Secure User Verification and Continuous Authentication via Earphone IMU

Jianwei Liu, *Student Member, IEEE,* Wenfan Song, *Student Member, IEEE,* Leming Shen, Jinsong Han, *Senior Member, IEEE,* and Kui Ren, *Fellow, IEEE*

Abstract—Biometric plays an important role in user authentication. However, the most widely used biometrics, such as facial feature and fingerprint, are easy to capture or record, and thus vulnerable to spoofing attacks. On the contrary, intracorporal biometrics, such as electrocardiography and electroencephalography, are hard to collect, and hence more secure for authentication. Unfortunately, adopting them is not user-friendly due to their complicated collection methods or inconvenient constraints on users. In this paper, we propose a novel biometric-based authentication system, namely *MandiPass*. *MandiPass* leverages inertial measurement units, which have been widely deployed in portable devices, to collect intracorporal biometric from the vibration of user's mandible. It provides not only one-time verification function but also continuous authentication function. Both the two functions are secure and user-friendly. We theoretically validate the feasibility of *MandiPass* and develop a series of deep learning techniques for effective biometric extraction. We also utilize a Gaussian matrix to defend against replay attacks. Extensive experiment results with 34 volunteers show that *MandiPass* can achieve low equal error rate, even under various harsh environments.

Index Terms—User Verification, Continuous Authentication, Biometrics, Inertial Measurement Units.

1 Introduction

User authentication plays an essential role in the security-relevant scenarios, such as access control and commercial transaction. With the prevalence of mobile computing, user authentication usually functions as the first defense for the device and system, e.g., unlocking a mobile phone. Prior works have widely adopted PIN-based [1] and pattern lock-based [2] mechanisms, which follow the principle of 'something a person has or knows' [3]. In this case, if someone has the credential, i.e., the 'something', s/he would be authenticated as the genuine user, no matter who s/he really is. Therefore, these approaches are vulnerable to many attacks, including the stealing, guessing, and shoulder-surfing attacks [4].

On the other hand, biometric-based authentication is known as 'something a person is or does' [3]. It shows advantages in terms of high security, convenience, non-transferability, and low possibility to be faked or stolen. However, existing pervasively adopted biometrics, including fingerprint, facial feature, and voice-print, are still prone to duplication attacks, because they are easily collected from body surfaces or remote positions. For example, fingerprint can be easily forged and is vulnerable to spoofing attacks [3]. FaceID adopts depth sensors like dot projector and infrared depth camera to improve its security, but it still could be deceived by 3D printed masks [5], [6]. Voices can be captured within a relatively large range. Thus, voice-based authentication is also vulnerable to replay attacks [7].

Recently, researchers exploit some 'unobtrusive' biometrics for authentication, such as brain waves [8], cardiac activities [5], and ear canal features [5]. These biometrics are more secure because they are usually collected from tissues and organs inside human bodies. Capturing, recording, or cloning them is extremely difficult. However, the collection of these biometrics is usually not user-friendly. For instance, users have to pose specific gestures for collecting the cardiac activities, e.g., measurements via electrocardiography (ECG) [5] and photoplethysmography (PPG) [9]. Meanwhile, extra sensing devices lead to inconvenience to users and hence impede adopting these intracorporal biometrics in authentication. For example, stable collection on the electroencephalograph (EEG) requires users to wear cumbersome sensing devices on their heads [8]. Collecting ear canal feature requires deploying extra hardware [5]. Even worse, some intracorporal biometrics are not stable, e.g., ECG and PPG are susceptible to human motion and emotion changes [9]. Therefore, utilizing the intracorporal biometric for authentication urgently requests stable, accurate, and easy-to-operate methods for the feature collection and extraction.

Recent years have witnessed the pervasive implementation of inertial measurement units (IMUs) in portable devices. Among them, earphone has become one of the most ubiquitous individual computing devices [10]. For instance, WT2 plus earbud [11] has integrated neural networks to realize real-time language translation. With these observations, we aim to explore a new biometric inside human body, which can be stably captured by the earphone's IMU, to achieve secure and accurate user authentication. Such an authentication system can also serve as a trusted portable login device to securely connect with other devices, such as smartphones, smart appliances, and autonomous vehicles.

- *Jianwei Liu, Jinsong Han (corresponding author) and Kui Ren are with Zhejiang University, China, ZJU-Hangzhou Global Scientific and Technological Innovation Center, China, and Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province, China. Email: hanjinsong@zju.edu.cn.*
- *Wenfan Song is with Zhejiang University, China, and ZJU-Hangzhou Global Scientific and Technological Innovation Center, China.*
- *Leming Shen is with Zhejiang University, China.*
- *Kui Ren is also with Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, China.*

In particular, it is well suitable for hands-free scenarios, e.g., driving and sports. Additionally, to enable secure communication in a session (e.g., online meeting) [12], we also attempt to leverage the new biometric to develop a continuous authentication function. Such a function is of significance because it can greatly enhance the security of an ongoing communication session. It is able to guarantee that each acoustic signal received by the microphone of the earphone originates from the legitimate user. This prevents the system from unpredictable attacks like voice command injection, which cannot be achieved by simple one-time verification [12].

However, to achieve these goals is non-trivial due to the following challenges. 1) It is difficult to find a brand-new biometric inside the human body to meet the user authentication requirements. 2) The sampling rate of common IMU is extremely low (not more than 500Hz [13]) and the raw IMU data contains too much noise, constraining the distinguishability of collected biometrics. 3) The computational capability of earphone is limited, but continuous authentication requires frequent calculations. This may cause users to be unable to communicate in a normal speed in conversation.

In this paper, we propose a new biometric-based authentication system, namely *MandiPass*. *MandiPass* is based on a new intracorporal biometric, termed as *MandiblePrint*, which is extracted from the vibration signal of user's mandible. *MandiPass* collects *MandiblePrint* via the IMU embedded in the earphone [14], [15], [16]. For device login, the user that wears the earphone only needs to make an 'EMM' sound for a very short time. The vibration generated by the throat will propagate through the mandible component, reach the ear, and finally be captured by the IMU for verification. In a session that needs continuous authentication, the user can speak normally. Each sentence will be authenticated quickly.

Specifically, to validate the feasibility of *MandiblePrint*, we build a one degree-of-freedom theoretical model and conduct a vibration propagation experiment. Moreover, to deal with the problem of low sampling rate and inferior quality of IMU data, we perform a series of denoising algorithms on raw IMU readings and leverage a two-branch deep neural network to extract high-distinguishability *MandiblePrint*. Finally, we leverage the knowledge distillation technique [17] to compress our deep neural network, so as to enable computation-friendly continuous authentication.

We invite 34 participants to carry out comprehensive experiments. The results show that *MandiPass* is highly accurate in one-time user verification and continuous authentication. The results also demonstrate the effectiveness and robustness of *MandiPass* in real-world scenarios, including eating food and performing different activities.

In summary, our contributions are as follows.

- We propose a secure and user-friendly biometric-based authentication system, *MandiPass*, which supports both one-time verification and continuous authentication. It leverages a brand-new intracorporal biometric, *MandiblePrint*, to identify individuals.
- We build a theoretical model to prove the validity of *MandiblePrint*. We also develop a series of deep learning techniques to extract high-distinguishability *MandiblePrint*.
- We implement a prototype of *MandiPass* and conduct experiments with 34 volunteers. The experimental results show that *MandiPass* is robust and secure, with a low equal error rate (EER).

2 Feasibility Study

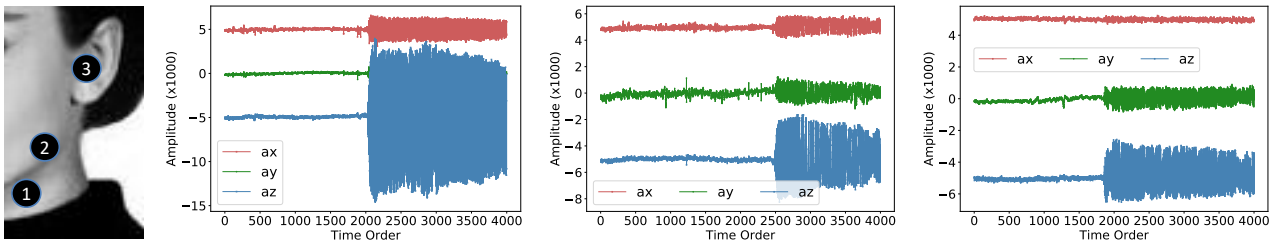
In this section, we first validate that the vibration produced by the throat can pass through the mandible before reaching the ear, which enables *MandiPass* to capture vibration signals containing mandible characteristics at the earphone. Then, a theoretical model is built to study the feasibility of extracting person-distinguishable biometrics from vibration signals.

2.1 Vibration Propagation Path

MandiPass employs IMUs to capture the desired biometric. A typical IMU contains two components, an accelerometer and a gyroscope. Each component has three axes (x, y, and z) of vibration information, which are time-series real numbers. The x-, y-, and z-axis of the accelerometer are represented by ax, ay, and az, respectively. Likewise, gx, gy, and gz respectively represent the x-, y-, and z-axis of the gyroscope. To validate that the vibration indeed propagates from the throat to the ear and can be eventually captured by an IMU, we conduct the following experiment. We first attach IMUs to three different locations on a volunteer's head, i.e., to a volunteer's throat, mandible, and ear (shown in Fig. 1(a)). Next, we ask the volunteer to keep silent for a while and then pronounce 'EMM' to collect the vibration signal. As shown in Fig. 1(b), the standard deviation of az is high at the throat location, indicating that the vibration is drastic at the throat. When the vibration propagates along the mandible, the standard deviation of az becomes lower, as shown in Fig. 1(c). At the location of the ear, as shown in Fig. 1(d), we find the lowest value of the standard deviation. The experiment results demonstrate that the vibration generated by the throat can propagate along the path 'throat-mandible-ear', although with a strength decay. During the propagation, the vibration first passes from the throat to the mandible, and then from the mandible to the ear. Moreover, since vibration fades slower in medium with larger density [18], and the density of bone is much larger than that of air and other tissues in human body, the collected vibration signals are mainly composed of vibration components propagating through the mandible. Therefore, the collected vibration signals contain the biometrical feature of the mandible, which is unique to a specific user.

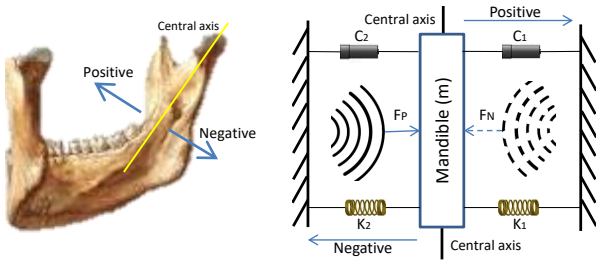
2.2 Theoretical Model

We model the mandible's vibration based on its physiological structure. When the mandible starts to vibrate, a vibration period can be divided into two phases according to the moving direction of the mandible: positive-direction vibration and negative-direction vibration. These two phases appear alternately. We illustrate our one degree-of-freedom vibration model in Fig. 2. To simplify the model, we neglect



(a) Three attaching locations. (b) Location 1: throat. The standard deviation of *az* is 3805. (c) Location 2: mandible. The standard deviation of *az* is 1050. (d) Location 3: ear. The standard deviation of *az* is 761.

Fig. 1. Standard deviations of vibration signals at three locations on user's face.



(a) (b)

Fig. 2. Vibration model of mandible component.

the procedure that the mandible moves from the outer vibration boundaries to the central axis (shown in Fig.2(b)).

In this model, m is the mass of the mandible. c_1 and c_2 are the damping factors of the two dampers. k_1 and k_2 are the two coefficients of elasticity of the two springs. The vibration resistance, i.e., the dampers and springs, is introduced by the tissues (*e.g.*, muscle and fat) surrounding the mandible. Apparently, the tissues on both sides of the mandible are not symmetrical, we thus have $c_1 \neq c_2$ and $k_1 \neq k_2$.

In the positive-direction phase, the two springs and damper c_1 hinder the positive-direction motion of the mass. Meanwhile, making the mandible vibrate is equivalent to applying a force on the mandible component. Suppose that the positive-direction force caused by the throat vibration is $F_P(t)$. According to Newton's second law, we have:

$$F_P(t) = mx^{''}(t) + c_1 x^{'}(t) + (k_1 + k_2)x(t), \tag{1}$$

where $x(t)$ is the positive-direction displacement of the mass. After performing Fourier transform and term transposition, we have:

$$X_P(w) = \frac{1 - e^{-iw\Delta t}}{-\frac{imw^3}{F_P(0)} - \frac{c_1 w^2}{F_P(0)} + \frac{i(k_1+k_2)w}{F_P(0)}}, \tag{2}$$

where w, $X_P(w)$, i, and $F_P(0)$ are the frequency component, the spectrum of the vibration signal, the imaginary component, and the constant positive-direction force induced by the positive-direction vibration of the throat, respectively.

If we denote the vibration propagation attenuation coefficient, the propagation distance from the throat to the ear, and the received positive-direction spectrum at the ear as α, d, and $Y_P(w)$ respectively, we obtain the following formula according to [19]:

$$Y_P(w) = X_P(w)e^{-\alpha d}. \tag{3}$$

Through replacing the term $X_P(w)$ in Eq. 3 with the right side of Eq. 2, we have:

$$Y_P(w) = \frac{e^{-\alpha d} - e^{-iw\Delta t - \alpha d}}{-\frac{imw^3}{F_P(0)} - \frac{c_1 w^2}{F_P(0)} + \frac{i(k_1+k_2)w}{F_P(0)}}. \tag{4}$$

Likewise, the received negative-direction spectrum can be formulated by:

$$Y_N(w) = \frac{e^{-\alpha d} - e^{-iw\Delta t - \alpha d}}{-\frac{imw^3}{F_N(0)} - \frac{c_2 w^2}{F_N(0)} + \frac{i(k_1+k_2)w}{F_N(0)}}. \tag{5}$$

Thus, the vibration of a complete period, which equals to $Y_P(w) \cup Y_N(w)$, can be formulated as:

$$\begin{aligned} Y(w) = &\frac{e^{-\alpha d} - e^{-iw\Delta t_1 - \alpha d}}{-\frac{imw^3}{F_P(0)} - \frac{c_1 w^2}{F_P(0)} + \frac{i(k_1+k_2)w}{F_P(0)}} \\ &\cup \frac{e^{-\alpha d} - e^{-iw\Delta t_2 - \alpha d}}{-\frac{imw^3}{F_N(0)} - \frac{c_2 w^2}{F_N(0)} + \frac{i(k_1+k_2)w}{F_N(0)}}, \end{aligned} \tag{6}$$

in which $\Delta t_1 + \Delta t_2$ equals to the time interval of a vibration period. m, c_1, $c2$, k_1, and k_2 vary among different persons [20]. Although $F_P(0)$, $F_N(0)$, Δt_1, and Δt_1 are identity-irrelevant noise components, they are relatively stable for a specific person when s/he only produces a single-tone sound 'EMM', because human's speaking habit and vocal frequency remain stable after puberty [21]. Hence, the received vibration signals, which record the characteristics of mandible, contain sufficient biometrics (*i.e.*, m, c_1, $c2$, k_1, and k_2) and are potential to be utilized to identify individuals. In this paper, we extract these biometrics, which are termed as *MandiblePrint*, both from positive-direction and negative-direction vibration signals to achieve accurate authentication.

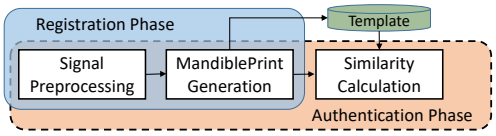


Fig. 3. Architecture of *MandiPass*.

3 System Design

In this section, we first introduce the overview of *MandiPass*, and then detail each module in *MandiPass*.

3.1 System Overview

As illustrated in Fig. 3, the architecture of *MandiPass* consists of two phases, i.e., registration phase and authentication phase. There are three modules in *MandiPass*: *signal preprocessing* module, *MandiblePrint generation* module, and *similarity calculation* module. The registration phase contains the first two modules, while the authentication phase contains all the three modules.

In the registration phase, *MandiPass*'s workflows for the login service and the continuous authentication service are the same. User needs to provide a segment of vibration signals to generate a cancelable template. Specifically, the user first pronounces 'EMM' for a short time to collect raw signals. Then, the identity-irrelevant components in the raw signals are removed by the *signal preprocessing* module. *MandiPass* obtains a 'clear' signal array from this module. Afterwards, the *MandiblePrint generation* module extracts a *MandiblePrint* vector from the signal array. The obtained *MandiblePrint* vector is then multiplied by a Gaussian matrix and becomes a cancelable one. Finally, the cancelable *MandiblePrint* vector is deemed as a *MandiblePrint* template and stored in the secure enclave [22] of the earphone.

In the authentication phase, the user can opt to use the login verification service or the continuous authentication service. In the former, the user initiates a verification request by pronouncing 'EMM' for a short time. Then, the collected raw signals are successively processed by the *signal preprocessing* module and the *MandiblePrint generation* module. After that, the obtained *MandiblePrint* vector and the *MandiblePrint* template stored in the secure enclave are utilized to calculate a similarity in the *similarity calculation* module. If the similarity is larger than a threshold we set in advance, the verification request will be accepted. Otherwise, the verification request will be regarded as an illegitimate one and thus rejected. In the continuous authentication service, the collected vibration signals also go through these modules successively. However, the inner algorithms (e.g., the neural network used for *MandiblePrint* extraction) of these modules in the continuous authentication service are different from that in the verification service. We will detail the differences in the following sections. Note that verification and continuous authentication will go through the same algorithms unless specifically stated.

3.2 Role of Module

The inner algorithms of the *signal preprocessing*, *MandiblePrint generation*, and *similarity calculation* modules (as shown in Fig. 4) are elaborated in this part.

Signal preprocessing: This module is used to remove identity-irrelevant components from raw signals. To this end, *MandiPass* needs to perform four operations. First, *MandiPass* detects the start/end timestamp of the vibration event. The signal of each axis is segmented based on these two timestamps. Then, the outliers caused by hardware imperfection and body motion are localized by *MandiPass*. These outliers will be replaced by the mean values of their adjacent normal values. After that, *MandiPass* leverages a high pass filter to remove the noise caused by human movements. Finally, the signal is normalized and the signal

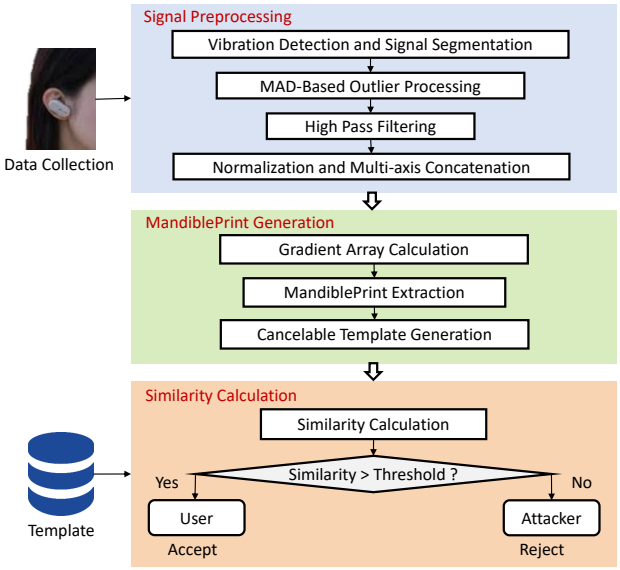


Fig. 4. Workflow of *MandiPass*.

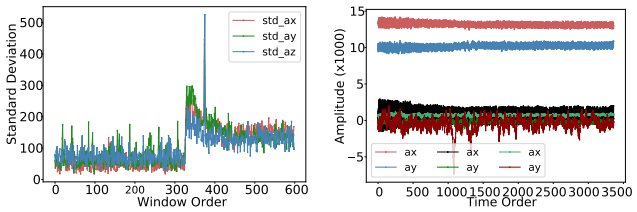
values in each axis are concatenated together to form a two-dimensional signal array. The details of each operation are introduced in Section 4.

***MandiblePrint* generation:** This module primarily contains three operations and *MandiPass* obtains a cancelable *MandiblePrint* vector after the three operations. First, *MandiPass* calculates gradients for each axis of signals in the signal array. A gradient array that contains both positive-direction and negative-direction vibration features is obtained through this operation. Afterwards, the gradient array is fed into a feature extractor (a deep neural network) and the biometric extractor outputs a vector, i.e., *MandiblePrint*. Finally, the *MandiblePrint* vector is multiplied by a Gaussian matrix to get a cancelable one. The design of our biometric extractor and the generation method of the cancelable *MandiblePrint* vector are elaborated in Section 5, 6, and 7.

Similarity calculation: In the verification service, *MandiPass* calculates the cosine distance [23], i.e., the similarity, between the cancelable *MandiblePrint* vector obtained from a verification request and the stored cancelable *MandiblePrint* template. If the similarity is larger than the threshold, it means that the verification request is initiated by the authentic user. The verification request is thus accepted. Otherwise, *MandiPass* rejects the verification request because it is likely to be initiated by an illegitimate user. In the continuous authentication service, *Mandipass* also calculates the similarity between the cancelable *MandiblePrint* template and the *MandiblePrint* vector from a signal sample (corresponding to an acoustic signal sample) collected in a conversation. If the similarity is larger than the threshold, it proves that the acoustic signal sample originates from the user. Otherwise, the acoustic signal sample is deemed from an attacker.

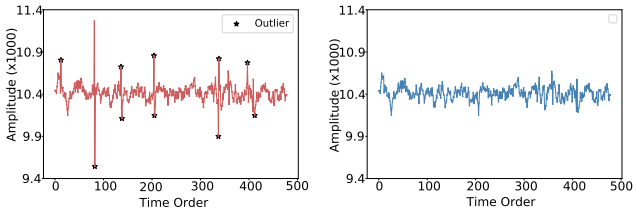
4 Signal Preprocessing

Vibration detection and signal segmentation: To obtain the signal segment that records mandible vibration, we need to find the start timestamp of the vibration in the raw signal. Since the mandible vibration would make the signal values



(a) Standard deviation becomes large when vibration starts. (b) Start values of different axes are different.

Fig. 5. Signal standard deviations and start values.

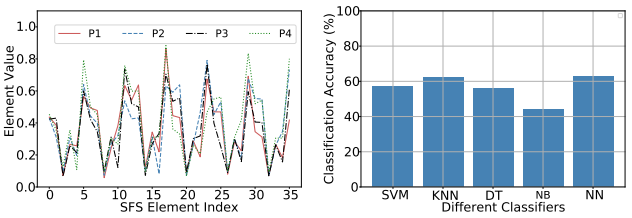


(a) All outliers are detected. (b) Outlier-replaced signal.

Fig. 6. All outliers are replaced with means.

(in each axis, each timestamp corresponds to a signal value) change drastically, which means that the standard deviation of a certain number of continuous signal values would become large, we determine the start timestamp according to the standard deviation. Specifically, we first divide the accelerometer signal values into windows and then calculate the standard deviation of each window. Each window has ten continuous signal values and the slide stride is also ten. As shown in Fig. 5(a), if the standard deviation of a window is larger than 250 and the standard deviations of the subsequent windows are not lower than 100, the vibration is regarded to start at this window. In particular, we consider the timestamp of the first signal value of this window as the start timestamp of the vibration event. The end timestamp is determined in a similar way. Next, if the user chooses the verification service, *MandiPass* selects n continuous signal values behind the start timestamp for each axis to get six signal segments. If the user chooses the continuous authentication service, for each axis, *Mandipass* will further divide the signal between the start timestamp and the end timestamp into k segments. Each segment has n continuous signal values and we totally obtain $6 \times k$ segments.

MAD-based outlier processing: Due to the hardware imperfection of IMU and motion noise (*e.g.*, 'walk'), the collected raw signals may have some values that are extremely large or small. These values should be regarded as outliers. To deal with these outliers, we first detect them by an MAD [24] algorithm, and then replace them with the means of normal values. To be specific, we first utilize the MAD outlier detection method to detect all outliers in each signal segment alternatively. As shown in Fig. 6(a), all outliers are found (marked by stars) in a segment, which demonstrates that the MAD algorithm is effective. Afterwards, in order to eliminate the impact of outliers, we perform a two-step mean-based outlier replacing on each signal segment, in which we replace each outlier with the mean of its two previous normal values and two subsequent normal values. The replacing result, shown in Fig. 6(b), proves that our two-step mean-based outlier replacing method is effective.



(a) SFSes of different users are similar. (b) Accuracy obtained by using statistical features with different classifiers.

Fig. 7. SFS can only achieve low classification accuracy.

High pass filtering: Since human activities may generate low-frequency components (LFCs) that are irrelevant to the *MandiblePrint*, we need to filter these LFCs out. According to the research in [19], the frequency components mostly are less than 10Hz during the body movements. Given the fact that normal people's fundamental frequency of vocal vibration varies from 100Hz to 200Hz [25], a high pass filter is needed to preserve the high-frequency components. Therefore, we utilize a high pass four-order Butterworth filter [26] with a cutoff frequency of 20Hz to remove the LPCs from each signal segment alternately.

Normalization and multi-axis concatenation: It is noteworthy that the start values of different axes are different, i.e., the elements of some axes oscillate around large values while that of other axes oscillate around small ones, as shown in Fig. 5(b). If we directly use un-normalized signals to extract *MandiblePrint*, the contribution of these axes (the values of which are small) would be concealed. Thus, we normalize the signal values through min-max normalization. For each signal segment, the normalized value x_n of each original value x_o can be calculated by:

$$x_n = \frac{x_o - x_{min}}{x_{max} - x_{min}}, \tag{7}$$

where x_{max} and x_{min} are the maximum and minimum values in this signal segment. Moreover, to make full use of captured signals of six axes and provide dimension-consistent input for our biometric extractor, we concatenate the signal segments of all axes. In this way, we obtain a signal array with dimensionality of $(6, n)/(6, k, n)$ in the verification/continuous authentication service. We empirically set n to 60 while k depends on the length of the corresponding acoustic signal.

5 *MandiblePrint* Extraction for Verification

In this section, we aim to extract person-distinguishable *MandiblePrint* from the signal array. However, our preliminary experiments show that calculating statistical features is infeasible to extract *MandiblePrint*. We thereby design novel deep learning models to extract high-quality *MandiblePrint*.

5.1 Infeasibility of Statistical Features

To extract *MandiblePrint*, traditional and intuitive solutions are to calculate some statistical features for each axis. Thus, we conduct a preliminary experiment to explore whether

the statistical features of different persons are distinguishable. Specifically, we first invite four volunteers and collect 500 signal arrays for each volunteer. In each signal array, we calculate six common statistical features (*i.e.*, mean, median, variance, standard deviation, upper quartile, and low quartile) for each axis. In this way, we obtain $6 \times 6 = 36$ statistical features for each signal array. Each set of 36 statistical features is called a statistical feature sample (SFS). We then randomly select a SFS for each volunteer and plot the selected four SFSes in Fig. 7(a), where one can find that it is hard to figure out the difference between different SFSes. Further, we label the four volunteers' SFSes by four integers from zero to three. By using 80% SFSes as the training set and the rest 20% ones as the testing set, we utilize four classic classifiers to perform classification: support vector machine (SVM), k-nearest neighbours (KNN), decision tree (DT), naive Bayes classifier (NB), and neural network (NN). The NN is composed of two fully-connected layers and two Sigmoid activation functions [27], where each fully-connected layer is followed by a Sigmoid function. The results in Fig. 7(b) indicate that even the highest classification accuracy is lower than 65%. Therefore, it is infeasible to use statistical features as the *MandiblePrint*.

5.2 Gradient Array Calculation

Since convolutional neural networks (CNNs) have shown excellent abilities of feature extraction [28], we attempt to design a CNN-based learning model to mine 'deep-hidden' *MandiblePrint* from signal arrays. Moreover, considering that different biometrics exist in positive-direction and negative-direction vibration signals (according to Eq. 6), we separately perform convolution on these two directions of signals.

To be specific, we first separate the positive- and negative-direction vibration signals by calculating gradients for each array in the signal sample. The i_{th} gradient of the j_{th} segment can be calculated by:

$$g_i^j = \frac{v_{i+1}^j - v_i^j}{|t_{i+1}^j - t_i^j|}, \quad i \in [1, n-1], \quad j \in [1, 6], \tag{8}$$

where v_i^j is the i_{th} signal value of the j_{th} segment, and $|t_{i+1}^j - t_i^j|$ is the normalized time interval between v_{i+1}^j and v_i^j. After calculating all gradients, we separate them according to their signs, i.e., the gradients that are larger than or equal to zero belong to the positive direction, and the rest gradients belong to the negative direction. In this manner, we obtain approximately $n/2$ gradients for each direction per segment. To provide dimension-consistent inputs for our CNN, we perform linear interpolation to make each direction have $n/2$ gradients. We finally obtain a gradient array with dimensionality of $(2, 6, n/2)/(2, 6, k, n/2)$ in the verification/continuous authentication service, where '2' means the two directions.

5.3 Biometric Extraction

After obtaining the gradient array, we design a two-branch CNN to extract *MandiblePrint* from the gradient array in the verification service. We noticed that the data structure of each axis is time-series values, thus it is reasonable

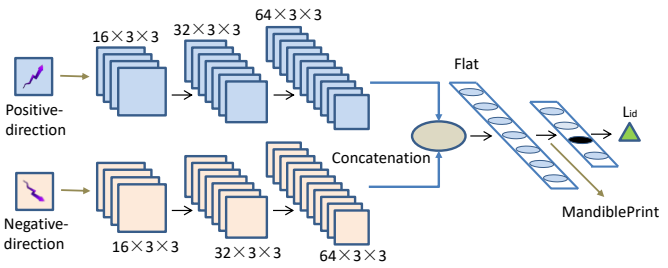


Fig. 8. Architecture of our biometric extractor for user verification.

to perform convolution on continuous gradients in each axis to extract temporal features. Meanwhile, since different axes contain different degree-of-freedom features, we also perform convolution among different axes to extract spatial features. Finally, the architecture of our biometric extractor is illustrated in Fig. 8. There are two convolutional branches responsible for extracting temporal-spatial features from the positive- and negative-direction gradients, respectively. Each convolution branch contains three convolutional layers and each of which is followed by a batch normalization (BN) function [29] and a rectified linear unit (ReLU) [30]. The size of each convolutional kernel is 3×3 and the stride size is 1×2. The BN is used to prevent data distribution from offset and the ReLU is used to decrease the inter-neuronal dependence. The BN and ReLU are simultaneously leveraged to improve the effectiveness and robustness of the biometric extractor. After the convolutional operation, we flatten the outputs of the two branches and concatenate them to obtain a feature vector. To prevent overfitting, a dropout layer [31] can be added here. The feature vector then passes through a fully connected layer and a Sigmoid function, and becomes *MandiblePrint*. The output of the Sigmoid function, i.e., *MandiblePrint*, is a biometric vector with dimensionality of $(1, 512)$. At last, a fully connected layer is used to project the biometric vector into different classes (*i.e.*, different person IDs), which enables us to train the biometric extractor through loss calculation and back propagation [32].

To make the biometric extractor learn to effectively extract *MandiblePrint*, we need to train it in a proper manner. However, it is noteworthy that users do not need to provide any vibration signal for the training process because the biometric extractor is trained by the service provider (SP) (*e.g.*, earphone manufacturer). To be specific, the SP can hire a large number of people to collect signal arrays. Then, these signal arrays are labeled and input into the biometric extractor in a unit of batch. The cross entropy [33] and Adam optimizer [34] can be utilized to calculate loss and update the parameters in the biometric extractor. Once the biometric extractor is well trained, it can be directly deployed on the earphone because it has possessed the ability of *MandiblePrint* extraction.

6 *MandiblePrint* Extraction for Continuous Authentication

To achieve continuous authentication in a conversation, we need to extract *MandiblePrint* from each gradient array corresponding to an acoustic signal sample. Intuitively,

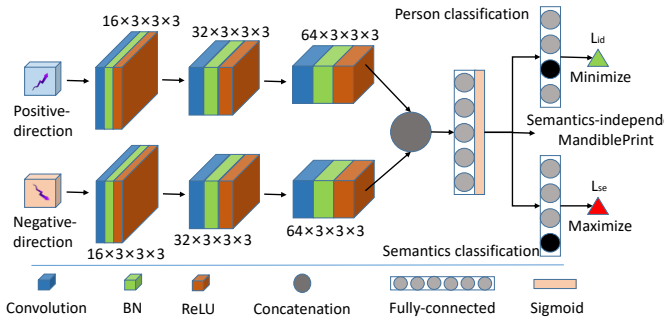


Fig. 9. Architecture of our three-dimensional adversarial network.

we can use the network designed in Sec. 5 to achieve this goal. However, there are two problems that prevent us from doing so. First, the dimensionality of a gradient array in continuous authentication service is $(6, k, n)$, which is three-dimensional. Two-dimensional CNN cannot be applied to extract features from three-dimensional input. Second, different from making an 'EMM' sound, the acoustic signal in a conversation has semantics and variety. Accordingly, the collected vibration is variable and semantics-specific. For example, the vibration introduced by speaking 'hello' is generally different from that introduced by speaking 'unlock the door'. Therefore, directly using CNN to extract biometrics will impact the person-uniqueness of *MandiblePrint*. In this section, we propose an adversarial network to extract semantics-independent *MandiblePrint* from three-dimensional gradient arrays. In addition, we utilize a knowledge distillation algorithm (i.e., model compression) to reduce the computational overhead used for *MandiblePrint* extraction. This guarantees that the delay induced by continuous authentication does not impact the normal conversation.

6.1 Three-dimensional Adversarial Network

To extract features from three-dimensional gradient arrays, we first replace all two-dimensional convolutional layers in Fig. 8 with three-dimensional ones. However, a new problem occurs when we try to flatten the output of the last convolutional layer. Since the second-dimension of the gradient array, i.e., k, is variable (varies with the time length the user spoke), the dimensionality of the output of the last convolutional layer is variable.as well. In this case, it is hard to fix the number of neurons of the first fully-connected layer. To solve this problem, we average the second dimension of the output of the third convolutional layer. With this treatment, the second dimension of the output can be fixed to one.

Network architecture: The architecture of our three-dimensional adversarial network is shown in Fig. 9. The kernel size and sliding stride of all convolutional layers are (3,3,3) and (1,2,2), respectively. The output of the last convolutional layer is flattened and fed into the first fully-connected layer. Before the first fully-connected layer, we can add a dropout layer to prevent overfitting. The output of the first fully-connected layer is then fed into two different classifiers, one for person classification and the other for semantics classification.

Training process: To train this adversarial network, we need to collect vibration signals of different persons and different semantics. For example, we can collect the vibration signals of ten persons when they are speaking ten different numbers from 'zero' to 'nine'. Accordingly, each gradient array extracted from the vibration signal sample will have two labels: an identity label and a semantics label. We still use cross entropy as the classification loss. The person classification loss can be formulated as:

$$\mathcal{L}_{id} = -\sum_{\mathrm{i=1}}^{M^{id}} y_i \log(P_i), \tag{9}$$

where y_i is the indication variable, P_i is the probability that the gradient array belongs to identity label i and M_{id} is the maximum of the identity labels. Similarly, the semantics classification loss can be formulated as:

$$\mathcal{L}_{se} = -\sum_{\mathrm{s=1}}^{M^{se}} y_s \log(P_s). \tag{10}$$

During training, we aim to minimize the person classification loss, so that the network can learn to extract person-distinguishable *MandiblePrint* from gradient arrays. On the contrary, for the semantics classification loss, we try to maximize it to make the network learn to discard semantics-relevant features, such that the extracted *MandiblePrint* is semantics-independent. As a result, the final loss L_f becomes:

$$\mathcal{L}_f = \alpha L_{id} - \beta L_{se}, \tag{11}$$

in which α and β are two hyper-parameters. Empirically, we set them to 0.8 and 0.2, respectively.

6.2 Network Compression

With the three-dimensional adversarial network, we can extract semantics-independent *MandiblePrint* from vibration signals collected in a conversation. Nevertheless, a continuous authentication service should have outstanding real-time performance; otherwise, the delay caused by continuous authentication will affect the normal communication. Unfortunately, our adversarial network is relatively deep (with relatively large computational overhead), so the extraction of *MandiblePrint* will introduce some delay indispensably. If we remove some layers from the network directly, the delay would be reduced. But the quality of extracted *MandiblePrint* is greatly impacted at the same time, which is unacceptable. To reduce the authentication delay without affecting the quality of the *MandiblePrint* too much, we leverage model compression algorithm to reduce the computational overhead.

To be specific, there are two networks in this algorithm, i.e., a teacher network and a student network. As shown in Fig. 10, the teacher network is composed of the biometric extractor and the identity classifier of the adversarial network. The student network only has one three-dimensional convolutional layer and two fully-connected layers. The teacher network will be trained at first. Then, with a special loss function, the teacher network can teach the student network to extract person-distinguishable and semantics-independent *MandiblePrint* from gradient arrays. The loss

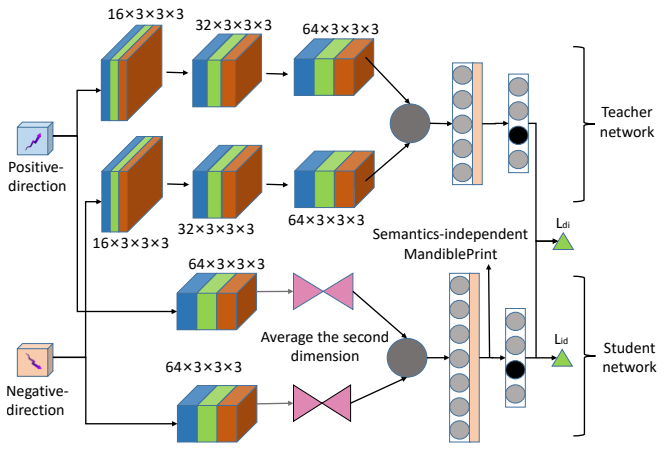


Fig. 10. Knowledge distillation: teacher-student network.

function, consisting of identity classification loss L_{id} and distillation loss L_{di}, can be formulated as:

$$L_{TS} = (1-\gamma)L_{id} + \gamma L_{di}, \tag{12}$$

where γ is a hyper-parameter (empirically set to 0.5). L_{di} can be calculated by:

$$L_{di} = KL(log_{softmax}(\frac{P_{stu}}{T}, softmax(\frac{P_{tea}}{T}))), \tag{13}$$

where $KL(\cdot)$, P_{stu}, P_{tea}, T are Kullback-Leibler divergence [35], student network's output, teacher network's output, and a hyper-parameter (empirically set to 2), respectively.

After the training of the teacher-student network, the convolutional layer and the first fully-connected layer can be employed as the final biometric extractor. This training process realizes the compression from the teacher network to the student network. The performance of the student will approach that of the teacher network.

7 Security Enhancement

It is critical to analyze the security of an authentication system. In this section, we first consider four main and potential attacks, and then discuss the defense methods against them.

7.1 Attack Model

Zero-effort attack: In this attack, we assume that the attacker has no awareness of *MandiPass*'s principle. The attacker steals the victim's earphone and attempts to use it to conduct authentication.

Vibration-aware attack: In this attack, our assumption is that the attacker knows the principle of *MandiPass*. The attacker attempts to produce a vibration signal to deceive *MandiPass*.

Impersonation attack: In this attack, we assume that the attacker first observes the verification process of the victim. Then, the attacker mimics the pronouncing manner of the victim to launch the impersonation attack.

Replay attack: Since the vibration propagates inside the human body, it is difficult for the attacker to eavesdrop on vibration signals. We assume that the replay attacker steals the *MandiblePrint* template stored in the secure enclave and exhibits it to *MandiPass* to launch the replay attack.

7.2 Defense

Zero-effort attack analysis: Since user needs to produce a short-time vibration to perform verification in *MandiPass*, the attacker who is not aware of this principle cannot provide signal array to *MandiPass*. Thus, the attacker cannot be successfully verified, which means that *MandiPass* is capable of defending against zero-effort attacks.

Vibration-aware attack analysis: In *MandiPass*, user is accepted if and only if her/is provided *MandiblePrint* is similar to the template stored in the secure enclave. The attacker is unable to provide such similar *MandiblePrint*, leading the attack to fail. Hence, *MandiPass* can defend against vibration-aware attacks.

Impersonation attack analysis: Even if the attacker is able to mimic the pronouncing manner of the victim, her/is *MandiblePrint* is still dissimilar to the victim's one, resulting in the calculated similarity smaller than the threshold. Therefore, the attack will fail and *MandiPass* is also able to defend against impersonation attacks.

Replay attack defense: To prevent *MandiPass* from replay attacks, we leverage a Gaussian matrix [9] to generate cancelable *MandiblePrint* template. Specifically, the *MandiblePrint* template is transformed by a Gaussian matrix before being stored in the secure enclave in the registration phase. The transformed *MandiblePrint* template is called cancelable *MandiblePrint* template. Let G be a Gaussian matrix and x be a *MandiblePrint* vector. A transformed *MandiblePrint* can be denoted by x' with $x' = x \times G$. In each verification/continuous authentication request, the new extracted *MandiblePrint* vector is also transformed as a cancelable one before similarity calculation. In this way, once the cancelable *MandiblePrint* template is stolen, the user can change Gaussian matrix used for transformation, such that the similarity between two *MandiblePrint* vectors transformed by different Gaussian matrices would be smaller than the threshold. The replay attacker, who does not know the changed Gaussian matrix, cannot pass the verification/continuous authentication when exhibiting the old template to *MandiPass*. Besides, the attacker cannot calculate the Gaussian matrix by only using the stolen template, which makes the transformation procedure secure. Meanwhile, legitimate authentication would not be impacted since the similarity between two *MandiblePrint* vectors transformed by the same Gaussian matrix is still high enough.

8 Evaluation And Result

We realize *MandiPass* with off-the-shelf devices and conduct extensive experiments to evaluate its performance in real-world environments.

Experiment setup: As shown in the left part of Fig. 11, we build a prototype of *MandiPass* on a Raspberry Pi [36]. This gadget allows us to access the IMU raw data. We use an Arduino Uno board [37] to control the signal collection. While collecting signals, the IMU is attached to the ear by adhesive tapes and covered by a normal earphone cover. We employ two types of IMU, i.e., *MPU-9250* and *MPU-6050*. to conduct experiments. In the default setting, we use *MPU-9250* IMU. The basic frequency of the Raspberry Pi CPU is 160Hz, which is the same as the one in WT2 earbuds and can be achieved by earphone mainboard. The *MandiblePrint*

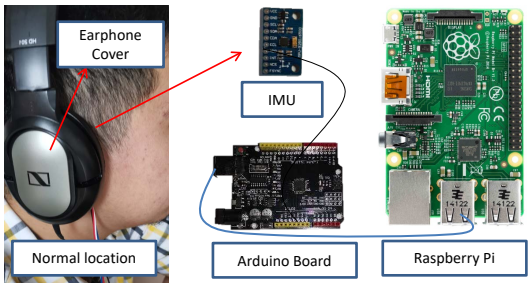


Fig. 11. Experiment setup for *MandiPass* evaluation.

extractors for user verification and continuous authentication are implemented based on *PyTorch* framework [38]. The batch size, learning rate, and number of epochs are set to 100, 0.001, and 10, respectively. All the experiments are conducted by adhering to the approval of our university's Institutional Review Board (IRB).

Data collection: To evaluate the performance of user verification, we totally invite 34 volunteers (28 males and 6 females) aged from 20 to 45 to participate in our experiments. We collect 23408 signal arrays for overall performance evaluation and each participant provides at least 500 signal arrays. We also collect over 11200 signal arrays in the extensive experiments to evaluate the robustness and security of *MandiPass*'s verification function. In the continuous authentication assessment, we invite 10 volunteers (6 males and 4 females) to collect 5000 (10 persons $\times$ 10 numbers $\times$ 50 instances) signal arrays when they are speaking different numbers from zero to nine. Since each instance lasts at least one second (at least five times that of a verification instance's duration), each volunteer provides the vibration signals associated with 500-second acoustic signals.

Metrics: To evaluate the authentication performance quantitatively, we define four metrics: false reject rate (FRR), false accept rate (FAR), EER, and verification success rate (VSR). FRR is the probability that a legitimate user is falsely rejected. It can be represented by the ratio between the number of falsely rejected signal arrays and the number of all signal arrays. The lower the FRR is, the better performance *MandiPass* has. FRR can be calculated by:

$$\frac{\sum_{i=0}^{V}\sum_{j=0}^{N_i-1}\sum_{k=j+1}^{N_i}\mathbb{1}_{sim(S_i^j,S_i^k)<t}}{\sum_{i=0}^{V}\sum_{j=0}^{N_i-1}\sum_{k=j+1}^{N_i}\mathbb{1}}, \tag{14}$$

where V, t, and N_i are the number of volunteers, the threshold, and the number of signal arrays of the i_{th} volunteer, respectively. $\mathbb{1}$ equals to one. $\mathbb{1}_{sim(S_i^j,S_i^k)<t}$ equals to one when the similarity between the *MandiblePrint* vectors extracted from S_i^j and S_i^k is less than t. Otherwise, it equals to zero. The FAR is the probability that an illegitimate user is falsely accepted. It can be represented by the ratio between the number of falsely accepted signal arrays and the number of all signal arrays. The smaller the FAR, the better *MandiPass* is. FAR can be calculated by:

$$\frac{\sum_{i=0}^{V-1}\sum_{j=0}^{N_i}\sum_{k=i+1}^{V}\sum_{l=0}^{N_k}\mathbb{1}_{sim(S_i^j,S_k^l)\geq t}}{\sum_{i=0}^{V-1}\sum_{j=0}^{N_i}\sum_{k=i+1}^{V}\sum_{l=0}^{N_k}\mathbb{1}}. \tag{15}$$

EER is the value of FAR or FRR when FAR equals to FRR. It can be obtained by altering the threshold. The lower the EER is, the better *MandiPass* is. VSR is the probability that a legitimate user is successfully accepted. Higher VSR means better *MandiPass*. It can be calculated by:

$$VSR = 1 - FRR. \tag{16}$$

8.1 Overall Performance

8.1.1 Performance of User Verification

We first evaluate the performance of our biometric extractor by comparing the classification accuracy of different classifiers, i.e., SVM, NB, DT, KNN, NN, and biometric extractor (BE). We randomly select 80% signal arrays as the training set and the rest 20% ones as the testing set. The classification experiment is performed ten times and we use the mean of ten accuracy as the final classification result. The experiment results are shown in Fig. 12. It can be observed that our biometric extractor outperforms other classifiers. It achieves the largest classification accuracy of 90.54%. Therefore, our biometric extractor can effectively extract person-distinguishable mandible biometrics from gradient arrays.

To extract *MandiblePrint*, we treat 33 volunteers' signal arrays as the training set of hired people and extract the rest volunteer's (plays the role of the user) *MandiblePrint* vectors. In this way, we extract *MandiblePrint* vectors of all the volunteers alternatively. We first calculate the mean similarity of the same user and different users. The results indicate that the mean similarity between different *MandiblePrint* vectors of the same user is 0.4884 while that of different users is 0.7032. We then increase the threshold from 0.5 to 0.6 to calculate FAR and FRR. The experiment results are shown in Fig. 13. It can be found that when the threshold is 0.5485, the FRR equals to FAR, where we obtain the EER, 1.28%. The low EER demonstrates that *MandiPass* performs significantly well in user verification. In the following experiments, we fix the threshold to 0.5485.

To explore if the verification performance is fair to different genders and users, we randomly select five males and five females and calculate their VSRs. The experiment results, shown in Fig. 14, indicate that *MandiPass*'s performance is fair to different genders as well as different users with the same gender.

As aforementioned, we use two types of IMUs for *MandiPass* evaluation, we find that the EERs of *MPU-9250* and *MPU-6050* are 1.28% and 1.29%, respectively. There is no apparent EER difference between the two types of IMUs, which shows that *MandiPass* has outstanding device scalability.

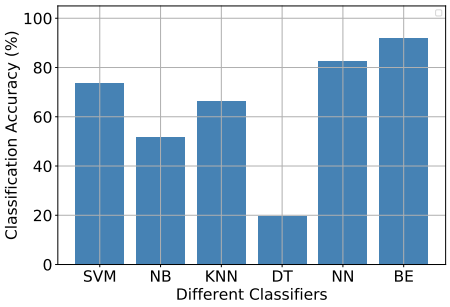


Fig. 12. Accuracy of different classifiers.

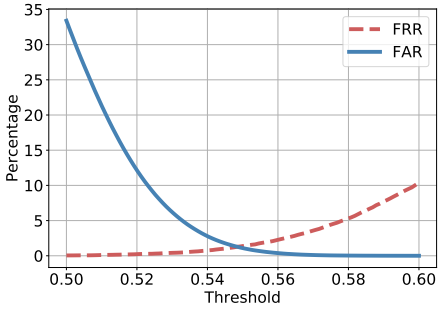


Fig. 13. FAR and FRR curves.

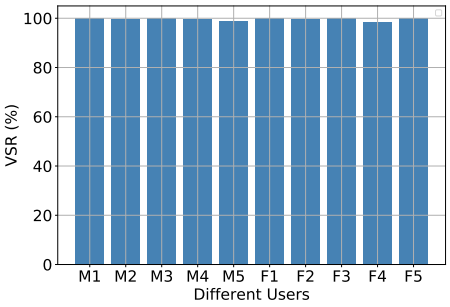


Fig. 14. VSRs of five males and five females.

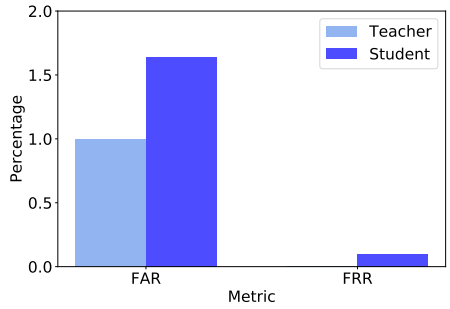


Fig. 15. FARs and FRRs of teacher and student networks.

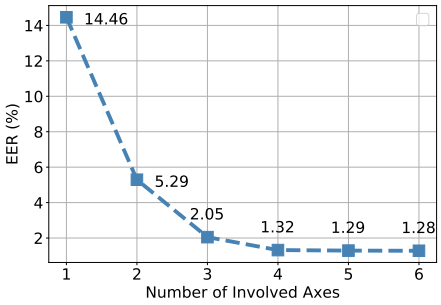


Fig. 16. Effect of number of axes on user verification.

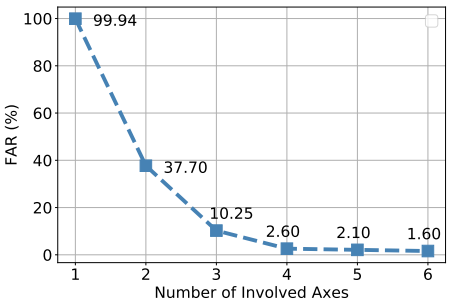


Fig. 17. Effect of number of axes on continuous authentication.

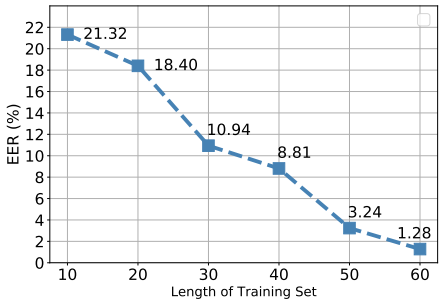


Fig. 18. Effect of training set length on user verification.

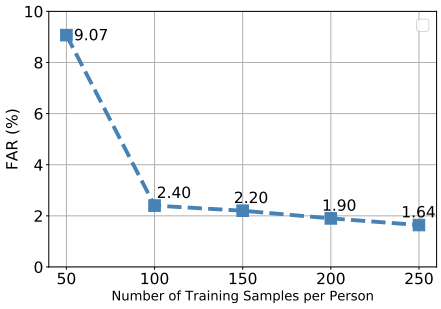


Fig. 19. Effect of training set size on continuous authentication.

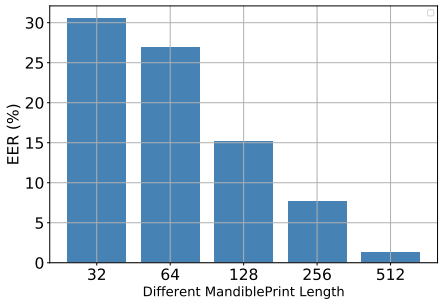


Fig. 20. Effect of length of *MandiblePrint* on user verification.

8.1.2 *Performance of Continuous Authentication*

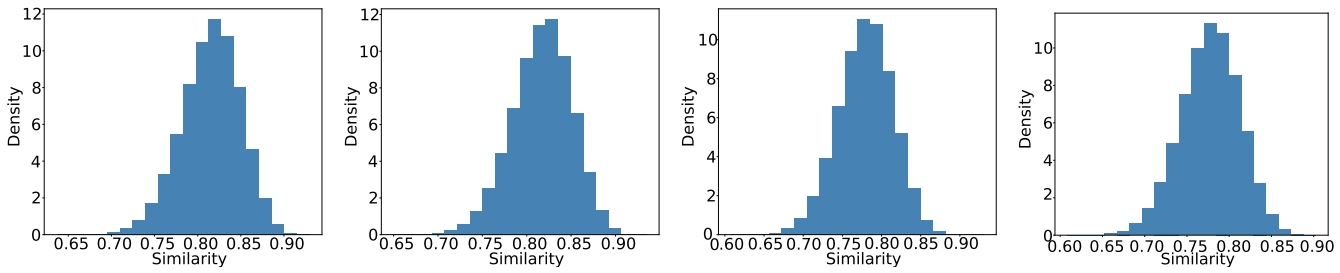
We first compare the classification accuracy before and after using the adversarial network in Fig. 9. We find that the accuracy is lower than 70% when we do not use the adversarial network. After using the adversarial network, the accuracy increases to 75.43%, which indicates that our adversarial network is effective in extracting semantics-independent *MandiblePrint*.

Then, with the feature extraction method used in the user verification experiment, we extract the semantics-independent *MandiblePrint* of all 10 volunteers. Specially, the signal arrays of zero to four are treated as training samples and that of six to nine are used as testing samples. In this way, it can be guaranteed that the semantics of the training set is different from that of the testing set. When the threshold is set to 0.5485, the FARs and FRRs of the teacher network and the student network (in Fig. 10) are shown in Fig. 15. It can be observed that the FAR and FRR of the teacher network are lower than that of the student network. However, the two FARs are close, so are the two FRRs. This demonstrates that our network compression method is effective. The student network can be used as the substitute of the teacher network to reduce the computational overhead.

8.2 Effect of System Settings

In this part, we evaluate the performance of *MandiPass* under different system settings, including the number of involved axes, the length of the training set, and the length of the *MandiblePrint* vector.

Effect of involved axes: In this experiment, we consider the axis order as '*ax*, *ay*, *az*, *gx*, *gy*, *gz*'. The involved axes are selected according to this order. For example, one axis means *ax*, two axes means '*ax*, *ay*', and so on. The verification results are shown in Fig. 16, which indicates that involving more axes can generate lower EER. Besides, using an accelerometer only can achieve an EER as low as 2.05%. Then, we explore the effect of the number of involved axes on continuous authentication. The experiment results are shown in Fig. 17. It can be observed that the FAR is as high as 99.94% when only one axis is used. Although the FRR



(a) Lollipop. (b) Water. (c) Walk. (d) Run.

Fig. 21. *MandiblePrint* similarity distributions when user is (a) eating lollipop, (b) drinking water, (c) walking, and (d) running.

is lower than 1% at this time, the FAR is too high to be acceptable. This result differs from that of the verification experiment. Fortunately, when the number of involved axes increases, the FAR decreases rapidly. When the three axes of the accelerometer are used only, the FAR decreases to 2.60%, making *MandiPass* secure. Thus, *MandiPass* can provide a good continuous authentication service when only accelerometer's measurements are available.

Effect of training set length/size: In user verification, the length of the training set is the time duration of collecting vibration signals for each hired person. We increase the length from 10 seconds to 60 seconds with stride of 10 seconds. As shown in Fig. 18, with the increase of the training set length, the EER keeps decreasing. When the length is 60 seconds, the EER achieves 1.28%. Therefore, collecting one-minute vibration signals for each hired person is sufficient to train the biometric extractor. As for continuous authentication, we vary the number of training samples per person from 50 to 250 and calculate the FAR and FRR. As a result, the FRR is always lower than 1% no matter how the number of training samples changes. By contrast, as shown in Fig. 19, the FAR decreases prominently with the increase of the training samples. When 100 training samples per person are available, the FAR is only 2.40%. Therefore, *MandiPass* can work well even if the training set size is small.

Effect of *MandiblePrint* length: It is worth noting that our default *MandiblePrint* length is 512. To explore if the *MandiblePrint* length affects *MandiPass*'s performance, we select other four commonly used biometric lengths: 32, 64, 128, and 256. The verification results shown in Fig. 20 indicate that the EER decreases with the increase of *MandiblePrint* length. When the length is 512, the EER is less than 1.5%. Thus, it is reasonable to set the length of *MandiblePrint* as 512. Similar to the verification experiment, we calculate the FAR and FRR of continuous authentication when the length of *MandiblePrint* changes, The FARs of 32, 64, 128, 256, and 512 are 24.25%, 27.67%, 12.12%, 9.50%, and 1.64%, respectively. In general, the FAR decreases with the increase of the *MandiblePrint* length. When such length is 512, *MandiPass* can provide the best continuous authentication service.

8.3 Impacts of Related Factors

We also consider the impacts of four factors from users' daily life. We categorize the factors into two groups: food and activity.

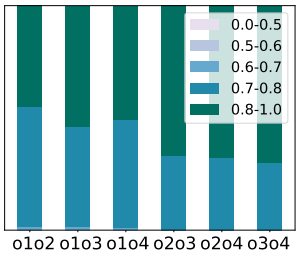


Fig. 22. Effect of IMU orientation on user verification.

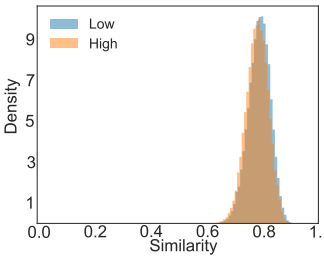


Fig. 23. Effect of pronouncing tone on user verification.

Food: We take the lollipop and water as the representatives of food since users may use *MandiPass* when they are eating food or drinking. We first conduct an extensive experiment with lollipops, in which we collect testing signal arrays with lollipops in users' mouths. The similarity distribution shown in Fig. 21(a) indicates that lollipop has negligible impacts on MandiPass, because all the similarity between the normal signal arrays (without lollipop) and the testing signal arrays (with lollipop) are larger than the threshold. Likewise, we conduct another extensive experiment with water. The similarity distribution shown in Fig. 21(b) proves that water also has negligible impact on *MandiPass* (the VSR is larger than 99%).

Activity: To assess the robustness of *MandiPass* towards human activity, we ask volunteers to walk or run while collecting testing signal arrays. We then calculate the similarity between the normal signal arrays (static) and the testing signal arrays (moving). The similarity distributions shown in Fig. 21(c) and Fig. 21(d) indicate that activity does not affect the performance of *MandiPass*. Thus, *MandiPass* is robust against motions.

8.4 Effect of Orientation and Tone

Since the orientation of the earphone and the tone of pronouncing may affect the performance of *MandiPass*, we also evaluate *MandiPass*'s performance with different orientations and tones.

Orientation of IMU: To explore the effect of orientation, we collect four groups of signal arrays and the gap between any two continuous groups is 90 degrees. We then calculate the similarity distributions of signal arrays between any two groups. The results are shown in Fig. 22, which indicate that the similarity of any two signal arrays with different orientations is still higher than the threshold. Therefore, *MandiPass* is robust against the orientation variation.

Tone of pronouncing: Even if we recommend users to produce 'EMM' sound naturally, users may change their tones unconsciously during authentication, which may further impact the EER of *MandiPass*. Hence, we ask volunteers to raise or lower their tones intentionally when collecting signal arrays in this experiment. Then, we calculate the similarity distributions between normal signal arrays (normal tone) and tone-changed ones (high or low tone). The results shown in Fig. 23 indicate that even with a high or low tone, users can still be successfully verified with a high similarity, which proves that *MandiPass* is robust against tone variation as well.

8.5 Overhead

Time cost: The time cost of *MandiPass* for processing an authentication request mainly comes from three components: vibration signal collection, signal preprocessing, and *MandiblePrint* extraction. First, user needs to pronounce 'EMM' for a short time to collect vibration signals, which costs 0.2 $(60 \div 350)$ seconds. Second, With the same CPU frequency of WT2 earbud, the signal preprocessing costs less than 0.01 seconds. Finally, with the WT2 earbud's CPU frequency also, the biometric extractor outputs a *MandiblePrint* vector within 1 second on average. Therefore, *MandiPass* can process an authentication request within 2 seconds and it has outstanding real-time performance.

Storage consumption: The storage consumption of *MandiPass* comes from two components: biometric extractor storage and cancelable *MandiblePrint* template storage. First, the biometric extractor requires approximately 5MB to store its parameters. Second, a cancelable *MandiblePrint* template consumes about 1.8KB storage space. Therefore, the total storage consumption is less than 6MB, which is acceptable to an authentication system.

Computational overhead: To show the effectiveness of our network compression method, we compare the number of floating-point operations (FLOPs) of the teacher network and the student network in Fig. 10. The FLOPs of the teacher network and the student network are 12.95M and 8.67M, respectively. It is apparent that the student network needs to perform fewer floating-point operations. Hence, our model compression method can reduce the computational overhead effectively.

8.6 Long-Term Observation:

To validate if *MandiPass* can still verify users with a high VSR after a long term, we randomly select six volunteers to conduct a validation experiment. Specifically, we first collect two batches of signal arrays at time t_1 and t_2, respectively. The time interval between t_1 and t_2 is two weeks. Then, we calculate the similarity between the *MandiblePrint* generated by signal arrays collected at t_1 and t_2. The experimental results show that the average VSR of these volunteers is larger than 99.5%. With the same experimental method but a different interval between t_1 and t_2, we collect the data of three volunteers to assess the long-term continuous authentication performance. The results show that the VSRs of all the volunteers are larger than 99% even after three months. Hence, *MandiblePrint* is stable and *MandiPass* is robust in long-term use.

TABLE 1
Comparing *MandiPass* with SkullConduct and EarEcho.

System	RTC$\leq$ 1s	FRR $\leq$ 2%	RARA	IAN
MandiPass	✓	✓	✓	✓
SKullConduct	✓	×	×	×
EarEcho	×	×	×	×

8.7 Security Assessment

As introduced in Section 7, we need to assess the security of *MandiPass* againsy four attack models. In the zero-effort attack experiment, we invite five volunteers (attackers) who do not know the principle of *MandiPass* to initiate authentication requests 20 times per attacker. As a result, the VSR for these attackers is 0%. In terms of the vibration-aware attack, the EER shows that the VSR for attackers is 1.28%. As for the impersonation attack, we first ask five volunteers (attackers) to observe the pronouncing manners of other five volunteers (victims). Then, we collect signal arrays with these attackers. After that, we calculate the similarity between attackers' *MandiblePrint* and victims' *MandiblePrint*. The experimental results show that the VSR for attackers is 1.30%. Finally, to assess the security of *MandiPass* against replay attacks, we calculate the similarity between cancelable *MandiblePrint* vectors transformed by different Gaussian matrices. The result, a VSR of 0.6%, indicates that nearly all replayed *MandiblePrint* vectors are rejected. Therefore, *MandiPass* can defend against these four types of attacks effectively.

8.8 Comparing with Existing Works

We compare *MandiPass* with two related works, i.e., SkullConduct [39] and EarEcho [5], in terms of the registration time cost (RTC), EER, replay attack resilience ability (RARA), and immunity against acoustic noise (IAN). SkullConduct is an acoustic signal-based authentication system collecting skull biometrics as the authentication credential, which can be deployed on GoogleGlass. EarEcho, a state-of-the-art earphone-based authentication system, collects ear canal biometrics to identify individuals. The comparing results are shown in Table 1. First, *MandiPass* and SkullConduct can finish the registration within one second, but EarEcho is unable to do that. Second, the FRR of *MandiPass* is lower than that of SkullConduct and EarEcho. Third, *MandiPass* can defend against replay attacks, while the other two systems cannot. Finally, *MandiPass* is immune to acoustic noise, but the other two systems are susceptible to acoustic noise. Thus, *MandiPass* outperforms SkullConduct and EarEcho.

9 Related Work

Authentications on wearable devices mainly fall into two categories: one-time verification [5], [9], [19], [40], [41] and continuous authentication [5], [12], [42], [43], [44]. The first category can be further divided into knowledge-based [40], [45], [46] and biometric-based [5], [9], [41] approaches. In the former, users usually need to remember some knowledge, such as password [45] and pattern [40]. For example, smartwatch can be unlocked through PIN input [45]. This category of authentication approaches are easy to operate,

yet vulnerable to shoulder-surfing attacks [4]. In consideration of the security and user-friendliness, biometrics are introduced into wearable device authentications, such as ECG [5], tapping characteristics [41], ear canal features [5], etc [47]. However, these biometrics are either unstable (e.g., susceptible to human emotion) or hard to collect (e.g., require extra hardware). Different from the previous biometrics, our proposed *MandiblePrint* is not only robust against outer factors like emotion, but also easy-to-collect. As for the continuous authentication, there are also non-biometric-based [12] and biometric-based methods [5], [42], [43], [44], [48], [49], [50], [51], [52]. The latter are relatively more secure because biometrics are relatively hard to be stolen/duplicated. But existing biometric-based continuous authentication methods have their respective drawbacks. For instance, leveraging ultrasound to probe ear canal information [42] could be impacted by acoustic noise. Continuously monitoring gait [49] consumes much power. Some in-body biometrics like PPG [44] are unstable. In this work, we propose *MandiPass* to overcome the above-mentioned shortcomings. *MandiPass* extracts stable *MandiblePrint* for authentication only when there is a security need, i.e., when an acoustic signal is received, so it saves power. Besides, it would not be impacted by acoustic noise.

IMU-based sensing on wearable devices is widely studied in recent years [46], [53], [54], [55], [56], [57], [58]. It has been implemented on various wearable devices (e.g., earphone [59], smartwatch [57], [60], etc [61]) to enable plentiful computer-human interaction tasks. For example, the IMUs on earbuds can be exploited for healthcare, such as cough detection [59]. In [62], the IMU is developed for motion measurement, e.g., localization. Similarly, the IMU on smartwatch can also be utilized for healthcare and motion measurement. The researchers in [60] leverage smartwatch IMU to guide the elder with dementia to do proper handwashing. In [58], the IMU on smartwatch helps the user to track her/is hand, achieving handwriting recognition. Different from previous works, we leverage earphone IMU to achieve both one-time verification and continuous authentication. To our best knowledge, this is the first work to extract biometric to realize continuous authentication using earphone IMU.

10 Conclusion

To realize secure and user-friendly biometric-based authentication, we propose *MandiPass*, which extracts biometrics from the vibration of user's mandible. The feasibility of *MandiPass* is validated via a rigorous theoretical model. We also introduce deep learning techniques to improve the efficiency and effectiveness of biometric extraction in both user verification and continuous authentication. The security of *MandiPass* is further enhanced via cancelable templates and transformation countermeasures. Extensive experiment results over 34 subjects indicate that *MandiPass* is highly accurate, robust, and secure in various environments.

Acknowledgement

This work is supported in part by National Key R&D Program of China (2020AAA0107705), National Natural Science Foundation of China under grant 62032021, 61872285, U21A20462, 61772236, and 61972348, Research Institute of Cyberspace Governance in Zhejiang University, Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005), Ant Group Funding No.Z51202000234, and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

References

[1] S. Rajarajan and P. Priyadarsini, "UTP: a novel PIN number based user authentication scheme," *International Arab Journal of Information Technology*, vol. 16, no. 5, pp. 904–913, 2019.

[2] P. Andriotis, G. C. Oikonomou, A. Mylonas, and T. Tryfonas, "A study on usability and security features of the android pattern lock screen," *Journal of Information and Computer Security*, vol. 24, no. 1, pp. 53–72, 2016.

[3] J. Liu, X. Zou, J. Han, F. Lin, and K. Ren, "BioDraw: Reliable multi-factor user authentication with one single finger swipe," in *IEEE/ACM International Symposium on Quality of Service (IWQoS)*, 2020.

[4] Y. Song, Z. Cai, and Z. Zhang, "Multi-touch authentication using hand geometry and behavioral information," in *IEEE Symposium on Security and Privacy (S&P)*, 2017.

[5] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "Earecho: Using ear canal echo for wearable authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 3, no. 3, pp. 81:1–81:24, 2019.

[6] W. Xu, J. Liu, S. Zhao, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "RFace: anti-spoofing facial authentication using cots rfid," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2021.

[7] K. N. R. K. R. Alluri and A. K. Vuppala, "IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermeasures challenge 2019," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[8] F. Lin, K. W. Cho, C. Song, W. Xu, and Z. Jin, "Brain password: A secure and truly cancelable brain biometrics for smart headwear," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2018.

[9] Y. Cao, Q. Zhang, F. Li, S. Yang, and Y. Wang, "PPGPass: Nonintrusive and secure mobile two-factor authentication via wearables," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2020.

[10] "Earphones: The next significant platform after smartphones," http://talks.cam.ac.uk/talk/index/128890, 2019.

[11] "Wt2 plus ai real-time translator earbuds," https://www.timekettle.co/products/wt2-plus, 2020.

[12] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2017.

[13] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer," in *Network and Distributed System Security Symposium (NDSS)*, 2020.

[14] APPLE, "The imu in airpods," https://www.apple.com.cn/airpods-pro/specs/, 2020.

[15] Cirmall, "The bma456 accelerometer in tws earphone," https://www.cirmall.com/articles/27711, 2020.

[16] HUAWEI, "The imu in huawei freebuds studio," https://consumer.huawei.com/cn/headphones/freebuds-studio/specs/, 2020.

[17] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[18] J. F. Hunt, H. Zhang, Z. Guo, and F. Fu, "Cantilever beam static and dynamic response comparison with mid-point bending for thin mdf composite panels," *BioResources*, vol. 8, no. 1, pp. 115–129, 2013.

[19] W. Chen, L. Chen, Y. Huang, X. Zhang, L. Wang, R. Ruby, and K. Wu, "Taprint: Secure text input for commodity smart wristbands," in *ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2019.

[20] S. WE, "The gross composition of the body," *Journal of Advances in Biological and Medical Physics*, vol. 4, no. 513, pp. 239–279, 1956.

[21] N. M. Meddy Fouquet, Katarzyna Pisanski and D. Reby, "Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood," *Journal of Royal Society Open Science*, vol. 3, 2016.

[22] J. POT, "What is apple's "secure enclave", and how does it protect my iphone or mac?" https://www.howtogeek.com/339705/what-is-apples-secure-enclave-and-how-does-it-protect-my-iphone-or-mac/, 2018.

[23] H. Zhang, X. Wang, and Z. He, "Weighted softmax loss for face recognition via cosine distance," in *Biometric Recognition - Chinese Conference (CCBR)*, 2018.

[24] Y. Li, Z. Li, K. Wei, W. Xiong, J. Yu, and B. Qi, "Noise estimation for image sensor based on local entropy and median absolute deviation," *Journal of Sensors*, vol. 19, no. 2, p. 339, 2019.

[25] M. Geiger, D. Schlotthauer, and C. Waldschmidt, "Improved throat vibration sensing with a flexible 160-ghz radar through harmonic generation," in *IEEE/MTT-S International Microwave Symposium*, 2018.

[26] "Butterworth filter: What is it? (design & applications)," https://www.electrical4u.com/butterworth-filter/.

[27] G. Mourgias-Alexandris, G. Dabos, N. Passalis, A. Tefas, A. Totovic, and N. Pleros, "All-optical recurrent neural network with sigmoid activation function," in *IEEE Optical Fiber Communications Conference and Exhibition (OFC)*, 2020.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[29] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: an empirical study of their impact to deep learning," *Multimedia Tools and Applications*, vol. 79, no. 19-20, pp. 12 777–12 815, 2020.

[30] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," in *6th International Conference on Learning Representations (ICLR)*, 2018.

[31] J. Liu, C. Xiao, K. Cui, J. Han, X. Xu, and K. Ren, "Behavior privacy preserving in rf sensing," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[32] A. Mukherjee, D. K. Jain, P. Goswami, Q. Xin, L. Yang, and J. J. P. C. Rodrigues, "Back propagation neural network based cluster head identification in MIMO sensor networks for intelligent transportation systems," *IEEE Access*, vol. 8, pp. 28 524–28 532, 2020.

[33] L. Li, M. Doroslovacki, and M. H. Loew, "Approximating the gradient of cross-entropy loss function," *IEEE Access*, vol. 8, pp. 111 626–111 635, 2020.

[34] X. Jiang, B. Hu, S. C. Satapathy, S. Wang, and Y. Zhang, "Fingerspelling identification for chinese sign language via alexnet-based transfer learning and adam optimizer," *Scientific Programming*, vol. 2020, pp. 3 291 426:1–3 291 426:13, 2020.

[35] ScienceDirect, "Kullback-leibler divergence," https://www.sciencedirect.com/topics/engineering/kullback-leibler-divergence, 2022.

[36] "Raspberry pi," https://www.raspberrypi.com/.

[37] "Arduino uno r3," https://docs.arduino.cc/hardware/uno-rev3.

[38] "Pytorch," https://pytorch.org/.

[39] S. Schneegass, Y. Oualil, and A. Bulling, "Skullconduct: Biometric user identification on eyewear computers using bone conduction through the skull," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.

[40] W. V. Vlaenderen, J. Brulmans, J. Vermeulen, and J. Schöning, "Watchme: A novel input method combining a smartwatch and bimanual interaction," in *ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*, 2015.

[41] H. Bi, Y. Sun, J. Liu, and L. Cao, "Smartear: Rhythm-based tap authentication using earphone in information-centric wireless sensor network," *IEEE Internet of Things Journal (IoTJ)*, vol. 9, no. 2, pp. 885–896, 2022.

[42] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 5, no. 1, pp. 39:1–39:27, 2021.

[43] T. Zhao, Y. Wang, J. Liu, and Y. Chen, "Your heart won't lie: Ppg-based continuous authentication on wrist-worn wearable devices," in *ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2018.

[44] 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 *IEEE Conference on Computer Communications (INFOCOM)*, 2020.

[45] "Set up a pin on huawei wearables to protect your privacy," https://consumer.huawei.com/eg-en/support/article/en-gb15867956/.

[46] A. Bianchi and I. Oakley, "Wearable authentication: Trends and opportunities," *Information Technology and Management*, vol. 58, no. 5, pp. 255–262, 2016.

[47] M. Shirvanian, S. Vo, and N. Saxena, "Quantifying the breakability of voice assistants," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2019.

[48] S. Mahto, T. Arakawa, and T. Koshinaka, "Ear acoustic biometrics using inaudible signals and its application to continuous user authentication," in *IEEE European Signal Processing Conference (EUSIPCO)*, 2018.

[49] W. Xu, Y. Shen, C. Luo, J. Li, W. Li, and A. Y. Zomaya, "Gait-watch: A gait-based context-aware authentication system for smart watch via sparse coding," *Ad Hoc Networks*, vol. 107, p. 102218, 2020.

[50] G. Cola, M. Avvenuti, F. Musso, and A. Vecchio, "Gait-based authentication using a wrist-worn device," in *ACM International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, 2016.

[51] A. H. Johnston and G. M. Weiss, "Smartwatch-based biometric gait recognition," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015.

[52] N. Al-Naffakh, N. L. Clarke, F. Li, and P. S. Haskell-Dowland, "Unobtrusive gait recognition using smartwatches," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2017.

[53] T. Hwang, A. O. Effenberg, and H. Blume, "A rapport and gait monitoring system using a single head-worn IMU during walk and talk," in *IEEE International Conference on Consumer Electronics (ICCE)*, 2019.

[54] I. C. Severin, D. M. Dobrea, and M. Dobrea, "Head gesture recognition using a 6dof inertial IMU," *International Journal of Computer Communication & Control*, vol. 15, no. 3, 2020.

[55] A. Gupta, I. Skog, and P. Händel, "Long-term performance evaluation of a foot-mounted pedestrian navigation device," in *Annual IEEE India Conference (INDICON)*, 2015.

[56] F. Abyarjoo, N. O.-Larnnithipong, S. Tangnimitchok, F. R. Ortega, and A. B. Barreto, "Posturemonitor: Real-time IMU wearable technology to foster poise and health," *Design, User Experience, and Usability: Interactive Experience Design*, vol. 9188, pp. 543–552, 2015.

[57] C. Luo, X. Feng, J. Chen, J. Li, W. Xu, W. Li, L. Zhang, Z. Tari, and A. Y. Zomaya, "Brush like a dentist: Accurate monitoring of toothbrushing via wrist-worn gesture sensing," in *IEEE Conference on Computer Communications (INFOCOM)*, 2019.

[58] W. Chen, L. Chen, M. Ma, F. S. Parizi, S. N. Patel, and J. A. Stankovic, "Vifin: Harness passive vibration to continuous micro finger writing with a commodity smartwatch," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 5, no. 1, pp. 45:1–45:25, 2021.

[59] S. Zhang, E. Nemati, M. Dinh, N. Folkman, T. Ahmed, M. M. Rahman, J. Kuang, N. Alshurafa, and A. Gao, "Coughtrigger: Earbuds IMU based cough detection activator using an energy-efficient sensitivity-prioritized time series classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[60] Y. Cao, H. Chen, F. Li, S. Yang, and Y. Wang, "Awash: Handwashing assistance for the elderly with dementia via wearables," in *IEEE Conference on Computer Communications (INFOCOM)*, 2021.

[61] T. N. Do and U.-X. Tan, "Novel velocity update applied for imu-based wearable device to estimate the vertical distance," in *International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, 2019, pp. 1–4.

[62] Z. Yang, Y. Wei, S. Shen, and R. R. Choudhury, "Ear-ar: indoor acoustic augmented reality on earphones," in *ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2020.

Jianwei Liu received the BS degree from Northwestern Polytechnical University in 2018. He received his Master degree from Xi'an Jiaotong University in 2021. He is working toward the Ph.D. degree at Zhejiang University. His research interests include RFID, mobile computing, and smart sensing. He is student member of the IEEE.

Wenfan Song received the BS degree from Sichuan University in 2019. She is currently working toward the Ph.D. degree at Zhejiang University. Her research interests include IoT security and wireless sensing.

Leming shen is working toward the BS degree at Zhejiang University. His research interests include IoT security and wireless sensing.

Jinsong Han received his Ph.D. degree from Hong Kong University of Science and Technology in 2007. He is currently a professor of the College of Computer Science and Technology, Zhejiang University. His research interests focus on IoT security, smart sensing, wireless and mobile computing.

Kui Ren received the Ph.D. degree from the Worcester Polytechnic Institute, Worcester, MA, USA. He is currently a professor of computer science and technology and the Director of the Institute of Cyberspace Research, Zhejiang University, Hangzhou, Zhejiang, China. His current research interests include cloud and outsourcing security, wireless and wearable system security, and artificial intelligence security. Dr. Ren is also a Distinguished Scientist and Fellow of the ACM. He was a recipient of the IEEE CISTC Technical Recognition Award 2017 and the NSF CAREER Award in 2011.