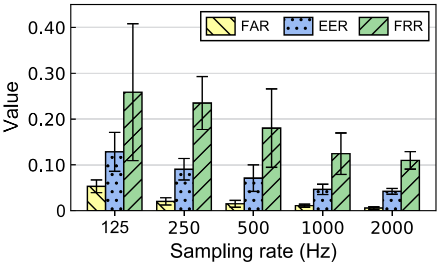


Fig. 19. System performance with different sampling rates.

10-min signal episodes. We then randomly segment each of them into a certain number of pieces with the sliding window. To determine an appropriate number of segments, we vary this parameter from 5 to 100 and test EarPrint's performance. The results are shown in Fig. 18. We can notice that the EER, FAR and FRR decrease from 26.4%, 19.5%, and 38.9% to 4.2%, 0.58%, and 10.9%, respectively, when the number of segments increases from 5 to 100. The reason is easy to understand. Extracting more segments from a given signal sequence provides more training data, which helps to train the model more optimally. This is more obviously when the number of segments is less than 10. Nevertheless, when this parameter exceeds 50, the performance improvement becomes slow. The reason is that segmenting a fixed-length signal sequence into an excessive number of pieces is more likely to produce overlaps which do not augment data diversity and boost system performance notably. Considering the tradeoff between authentication performance and training overhead, we set this parameter to be 100 in EarPrint.

6) Impact of Sampling Rate: At last, we also evaluate the impact of sampling rate of microphones. To do this, we perform downsampling processes on the raw data and reduce the sampling frequency from 2000 to 1000 Hz, 500, 250 and 125 Hz. The corresponding performance are shown in Fig. 19. It is obvious that a higher sampling rate results in better performance, as body sounds span a relatively wide frequency range and a high sampling rate is helpful for obtaining richer information. But as we can also see, the EER, FAR, and FRR decrease slowly when the sampling rate is higher than 1000 Hz. This may be because most useful information is contained in a frequency band below 1000 Hz. To achieve better performance, we set this parameter to be 2000 Hz in the system implementation. In contrast with active-sensing methods [8], [20], our sampling rate is more than 20 times lower (i.e., 2000 versus 44 100 Hz), which reduces computation overhead and energy consumption resource-constrained smart devices.

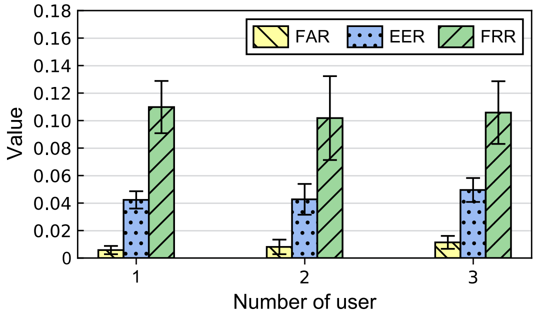


Fig. 20. System performance with multiuser authentication.

TABLE II
RESULTS OF ABLATION STUDIES

Metric Item	EER (%)	FAR (%)	FRR (%)
Without X_b	5.69±0.67	3.00±1.64	8.88±2.09
Without X_p	24.14±2.05	4.38±1.35	67.47±7.50
Without attention	16.78±1.69	2.78±1.21	56.27±2.09
With classifier	-	3.38±1.22	4.89±2.09
Ours	**4.23±0.63**	**0.58±0.30**	**10.98±1.90**

G. Multiuser Authentication Performance

In addition to the conventional identity authentication task, our system is also designed to simultaneously achieve identity recognition, meaning that it can handle multiple legitimate users' feature vectors concurrently. This capability is particularly meaningful in home scenarios where family members may share each other's devices. To assess EarPrint's performance with multiple users, we reorganize the testing set, modifying the proportion of legitimate users over illegitimate users. We perform individual result calculations and compute averages and variances for each legitimate user. The results are shown in Fig. 20. It is evident that when the number of legitimate users is 1, 2, and 3, the corresponding EER is 4.23%, 4.27%, and 4.95%, respectively. This suggests that as the number of users increases, the system performance tends to degrade. This is reasonable because having more users leads to additional clusters in the vector space, resulting in increased instances of occasional sample overlap between these clusters. However, we observe that with three users, the EER only increases by 0.72%. Hence, it can be concluded that, when the number of legitimate users is no more than three, the system performance remains relatively stable. Considering that multiuser authentication is not common in real-world scenarios, our system is configured with one legitimate user by default.

H. Ablation Studies

To valid the model design, we conduct ablation studies by removing or changing a certain part at one time. As shown in Table II, Without X_b and Without X_p represent that behavioral or physiological channel signals are not utilized. Without attention represents both channels are used

TABLE III
COMPARISON WITH OTHER EARPHONE-BASED AUTHENTICATION METHODS

Method	Data	Algorithm	Retraining cost	Multi-user	Latency (s)	FAR (%)	FRR (%)	EER (%)
Takashi et al [22].	EEG	Cosine dist.	No	No	60	2.3	32.2	-
EarEcho [8]	Acoustic (echo)	SVM	Yes	No	1	4.8	6.2	-
EarGate [20]	Acoustic (gait)	Bi-SVM	Yes	No	0.5	3.2	2.25	-
HeartPrint [19]	Acoustic (implicit)	Classifier	Yes	No	-	1.6	1.8	-
EarID [27]	Acoustic (implicit)	Classifier	Yes	Yes	2.2	3.38	4.89	-
BreathSign [21]	Acoustic (implicit)	DML	No	Yes	-	-	-	-
EarPrint	**Acoustic (implicit)**	**DML**	**No**	**Yes**	**2.2**	**0.6**	**11.0**	**4.23**

TABLE IV
REAL-TIME RUNNING PERFORMANCE OF EarPrint

Tester ID Testing items	1	2	3	4	5	6	7	8	9	10	Average	STD
Authentication latency (ms)	2218	2601	2186	2075	2134	2503	2480	2114	2127	2113	**2255.1**	**184.8**
Memory (MB)	100.4	100	88	96.7	91.3	100.3	92.1	106.3	101.3	85.5	**96.2**	**6.3**
CPU (%)	41	40	50	40	45	42	40	49	39	38	**42.4**	**3.98**

but without attention mechanism. With classifier means that we remove semi-hard sampling and triplet loss components as shown in Fig. 4, but add a fully-connected layer to do multiclass classification. Ours means the complete network design as introduced in this article. From the table, we can obtain several conclusions as follows. First, both channels contribute to authentication performance but weighs differently. It is obvious that *physiological acoustics* are more important. Second, the attention mechanism is effective in assigning different weights to both channels adaptively. Notably, ours method substantially improves FAR. Third, when we follow a classification approach, the system performance goes down in terms of FAR. What is more, this approach requires to retrain the model when unseen persons use the system.

I. Comparison With Other Methods

We have also compared our work with existing earable authentication systems, including Nakamura et al. [22], EarEcho [8], EarGate [20], HeartPrint [19], EarID [27], and BreathSign [21], from different aspects of sensing signals, authentication algorithms, necessity of retraining with new users, whether supports multiuser authentication, response time, FAR, FRR, and EER. Among them, necessity of retraining means whether the model needs to be retrained when new valid users are involved. Multiuser authentication represents that the model is able to identify multiple users. The comparison results are shown in Table III. As we can see, compared with existing earable authentication methods, EarPrint shows superiority in most of comparison aspects, including zero retraining cost, supporting multiuser authentication, much lower FAR, and EER. Compared with our previous work [27], EarPrint does not need model retraining for unseen users and support multiuser authentication, which makes it more practical in real-world application scenarios. Nevertheless, EarPrint also has limitations in FAR and latency. In the future work, we shall focus on improving our system by boosting FAR and real-time response.

J. System Running Performance

1) Computing Resources Occupation: Moreover, we evaluate EarPrint's real-time running performance, including authentication latency, memory occupation, and energy consumption on a mobile end. The latency represents time period of a complete authentication process. To evaluate this, we insert a piece of codes in the Android software to measure the time duration of an complete authentication process. Meanwhile, we record memory and CPU occupation of EarPrint application with Android Studio Profiler. To do this, we turn off all the third-party applications, such as localization and networking. The screen lightness is set to be the lowest level. Ten participants are involved in this experiment, each of whom tries EarPrint for twenty times. Table IV shows the average authentication latency, memory and CPU occupation according to our experiment. We can see that EarPrint's average latency is about 2.3 s with a standard deviation of 0.18 s which is composed of data sampling time (i.e., window size discussed in Section VI-F4) and inference time. It indicates that even though a larger window size benefits to improving authentication performance, it results in larger running latency yet. The actual model inference time is only about 0.3 s which is comparable with commercial fingerprinting-based scheme. Furthermore, according to our experiments, the average occupation of memory and CPU of EarPrint are 96.2 MB and 42.4%, respectively.

2) Energy Consumption: In addition, we measure the energy consumption of EarPrint software on a smartphone. Similar to the above, we shut down all the other applications except EarPrint and Bluetooth connection, and kill background services, including networking, localization, etc. The screen lightness is set to be the lowest level as well. After that, we request participants to wear earphones, turn on EarPrint system, and monitor the battery level of the smartphone every 5 min with Android API. In this experiment, EarPrint has run 3 h finally and finishes a total number of 4715 authentication trials. The recorded battery level is shown by the blue line in Fig. 21. As we can see, the battery level decreases linearly

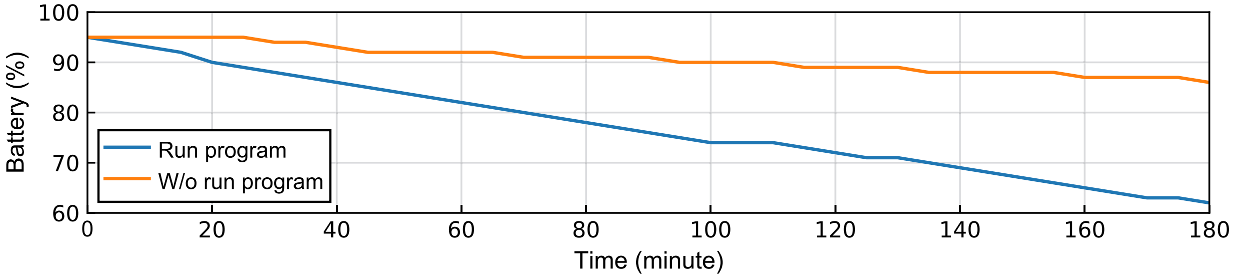


Fig. 21. Energy consumption of EarPrint application on a smartphone.

with time from 95% to 62% after 3 h. In order to know more clearly about EarPrint's power consumption, we conduct a comparative experiment as follows. We shut down all the applications, including EarPrint and only keep Bluetooth connection with earphones without data transmission. In this case, the battery level of the same smartphone varies as shown by the red line in Fig. 21. As we can see, when EarPrint runs, the battery level decreases by 33% in 3 h which is 24% larger than the results of comparative experiment. We believe that EarPrint's energy consumption performance can be further improved in the future by making use of ultralow power Bluetooth 5.0.

Moreover, we also measured the energy consumption of the smart earphones. We fully charged the earphones with a 500 mAh battery, and after connecting it to the EarPrint, we kept it in the state of collecting and transmitting data. In this experiment, the smart earphones has run 3.86 h. Since the time required for a user authentication is 2.3s, the average power consumption of smart earphones each time is less than 0.1mAh, which is negligible compared to the power of smart earphones.

VII. Discussions and Future Work

Although EarPrint achieves high authentication performance, it still has the following limitations to be dealt with in the future work.

Sensitivity to Motions: EarPrint is easy to be affected by human motions according to the results shown in Fig. 9. Although it can be mitigated by collecting new samples to retrain the model, the performance is still not very good according to Fig. 11. As a result, a possible way to deal with this is to train a different model with data collected when users are in motion states as done in the related work [20]. However, to determine which authentication model should be used, it is necessary to design a method to identify whether a user is static or moving. This shall be easy to be achieved since human activity recognition has been a research hotspot for a long time.

Relatively High FRR: Present version of EarPrint has relatively high FRR due to the tradeoff between FAR and FRR. By changing the similarity threshold δ, it is possible to decrease FRR but with FAR increasing at the same time. Another solution is to perform repetitive authentication trials which only increases latency but does not require a user to participate in the process consciously. In practice, a user can set the similarity threshold according to specific applications.

Limitations in Practice: At present, EarPrint system conducts user authentication without playing any other sounds, such as music. But a previous work [50] has proposed a method to separate heartbeat sounds from music sounds captured by an in-ear microphone, so as to extract accurate heart rate. With such a method, it is possible to achieve simultaneous user authentication while playing music. Since the focus of our paper is the authentication technique, we leave this part as one of the future work. In addition, since our work is on the basis of occlusion effect, that is, the earphones should form a closed cavity with the ear canal and eardrum, in order to capture high-SNR heartbeat sounds. However, the semi-in-ear headsets can not strictly satisfy such a requirement. Therefore, our proposed method is not suitable for semi-open headphones.

VIII. Conclusion

In this work, we propose an earphone-based implicit authentication system called EarPrint which takes advantages of *behavioral* and *physiological* acoustics during using earphones with embedded microphone sensors. To ensure the unobtrusiveness, convenience and high performance, we borrow the idea of channel and spatial attention mechanism to design an embedding learning network which can extract intrinsic features of two channel signals. We have designed a low-cost wireless earphone system and developed an Android mobile application to verify its usability in practice. We have also evaluated its performance with extensive experiments under various settings. Our results show that EarPrint can accomplish high authentication performance with rather low EER and FAR.

References

[1] A. Kumar and C. Kwong, "Towards contactless, low-cost and accurate 3D fingerprint identification," in *Proc. IEEE CVPR*, 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. 2012.

[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*. Amsterdam, The Netherlands: Elsevier, 2009, pp. 715–739.

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

[7] Y. Zou, M. Zhao, Z. Zhou, J. Lin, M. Li, and K. Wu, "BiLock: User authentication via dental occlusion biometrics," *Proc. ACM Interact., Mobile, Wearable Ubiquit. Technol.*, vol. 2, no. 3, pp. 1–20, 2018.

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

[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] D. Li, S. Cao, S. I. Lee, and J. Xiong, "Experience: Practical problems for acoustic sensing," in *Proc. ACM Mobicom*, 2022, pp. 381–390.

[12] T. L. Wiley, K. J. Cruickshanks, D. M. Nondahl, T. S. Tweed, R. Klein, and B. E. Klein, "Aging and high-frequency hearing sensitivity," *J. Speech, Lang., Hear. Res.*, vol. 41, no. 5, pp. 1061–1072, 1998.

[13] 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*, 2012, pp. 987–996.

[14] 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 IPCCC*, 2014, pp. 1–8.

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

[16] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. London, U.K.: Springer, 2009.

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

[18] T. Zhao, Y. Wang, J. Liu, Y. Chen, J. Cheng, and J. Yu, "TrueHeart: Continuous authentication on wrist-worn wearables using ppg-based biometrics," in *Proc. IEEE Infocom*, 2020, pp. 30–39.

[19] Y. Cao, C. Cai, F. Li, Z. Chen, and J. Luo, "HeartPrint: Passive heart sounds authentication exploiting in-ear microphones," *Heart*, vol. 50, no. S1, p. S2, 2023.

[20] A. Ferlini, D. Ma, R. Harle, and C. Mascolo, "EarGate: Gait-based user identification with in-ear microphones," in *Proc. ACM Mobicom*, 2021, pp. 337–349.

[21] F. Han, P. Yang, S. Yan, H. Du, and Y. Feng, "BreathSign: Transparent and continuous in-ear authentication using bone-conducted breathing biometrics," in *Proc. IEEE INFOCOM*, 2023, pp. 1–10.

[22] T. Nakamura, V. Goverdovsky, and D. P. Mandic, "In-ear EEG biometrics for feasible and readily collectable real-world person authentication," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 648–661, 2017.

[23] M. T. Curran, J.-K. Yang, N. Merrill, and J. Chuang, "Passthoughts authentication with low cost earEEG," in *Proc. IEEE EMBC*, 2016, pp. 1979–1982.

[24] S. Chang, X. Hu, H. Zhu, W. Liu, and L. Yang, "VOGUE: Secure user voice authentication on wearable devices using gyroscope," in *Proc. IEEE SECON*, 2022, pp. 361–369.

[25] Y. Jiang, H. Zhu, S. Chang, and B. Li, "MAUTH: Continuous user authentication based on subtle intrinsic muscular tremors," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1930–1941, Feb. 2024.

[26] H. Zhu, J. Hu, S. Chang, and L. Lu, "Shakein: Secure user authentication of smartphones with single-handed shakes," *IEEE Trans. Mobile Comput.*, vol. 16, no. 10, pp. 2901–2912, Oct. 2017.

[27] Y. Zou, H. Lei, and K. Wu, "Beyond legitimacy, also with identity: Your smart earphones know who you are quietly," *IEEE Trans. Mobile Comput.*, vol. 22, no. 6, pp. 3179–3192, Jun. 2023, doi: 10.1109/TMC.2021.3134654.

[28] 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. 2018.

[29] 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.

[30] N. Bui et al., "eBP: A wearable system for frequent and comfortable blood pressure monitoring from user's ear," in *Proc. ACM Mobicom*, 2019, pp. 1–17.

[31] 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. 23402–23417, 2015.

[32] C. Athavipach, S. Pan-Ngum, and P. Israsena, "A wearable in-ear eeg device for emotion monitoring," *Sensors*, vol. 19, no. 18, p. 4014, 2019.

[33] 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. Thorac. Soc.*, vol. 13, no. 12, pp. 2229–2233, 2016.

[34] 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. ACM SenSys*, 2016, pp. 230–244.

[35] C. Min, A. Mathur, and F. Kawsar, "Audio-kinetic model for automatic dietary monitoring with earable devices," in *Proc. ACM Mobisys*, 2018, pp. 517–517.

[36] S. Bi et al., "Auracle: Detecting eating episodes with an ear-mounted sensor," *Proc. ACM IMWUT*, vol. 2, no. 3, pp. 1–27, 2018.

[37] 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.

[38] 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 *Proc. Adjun. Proc. UBICOMP/ISWC*, 2015, pp. 1333–1338.

[39] T. Amesaka, H. Watanabe, and M. Sugimoto, "Facial expression recognition using ear canal transfer function," in *Proc. ACM ISWC*, 2019, pp. 1–9.

[40] K. Carillo, O. Doutres, and F. Sgard, "Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation," *J. Acoust. Soc. Amer.*, vol. 147, no. 5, pp. 3476–3489, 2020.

[41] 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 Infocom*, 2015, pp. 2686–2694.

[42] 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.

[43] W.-H. Lee, X. Liu, Y. Shen, H. Jin, and R. B. Lee, "Secure pick up: Implicit authentication when you start using the smartphone," in *Proc. 22nd ACM Symp. Access Control Models Technol.*, 2017, pp. 67–78.

[44] 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.

[45] A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms," in *Proc. IEEE EMBC)*, 2018, pp. 1–4.

[46] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE CVPR*, 2015, pp. 815–823.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[49] "Specifications of max9814." Adafruit. Accessed: Jun. 23, 2022. [Online]. Available: https://learn.adafruit.com/adafruit-agc-electret-microphone-amplifier-max9814

[50] S. Nirjon et al., "MusicalHeart: A hearty way of listening to music," in *Proc. ACM SenSys*, 2012, pp. 43–56.

Yongpan Zou (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2017.

He is currently an Associate Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interest includes ubiquitous sensing, mobile computing, and human–computer interaction.

Jianhao Weng received the master's degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2024.

He is currently a Software Engineer with Shenzhen Futu Network Technology Company Ltd., Shenzhen. During his postgraduate stage, his research interest covers mobile computing and ubiquitous sensing.

Haibo Lei received the master's degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2022.

He is currently an Algorithm Engineer with Tencent, Shenzhen. During the postgraduate stage, his research interest covers wearable computing and HCI.

Danyang Wang received the master's degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2023.

She is currently a Software Engineer with Hisense, Qingdao. During the postgraduate stage, her research interest covers ubiquitous computing and smart sensing.

Victor C. M. Leung (Life Fellow, IEEE) received the B.A.Sc. (Hons.) and Ph.D. degrees in electrical engineering from UBC in 1977 and 1981, respectively.

He is currently the Dean of the Artificial Intelligence Research Institute and a Professor of Engineering with Shenzhen MSU-BIT University, Shenzhen, China. His published works have together attracted more than 60 000 citations. His research is in the broad areas of wireless networks and mobile systems.

Dr. Leung has received many academic awards, such as the 1977 APEBC Gold Medal, 1977–1981 NSERC Postgraduate Scholarships, IEEE Vancouver Section Centennial Award, 2011 UBC Killam Research Prize, and 2017 Canadian Award for Telecommunications Research.

Kaishun Wu (Fellow, IEEE) received the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, in 2011.

He is currently a Professor with the Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. His research interests include wireless communications and mobile computing.

Prof. Wu won several best paper awards of international conferences, such as the IEEE Globecom 2012 and the IEEE MASS 2014.