Sound & Vibration

Tech Science Press

DOI: 10.32604/sv.2024.045193

ARTICLE

GestureID: Gesture-Based User Authentication on Smart Devices Using Acoustic Sensing

Jizhao Liu[1,2], Jiang Hui[1,2,*] and Zhaofa Wang[1,2]

[1]School of Computer Science, Zhongyuan University of Technology, Zhengzhou, 450007, China

[2]Henan Development and Innovation Laboratory of Industrial Internet Security Big Data (Zhongyuan University of Technology), Zhengzhou, 450007, China

[*]Corresponding Author: Jiang Hui. Email: hui.jiang@zut.edu.cn

Received: 20 August 2023 Accepted: 26 December 2023 Published: 19 March 2024

ABSTRACT

User authentication on smart devices is crucial to protecting user privacy and device security. Due to the development of emerging attacks, existing physiological feature-based authentication methods, such as fingerprint, iris, and face recognition are vulnerable to forgery and attacks. In this paper, GestureID, a system that utilizes acoustic sensing technology to distinguish hand features among users, is proposed. It involves using a speaker to send acoustic signals and a microphone to receive the echoes affected by the reflection of the hand movements of the users. To ensure system accuracy and effectively distinguish users' gestures, a second-order differential-based phase extraction method is proposed. This method calculates the gradient of received signals to separate the effects of the user's hand movements on the transmitted signal from the background noise. Then, the second-order differential phase and phase-dependent acceleration information are used as inputs to a Convolutional Neural Networks-Bidirectional Long Short-Term Memory (CNN-BiLSTM) model to model hand motion features. To decrease the time it takes to collect data for new user registration, a transfer learning method is used. This involves creating a user authentication model by utilizing a pre-trained gesture recognition model. As a result, accurate user authentication can be achieved without requiring extensive amounts of training data. Experiments demonstrate that GestureID can achieve 97.8% gesture recognition accuracy and 96.3% user authentication accuracy.

KEYWORDS

Acoustic sensing; hand gesture; user authentication

1 Introduction

User authentication on smart devices is crucial for many everyday applications. Several user authentication technologies, such as fingerprint, iris, and face recognition, have been widely used. The accuracy and convenience of biometric authentication have met people's daily needs, but there are some limitations. Fingerprint-based user authentication methods, such as Apple's TouchID [1], identify users by recognizing their unique finger texture. However, this method is vulnerable to spoofing with fingerprint films. Iris-based recognition can achieve high accuracy [2], but it requires expensive, specialized sensors. Two-dimensional face recognition is prone to replay attacks on the user's facial image or video [3], while

three-dimensional face recognition systems, such as FaceID [4], require special sensors and are affected by environmental factors such as light and angle.

Acoustic sensing techniques have been extensively studied with the development of pervasive computing and the Internet of Things (IoT). These techniques enable various applications, including coarse-grained motion sensing (e.g., gesture recognition [5–7] and gait recognition [8]), as well as fine-grained activity sensing (e.g., tracking [9–11], lip reading [12–14], breath monitoring [15,16]), and user authentication (e.g., liveness detection [17,18], facial authentication [19,20], lip authentication [21,22]). Acoustic sensing for smart device authentication provides convenience, low cost, non-intrusiveness, and accurate authentication.

VoiceLive [17] and VoiceGesture [18] proposed a liveness detection mechanism using a voice-based authentication method. It achieves over 99% detection accuracy by utilizing time-difference-of-arrival (TDoA) and Doppler frequency shifts of the received signal. However, both VoiceLive and VoiceGesture require that the distance between the user's mouth and the phone must be within 3 cm. This limitation can result in a poor user experience. LipPass [21] and SilentKey [22] proposed a lip-reading-based user authentication system that utilizes Doppler shifts and the signal envelope of the acoustic signal to derive distinct behavioral traits of users' speaking lips for authentication. However, the proposed method lacks the ability to eliminate the interference of multipath and unpredictable phase, and relies on a quantity of training samples. HandLock [23] proposed a gesture authentication method that relied on acoustic sensing. They employed a supervised machine learning approach to achieve an average true-positive-rate (TPR) of 96.51%. However, the proposed method requires the manual extraction of features to serve as input for the machine learning model. The ideal acoustic authentication mechanism should (1) use existing sensors without extra equipment, (2) be more convenient and flexible to use, (3) require a small number of samples for per user, and (4) achieve higher authentication accuracy [24].

To address these issues, we propose GestureID, a user authentication system based on hand motion sensing. It utilizes acoustic sensing technology to capture unique behavioral features of users' hand movements. GestureID employs a speaker for transmitting multiple-frequency ultrasonic waves and a microphone for receiving echoes affected by reflections caused by the user's hand movements. After pre-processing, which includes I/Q modulation, denoising, and segmentation, the CNN-BiLSTM model is employed for classification. Compared to existing work [21,23], the method proposed in this paper has the following advantages: (1) During the signal processing stage, we transmit continuous waves of various frequencies and extract phase and acceleration information from each frequency waveform to capture the hand motion features of users. (2) To improve the signal sensing capability, we calculate the gradients of the received signals to eliminate static components that are unrelated to hand movements. (3) To train the user authentication model, we employ a transfer learning approach, which enhances the accuracy and generalization capabilities of authentication without requiring extensive training data.

In summary, this paper makes the following main contributions:

- The feasibility of using acoustic signals from smart devices to extract hand movement features for user authentication is investigated. The proposed method can achieve convenience, low cost, non-intrusiveness, and high authentication accuracy.
- A method is proposed for phase extraction that is based on second-order differences. This method aims to improve the sensing capability of the signal by calculating the gradient of the received signals. The objective is to eliminate the static components from the signal.
- A CNN-BiLSTM network is used to train a gesture recognition model. Then, the pre-trained model is fine-tuned using transfer learning to create a user authentication model for each gesture. This approach allows us to achieve higher authentication accuracy and generalization capability, even with limited training data.

- A GestureID prototype is developed on a smartphone, and experiments are performed in a real-world environment. Experiments reveal that GestureID achieves 97.8% accuracy in gesture recognition accuracy and 96.3% accuracy in user authentication.

The rest of the paper is organized as follows. In Section 2, we explore related work. Section 3 describes the comprehensive architecture of GestureID. Section 4 presents the evaluation results of GestureID. Section 5 concludes the paper.

2 Related Work

2.1 Acoustic-Based Applications

Acoustic-based sensing methods have been widely studied and have been used in many fields, such as activity and gesture recognition, localization and tracking, and lip reading, due to their popularity, convenience, low cost, non-intrusiveness, and high perceptual accuracy. Both SoundWave [5] and AudioGest [6] used the Doppler effect of acoustic signals to recognize different hand gestures. RobuCIR [7] combined frequency-hopping and CIR information to reduce frequency selective fading and prevent signal interference for precise and reliable contactless gesture recognition in various scenarios. CAT [10] utilized a distributed Frequency Modulated Continuous Waveform (FMCW) and Doppler effect to achieve sub-millimeter accuracy in tracking cell phone motion. LLAP [11] utilized the phase change of the received baseband signal due to finger motion for finger movement tracking. SoundLip [14] implemented an end-to-end lip-sync interaction system using the Doppler features of multi-frequency ultrasonic sine wave signals. Each of these studies employs acoustic sensing to identify user activity.

2.2 Gesture-Based Authentication

User gesture movements have unique behavioral features that can be utilized for authentication. Gesture-based authentication and recognition have been extensively studied. Hong et al. [25] proposed WA, a system for motion gesture authentication that relies on accelerometers. This system utilized eight identification features that were hidden within the acceleration trajectory of motion gestures. The classification of these features is performed using a single-class support vector machine. WiID [26] extracts velocity time series features from WiFi channel state information (CSI) to identify users who perform redefined gestures. FingerPass [27] utilized the CSI phase of WiFi signals to capture and differentiate the unique behavioral features of various users and authenticate them consistently during each finger gesture interaction. However, WiFi signals can be affected by environmental interference and require users to be within a WiFi-covered environment, thereby restricting their usefulness. Au-Id [28] proposed an RFID-based authentication method performed through continuous daily activities. Based on the correlation between RFID tags and infrastructure, Au-Id stacks a CNN with long short-term memory (LSTM) to automatically tag different activities. The data is tagged and then inputted into another CNN for user identification, resulting in an average identification accuracy of 97.72%. However, the RFID-based approach requires additional dedicated equipment for the user. Acoustic signals have an advantage over WiFi and RFID because they do not need extra sensors. Good authentication can be achieved using only the smart device's built-in speaker and microphone.

2.3 Acoustic-Based Authentication

In addition to the user authentication methods mentioned above, acoustic sensing techniques for user authentication have gained significant attention. Many excellent works have been proposed in this field. These techniques can be categorized into physiological biometric-based authentication and behavioral biometric-based authentication, depending on the type of identity information used for authentication.

Physiological biometric-based authentication uses physiological features such as the user's face or vital signs to authenticate. EchoPrint [19] employed the FMCW method to identify the depth information of faces

for two-factor user authentication. However, it lacks resistance against 3D face forgery attacks. BreathPrint [29] used human breathing to generate audible sounds and extract acoustic features. These features include the signal strength of acoustic signals and MFCC. Machine learning methods are then applied for authentication. However, as the breathing sound falls within the audible frequency range, the method is vulnerable to interference from surrounding environmental noise. Additionally, the breathing sound may change greatly when the user performs intense exercise, thereby significantly affecting the authentication performance.

Behavioral biometric authentication employs acoustic signals to detect user behaviors, including lip movements and gestures, for the purpose of authentication. VoiceLive [17] utilized TDoA of the received signal to detect user liveness. However, TDoA measures absolute distance, so users must keep the phone in a consistent position for each use. VoiceGesture [18] proposed to use the Doppler shift of acoustic signal to model users' mouth movements during speech for identity authentication. However, all these works suffer from a limited effective authentication distance (within 3 cm) and poor user experience. LipPass [21] used a three-layer autoencoder network to extract the acoustic Doppler effect from lip movements. It then employed SVM for user identification. SilentKey [22] also employed an acoustic Doppler shift for a lip-reading-based user authentication system. However, both works authenticate users based on the mouth state during speech, but only in a password-related way, and require a significant amount of training samples. HandLock [23] converted the acoustic phase time series into velocity and acceleration series for user authentication using a machine learning model. However, the proposed method requires manual extraction of features as input to the machine learning model.

To address the limitations of the above work, our work aims to use the unique behavioral features of the user's hand movements to authenticate the user's identity. By using only the smart device's built-in speaker and microphone, we can send ultrasonic waves of various frequencies through the speaker. The microphone then picks up the echoes, which are influenced by the reflection of the user's hand movements. To remove the static components of the received signal that are not related to hand movements, we use a second-order difference-based phase extraction method. This method enhances the sensing capability by calculating the gradient of the received signals. The extracted second-order differential phase and phase-dependent acceleration information are then utilized as inputs to a CNN-BiLSTM model for the identification of various hand gestures. Finally, the user authentication model is trained for each gesture by fine-tuning the pre-trained gesture recognition model for verification users using a transfer learning approach.

3 System Design

This section describes the design of GestureID, a user authentication system that utilizes hand motion sensing. It implements user authentication based on gesture recognition on commercial smartphones by utilizing the phase information of acoustic signals.

3.1 System Overview

GestureID is comprised of four modules, as shown in Fig. 1: signal collection, signal pre-processing, model construction and verification. During the signal collection stage, the smartphone emits inaudible near-ultrasonic sound between 17–23 kHz using its built-in speaker. This frequency range is supported by most smart devices for sending sound waves. Sound above 17 kHz are typically inaudible to most people [30] and do not cause any disturbance. The microphone receives the echoes affected by the reflection of users' hand movements, carrying rich information about users' hand movements in the received signal. During the signal pre-processing stage, the received signal is subjected to I/Q modulation to extract the hand movement-induced changes. In order to enhance the sensing capability and remove non-hand motion-related static components of the received signal, we employ a second-order difference-based phase extraction method. This method involves calculating the gradient of the signals. The model

construction module constructs a gesture recognition model using a CNN-BiLSTM network. Then, a transfer learning approach is used to build a user authentication model for each predefined gesture. This is done by fine-tuning the pre-trained gesture recognition model. The gesture recognition module detects and recognizes a gesture. It then sends the gesture to the user authentication model for authentication. The trained models are stored in a database for further deployment when needed. The model design details are outlined in the following sections.

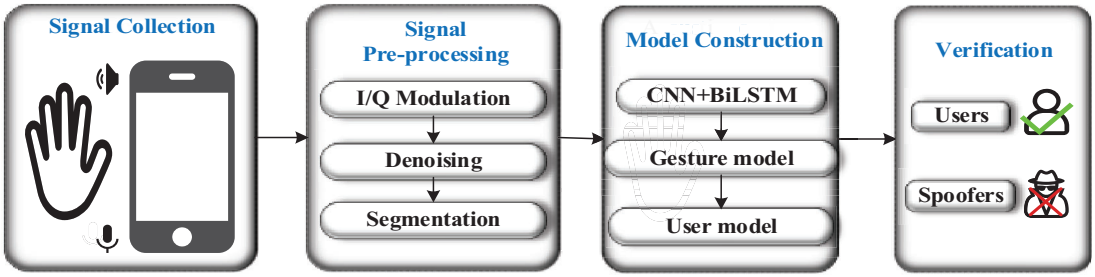


Figure 1: Overview of GestureID

3.2 Signal Collection

During signal collection, commercial smart devices utilize speakers and microphones as active sonar to sense the surroundings. GestureID utilizes inaudible near-ultrasound with frequencies ranging from 17–23 kHz. This frequency range corresponds to acoustic wavelengths of less than 2 cm. Consequently, even slight movements of a few millimeters can significantly change the phase of the received sound waves [11]. In this paper, eight frequency signals are used, each with a frequency interval of 700 Hz. The signals of eight frequencies are summed and normalized to $A\sum \cos 2\pi ft$, where A is the amplitude of signals, the eight frequencies start at 17350 Hz and end at 22250 Hz. The speaker sends acoustic signals, and the microphone receives echo signals at a sampling rate of 48 kHz. The transmitter and receiver have no carrier frequency offset (CFO) because they use the same device to send and receive signals.

3.3 Signal Pre-Processing

3.3.1 Limitations of Doppler Shift

Several studies utilize the Short-Time Fourier Transform (STFT) to calculate the Doppler shift and estimate the condition of surrounding reflectors. However, the resolution of STFT is limited by the constraints of time-frequency analysis [14]. For instance, with a segment size of 2048 samples and a sampling rate of 48 kHz, the STFT has a frequency resolution of 23.4 Hz. This corresponds to a movement speed of 0.2 meters per second when the sound wave has a frequency of 17 kHz. In other words, the minimum detectable speed of hand movement using the STFT approach is 20 cm/s [14]. Therefore, the Doppler shift can only provide a coarse-grained measurement of the speed or direction of hand/finger movement [11]. Fig. 2a shows the STFT result of a moving hand when making the gesture "W". The frequency of the signal is 21.55 kHz. The figure only provides a rough observation of hand movements, whereas our method can capture more detailed time-series information such as amplitude, phase, and acceleration, as shown in Fig. 2b.

3.3.2 Signal I/Q Modulation

To overcome the limitations of the Doppler shift, we utilize the amplitude, phase, and acceleration information of the received signal to accurately analyze subtle hand movements. To extract the phase and amplitude from the received signal, we initially down-convert the received signal. This conversion aims to convert high-frequency signals into low-frequency signals, resulting in a significant reduction in data

size and an enhancement in signal processing efficiency. The Doppler shift caused by hand movement is below 100 Hz [23]. Thus, we use a fifth-order Butterworth Band-Pass Filter [31] to extract the target frequency band from the received signals, which is $[f_0 - 100, f_0 + 100]$ Hz. Then, we utilize the conventional coherent detector structure depicted in Fig. 3 to convert the received signal from passband to baseband signal [32]. The received signal is divided into two identical copies, which are multiplied with the transmitted signal $\cos 2\pi ft$ and its phase shift version $-sin2\pi ft$. Then, we use a fifth-order Butterworth Low-Pass Filter (LPF) [31] with a stop frequency at 100 Hz to eliminate high-frequency components and obtain the in-phase (I-component) and quadrature (Q-component) components of the baseband signal. The in-phase and quadrature components are two-dimensional projections of a signal, used to drive the signal's phase and amplitude. The phase of the received signal is directly related to the length of the acoustic propagation path. Hand movement can change the path length, leading to phase fluctuations. The speed of motion is proportional to the change in phase, and the acceleration related to the phase change can be obtained by taking the derivative of the phase change. Fig. 4b shows the amplitude, phase, and acceleration profiles obtained from the same gesture record that generates the spectrogram in Fig. 4a. The patterns caused by hand movements can be clearly observed.

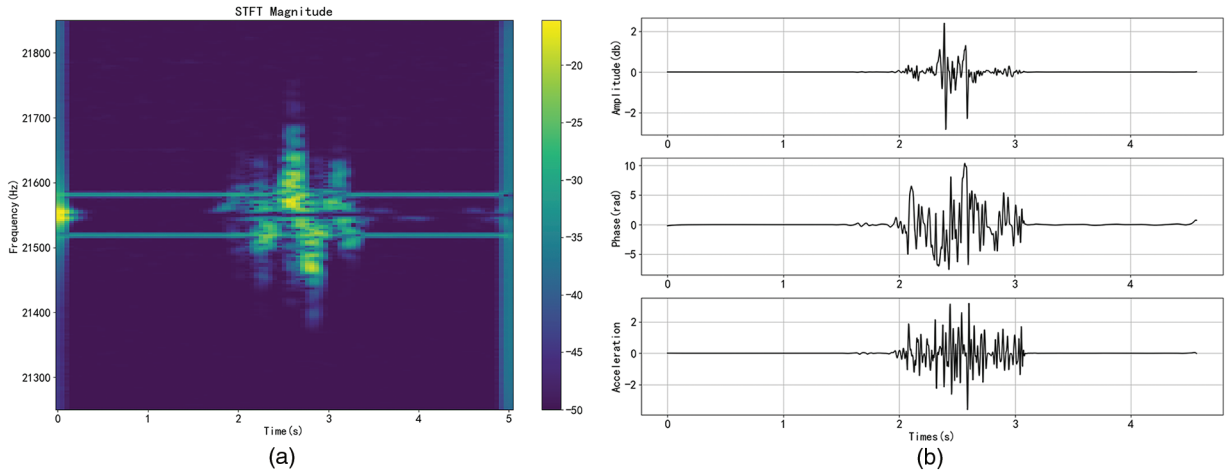


Figure 2: Acoustic signals of hand movements. (a) Doppler shift. (b) Amplitude, phase, and acceleration

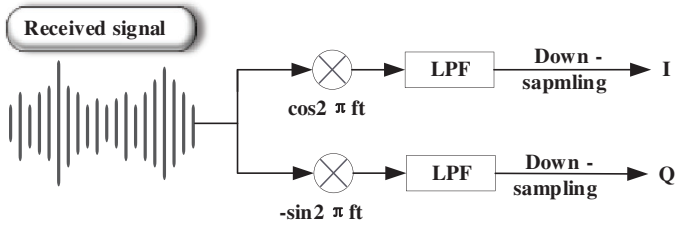


Figure 3: I/Q modulation process

3.3.3 Signal Denoising

The signal received by the microphone is a mixed signal with multipath propagation. In addition to dynamic signals caused by hand movements, there are also static signals. These include LOS signals (i.e., signals propagating directly from the speaker to the microphone) and signals reflected from the surroundings (e.g., obstacles and the body). In addition, the static signal may also change slowly with the

movement of the user's hand or body. Fig. 4 shows the amplitude, phase, and acceleration curves of the gesture "W" with a frequency of 17350 Hz. It can be seen that after 3.1 s (the end time of the gesture movement), the phase is still slowly changing and there is noise. To remove irrelevant data, we compute the gradient of the received signals [14]. This gradient indicates the phase (amplitude) difference between two consecutive samples at times t-1 and t. Wavelet-based denoising techniques are then employed. The objective of wavelet denoising [33] is to eliminate noise from signal. Figs. 4a and 4b show the signals before and after denoising, respectively. After denoising, it is evident that the amplitude, phase, and acceleration become almost zero when there is no gestural motion. This confirms that the static signals are effectively reduced.

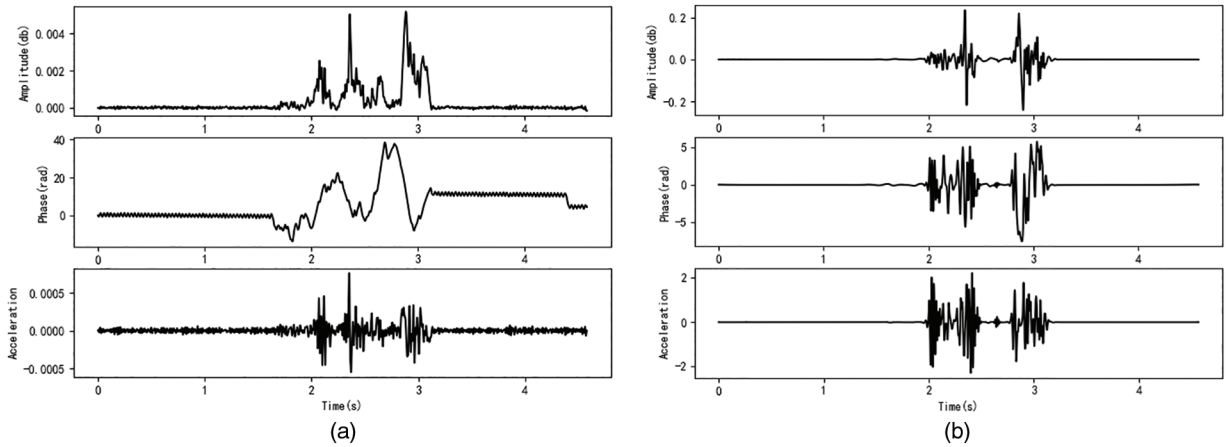


Figure 4: Amplitude, phase, and acceleration of the gesture "W". (a) Without denoising. (b) With denoising

3.3.4 Signal Segmentation

After filtering, the start time and end time of the gesture need to be detected to extract the part of hand movement. Given the use of an eight-frequency signal, eight phase curves are generated. Additionally, eight differential phase curves are utilized for hand motion segmentation. Specifically, assume that $\theta_i(t) = \{i : i = 0, 1, ..., 7\}$ is the differential phase value of eight-frequency signal at moment t, we take the variance of the eight-frequency phase values at moment t:

$$Var_t = \frac{1}{8}\sum_{i=1}^{8} (\phi_i - \bar{\phi})^2 \tag{1}$$

where $\bar{\phi}$ represents the mean of eight-frequency phase values at moment t, i.e.,

$$\bar{\phi} = \frac{1}{8}\sum_{i=1}^{8} \phi i \tag{2}$$

Fig. 5a shows the variance curve of four gesture "W" samples. It can be seen that the variance is zero in the absence of hand movement.

After the variance curves are obtained, the first step is to calculate the standard deviations of the variance curves. After experimenting with different sliding windows, we chose a sliding window of size 200 with a step size of 10 (overlap rate of 5%) to obtain a series of standard deviations. After obtaining the standard deviation curve, we employ a dynamic threshold selection scheme using the criteria mentioned in [34]. Based on our experiment, we observed that the standard deviation curve is close to zero without any hand movement. However, in the presence of hand movements, it becomes non-zero, as shown in

Fig. 5b. Therefore, we determine the threshold σ by the maximum and minimum values of the standard deviation curve of the signal within the frame. The formula for selecting the threshold is as follows:

$$\sigma = min + 0.1 \times (max - min) \tag{3}$$

where *min* and *max* are the minimum and maximum of the variance curve in the current frame, respectively. The points larger than the threshold value indicate the gesture motion part with subscripts $i_1, ..., i_n$. We use these subscripts to calculate the start frame t_{start} and end frame t_{start} of hand motion can be calculated as follows:

$$\begin{cases} t_{start} = i_{start} \times l \times (1 - p) + \frac{l}{2}, \\ t_{end} = i_{end} \times l \times (1 - p) + \frac{l}{2}, \end{cases} \tag{4}$$

where l and p represent the size of the sliding window and the overlap rate, respectively. Fig. 5a shows the variance curve of four gestures "W" with the start (marked as blue line) and end (marked as red line). Fig. 5b shows the standard deviation curve of the variance curve in Fig. 5a and the corresponding threshold value. Fig. 6 shows the differential amplitude, phase, and acceleration curves of the segmented eight-frequency signal.

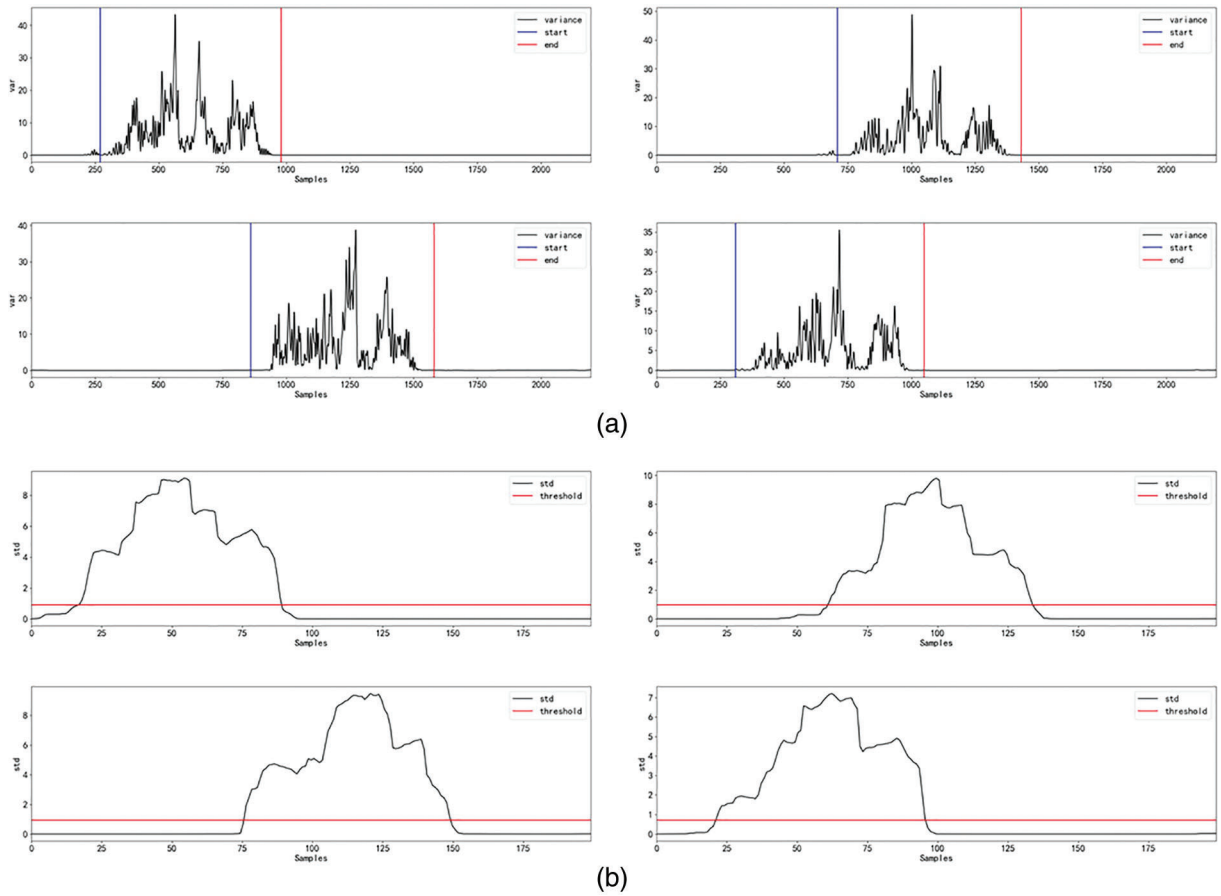


Figure 5: Detection of the start and end points of a gesture within a signal. (a) Variance of multi-frequency phase. (b) Sliding standard deviation of the variance curve

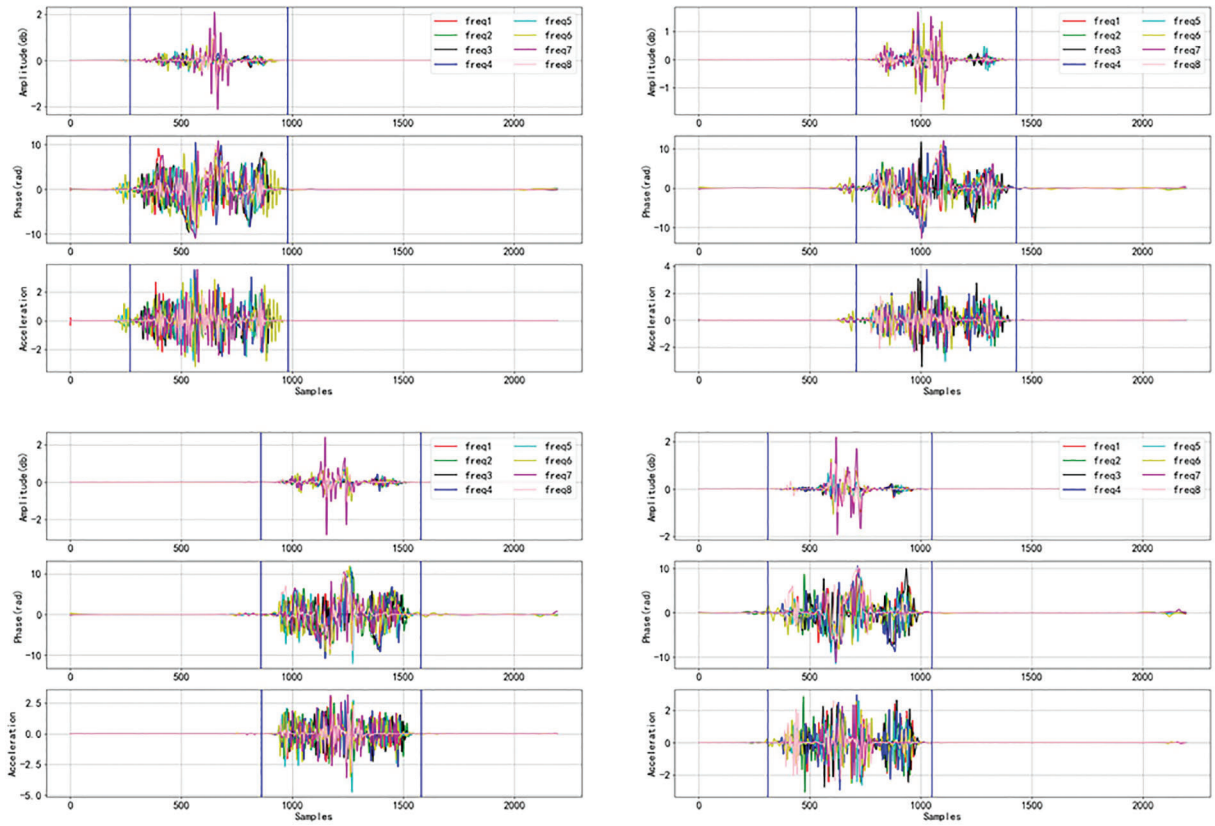


Figure 6: Amplitude, phase, and acceleration signals after segmentation

Since different users take different time to complete a gesture, and even the same user takes different times to complete a gesture, the speed variation of the gesture movement must be handled. The average duration of a gesture in our dataset is approximately 1.5 s, with a sampling frequency of 480 Hz. To ensure equal sample sizes for each gesture data, the data is resampled to 720 samples. Specifically, for a gesture segment containing M data points, GestureID up-samples θ_i to 720 points if M < 720, while it down-samples θ_i to 720 points if M > 720. Then, we perform maximum-minimum normalization on the resampled time-series, ensuring that all values are mapped between −1 and 1. This process ensures that the range of the same gesture action is consistent, making it easier to recognize and authenticate. It can be calculated as follows:

$$X_{norm} = \frac{2 \times (X - X_{min})}{X_{max} - X_{min}} - 1 \tag{5}$$

where X is the original gesture data, X_{min} and X_{max} are the minimum and maximum values in the original gesture data set, respectively. X_{norm} is the normalized gesture data value at the current moment. The differential amplitude, phase, and acceleration curves of resampled and normalized eight-frequency signals are shown in Fig. 7.

3.4 Model Construction

Feature vectors. In this paper, we consider the proportional relationship between the speed and acceleration of hand motion and the phase change and its acceleration. Therefore, we use phase and acceleration sequences as inputs for the neural network. Additionally, we incorporate the characteristics of

acoustic signals into the design of the network model. First, the phase profiles are time-series signals with eight carrier frequencies. To calculate their delta features, we apply the concept of speech recognition and compute the second-order difference of phase [35]. The acceleration of the second-order differential phase is then computed to extract hand motion acceleration features. These 16 feature dimensions are used as inputs to the neural network, which include eight second-order differential phase dimensions and eight acceleration dimensions.

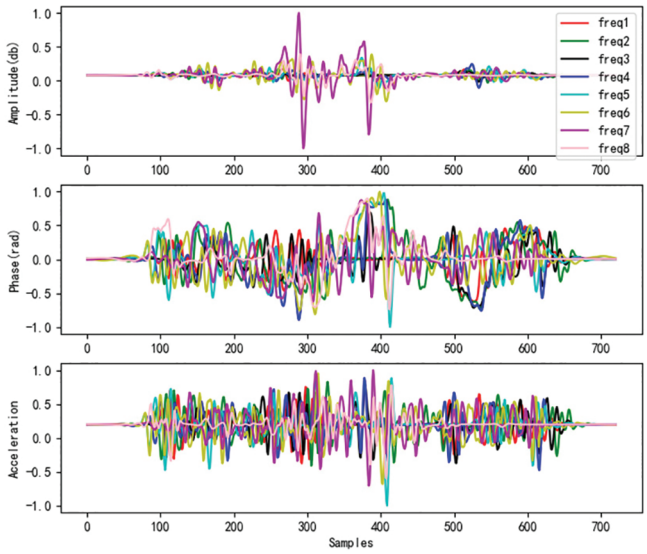


Figure 7: Amplitude, phase, and acceleration signals after resampling and normalization

Model structure. The neural network model must combine the phase and acceleration data. This is because the phase and acceleration data for the eight frequencies are one-dimensional vectors with a time sequence length of 720. To accomplish this, we utilize a one-dimensional convolutional neural network (1D-CNN) [36]. This type of network, utilizing weight sharing through one-dimensional convolutional kernels, enables the extraction of features from each dimension of the one-dimensional vector, serving as a more comprehensive feature information extraction network. In a one-dimensional CNN, it is required to predefine one-dimensional convolutional kernels. Subsequently, information is extracted from the one-dimensional vector using the specified stride. This process captures information from dimensions and extracts inter-dimensional connection information. Due to its structural characteristics, recurrent neural networks (RNNs) maintain memory based on historical information, making them suitable for processing sequential data. Long Short-Term Memory (LSTM) networks are explicitly designed to address long-term dependencies by using specialized memory units, demonstrating superior performance in longer sequences. Bidirectional LSTMs (BiLSTM) take into account the comprehensive impact of preceding and succeeding action information on the current moment. In terms of temporal signals, BiLSTM [37] has a more powerful capability for extracting representations compared to LSTM. It achieves this by considering both previous and subsequent information, while also addressing problems like gradient vanishing and exploding. Therefore, we developed a CNN-BiLSTM deep neural network model (Fig. 8) to extract hand motion features from multidimensional temporal data and perform gesture classification and user authentication. The model is composed of four parts: the input layer, CNN layer, BiLSTM layer, and output layer. The input data of the model is the multidimensional time series obtained in Section 3.3. Therefore, we first use three layers of one-dimensional convolution to extract the local features of the

time series, followed by two BiLSTM layers to model the sequence and capture its temporal information. The output layer of the model is a fully connected layer. The probability prediction of each category is obtained using a softmax function. For all layers, ReLU is used as the activation function. L2 regularization methods are used at each CNN and BiLSTM layer to mitigate overfitting. Moreover, dropout layers are added after each layer in the model with a dropout rate of 0.2. Table 1 shows the details of the CNN-BiLSTM model.

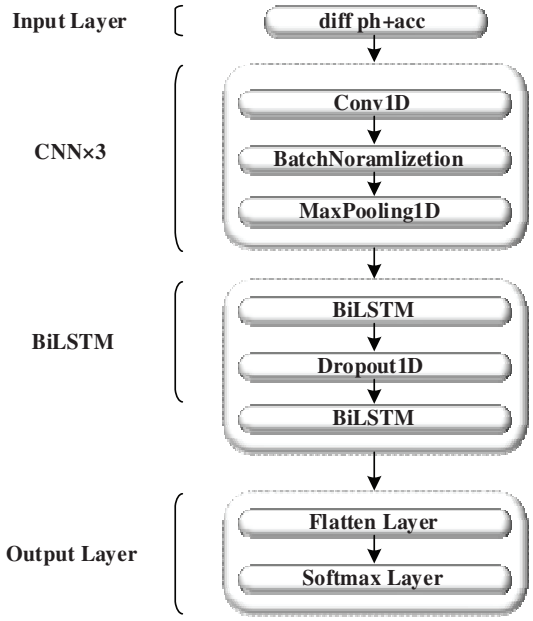


Figure 8: The architecture of CNN-BiLSTM model

Table 1: Details of CNN-BiLSTM model

Layer	Layer type	Output shape	# Param
1	Conv1D + ReLU	(1,720,128)	18560
2	MaxPooling1D	(1,360,128)	
3	Conv1D + ReLU	(1,360,128)	147584
4	MaxPooling1D	(1,180,128)	
5	Conv1D + ReLU	(1,180,256)	295168
6	BiLSTM	(1,180,512)	1050624
7	BiLSTM	(1,180,512)	1574912
8	Dropout1D	(1,180,512)	
9	Dense + ReLU	(1,180,128)	65664
10	Flatten	(1,23040)	
11	Dense + Softmax	(1,5)	115205

1D-CNN. Given that the input signal is a time series, the 1D convolutional neural network can be employed to extract features in each dimension of the 1D vector using a 1D convolutional kernel with shared weights. The CNN network in this paper comprises three one-dimensional convolutional layers and two max pooling layers. The convolution kernel extracts the hand motion features of the time series, and the maximum pooling layer down-samples the feature maps after the 1D convolution. Batch normalization (BN) is used in every convolutional layer for faster and more stable training [38]. BN also aids in preventing parameters from falling into poor local minima. The output of the CNN network is a 1 × 180 × 256 feature map, which is used as input to the BiLSTM network.

BiLSTM. The BiLSTM layer consists of a forward LSTM and a backward LSTM together. BiLSTM is a variant of RNN. It has a stronger ability to extract forward and backward information representations than LSTM, especially for temporal data. In this study, we employ a two-layer BiLSTM with 256 hidden units in each layer. With the bi-directional LSTM, the information of the input sequence can be learned more comprehensively, including its order and context.

Output layer. In the output layer of the model, we use the softmax activation function as the classifier. The final output layer of the gesture recognition network uses a softmax function to output the probability of five gestures, denoted as $p = p_i(p_1, p_2, ..., p_5)$, where p_i denotes the probability that the input gesture belongs to the i-th gesture.

Transfer learning. Retraining the model when a new registered user joins is time and resource consuming, and the feature extraction capability is insufficient. Therefore, for user authentication, we utilized a transfer learning [39] approach with pre-training and fine-tuning [40]. Each gesture has its own user authentication model, which is trained using a pre-trained gesture recognition model. The pre-trained model is truncated at the softmax layer and replaced with a sigmoid function for binary classification. It identifies if the user is registered or a spoofer. The objective of the user authentication network is to improve authentication accuracy and generalization capability using a limited number of training samples.

4 Evaluation

4.1 Experimental Setup

Experimental Equipment. We utilized a HUAWEI Mate40 smartphone with HarmonyOS 3.0 operating system and a Dell Inspiron 5577 laptop for data collection and processing. Keras [41] with a TensorFlow [42] backend was employed to construct and train the neural network offline. The CNN-BiLSTM models were trained offline on a server with 32 G of RAM and an Intel i7-8700k@3.7 GHz processor, along with a GeForce RTX 2080 graphics card.

Data Collection. Due to the absence of a publicly available dataset for researchers in the field of acoustic sensing, both domestically and internationally, we obtained real data on gestures from multiple volunteers and formed a dataset. Nine volunteers (four males and five females, aged 18–28 years) were invited to participate in the data collection. Before the experiment, they were informed about the purpose and process of this study. Data were collected in three environments: laboratory, bedroom, and living room. To evaluate the effectiveness of our gesture recognition and user authentication models, we referred to existing literature [23] and focused on five common gestures: “+”, “O”, “W”, “X”, and “Z” (shown in Fig. 9). Before collecting the data, the volunteer was asked to practice the gestures several times so that he or she could understand the data collection process. Each participant was instructed to perform a given gesture repeatedly, with a brief pause between each gesture. The participants were required to place their hands within a range of 5~30 cm from the smartphone. The collection process relied solely on the user’s habits. Five popular gestures were repeated 50 times by nine volunteers, and the collection process lasted for one month. In total, we collected 2250 gesture samples from various ages, genders, and environments.

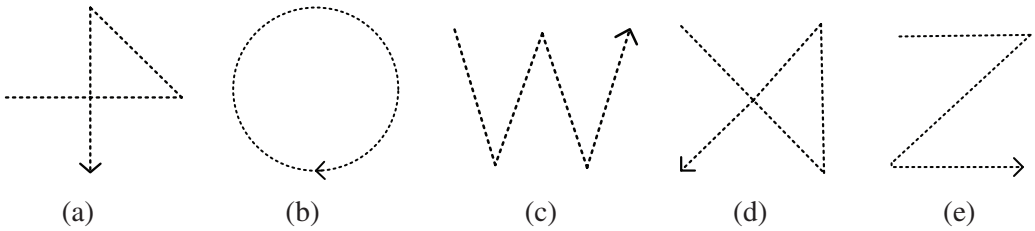


Figure 9: Different types of gestures evaluated

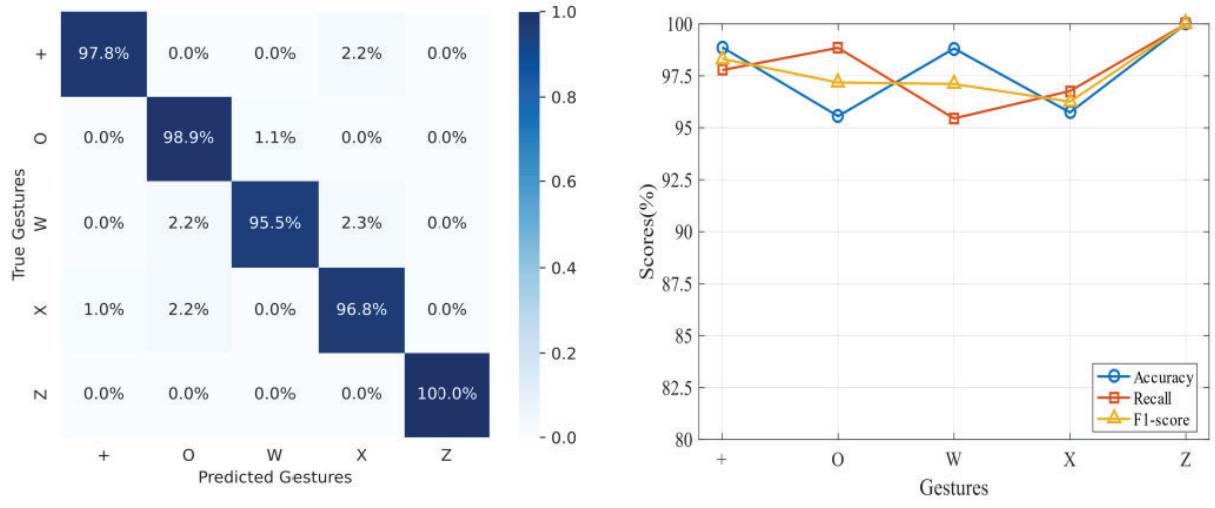
Evaluation Protocol. To comprehensively evaluate the performance of system, we use the following metrics:

- Confusion Matrix: Each row of the matrix represents the true label (ground truth), while each column represents the classification result of the system.
- Accuracy (Ac = $TP + TN/TP + TN + FP + FN$): the ratio of the number of all correctly classified samples to the total number of samples.
- Recall rate ($Re = TP/FN + TP$): percentage of true positive classifications in all target class instances.
- F1-Score ($F1 = 2 \times PR \times RE/PR + RE$): the weighted summed average of Precision and Recall.

4.2 Overall Performance

In this section, we evaluate the performance of GestureID in the gesture recognition phase and the user authentication phase.

First, to ensure model generalization, we employ the k-fold cross-validation to assess gesture recognition performance. Specifically, the training and test sets are divided in a ratio of 8:2. Then, the original training set is randomly divided into k mutually non-overlapping subsets. Out of these subsets, k − 1 subsets are sequentially used as the training set, while the remaining one is used as the validation set. In this paper, we use ten-fold cross-validation, i.e., k = 10. Fig. 10 shows the confusion matrix, accuracy, recall, and F1 score of the gesture recognition model. The recognition accuracy of each gesture is above 95%. The average recognition accuracy of GestureID for five gestures is 97.8%. The accuracy, recall, and F1 score are above 95%, proving the good performance of GestureID in the gesture recognition stage.



(a) **(b)**

Figure 10: Performance on gesture recognition. (a) Confusion matrix of gesture recognition. (b) accuracy/recall/F1 score of gesture recognition

To train user authentication models for the five predefined gestures, we adopt a transfer learning approach. This involves utilizing a pre-trained gesture recognition model. This approach is also effective in decreasing the time and resources needed to retrain the model when a new user is added, while enhancing the model's ability to extract features. We employ the K-fold cross-validation approach to train the user authentication model, evaluating each gesture. In the user authentication model, we employ five-fold cross-validation to split the training set and test set in an 8:2 ratio. Five users were randomly selected to create the registered users, while the remaining four users were designated as spoofers. Fig. 11 shows the accuracy of user authentication for the five gestures. The average authentication accuracy of GestureID is 96.3%, and the authentication accuracy of all five gestures is above 94%. GestureID's effectiveness lies in its ability to differentiate between registered users and unknown spoofers, thereby enhancing the accuracy of user authentication.

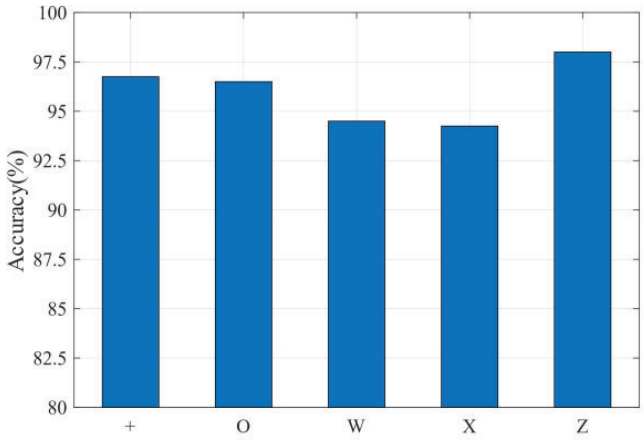


Figure 11: Authentication accuracy of GestureID

4.3 Impact of Signal Processing

Signal processing methods play a key role in the authentication task. In this study, we utilize phase and acceleration information (ph and acc) as an alternative to coarse-grained Doppler frequency shifts for capturing hand motion. To eliminate static signals in the received signal that is not related to the hand motion, we use a second-order differential-based phase extraction method that further enhances the sensing capability of the signal by calculating the gradient of the received signals. The extracted second-order diff phase and the phase-dependent acceleration info (diff ph+acc) are inputs to the CNN-BiLSTM model. The accuracy of different pre-processing methods is presented in Fig. 12. In particular, ph+am, ph +acc, diff ph+am and diff ph+acc are the four types of input features. The last one, diff ph+acc, is the technique employed in this paper. In Fig. 12, it is evident that every step of the signal processing process in this paper enhances authentication performance to some extent. The result after taking the second-order difference is much better than the other two features.

4.4 Impact of Transfer Learning

To reduce the time and resource consumption of retraining the model during new user registration and to enhance the performance of user authentication, a pre-trained and fine-tuned transfer learning approach is used to train the user authentication model. In practice, using the transfer learning approach can reduce the sample collection time for new user registration as well as the training time of the network, thus effectively improving the usability of GestureID in real-world scenarios. In this study, we compared the user authentication model trained using the original CNN-BiLSTM network with the model trained using

a pre-trained gesture recognition model through transfer learning. Each network is tuned for 100 epochs to ensure convergence. Figs. 13a and 13b show the accuracy and training time of the two model training methods. It is evident that utilizing the transfer learning approach significantly decreases the model training time and enhances the model performance.

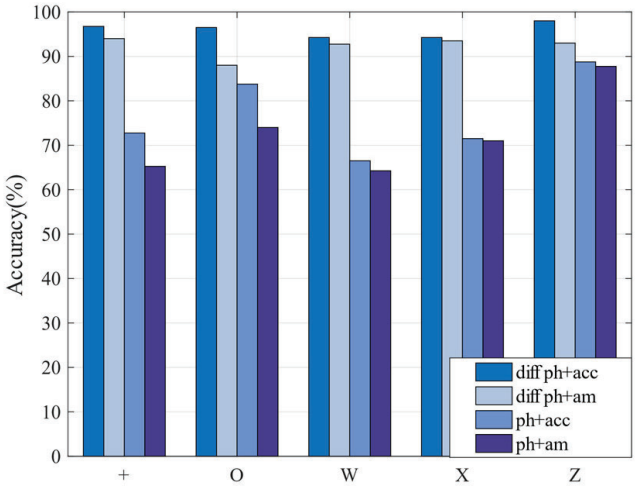


Figure 12: Accuracy of different signal processing methods

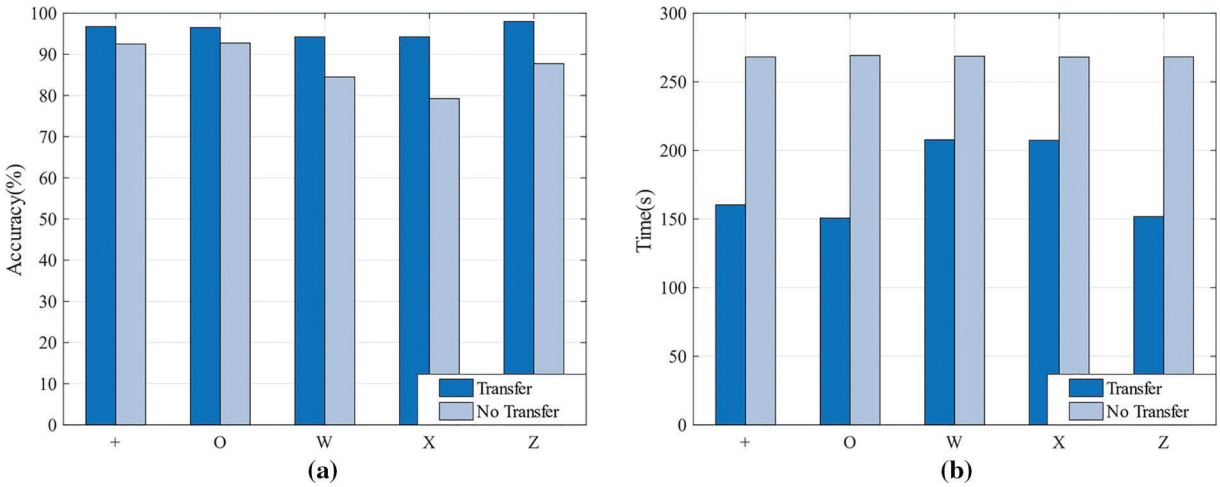


Figure 13: Impact of transfer learning on performance. (a) Impact of transfer learning on accuracy. (b) Impact of transfer learning on training time

4.5 Impact of Training Scale

In practice, the size of the data used to train the network is an important influencing factor. The size of the training set affects the network's performance. Typically, a large training set aids in mitigating overfitting and enhancing the mode's generalization capability. However, users often expect quicker data collection. Therefore, we performed experiments to investigate the effect of different training set sizes on user authentication performance. Specifically, we train the network with varying proportions (0.1 to 0.8) of training set samples, and the remaining samples are used for testing. The user authentication accuracy with the proportion of the training set is plotted in Fig. 14, and the results show that the user

authentication performance shows an increasing trend as the size of the training set increases. After using 50% of the dataset for training, the accuracy of user authentication can exceed 93%. This indicates that it does not require too many training samples to build a good authentication model.

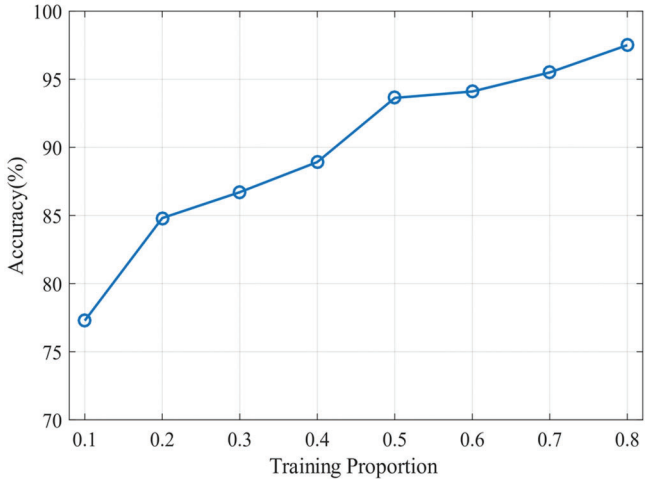


Figure 14: Impact of training set proportions

4.6 Impact of Distance

To evaluate the impact of hand-to-device distance on authentication performance, we tested the authentication model for the gesture "+" at different distances from user to device. Volunteers are asked to perform the "+" gesture at distances from 5 to 40 cm. Fig. 15 shows the accuracy rate at different distances. The authentication accuracy is at least 94% within a 20-cm range. When the distance exceeds 30 cm, the accuracy drops to approximately 89%. The results indicate that GestureID's performance decreases with greater authentication distance. Additionally, the authentication distance is constrained by the signal strength of hand motion sensing. This implies that the sensing range is limited by the power of the transmitted signal. Therefore, it is best for users to keep the distance within 30 cm when using GestureID.

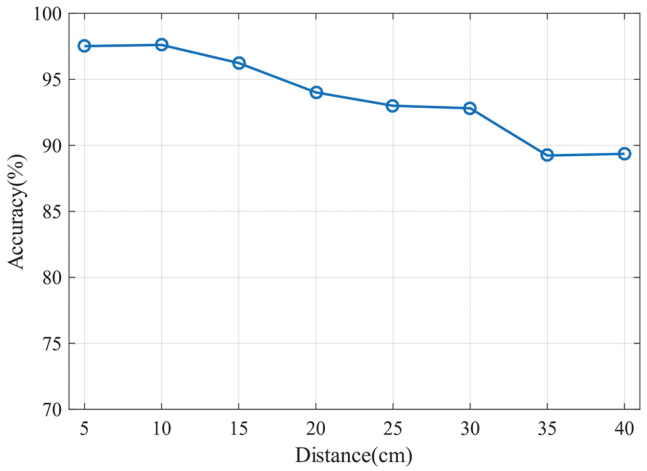


Figure 15: Accuracy of authentication at different distances

5 Conclusions

This paper proposes GestureID, a user authentication system based on hand motion sensing. User authentication is performed by extracting unique behavioral features of users' hand movements using acoustic sensing technology. First, we propose a method for phase extraction that uses second-order differentials to remove static components from received signals. This is performed by calculating the signal gradient of the received signals. Then, the second-order differential-based phase and phase-dependent acceleration information are used as the input of the CNN-BiLSTM model to model the hand motion features. Finally, transfer learning is used to build a user authentication model by utilizing the pre-trained gesture recognition model. This allows us to achieve accurate user authentication without requiring extensive training data. Experiments show that GestureID can achieve 97.8% gesture recognition accuracy and 96.3% user authentication accuracy. In the future, we plan to investigate the adaptability of GestureID to attacks in order to validate the system's security in realistic scenarios.

Acknowledgement: The authors would like to express their gratitude to the editors and anonymous reviewers for their comments and suggestions.

Funding Statement: This research was funded by the Science and Technology Research Program of Henan Province of China (No. 182102210130), "Research on Key Technologies of Practical Quantum-Resistant Authenticated Key Agreement Protocols".

Author Contributions: The authors confirm their contributions to the paper as follows: study conception and design: Liu Jizhao, Jiang Hui, Wang Zhaofa; data collection: Jiang Hui, Wang Zhaofa; analysis and interpretation of results: Liu Jizhao, Jiang Hui; draft manuscript preparation: Liu Jizhao, Jiang Hui. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The experimental data used to support the funding of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Apple. Use touch ID on iPhone and iPad. https://support.apple.com/en-us/HT201371/. (accessed on 15/12/2023).
2. Latman, N. S., Herb, E. (2013). A field study of the accuracy and reliability of a biometric iris recognition system. *Science & Justice, 3(2),* 98–102.
3. Duc, N. M., Minh, B. Q. (2009). Your face is not your password face authentication bypassing lenovo-asus–toshiba. *Black Hat Briefings, 4,* 158.
4. Apple. About face ID advanced technology. https://support.apple.com/en-us/HT208108/. (accessed on 15/12/2023).
5. Gupta, S., Morris, D., Patel, S., Tan, D. (2012). Soundwave: Using the doppler effect to sense gestures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1911–1914. New York, USA.
6. Ruan, W., Sheng, Q. Z., Yang, L., Gu, T., Xu, P. et al. (2016). AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 474–485. Heidelberg, Germany.
7. Wang, Y., Shen, J., Zheng, Y. (2020). Push the limit of acoustic gesture recognition. *IEEE Transactions on Mobile Computing, 21(5),* 1798–1811.
8. Wang, Y., Chen, Y., Bhuiyan, M. Z. A., Han, Y., Zhao, S. et al. (2018). Gait-based human identification using acoustic sensor and deep neural network. *Future Generation Computer Systems, 86,* 1228–1237.

9. Yun, S., Chen, Y. C., Qiu, L. (2015). Turning a mobile device into a mouse in the air. *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 15–29. Florence, Italy.
10. Mao, W., He, J., Zheng, H., Zhang, Z., Qiu, L. (2016). High-precision acoustic motion tracking. *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pp. 491–492. New York, USA.
11. Wang, W., Liu, A. X., Sun, K. (2016). Device-free gesture tracking using acoustic signals. *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pp. 82–94. New York, USA.
12. Zhang, Y., Chen, Y. C., Wang, H., Jin, X. (2021). CELIP: Ultrasonic-based lip reading with channel estimation approach for virtual reality systems. *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pp. 580–585. Virtual, USA.
13. Gao, Y., Jin, Y., Li, J., Choi, S., Jin, Z. (2020). EchoWhisper: Exploring an acoustic-based silent speech interface for smartphone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 4, no. 3, pp. 1–27.
14. Zhang, Q., Wang, D., Zhao, R., Yu, Y. (2021). Soundlip: Enabling word and sentence-level lip interaction for smart devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 5, no. 1, pp. 1–28.
15. Wang, T., Zhang, D., Zheng, Y., Gu, T., Zhou, X. et al. (2018). C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 1, no. 4, pp. 1–20.
16. Wang, T., Zhang, D., Wang, L., Zheng, Y., Gu, T. et al. (2018). Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices. *IEEE Internet of Things Journal, 6(2),* 2959–2973.
17. Zhang, L., Tan, S., Yang, J., Chen, Y. (2016). Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1080–1091. Vienna, Austria.
18. Zhang, L., Tan, S., Yang, J. (2017). Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 57–71. Dallas, USA.
19. Zhou, B., Lohokare, J., Gao, R., Ye, F. (2018). EchoPrint: Two-factor authentication using acoustics and vision on smartphones. *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pp. 321–336. New Delhi, India.
20. Chen, H., Wang, W., Zhang, J., Zhang, Q. (2019). Echoface: Acoustic sensor-based media attack detection for face authentication. *IEEE Internet of Things Journal, 7(3),* 2152–2159.
21. Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y. et al. (2018). LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals. *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 1466–1474. Honolulu, USA.
22. Tan, J., Wang, X., Nguyen, C. T., Shi, Y. (2018). SilentKey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 2, no. 1, pp. 1–18.
23. Zhang, S., Das, A. (2021). Handlock: Enabling 2-fa for smart home voice assistants using inaudible acoustic signal. *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses*, pp. 251–265. San Sebastian, Spain.
24. Shao, Y., Yang, T., Wang, H., Ma, J. (2020). AirSign: Smartphone authentication by signing in the air. *Sensors, 21(1),* 104.
25. Hong, F., Wei, M., You, S., Feng, Y., Guo, Z. (2015). Waving authentication: Your smartphone authenticate you on motion gesture. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 263–266. Seoul, Korea.
26. Shahzad, M., Zhang, S. (2018). Augmenting user identification with WiFi based gesture recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 2, no. 3, pp. 1–27.

27. Kong, H., Lu, L., Yu, J., Chen, Y., Kong, L. et al. (2019). Fingerpass: Finger gesture-based continuous user authentication for smart homes using commodity wifi. *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 201–210. Catania, Italy.
28. Huang, A., Wang, D., Zhao, R., Zhang, Q. (2019). Au-Id: Automatic user identification and authentication through the motions captured from sequential human activities using rfid. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 3, no. 2, pp. 1–26.
29. Chauhan, J., Hu, Y., Seneviratne, S., Misra, A., Seneviratne, A. et al. (2017). BreathPrint: Breathing acoustics-based user authentication. *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 278–291. New York, USA.
30. Rodríguez Valiente, A., Trinidad, A., García Berrocal, J. R., Górriz, C., Ramírez Camacho, R. (2014). Extended high-frequency (9–20 kHz) audiometry reference thresholds in 645 healthy subjects. *International Journal of Audiology, 53(8),* 531–545.
31. Selesnick, I. W., Burrus, C. S. (1998). Generalized digital Butterworth filter design. *IEEE Transactions on Signal Processing, 46(6),* 1688–1694.
32. Tse, D., Viswanath, P. (2005). *Fundamentals of wireless communication*. UK: Cambridge University Press.
33. Sardy, S., Tseng, P., Bruce, A. (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing, 49(6),* 1146–1152.
34. Li, Q. (2020). *Research and implementation on gesture recognition method based on high frequency acoustic of smartphone (Master Thesis)*. Northwest University, China.
35. Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B. et al. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, pp. 2613–2617. Graz, Austria.
36. Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M. et al. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing, 151,* 107398.
37. Graves, A., Jaitly, N., Mohamed, A. R. (2013). Hybrid speech recognition with deep bidirectional LSTM. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278. Olomouc, Czech Republic.
38. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A. (2018). How does batch normalization help optimization?. *Processings of the Advances in Neural Information Processing Systems*, pp. 2483–2493. Montreal, Canada.
39. Pan, S. J., Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22(10),* 1345–1359.
40. Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks?. *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3320–3328. Montreal, USA.
41. Chollet, F. Keras: the python deep learning api. https://keras.io/. (accessed on 15/12/2023).
42. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A. et al. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 16)*, pp. 265–283. Savannah, GA, USA.