Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2134493, 14 pages
https://doi.org/10.1155/2022/2134493

Research Article

SnapUnlock: A Contrastive Learning-Based Contactless Authentication via Heterogeneous Sensors

Mengqi Chen,[1] **Jiawei Lin**,[1] **Wanlong Li**,[2] **Yongpan Zou**,[1] **and Kaishun Wu**[1]

[1]*College of Computer Science and Software Engineering, Shenzhen University, China*
[2]*Huizhou Institute of Space Information Technology, China*

Correspondence should be addressed to Mengqi Chen; chenmengqi2017@email.szu.edu.cn

Received 21 May 2022; Accepted 2 September 2022; Published 30 September 2022

Academic Editor: Ning Wang

Contactless authentication is crucial to keep social distance and prevent bacterial infection. However, existing authentication approaches, such as fingerprinting and face recognition, leverage sensors to verify static biometric features. They either increase the probability of indirect infection or inconvenience the users wearing masks. To tackle these problems, we propose a contactless behavioral biometric authentication mechanism that makes use of heterogeneous sensors. We conduct a preliminary study to demonstrate the feasibility of finger snapping as a natural biometric behavior. A prototype-SnapUnlock system was designed and implemented for further real-world evaluation in various environments. SnapUnlock adopts the principle of contrastive-based representation learning to effectively encode the features of heterogeneous readings. With the representations learned, enrolled samples trained with the classifier can achieve superior performances. We extensively evaluate SnapUnlock involving 50 participants in different experimental settings. The results show that SnapUnlock can achieve a 4.2% average false reject rate and 0.73% average false accept rate. Even in a noisy environment, our system performs similar results.

1. Introduction

Password, fingerprint, and face recognition have proved their commercial successes in the user authentication field. However, even though these techniques are widely deployed in commercial terminals, they still have limitations in our daily use. For example, passwords and PINs have contradictory issues: short passwords are insecure (e.g., smudge attacks [1] and shoulder surfing [2]), and long passwords are cumbersome and hard to remember. Biometric authentications, such as fingerprint and facial recognition, are more secure and user-friendly choices. Nevertheless, fingerprint authentication devices, such as digital door locks, are usually used by more than one person. It is a health and safety issue that people verify their fingerprints on the same panel. Furthermore, the skin condition of fingers also makes authentication a challenge. For example, after hand washing, moist fingers lead to a high false reject rate. As to face recognition, wearing a mask is an effective measure to reduce the risk of coronavirus infection during the COVID-19 epidemic. However, it brings difficulty to face recognition.

Recently, behavioral biometric authentication has been a popular research topic due to its inimitable property. For example, gait-based approaches [3, 4] use wireless signals or cameras to capture the motion posture for authentication. Brain waves [5–8] are more trusted mechanisms since it is controlled by the human's unique brain. Beyond these, we raise this question: can we explore a new contactless behavioral biometric scheme that weighs security and convenience?

To address the above question, this paper proposes SnapUnlock, a novel biometric user authentication scheme based on finger snapping. SnapUnlock leverages a smartwatch to capture acoustic and hand motion while the user snaps her/his fingers. Then, the smartwatch is acted as a transmitter. It transmits the detected finger-snapping event to the cloud server for authentication. Based on the authenticated results, the target device decides whether to unlock (see Figure 1). In principle, SnapUnlock extracts unique features from the sounds and the motion generated by a user's finger-snapping gesture. The key idea is based on the observation that sound and vibration patterns produced by finger

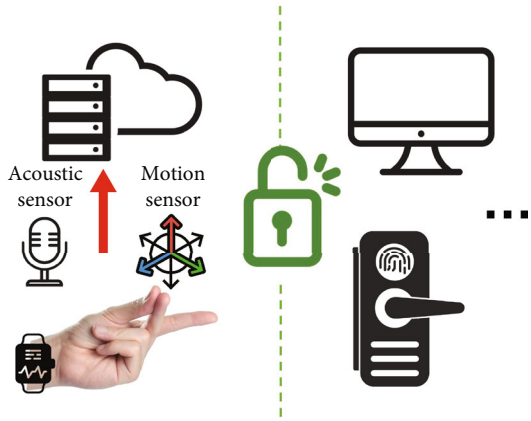


Figure 1: SnapUnlock leverages both the acoustic and motion sensors of a smartwatch to extract acoustic and hand motion features while the user snaps her/his fingers. These features are inputted into a classifier for user authentication. Based on the results of the classification, the target device will decide whether to unlock.

snapping can serve as a unique signature for a person. To evaluate the feasibility of such a unique biometric feature, we first conduct a preliminary study to clarify the uniqueness across individuals. After scrutinizing feasibility, we leverage the heterogeneous sensors (microphone, IMU) to capture experiment data. The wide deployment of them on smartwatches makes our system easy to implement and ubiquitous. In addition, heterogeneous sensors compensate for each other's weaknesses and enhance the noise tolerability of the system. To process the unaligned readings from sensors, we design a fusion and normalization approach; the raw readings are properly preprocessed before subsequent operation. After, we design a contrastive pretraining workflow for the signature extractor and a supervised learning phase for authentication.

SnapUnlock brings about the following advantages. (1) Ubiquity: it relies on a microphone and accelerometer which are readily integrated into the smartwatch. (2) Usability: finger snapping as a behavior has been proved its socially acceptable and user-friendly properties in previous research [9]. (3) Security: compared with depending on a single voiceprint feature, the combination of motion and sound is more resilient to replay attacks. The smartwatch only serves as a collection and transmission device; it does not store the user's authentication information. Therefore, even if the device is stolen, the attacker cannot pass the authentication.

We implement the SnapUnlock prototype system on the Android platform. We collected more than 2500 finger-snapping data from 50 volunteers. The evaluation shows that SnapUnlock achieves 3% average false reject rate and 0.95% average false accept rate. In summary, this paper makes the following contributions:

(i) This paper proposes a contactless biometric authentication system to leverage sound and motion performed by finger snapping. We demonstrate the diversity of finger snapping captured by heterogeneous sensors across individuals and their consistency for the same individual. The measurements serve as an empirical feasibility study of finger snapping as a new biometric modality for authentication or access control

(ii) We design and implement an authentication system that fuses readings from heterogeneous sensors, extracts acoustic and motion features, and accurately classifies the features. We also implement a model adaptation scheme to stabilize the performance over a period of time

(iii) We build a prototype of SnapUnlock on the android platform. It takes us one and a half months to collect more than 2500 finger-snapping samples from 50 users. Extensive experiments were conducted to evaluate the performance in various situations. The result shows that SnapUnlock is robust against different environments

The rest of this paper is organized as follows. In Preliminary Study, we conduct a preliminary study to demonstrate that it is feasible to use finger snapping as an authentication feature. In System Design, we introduce the detailed design of our SnapUnlock system. In Evaluation, we describe the experiment of data collection and evaluate the proposed system from different perspectives. Related Work and Discussion provide the related work and discuss the limitations and future work of our system. Finally, we summarize this paper in Conclusion.

2. Preliminary Study

2.1. Physiological Mechanism of Finger Snap. Finger snap is the act of creating a clicking sound with one's fingers. While snapping fingers, we slide one finger against another with force; then, the middle finger gains more momentum and strikes the palm surface. The collision and friction will produce a sound and an arm vibration. According to the diversity of hand geometry biometrics (e.g., area/size of the palm, length, and width of fingers), this effect has a variance across different individuals. Prior works have employed static features as a kind of authentication feature [10, 11]. Moreover, the sound raised by the action of sliding finger surfaces against one another is called distinct fingerprint-induced sonic effect [12]. In Rathore et al.'s [13] work, they have proved that this effect is reliable in authentication.

2.2. Data Collection. For data collection, we develop an application for the Android Wear OS 2.0 operating system. As shown in Figure 2, we can select one or more sensors to record at the same time in the "select sensor" page (Figure 2(b)) in the application. Then, return to the main screen (Figure 2(a)) and click "record" to start recording. In this data collection experiment, we select a built-in microphone and accelerometer to measure the sound and the vibration. The device we used is Huawei Watch 2. It contains a 1.1 GHz quad-core CPU and 768M RAM. The sample rate of the microphone and IMU is 48 KHz and 100 Hz, respectively.

(a) Main page (b) Select sensor page

Figure 2: Data collection application UI.

We recruit 25 volunteers (labeled as $V_1 - V_{25}$) for data collection in an indoor environment. Among them, 8 volunteers were female, and 17 were male. Their age ranges from 19 to 35. The whole data collection takes about one month and a half. Before the data collection experiment, we explain the purpose and usage of SnapUnlock to volunteers. After that, each volunteer was asked to practice finger snap until he/she becomes skilled. In the data collection phase, each volunteer is asked to wear a smartwatch when they are snapping their fingers. We did not limit the volunteer's posture and only asked them to behave as naturally as they do in daily life. Each volunteer was asked to provide 25 samples under a 30 ~ 40 dB noise level in each session. The experiment will be conducted twice each day. Moreover, to analyze whether the time duration will affect the result, intervals between each data collection day are varied. We label the experiment on the first day as Session 0 (S0). The volunteer will participate in the experiment again after 1 day (S1, S2), 3 days (S3, S4), 7 days (one week) (S5, S6), 30 days (one month) (S7, S8), and 45 days (a month and a half) (S9, S10), respectively. Note that every two sessions are conducted in the morning and afternoon, respectively (e.g., S1 and S2 are conducted in the morning and afternoon of day 1, respectively; S3 and S4 are conducted in the morning and afternoon of day 7, respectively). In summary, we collected a total number of 25 (number of samples for each volunteer) × 25 (number of volunteers) × 10 (number of sessions) = 6250 samples in the data collection experiment.

2.3. Data Analysis. In this section, we present data analysis to validate the following assumption.

(i) For the same person, finger-snapping signatures will not vary too much after a long term

(ii) For different people, finger-snapping signatures show different signal patterns

Based on the above assumptions, we conduct intrauser and interuser analyses.

2.3.1. Intrauser Analysis. In this section, we analyze the time variation of finger snap in intrauser. Two volunteers (denoted as A and B) were randomly selected as analyzed targets from the collected dataset. Then, we use data in S0 as the target to calculate the average PSD (Power Spectrum Density) across the other 10 sessions. In each session, 25 instances were randomly selected to calculate with the instances in S0. The results of volunteer A are showed in Figures 3(a)–3(d), and the results of volunteer B are showed in Figures 3(e)–3(h).

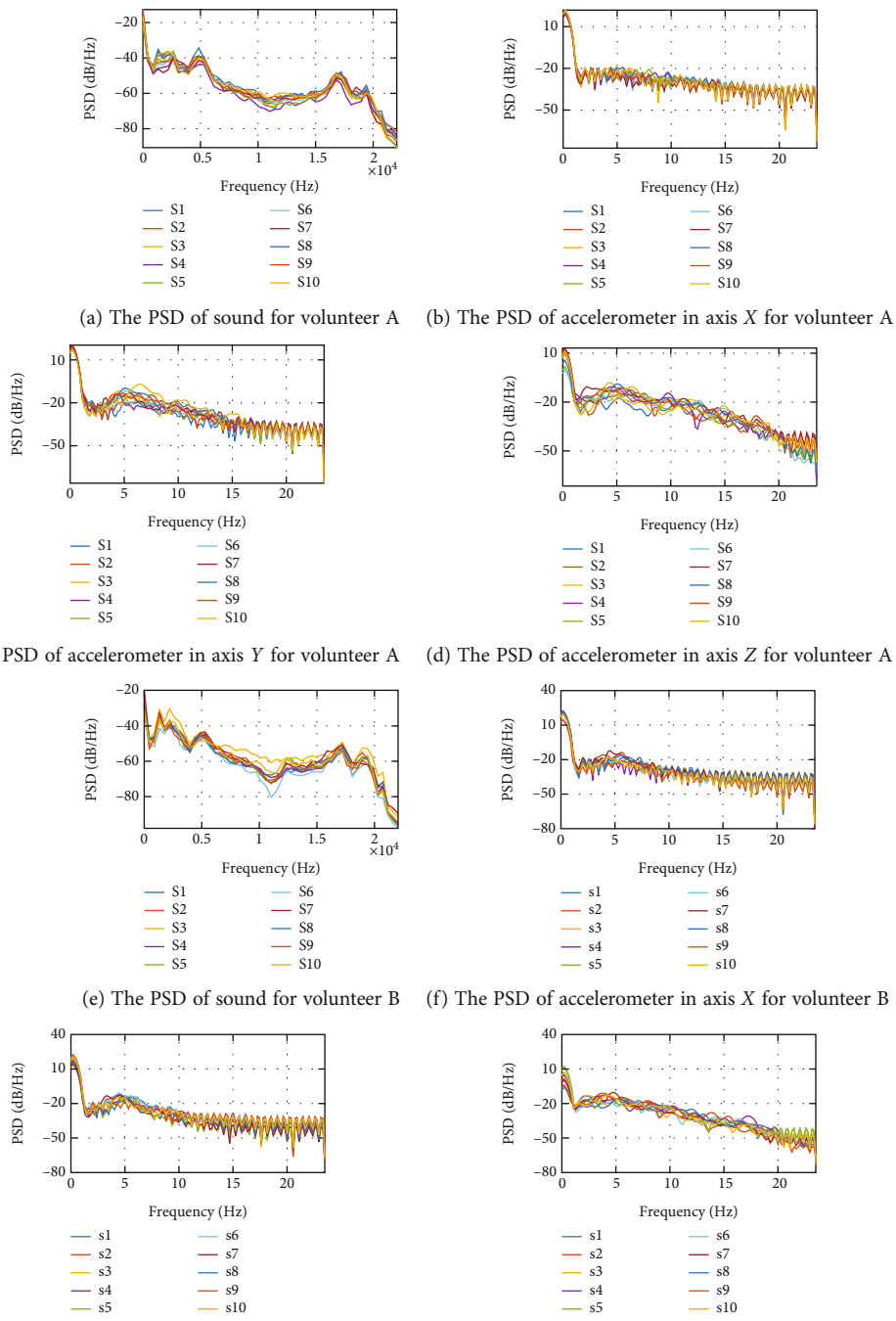
Figures 3(a) and 3(e) are the average PSD of sound signals. We observe that even after a month had passed, people have consistent patterns of finger-snapping sounds. Figures 3(b)–3(d) and 3(f)–3(h) show the average PSD of accelerometer readings on three axes. We can see that the axis X and axis Y patterns remain highly consistent, but there is a decrease in the Z-axis.

To further quantify the similarity of the PSD curve, we calculate the correlation coefficients for volunteer A in each session, and the results are shown in Figure 4. The average correlation coefficients of sound on all the sessions are above 0.9, and the accelerometer's coefficients are relatively low, but they are all above 0.85. In summary, this result validates our assumption that finger-snapping signature will not vary significantly across different times during a long term for the same person.

2.3.2. Interuser Analysis. After intrauser analysis, we calculate the correlation coefficients of PSD curves for volunteers A and B in different sessions to validate their similarity. As shown in Figure 5, the PSD correlation coefficients of sound between volunteers A and B are below 0.5, which is lower than Figure 4(a). In addition, the data from the accelerometer on three axes (Figures 5(a)–5(d)) also show similar results. In summary, this noticeable difference validates our assumption that finger snap shows different signal patterns for different people.

3. System Design

The system architecture of SnapUnlock is shown in Figure 6. There are two major parts: user enrollment and user authentication. During the user enrollment phase, the user needs to provide a number of finger-snapping samples. A pretrained LIMU-BERT model is applied to extract joint features from acoustic and accelerometer data. Finally, the joint features



(a) The PSD of sound for volunteer A

(b) The PSD of accelerometer in axis X for volunteer A

(c) The PSD of accelerometer in axis Y for volunteer A

(d) The PSD of accelerometer in axis Z for volunteer A

(e) The PSD of sound for volunteer B

(f) The PSD of accelerometer in axis X for volunteer B

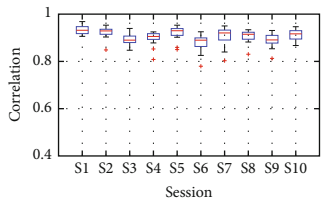
(g) The PSD of accelerometer in axis Y for volunteer B

(h) The PSD of accelerometer in axis Z for volunteer Z

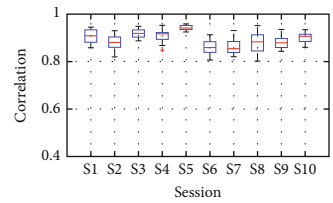
Figure 3: The PSD curve of sound and motion data for volunteers A and B in different sessions.

will be fed into a fully connective layer for prediction. We follow a non-end-to-end training strategy. In the pretraining phase, the SnapUnlock takes the size of 70% targets in the dataset as a pretraining set. During user authentication, we keep the signature extractor parameters the same and only train the classifier with the target's reference samples.

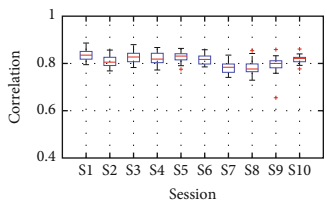
3.1. Finger-Snapping Event Detection. SnapUnlock targets real-time detection of the finger-snapping event even in a noisy environment. At the same time, event detection should be lightweight enough to run on wearable devices since they have limited computation capability and battery capacity. However, it is not straightforward to correctly detect events using the conventional thresholding method due to the complex environment. In this paper, we make use of the Constant False Alarm Rate (CFAR) [14] method to detect the start of each event. Essentially, CFAR is an energy-based adaptive thresholding method which adapts threshold value according to levels of external interference. We apply the detection algorithm to the IMU readings except for the microphone due to the fact that the sample rate (200 Hz) of IMU is much lower than the microphone (48000), and



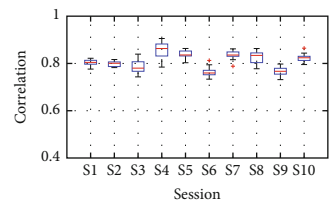
(a) The PSD correlation coefficients of sound



(b) The PSD correlation coefficients of accelerometer in axis X



(c) The PSD correlation coefficients of accelerometer in axis Y



(d) The PSD correlation coefficients of accelerometer in axis Z

Figure 4: The PSD correlation coefficients of volunteer A in different sessions.

noise barely affects IMU. Assuming that the environment noise follows Gaussian distribution, we first use a sliding window of width W to calculate the average energy of noise. Let $\mu(t)$ and $\sigma(t)$ denote the average energy and its standard deviation at time t, respectively. They can be formulated as

$$\begin{aligned}\mu(t) &= \frac{1}{W}A(t) + \left(1 - \frac{1}{W}\right)\mu(t-1),\\ \sigma(t) &= \frac{1}{W}B(t) + \left(1 - \frac{1}{W}\right)\sigma(t-1),\end{aligned} \tag{1}$$

where A is the accumulated energy and B is the overall standard deviation of signal power within a slide. Initially, $\mu(0) = 0$ and $\sigma(0) = 0$. A and B are formulated as

$$\begin{aligned}A(t) &= \frac{1}{W}\sum_{k=t}^{W+t}\left|S(k)^2\right|,\\ B(t) &= \sqrt{\frac{1}{W}\sum_{k=t}^{W+t}\left(|S(k)|^2 - A(k)\right)^2},\end{aligned} \tag{2}$$

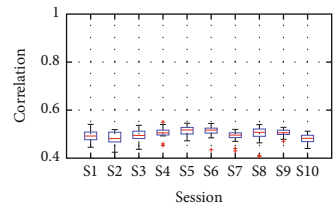
where $S(t)$ is the raw reading of IMU. Based on the above equations, a potential start point $S(t)$ can be determined when $S(t)$ meets the following condition:

$$|S(t)|^2 > \mu(t) + \lambda_1\sigma(t), \tag{3}$$

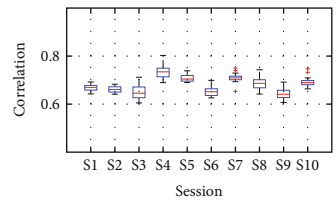
where λ_1 is a constant which is independent of the noise level. From observation, we empirically set W, λ_1 as 1000 and 18, respectively. Even a higher lambda may cause more queried samples for authentication; it also means that the snapping finger event has a higher probability of being captured as a query sample. Because the classifier can distinguish which samples are invalid, a higher lambda does not influence the authentication performance and increases the user experience. After detecting a start point in IMU readings, we project the point-to-sound reading according to the sample rate ratio. Because both acoustic and motion signals are collected simultaneously, their starting point position should well synchronize. Figure 7 shows some acoustic samples detected under different noise levels. We observe that the duration of each event is around 0.42 seconds. Therefore, we cut the readings by a fixed window $\mathbf{S} \in \mathbb{S}^{S_{\text{dim}} \times L}$ after the start point, where L is the length of readings in each window and S_{dim} is the dimension of sensors. In our work, S_{dim} is 1 and L is 20000 when S is collected from the microphone. S_{dim} is 3 and L is 42 when S is collected from the accelerometer, ensuring the alignment of each sample. By applying the algorithm, SnapUnlock can well extract the target event at different noise levels.

3.2. Background Noise Removal and Normalization. Background noise and the distribution variation of heterogeneous sensors are critical factors that impact the performance of the model. To address this, we first use a Butterworth filter with a passband of $[10, 20000]$ Hz to remove background noises since we find most of the energy exists in the frequency band of $10 - 20$ kHz in Figures 3(a) and 3(e). In addition, we keep the IMU readings the same since background noise does not affect the IMU. Second, SnapUnlock should handle multiple sensor data. But the readings of IMU sensors and microphone have different distributions (e.g., the amplitude of sound readings ranges from -1 to 1, and accelerometer readings range from -20 to 20). Such differences would affect the model performance in the training phase. Features with a large distribution range will play a decisive role to dominate the descending gradient, while small distribution features may be neglected. It leads to an irregular contour plot of the loss function and thus slows the converges during training. Therefore, the sensor reading needs to be properly normalized before training. We adopt the min–max scaling method for normalization. The equation can be formulated as follows:

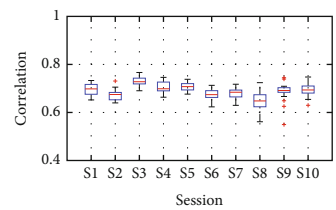
$$N_{\text{minmax}}(\mathbf{S}) = \frac{\text{clip}(\mathbf{S}, \max(\mathbf{S}), \min(\mathbf{S}))}{\max(\mathbf{S}) - \min(\mathbf{S})}. \tag{4}$$



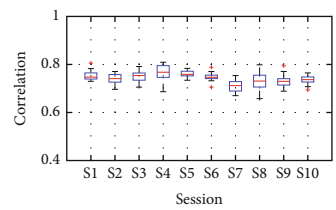
(a) The PSD correlation coefficients of sound between volunteers A and Y



(b) The PSD correlation coefficients of accelerometers in axis X between volunteers A and Y



(c) The PSD correlation coefficients of accelerometers in axis Y between volunteers A and Y



(d) The PSD correlation coefficients of accelerometers in axis Z between volunteers A and Y

Figure 5: The PSD correlation coefficients between volunteers A and Y in different sessions.

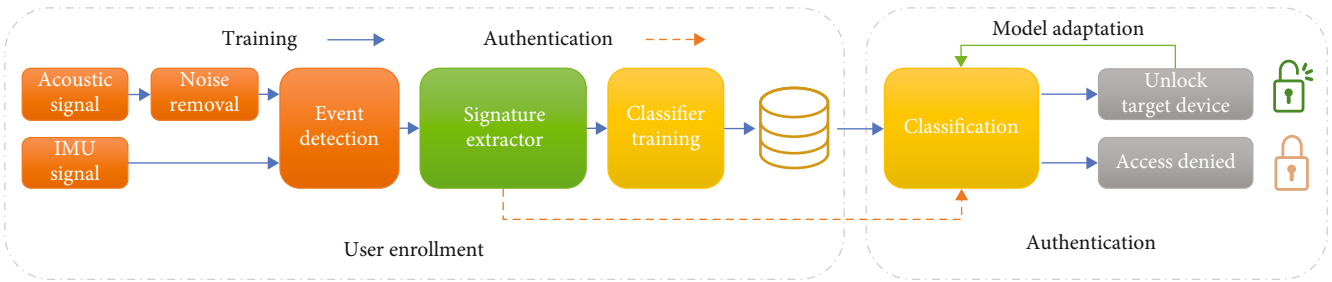


Figure 6: System architecture.

By applying background removal and normalization, the neural net can learn parameters better.

3.3. Signature Extractor. We expect our system to fuse the data stream from heterogeneous sensors and model the signature dependencies by a general feature representation. This feature representation should extract the signature of unseen users (none of their data is included in the extractor training phase). In another word, the feature modeled by the signature extractor should only relate to the signature for authentication. To this end, we design a contrastive learning flow which is an effective way to model default input as representations [15]. Figure 8 shows the contrastive learning phase.

Specifically, in this phase, a pretraining model learns representations by optimizing the principles min $g(x_i, x_i^+)$ and max $g(x_i, x_i^-)$, where the anchor instance x_i is from raw data in the pretraining set, positive instance x_i^+ is

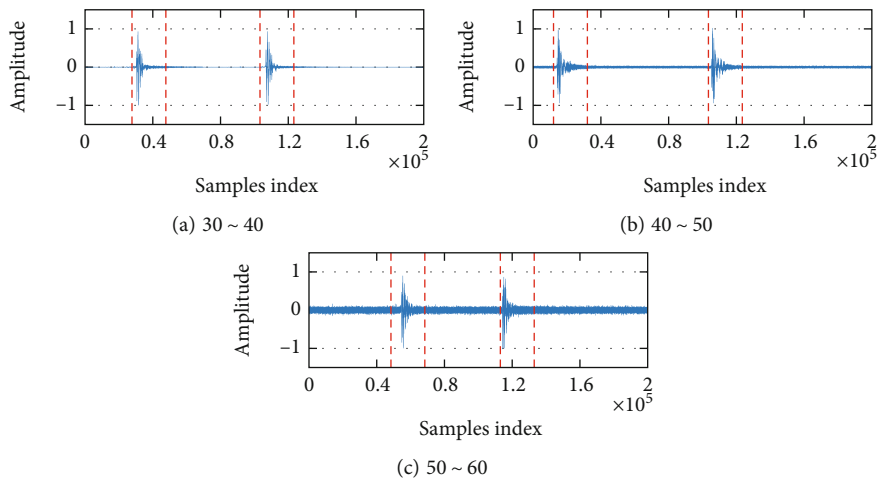


Figure 7: Example of event detection under three noise levels.

randomly sampled from the same participant's reference sample, and negative instance x_i^- is randomly sampled from other participants in the same training batch. In other words, the goal is to close semantically relative instances (positive pairs) and to take apart nonrelatives (negative pairs). Figure 9 shows the detail of the contrastive learning phase. Let h_i, h_i^+, and h_i^- denote the representations of x_i, x_i^+, and x_i^-, respectively. They are 4 outputs from the same signature extractor. We use LIMU-BERT [16] as a signature extractor and multilayer perceptron (MLP) as the final layer for prediction. The key operation is that we place independently sampled dropout masks on attention probabilities (default $p = 0.1$) and MLP with identical positive pairs and negative. The MLP will predict true or false according to the pair type. We use infoNCE [17] as the loss function:

$$\mathbf{loss}_i = -\log \frac{e^{\mathrm{sim}\left(h_i, h_i^+\right)/\tau}}{\sum_{j=1}^{N}\left(e^{\mathrm{sim}\left(h_i, h^+\right)/\tau} + e^{\mathrm{sim}(h_i, h^-)/\tau}\right)}, \tag{5}$$

where N is minibatch size, τ is a temperature hyperparameter, and $\mathrm{sim}(h_1, h_2)$ is the cosine similarity $h_1^{\mathrm{T}} h_2 / \|h_1\| \cdot \|h_2\|$.

Before pretraining, we apply the MFCC (Mel-Frequency Cepstral Coefficients) on sound readings. As shown in Figures 3(a) and 3(e), most of the snapping sound energy is concentrated in the range that is audible (0 ~ 6 kHZ). We chose to use the MFCC because it simulates the human auditory system by spacing the frequency bands, and it makes the signal less time-sensitive and easier to obtain valuable information. We first calculate MFCC from each segment with a Hamming window. The size of Hamming windows is 1440, and there will be 960 samples overlapping between each adjacent segment. The output size of the MFCC function on the sound readings is 14×42. We concatenate preprocessed sound and IMU readings as a vector for pretraining. The output of the extractor is empirically set to 32 according to the experiment in Impact of Representation Dimensionality. After the contrastive pretraining phase, the trained extractor is able to extract authentication-related representation. We use a trained extractor as the initial parameters in the supervised learning phase. It connects to a new classifier for a new user.

3.4. Classifier. As shown in Figure 8, we adopt another MLP as the final classifier in the supervised learning phase and testing phase for each new user. In the supervised learning phase, we train the MLP with the reference data, and the signature extractor is achieved from the contrastive pretraining phase. The function of the MLP is quite simple: it learns the boundary of reference representation. A questioned sample is projected to a high-dimensional space by the extractor, and the MLP decides whether it is true or false according to its representation in range.

3.5. Model Adaptation. As the human body changes with age, the shape of the palms will change gradually. The evaluation result in Section Effect of Model Adaptation also meets our assumption. Therefore, we design a model adaptation algorithm to update the model periodically. Considering the uncertainties associated with hand shape changes in each individual at different moments of growth, we design a confidence ranking algorithm that can select a legitimate sample with a high confidence level at the end of each authentication. The algorithm is listed in Algorithm 1. Selected high confidence legitimate samples will be added to the training pool. The model will be retrained at regular intervals. This eliminates the need for users to retrain and automatically adds up-to-date legitimate samples to the model.

4. Evaluation

This section discusses the performance results of SnapUnlock from different perspectives.

4.1. Experiment Setup. We implement the SnapUnlock prototype on the Android platform (Huawei Watch 2) and a desktop. The overall authentication flow is trained off-line on a desktop with four-core Intel® Xeon® E3-1231 CPU, 16 G RAM, and RTX 2070s GPU which is running Windows 10 with Matlab R2020a software. We conduct experiments on real-world datasets which are mentioned in Data

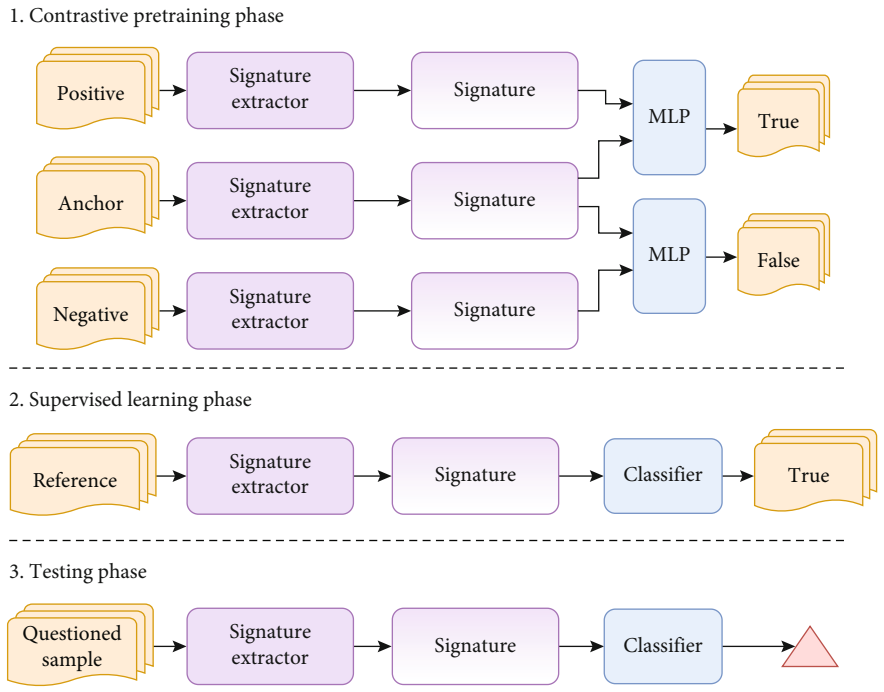


Figure 8: Illustration of model learning workflow. It contains three phases: contrastive pretraining phase, supervised learning phase, and testing phase.

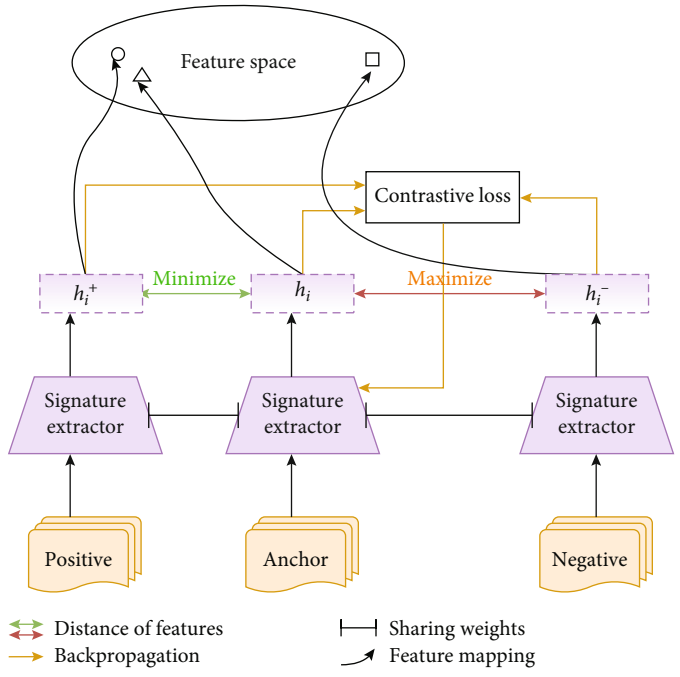


Figure 9: The detail of contrastive learning phase.

Collection. There are two tasks for biometric security systems. One is the recognition of unauthorized users, and the other is incorrectly identifying an authorized person. Therefore, we use FAR and FRR as metrics to quantify the performance of SnapUnlock.

(i) FAR is short of false accept rate. It shows the likelihood that an authentication system incorrectly accepts an unauthorized user. An authentication system with a higher FAR is more secure than a lower one

Input: $CorrectSample(CS)$, $Threshold(T)$
Output: $HighConfidenceSample(HCS)$
1 *load TrainedModel*;
  // $TrainedModel = MLP(TariningSet)$;
2 $LenOfCS = length(CS)$;
3 $D = zeros(LenOfCS, 2)$;
4 for $i = 1$ to $LenOfCS$ **do**
5 $D(i, 1) = \text{sgn}(w * CS(i) + b)/w$; ;
  // *CalculatedDistance*
6 $D(i, 2) = i$;
7 end
8 $HD = Confidence(Sort(D))$;
  // *AchieveHighConfidencePoint*
9 for $i = 1$ to $length(HD)$ **do**
10 **if** $HD(n, 1) > T$ **then**
11 $HCS = [HCS; CS(HD(n, 2))]$;
12 **end**
13 end
14 return HCS

Algorithm 1: Confidence ranking algorithm.

(ii) FRR, the short for false reject rate, is the ratio between incorrectly rejected attempts of legitimate users and all rejected samples. It depicts how user-convenient an authentication system is. An authentication system with a higher FRR means legitimate users need to pay more trials to gain access

4.2. Overall Performance. Since SnapUnlock leverages heterogeneous sensors, it is necessary to evaluate the overall performance across different sensors. Figure 10 illustrates the comparative performances of SnapUnlock trained and tested with a single and combination of sensors. 30 participants are involved in the pretraining phase, and 20 participants are in the training and testing phases. For evaluation, we randomly divide the sample from each participant into the training (70%) and test (20%) sets. Each positive sample pairs with a randomly sampled native from other participants in pretraining and training. We make sure that all participants in the training and testing phases are unseen in pretraining. According to the results, SnapUnlock with heterogeneous sensors outperforms others by the largest margin of 5.4% between heterogeneous and motion cases, with the effectiveness of heterogeneous sensors. Specifically, heterogeneous sensors achieve 0.73% FAR and 4.2% FRR. On the other hand, the FAR and FRR are (2.3%, 6.3%) for the microphone and (4.1%, 7.1%) for the accelerometer, respectively. We suspect that the heterogeneous readings contain more information related to snapping behavior than single.

In summary, the performance of heterogeneous sensors is significant. The authentication system can gain improvement not only from acoustic but also from motion sensors.

4.3. Impact of Reference Number. Reference refers to the samples that are registered as templates in the authentication system before being first used for a new user. The growth in the number of reference samples improves accuracy and

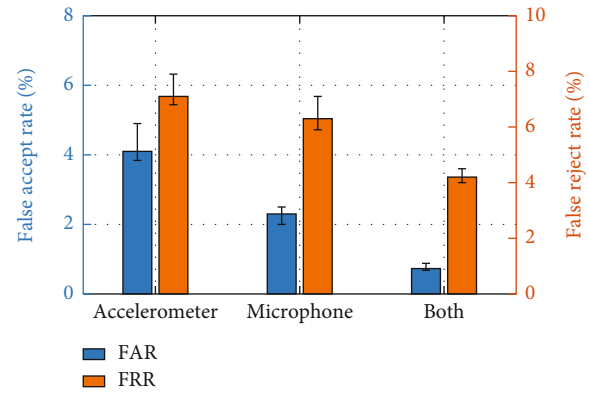


Figure 10: Overall performance of SnapUnlock.

reduced user experience. We keep the pretraining setting the same, and the reference number in training varies from 3 to 30. Figure 11 depicts the results of FAR and FRR. As shown clearly, SnapUnlock is able to achieve higher accuracies as the reference number increases. The FRR rapidly decreases from 97.7% to 3.2%, and FAR slightly increased from 0.4% to 1.1%. We also observe that the slopes of FAR and FRR start to slow down when the reference sample increases from 23 to 30. Therefore, we suggest setting a reference number higher than 23 in the application.

4.4. Impact of Participant Number. Participant numbers influence the diversity of the pretraining set. With the growth of diversity, a pretrained signature extractor is able to extract a more appropriate representation for authentication. To investigate the impact of participant numbers in representation learning, we keep the pretraining setting the same as in the previous section, and the participant number in pretraining varies from 1 to 40. When the participant number increases, we decrease the training pair amount of each participant so that the total amount of the pretraining set is consistent. Figure 12 depicts the results of FAR and FRR. With the increment of participants, the FRR rapidly decreases from 87.5% to 3.4%, and FAR decreases from 13.6 % to 0.6%. When the number of participant number reaches 30, it stabilizes. Overall, the results show that signature learning significantly benefits from the diversity of participants, and SnapUnlock needs about 30 participants for pretraining.

4.5. Impact of Training Pair Number. We have demonstrated that increasing pretraining set diversity is able to improve performance. In this section, we examine how training pair numbers influence our model with the same diversity in the pretraining phase. In the pretraining phase, we ensure the participant number is 30 and vary the training pair ratio of each participant from 10% to 30%. Figure 13 depicts the results of FAR and FRR. The results show that increasing the training pair number obtains lower FRR. The FRR decrease from 8.6% for 20% to 3.4% for 80%. However, the FAR slightly decreased from 0.6% to 0.9%. Combined with the result in Impact of Participant Number, it is suggested to vary data diversity than increase the pretraining set of each participant.

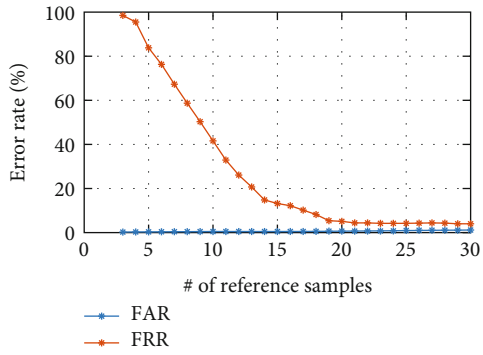


Figure 11: Impact of reference number.

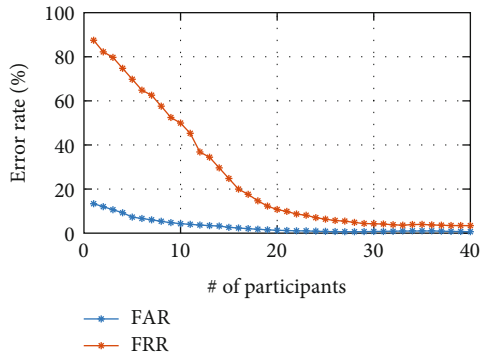


Figure 12: Impact of participant number in pretraining phase.

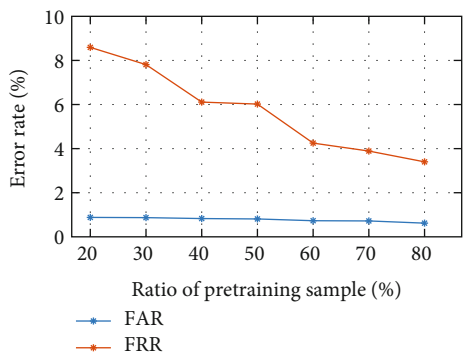


Figure 13: Impact of training pair number in pretraining phase.

4.6. Effect of Model Adaptation. We evaluate the effect of model adaptation by varying the time. As mentioned in Data Collection, we have collected data with day variance. In this experiment, we train the classifier with the data collected on the first day (S0). The test data are split as S1, S2 (after 1 day), S3, S4 (after 3 days), S5, S6 (after 1 week), S7, S8 (after 30 days), and S9, S10 (after a month and a half), respectively. We show the result in Figure 14. As we can see, the FRR increase from 4.2% to 13.5% as the day passes without model adaptation, and the FAR stabilizes in the range of 0.38% to 0.79%. We assume that the human skeleton grows or shrinks over time. Therefore, SnapUnlock tends to distrust the long-term sample. With the help of model adaptation, SnapUnlock gains the ability, which learns the variance of snapping behavior. It greatly stabilizes the performance of SnapUnlock.

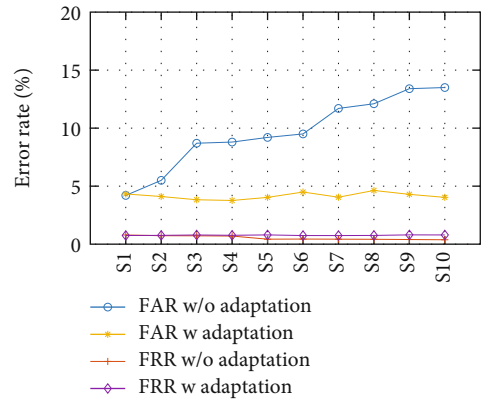


Figure 14: Impact of day changing.

4.7. Impact of Noise. This section analysis how SnapUnlock performs under different noise levels. Similar to Overall Performance, we keep the participant's number the same in each phase, pretrain and train the model with data collected under 30 db, and test in three noise levels (e.g., 30db, 45db, and 60db). The results are illustrated in Figure 15. Compared to the model trained with microphone data, the model trained with heterogeneous data performs better as the noise level increases. Even when the noise level reaches 60 db, the performance with heterogeneous data degrades slightly by the margin of 0.16% for FAR and 0.46% for FRR. The degraded margins with microphone are 3.5% for FAR and 1.8% for FRR. Overall, SnapUnlock reinforces the resistance to environmental noise with the assistance of an accelerometer.

4.8. Impact of Hand Surface State. In this section, we evaluate the impact of the hand surface state. Moisture and dryness on the skin change the friction of the skin when the finger is snapped, leading to changes in the acoustic signal. To evaluate this, we compare the performance in both wet and dry hand surface states. Each participant's authentication model is trained with the standard setting, but the testing dataset is collected under different hand surface states (e.g., wet, dry, and normal). In the wet situation, we collect data after the participants washed their hands. Then, we let the participants dry their hands with a hand dryer, ensuring no other substances were on their hands. The data collected after this phase are marked as dry. The normal data are randomly selected from the dataset as mentioned in Data Collection. Figure 16 shows the results. There is no significant difference between the normal and dry states. However, while the hand surface state is wet, the performance will decrease. One possible reason is that the acoustic feature's weight in the joint feature is higher than the motion feature. Therefore, we recommend that users do not get their hands wet when using the SnapUnlock system.

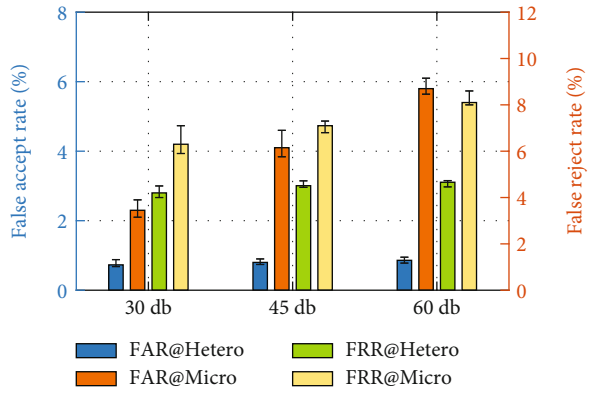


Figure 15: Impact of noise.

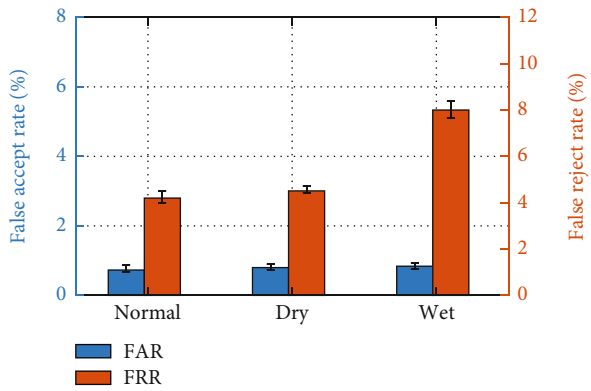


Figure 16: Evaluation of the impact of hand surface state.

4.9. Impact of Representation Dimensionality. In this section, we investigate how the representation dimensionality affects the model performance. Figure 17 illustrates the FAR and FRR of the feature extractor with the representation dimensionality varying from 8 to 64 while the other optimal hyperparameters are unchanged. It is easy to observe that the best performance appeals at dimension 32. And, when the dimension is higher than 32, the performances decrease. The result demonstrates that a large representation dimensionality does not benefit model improvement. Considering that the best result was achieved with dimension 32, we use 32 as the output dimension.

4.10. User Experience Study. In addition to validating SnapUnlock effectiveness, we also assess its users' experience. We conduct a user experience survey on 50 participants. All of them have previously used passwords and fingerprints as authentication schemes. Out of the 50 participants, 25 are involved in previous experiments. The rest of the 25 users use SnapUnlock for the first time. Guided by the tutorial, we found that all 25 new users mastered the finger-snapping technique within an hour. We first informed all the participants of the aim of the study and showed them how to use SnapUnlock. They were asked to install the software on a smartwatch and use SnapUnlock for authentication in unlocking personal computer scenarios. After the

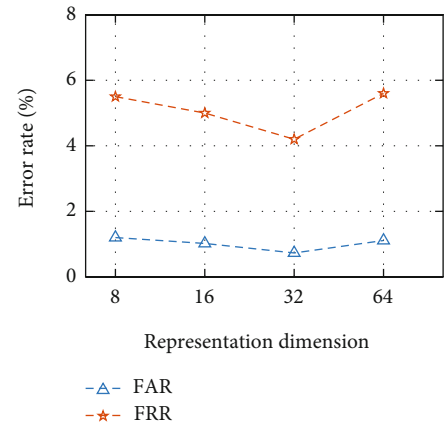


Figure 17: FAR and FRR comparison of representation dimensions.

trial, we distributed a questionnaire to all 50 participants and collected feedback from each participant. The questionnaire consists of the following questions.

(i) By jointly considering the accuracy, robustness, and usability, please rate an overall score (from 0 to 5) to these three unlocking schemes (SnapUnlock, password, fingerprint, and faceID) (0 means worst; 5 means best)

(ii) Are you willing to use the three authentication schemes (SnapUnlock, password, fingerprint, and faceID) daily? Please rate a score from 0 to 5 (0 means I never want to use it daily; 5 means I would certainly use it in public)

(iii) How difficult learning snap finger is do you think? Please rate a score from 0 to 5 (0 means easy; 5 means difficult)

Figure 18 shows the results. The average overall scores are 3.8, 2.2, 3.1, and 3.7 for the SnapUnlock, password, fingerprint, and faceID, respectively. The average willingness ratings are 3.3, 2.6, 3.1, and 3.3 for the SnapUnlock, password, fingerprint, and faceID, respectively. It shows that SnapUnlock is more acceptable than other methods. In the end, the average difficulty ratings are 3.1, 3.2, 2.1, and 0.8 for the SnapUnlock, password, fingerprint, and faceID, respectively. Even though our method requires a period of practice for new users, the results demonstrate that it is still slightly better for passwords. Most of the participants say this is due to the need to remember complex password combinations for password security.

5. Related Work

We categorize SnapUnlock's research into three subsections: physiological biometric authentication, behavior biometric authentication, and nonspeech body sound sensing.

5.1. Physiological Biometric Authentication. Physiological biometric features can be easily quantified into digital data

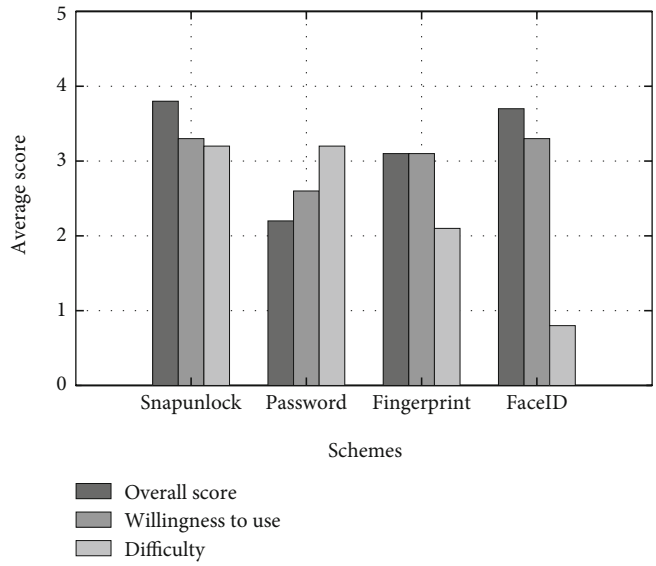


Figure 18: The users' overall rating of different methods, the willingness to daily use, and the difficulty of different methods.

by different types of sensors embedded in mobile or wearable devices. Prior works [18–21] explore the possibility of applying in authentication by proving their uniqueness. Fingerprints [18] can be easily captured by fingerprint sensors, so it is widely used in mobile devices. However, due to the shortcomings of the fingerprint sensors, such as squeezing screen space and failure in case of water on fingers, fingerprint authentication has been abandoned by some mobile devices (e.g., iPhone [22]). Face recognition [19–21] also is one of the most popular authentication approaches. However, the camera and other relative sensors cut the screen into a notch shape, bringing a negative user experience to customers [23]. On some special occasions (e.g., wearing a protective mask), face recognition may fail to verify the legitimate user. Compared with the physiological biometric feature, SnapUnlock requires no additional sensor costs and achieves satisfactory performance.

5.2. Behavior Biometric Authentication. Behavioral biometric authentication has attracted more attention in recent years. Gesture-based [24–26], keystrokes-based [27, 28], and gait-based [29–31] authentication is a popular topic in recent years. Their efforts exploit including but are not limited to accelerometer, keyboard, and touch-screen to extract unique features such as movement speed, rhythm, and other properties from particular behaviors. Besides, some researches [32–36] focus on exploring other new behavior biometric features. BreathPrint [32] records the sound of breath from the user's nose to verify legitimate users. BreathLive [33] extracts features from both sounds and motion caused by deep breathing to realize a reliable authentication system. Wang et al. [34] press the phone on the user's chest, and it measures the heartbeat signal using the inertial accelerometer in smartphones to perform authentication. Bilock [35] innovatively leverages the sound of dental occlusion (i.e., tooth click) to authenticate the user. Brain waves [36] also show their potential in the authentication field. SnapUnlock introduced a new behavioral biometric mechanism, finger-snapping, which is usable and resilient to replay attacks and can be easily captured by an accelerometer and microphone in a commodity smartwatch to build a stable biometric authentication system.

5.3. Nonspeech Body Sound Sensing. According to the type of nonspeech body sound, we can utilize it in the field such as monitoring health status or activity recognition. Prior works [37–41] try to extract useful information from these nonspeech body sounds for different purposes. Bodyscope [37] develops an acoustic-based wearable system to record the sound by placing a custom Bluetooth headset in the area of the user's throat. This system can classify different types of nonspeech body sounds, such as eating, drinking, breathing, speaking, laughing, and coughing at around 71.5% accuracy. Bodybeat [38] is a mobile sensing system that can capture a diverse range of nonspeech body sounds to recognize physiological reactions. Similar to Bodyscope [37], it places a custom piezoelectric microphone near the user's throat. They also develop a body sound classification algorithm to distinguish different sounds of human behavior with about 71.2% accuracy. SymDetector [39] utilizes the built-in microphone on a smartphone to continuously monitor and detect four types of respiratory sounds (i.e., sneeze, cough, sniffle, and throat clearing). SleepHunter [41] and iSleep [40] both monitor the sleep quality using the microphone of the off-the-shelf smartphone. All of these works are aimed at using nonspeech sounds in the healthcare field. But SnapUnlock is interested in using nonspeech sound (finger-snapping) for user authentication on the smart device due to its unique property.

6. Discussion

In this section, we mainly discuss the limitations of SnapUnlock. An inherent limitation of our system is that

users need to learn to snap their fingers. After observing those who did not produce sound due to incorrect posture, we found that they could still generate wrist vibration. Therefore, an alternative solution is to collect more data from those who are unfamiliar with finger-snapping. This way will increase the motion feature's weight in the training phase, allowing motion features to have a higher weight in the joint feature during decisions.

7. Conclusion

In this paper, we present SnapUnlock, a touchless authentication approach that can unblock devices by leveraging finger-snapping gestures. We utilize the inherent correlation between wrist motion and sound caused by finger-snapping action and seamlessly integrate contrastive learning techniques into signature extractor learning to realize a reliable authentication system.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China Grant (No. 62172286) and Natural Science Foundation of Guangdong Province Grant (No. 2022A1515011509).

References

[1] F. Schaub, R. Deyhle, and M. Weber, "Password entry usability and shoulder surfing susceptibility on different smartphone platforms," in *Proceedings of the 11th international conference on mobile and ubiquitous multimedia*, p. 13, ACM, 2012.

[2] F. Tari, A. Ozok, and S. H. Holden, "A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords," in *Proceedings of the second symposium on Usable privacy and security*, pp. 56–66, ACM, 2006.

[3] M. Mohamed and M. Cheffena, "Received signal strength based gait authentication," *IEEE Sensors Journal*, vol. 18, no. 16, pp. 6727–6734, 2018.

[4] V. Toral-Alvarez, C. Alvarez-Aparicio, A. M. Guerrero-Higueras, and C. Fernandez-Llamas, "*Gait-Based Authentication Using a rgb Camera*," in *Computational Intelligence in Security for Information Systems Conference*, pp. 126–135, Springer, 2021.

[5] Y. S. Soni, S. B. Somani, and V. V. Shete, "Biometric user authentication using brain waves," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, pp. 1–6, Coimbatore, India, 2016.

[6] E. Grace Mary Kanaga, R. Muthu Kumaran, M. Hema, R. Gowri Manohari, and T. A. Thomas, "An experimental investigations on classifiers for brain computer interface (bci) based authentication," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, pp. 1–6, Tirunelveli, India, 2017.

[7] N. Tran, D. Tran, S. Liu, W. Ma, and T. Pham, "Eeg-based person authentication system in different brain states," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 1050–1053, San Francisco, CA, USA, 2019.

[8] G. A. Vadlamudi and K. H. Kishan, "Security authentication using brain waves," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 104–107, Coimbatore, India, 2021.

[9] M. Nielsen, M. Storring, T. B. Moeslund, and E. Granum, "A procedure for developing intuitive and ergonomic gesture interfaces for hci," in *International Gesture Workshop*, pp. 409–420, Springer, 2004.

[10] A. Kumar, D. C. M. Wong, H. C. Shen, and A. K. Jain, "*Personal verification using palmprint and hand geometry biometric*," in *International conference on audio-and video-based biometric person authentication*, pp. 668–678, Springer, 2003.

[11] G. Fouquier, L. Likforman, J. Darbon, and B. Sankur, "The biosecure geometry-based system for hand modality," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, p. I–801, Honolulu, HI, USA, 2007.

[12] A. Akay, "Acoustics of friction," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1525–1548, 2002.

[13] A. S. Rathore, W. Zhu, A. Daiyan et al., "Sonicprint: a generally adoptable and secure fingerprint biometrics in smart devices," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pp. 121–134, 2020.

[14] M. A. Richards, *Fundamentals of Radar Signal Processing*, McGraw-hill Education, 2014.

[15] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pp. 1735–1742, New York, NY, USA, 2006.

[16] X. Huatao, P. Zhou, R. Tan, M. Li, and G. Shen, "Limu-bert: unleashing the potential of unlabeled data for imu sensing applications," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pp. 220–233, 2021.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," https://arxiv.org/abs/1807.03748, 2018.

[18] N. Shabrina, T. Isshiki, and H. Kunieda, "Fingerprint authentication on touch sensor using phase-only correlation method," in *2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, pp. 85–89, Bangkok, Thailand, 2016.

[19] M. E. Fathy, V. M. Patel, and R. Chellappa, "Face-based active authentication on mobile devices," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1687–1691, South Brisbane, QLD, Australia, 2015.

[20] G. Guo, L. Wen, and S. Yan, "Face authentication with makeup changes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 814–825, 2014.

[21] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pp. 321–336, 2018.

[22] "iPhone X, Apple. Inc," https://en.wikipedia.org/wiki/IPhone_X.

[23] "Display notches: The good, the bad, and the (very) ugly," https://www.androidpolice.com/2018/04/20/display-notches-good-bad-ugly/.

[24] M. Shahzad, A. X. Liu, and A. Samuel, "Behavior based human authentication on touch screen devices using gestures and signatures," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2726–2741, 2017.

[25] A. Buriro, R. Van Acker, B. Crispo, and A. Mahboob, "Airsign: a gesture-based smartwatch user authentication," in *2018 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–5, IEEE, 2018.

[26] 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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 987–996, 2012.

[27] C. Liu, G. D. Clark, and J. Lindqvist, "Where usability and security go hand-in-hand: robust gesture-based authentication for mobile systems," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 374–386, 2017.

[28] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: towards finger-input authentication on ubiquitous surfaces via physical vibration," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 73–87, 2017.

[29] A. Jain and A. Kumar, *Biometric Recognition: An Overview, Second Generation Biometrics: The Ethical, Legal And Social Context, E. Mordini and D. Tzovaras*, 2012.

[30] D. Gafurov, K. Helkala, and T. Sondrol, "Biometric gait authentication using accelerometer sensor," *JCP*, vol. 1, no. 7, pp. 51–59, 2006.

[31] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell phone-based biometric identification," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–7, IEEE, 2010.

[32] J. Chauhan, H. Yining, and S. Seneviratne, "*Breathprint: Breathing Acousticsbased User Authentication*," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 278–291, 2017.

[33] C. Huang, H. Chen, L. Yang, and Q. Zhang, "BreathLive," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–25, 2018.

[34] L. Wang, K. Huang, K. Sun et al., "Unlock with your heart," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–22, 2018.

[35] Y. Zou, M. Zhao, Z. Zhou, J. Lin, M. Li, and K. Wu, "BiLock," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–20, 2018.

[36] J. Chuang, H. Nguyen, C. Wang, and B. Johnson, "*I think, therefore i am: usability and security of authentication using brainwaves*," in *International Conference on Financial Cryptography and Data Security*, pp. 1–16, Springer, 2013.

[37] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 341–350, 2012.

[38] T. Rahman, A. T. Adams, M. Zhang et al., "Bodybeat: a mobile system for sensing non-speech body sounds," *MobiSys*, vol. 14, pp. 2594368–2594386, 2014.

[39] X. Sun, L. Zongqing, H. Wenjie, and G. Cao, "Symdetector: detecting sound-related respiratory symptoms using smartphones," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 97–108, 2015.

[40] T. Hao, G. Xing, and G. Zhou, "Isleep: unobtrusive sleep quality monitoring using smartphones," in *Proceedings of the 11th ACM Conferenceon Embedded Networked Sensor Systems*, pp. 1–14, 2013.

[41] T. Rahman, A. T. Adams, M. Zhang et al., "Intelligent sleep stage mining service with smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 649–660, 2014.