Combating Voice Spoofing Attacks on Wearables via Speech Movement Sequences

Shan Chang, Luo Zhou, Wei Liu , Hongzi Zhu, Xinggang Hu and Lei Yang

Abstract—Voice assistants, increasingly integrated into wearable devices with limited human-computer interaction modalities, are susceptible to voice spoofing attacks. Such attacks exploit pre-recorded or synthesized voice commands to trick the assistants into executing actions unauthorized by legitimate users. In this work, we propose GyroTalk, a novel approach extracts individual and reliable features from speech movement sequences of users, using built-in gyroscopes in wearables, to differentiate between legitimate users and malicious attackers. GyroTalk is inspired by two critical insights. Firstly, speech, as a highly intricate motor task, necessitates the synchronized coordination of multiple respiratory, laryngeal, lingual and mandibular muscles. These collective muscle movements propagate throughout the body, providing unique movement signatures. Secondly, the distinctive speech movement sequences of individual speakers, essential for generating specific words, can be grabbed by embedded IMU of wearables. We conduct a comprehensive evaluation of GyroTalk across various COTS Android devices, including smart phones, watches and glasses. Our experimental results demonstrate that GyroTalk can achieve a mean FAR of 2.23% and a FRR of 2.48%, even in the face of complicated voice spoofing attacks.

Index Terms—voice assistant, spoofing attack, speech movement sequence, gyroscope, wearable

1 Introduction

Over the past decade, the prevalence of wearables has significantly increased, ranging from smart glasses, watches, and headphones to wrist-bands and jewelries. However, such wearable devices often lack adequate human-computer interfaces (i.e., limited display size and restricted input methods). As a result, traditional keyboard- or touchscreen-based interactions, typically used in PC and mobile scenarios, respectively, are less suitable for wearables. Conversely, voice-based interaction, which utilizes voice assistant apps (e.g., Google Assistant, Siri, Alexa, and Cortana) to make phone calls, send messages, and play music without typing commands, delivering a more user-friendly experience.

However, voice-based interactions are highly vulnerable to spoofing attacks due to the open sound channel, making it easy to record voice activities without being detected [1]. The widespread availability of high-quality and affordable recording/replaying devices, such as smartphones and digital voice recorders, has simplified the launch of replay attacks. Additionally, attackers can fabricate voice commands by splicing multiple voice fragments of a victim [2] [3]. Existing research indicates that, with only one-minute voice training data, attackers can build a fine-tuned model, which is capable of producing voice commands highly similar to the original voice of the victim. These vivid voice commands can be, afterwards, surreptitiously injected into the sound channel [4] [5] [6] [7]. Such attacks, which can lead to serious consequences such as making fraudulent phone calls, unlocking devices and authorizing online payments, have raised significant concerns.

In academic research, liveness detection is employed to combat voice spoofing attacks by distinguishing authentic human voice samples from machine-generated ones. Current liveness detection methods are categorized into three types. Firstly, Feng *et al.* utilize a high-sampling-rate (11kHz) vibration sensor to detect vibrations from vocal cords and match them with the corresponding microphone-captured speech signals [8]. However, the sensor is regraded as a hardware extension for wearable device and needs to be placed near the throat ofwawve speaker. Secondly, use a camera to capture lip movements [9] or extract reliable facial features [10]. Nevertheless, such schemes are subject to lighting conditions, meanwhile, affordable wearable products often lack cameras. Thirdly, ultrasonic-based approaches leverage the built-in speaker of a smartphone to emit high-frequency sound waves, then detect Doppler shift in the reflected waves while a user utters voice commands, which are caused by articulatory gestures [11] [12] [13]. But the user has to hold the smartphone right in front of his/her face within an extremely short distance (2-4cm), making it impractical for wearable devices like glasses, necklaces, and headphones.

In this work, we introduce GyroTalk, a novel liveness detection methodology that utilizes *Speech Movement Se-*

- *Shan Chang, Luo Zhou and Xinggang Hu are with School of Computer Science and Technology, Donghua University, Shanghai 201620, China. E-mail: changshan@dhu.edu.cn, {1229162, hxg}@mail.dhu.edu.cn.*
- *Wei Liu is with School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China. E-mail: liuwei628@aufe.edu.cn*
- *Hongzi Zhu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: hongzi@cs.sjtu.edu.cn*
- *Lei Yang is with the Department of Computing, Hong Kong Polytechnic University, Hongkong 100872, China. E-mail: tagsysx@gmail.com*

This work was supported in part by the Natural Science Foundation of Shanghai (Grant No. 22ZR1400200), the Fundamental Research Funds for the Central Universities (No. 2232023Y-01), the National Natural Science Foundation of China (Grant No. 62202001), and the University Grant Committee of Hongkong (Grant No. 15204820).
(Corresponding authors: Wei Liu, Hongzi Zhu)

quences (SMSs). The key concept of GyroTalk is to leverage the embedded gyroscopes in wearables to record individual SMSs when a speaker utters a vocal command. The recorded sequences are then used to verify user authenticity. GyroTalk is driven by two fundamental observations. First, voice is a task-oriented motor behavior that necessitates coordinated efforts from multiple joints and muscles. Consequently, functional motions of the jaw and head-neck regions arise from the coordinated activation of muscles in both the jaw and neck, resulting in joint motions in the temporomandibular, atlantooccipital, and cervical spine regions. The synchronized motions of jaw and head-neck parts also extend to other distant body segments, facilitated by the interlinked human skeleton and interconnected muscle fibers. Second, even when uttering the same vocal command, individuals exhibit diversity in SMSs due to variations in muscle size, shape, and strength, as well as differences in habitual patterns of muscle force exertion on bones and joints. Such unique features can be used to catch an malicious attacker who attempts to mimic the SMS of a legitimate user.

GyroTalk provides offline user enrollment, which allows the user to repeatedly utter preselected voice commands and records the corresponding movement sequences as training samples. In online authentication, the user only needs to utter the pre-registered voice commands once, then the system will compare the corresponding movement sequences with the voice samples. However, GyroTalk currently faces three main challenges. First of all, accurately perceiving tiny movements induced by speech is not an easy task. By analyzing the perceptual data from common embedded motion sensors, we find that speech movement usually manifests itself as a slight acceleration, but exhibits relatively large rotations on body parts. Therefore, we leverage a gyroscope to capture SMSs.

Secondly, it is hard to choose movement features that adequately represent the characteristics of a speaker. Despite the widespread adoption of statistical features, which are often insufficient in accurately representing the comprehensive trend of the entire movement. In addition, directly feeding movement sequences into the classifier may lead to higher computational costs and potential inaccuracies. To solve this problem, we choose to analyze the similarity between registered samples and test samples. To enhance the robustness of feature vectors based on similarity, we compare the test vocal commands with top-K samples.

Finally, due to the difficulty of eliminating variances induced by different speaking rhythms, it is hard to calculate the similarity between two motion signals. There are two aspects to this problem. In instances where the same speaker utters two passphrases, varying speaking rhythms can cause corresponding syllable movements to appear at different points in the gyroscope signals. This discrepancy often results in a mistakenly low similarity. To tackle this problem, we introduce *Speech Aligned Dynamic Time Warping* (SA-DTW), a method that adjusts gyroscope signals to align with time-synchronized microphone signals for enhanced similarity comparison. When two passphrase phrases are uttered by different speakers, it is essential to avoid erroneous high similarity due to incorrect alignment of different syllabic actions. We observe that, in the best match to the template sequences, SMSs from real users should have small aberrations. Thus, we define *Warping Score* to measure the similarity and SA-DTW distance between SMSs.

We conduct extensive real-world experiments with 15 enrolled participants on three commercially available off-the-shelf (COTS) Android devices, including a watch, a phone, and a pair of smart VR glasses. GyroTalk is capable of accurately separating authorized users from adversaries (whether human or machine) of voice assistant systems. This innovative technology can be utilized on wearable products that are worn on different parts of the body without any complicated procedures and hardware extensions. Experimental results demonstrate that GyroTalk can achieve a low mean false acceptance rate (FAR) of 2.23% and a false rejection rate (FRR) of 2.48%, even against complicated voice spoofing attacks.

Contributions. This article makes the following contributions.

- We show that the SMSs of speaking are resulted from the coordinated movements of several muscles, which can be captured by the built-in motion sensors in wearable devices. It lays the foundation of our GyroTalk design.
- Based on the above observation, we develop GyroTalk, which extracts the user-specific features of SMSs when speaking a passphrase for liveness detection. GyroTalk relies on the speech movement control system of a speaker and requires no complex operations or additional hardware, just a gyroscope, which is commonly embedded in wearable devices.
- We evaluate GyroTalk on 3 COTS Android devices, including a pair of VR glasses, a phone and watch. We also recruit 15 volunteers to conduct extensive real-world experiments and the experimental results show that even when facing advanced voice spoofing attacks, GyroTalk can achieve a mean FAR of 2.23% and FRR of 2.48%. These findings demonstrate the robustness and easy deployment of GyroTalk.

Organization. The rest of the paper is organized as follows. Section 2 reviews two main approaches for detecting voice spoofing attacks and discusses the corresponding limitations. Section 3 introduces our GyroTalk system and attack model. Section 4 details the design scheme of GyroTalk. Section 5 describes the GyroTalk implementation and reports the evaluation results of GyroTalk in extensive experiments. Section 6 discusses some limitations of GyroTalk. At last, Section 7 concludes the paper.

2 Related Work

Existing research on detecting voice spoofing attacks focuses on two main strategies: liveness detection and channel characteristic analysis.

2.1 Liveness Detection

Liveness detection systems are intended to distinguish authentic human voice samples from machine-generated ones, and these systems can be divided into three categories.

2.1.1 Vibration-based Detection Methods

The high-frequency speech signal is assessed through the utilization of a designed acceleration sensor, which is then juxtaposed with voice signals captured by the microphone. The detection of a voice attack is triggered if the correlation between these two signals falls below a specified threshold [8]. Drawbacks of this approach lie in the necessity for an additional sensor, and requiring placement in close proximity to the throat of speakers to capture the vocal cord vibrations. Wang *et al.* find that when a user approaches a smartphone microphone to utter a voice command, breathing can make the microphone vibrate and its vibration is recorded as a pop noise, which is hard to be replayed or imitated by attackers. Such characteristic can be regraded as a user-specific liveness feature [14]. However, this method is susceptible to ambient noise, and the user has to hold the microphone in close proximity to the mouth (i.e., 2-4cm). Liu *et al.* introduce a multi-modal recognition system integrating mmWave vibrations and audio signals, facing hardware dependency and potential user inconvenience compared to GyroTalk [15]. Han *et al.* employ accelerometer data for speech vibrations, emphasizing robust liveness detection, albeit sensitive to ambient noise and hardware calibration [16]. Embedded motion sensors in mobile devices are used to extract, classify and verify subtle muscle tremors that occur naturally in the body during use of the device [17], or capture biometrical features when a user shaking the phone in his customized way [18].

2.1.2 Conversation-face based Detection Methods

Detection methods based on facial and lip motion analysis are used to identify the speakers. These methods calculate the similarity between recorded voice and facial information to prevent attacks [10] [19] [20] [21]. Nevertheless, the effectiveness of these schemes can be influenced by illumination. On the other hand, accessing the camera has the potential to compromise sensitive user privacy. To overcome the disadvantages of camera-based approaches, Xu *et al.* utilize a commercially available mmWave radar, moving along a set path, to scan the human face, and the reflected signals can capture facial biometric and structural features [22]. However, this method requires the devices to be directly in front of human face within a range of angles.

2.1.3 Ultrasound-based Detection Methods

Upon a passphrase spoken by a user, the built-in speaker emits ultrasonic waves. The microphone then captures the Doppler shift in these waves, reflected by various articulators of the sound system [11] [12]. The disadvantages of these methods are that the device must be placed near the mouth of a user and certain posture is required to ensure accurate capture of the reflected signal. Wang *et al.* employ Wiener deconvolution preprocessing and device-specific convolution to boost ultrasound-based anti-spoofing systems [23]. However, it mandates collecting pulse response data from target devices and conducting data augmentation and training for each unseen device, potentially requiring extra equipment and data collection, thereby hampering its universality and quick deployment on new devices. Lee *et al.* propose a sonar-based detection scheme. It verifies sound source consistency by emitting an inaudible sound to identify the relative location of the user via a smart speaker [24]. Unfortunately, this method consumes too much power for wearable devices. Jiang *et al.* present a software-based anti-spoofing approach for voice authentication on smartphones. It leverages unique sequences to describe the relationship between pop noises and phonemes, then calculates the similarity between estimated and actual pressure signals [25]. However, their scheme requires the smartphone right in front of the mouth of the speaker within a short distance (i.e., 2-4cm), in order to detect the ultrasonic signal effectively.

2.2 Channel Characteristic Detection

Channel characteristic-based schemes detect differences in data between Hi-Fi recordings and original sound channels. These schemes use the Universal Background Model (UBM) to model the channel, and analyze the silent segments of voice data to verify if the authenticated channel matches the training voice channel [26]. While these methods provide valuable insights into detecting voice attacks, it is important to note that due to the lower amplitude and susceptibility to noise pollution of the silent segments, additional measures are required for comprehensive protection. In recent years, Meng *et al.* propose a multi-channel audio-based scheme which leverages the layout characteristics of microphone arrays in smart speakers and the channel properties of sound propagation to achieve the discrimination between genuine human and replayed speech [27] [28]. Yang *et al.* leverage differences between multiple microphone channels to extract acoustic dynamics features, distinguishing between genuine voice commands and spoofing commands played through speakers [29]. However, our GyroTalk is primarily intended for wearable devices, which typically do not feature such microphone arrays. Moreover, human breath will leave a trail in the WiFi Channel State Information (CSI) during the course of speaking commands, PRADHAN *et al.* use changes in the CSI traces to detect breath to verify human presence [30], but this approach is restricted to scenarios where WiFi is present.

Overall, our GyroTalk is a new form of liveness detection method for wearable devices, which can be worn in flexible ways (e.g., hand-held, wrist-, and head-worn). Meanwhile, GyroTalk does not require additional hardware extension, and there is no risk of leaking user privacy.

3 Preliminaries

3.1 System and Attack Model

GyroTalk aims to ensure the safety of speaker recognition tasks, that is, text-prompted voice authentication. In enrollment, more specifically, a user is requested to utter a prompted phrase predefined by the provider several times. In future verification, the user attempts to get a pass by speaking the enrolled phrase. GyroTalk can be deployed on a variety of COTS wearables such as necklaces, glasses, bracelets, helmets, and mobile phones. These devices must have gyroscopes and computing capability. Additionally, they should be carried or worn by users to allow the inertial sensor to detect micro-movements induced by speaking.

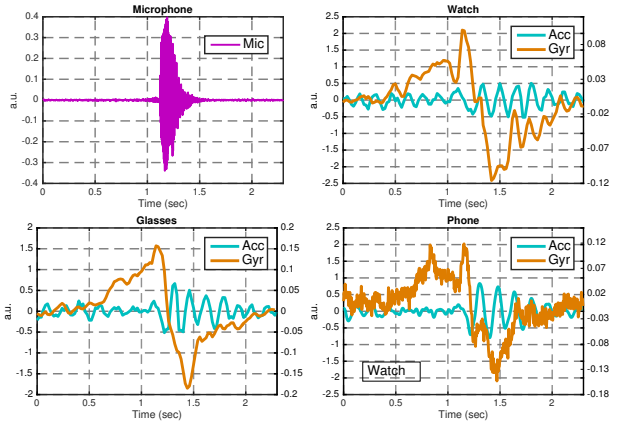


Fig. 1: When a speaker pronounces "*home*", the microphone picks up the raw audio data; meanwhile, the accelerometer and gyroscopes of smartwatches, glasses and smartphones record the corresponding motion sequences.

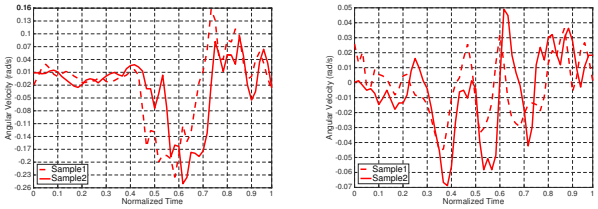


Fig. 2: Gyroscope signals of a smart watch sampled during volunteer $\mathcal{A}$ (left) and $\mathcal{B}$ (right) pronouncing "*America*" twice.

Suppose an attacker who seeks to deceive the voice authentication system of the victim by using inveracious passphrases. The attacker has two options to carry out spoofing attack: a replay attack (i.e., recording the victim's voice beforehand and playing it back during authentication) or a voice impersonation attack (i.e., injecting a modulated or imitative vocal command into the audio channel). *Furthermore, we assume that a malicious attacker can obtain (i.e., have physical access to) the victim's device and master the detail of GyroTalk.* Specifically, we consider two attack scenarios as follows:

- *Scenario-A:* **Spoofing Attack without Priori Knowledge.** In this scenario, the attacker wears the victim's device and silently mouthing the targeted vocal command while replaying or broadcasting it without making audible sound. Besides, the attacker has no knowledge of how victim's body moves during passphrase articulation.
- *Scenario-B:* **Spoofing Attack with Priori Knowledge.** This attack scenario follows the same procedure as a spoofing attack without priori knowledge, but with an additional step of the attacker first observing how the victim delivers a voice command, i.e., filming a clip of the victim and rehearsing before executing the attack.

3.2 SMS

Typically, the activity of speaking involves the utilization of air pressure generated by the lungs to produce sound through the laryngeal valve. Subsequently, the sound is transformed into various consonants and vowels as it is modified by the vocal tract. Generating specific syllables involves the recruitment of a set of predominantly unvarying muscles, which triggers synchronized gradual motions of diverse body regions, spanning from the diaphragm to the lips [31].

Definition 1. A *Speech Movement Sequence (SMS)* is defined as a sequence of intricate bodily motions that occur during speaking, which involves:

- *Immediate Synchronous Motions:* the simultaneous contraction of more than 100 muscles during sound producing. These muscles include the septum, pectoralis major, jaw muscles, cheek muscles, and pectoral muscles, which work together on different sounds producing organs (e.g., lungs, throat, tongue, lips, etc.) to generate precise sound. The coordination of these muscles is an unconscious response of the autonomic nervous system and the brain does not need to consciously control it. When we speak, these muscles work in coordination through rapid signaling. In order to produce the desired sound, the brain sends signals to the muscles and tells them which contraction and relaxation to perform.
- *Traction Induced Cascades:* the interactions and mechanical properties between tissues. When a muscle group produces motion through traction, it can cause remote muscle groups and body parts to move and shift due to the interconnection of muscle fibers and attachment to the skeletal system. In the human skeletal systems, there exist complex mutual interactions among muscles, bones, and other soft tissues. Each muscle is composed of multiple muscle fibers that are interconnected by inherent proteins. Meanwhile, the muscle tissue adheres tightly to the skeleton and is fixed to it via tendons (rope-like substances of connective tissue). Therefore, when a muscle group shortens or stretches, it applies force and stretches adjacent muscle and bone tissue.

3.3 Capturing Diversity of SMS

We hypothesize that the contraction of specific muscle groups and body segment motions involved in voice production can be captured using motion sensors. Therefore, to capture the body motions of a speaker, we utilize both the accelerometer and gyroscope sensors commonly furnished in wearable products, which allows us to accurately track their physical movements. Based on our research, we determine that the gyroscope is a more suitable sensor for speech movements capture due to its higher sensitivity than accelerometer. This insight indicates the gyroscope is a better choice for tasks that require precise measurement of motions during speaking, which may be due to the fact that speech movements tend to involve relatively small accelerations and larger rotations of body segments. Through detecting the output signals from a gyroscope, then conducting preprocessing and calculation, we can derive the moving direction and attitude information of a user. Figure 1 demonstrates the original audio data recorded from a microphone positioned at the top left corner while a speaker utters the word "*home*". Additionally, it exhibits movement sequences collected from accelerometers and gyroscopes of

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3418908

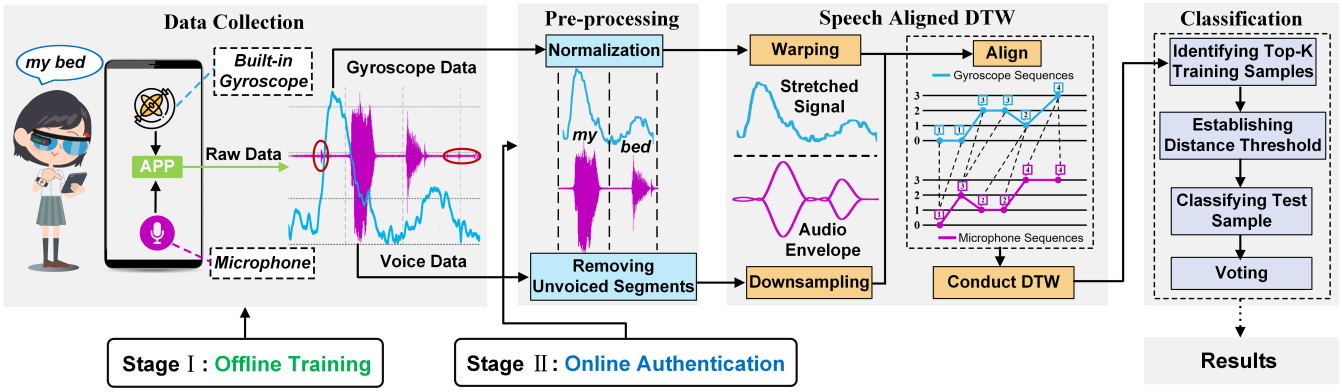


Fig. 3: The overall architecture of GyroTalk.

different wearable devices, including a pair of VR glasses, a smart watch, and a phone. These movement sequences are corresponding to the uttered word and recorded with the audio data simultaneously. The figure demonstrates that speech movements in gyroscope data can generate more significant fluctuations, even when measuring remote body parts far from the vocal tract (i.e., hand and wrist).

In addition, we believe that, when a person utters a specific voice command, the muscles of the larynx and mouth will move in concert to produce a series of SMSs, which exhibit spatial-temporal characteristics that are not only unique among each individual, but are also relatively consistent when the same person speaks. Specifically, this is a result of individual differences in the size, shape, and strength of muscles, bones, and joints, as well as variations in the way muscles exert force in different individuals. The uniqueness and consistency of these movement sequences also mean that they can be regarded as biometric features to identify legitimate speakers. Figure 2 displays the time series of gyroscope data from a smart watch recorded among two volunteers, $\mathcal{A}$ (left) and $\mathcal{B}$ (right), each pronounces *"America"* twice. The result demonstrates high intra-sample similarity from the same speaker, but significant differences between samples from different speakers.

4 Design of GyroTalk

4.1 Overview

The core idea of GyroTalk is to utilize the built-in gyroscopes of COTS wearable products to capture and identify unique and robust features in the SMSs of users. When a speaker utters a vocal command into a wearable device, both the voice and body movements are recorded by its microphone and gyroscope, respectively. After passing voice authentication via the microphone signal, the corresponding gyroscope data will undergo further verification by GyroTalk.

Figure 3 shows the overall architecture of GyroTalk. The GyroTalk authentication process includes two stages: offline training and online authentication. During offline training, a user utters a vocal command multiple times to capture voice and movement signal templates. These templates are then utilized for classifier construction. Subsequently, in online authentication, test inputs will be fed into the well-trained classifier, and then compared with the established templates to verify whether it originates from a registered user.

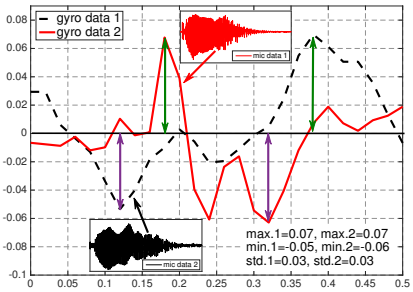


Fig. 4: A comparison of two segments of gyroscope data from different speakers, with different shapes but similar *maximums*, *minimums*, and *standard deviations*.

There are two reasons why original samples are unsuitable for classifier training. Firstly, the length of samples may differ due to variable pronouncing speed among speakers. Secondly, samples in GyroTalk may contain a substantial amount of data points, particularly for longer vocal commands. This often requires numerous training samples to achieve optimal accuracy, which could place an additional registration workload on users. An optimal solution is to derive feature vectors by extracting effective features from original signals. However, this approach is non-trivial. The conventional statistical features commonly employed in time series analysis, such as mean, standard deviation, and extreme values, are insufficient for classification purposes because the trend in the time series of gyroscope data, which is crucial for classification, cannot be effectively captured using these features. Stated differently, features in a data sequence are expected to appear at specific positions. As an illustration, Figure 4 depicts two segments of gyroscope data collected from glasses worn by different speakers uttering *"do"*. Despite their similar maximum and minimum values and standard deviations, the two segments exhibit different shapes, emphasizing that relying solely on statistical features is not sufficient.

In this paper, we present the Speech Aligned DTW (SA-DTW) algorithm, which facilitates the comparison of shape between two movement sequences that are aligned with speech. When an individual gives two vocal commands, the resulting movement sequences will carry a small SA-DTW distance and a short warping path. SA-DTW is utilized to extract distance vectors from SMSs pairs. During the classifier training process, we utilize both SA-DTW distance and warping score.

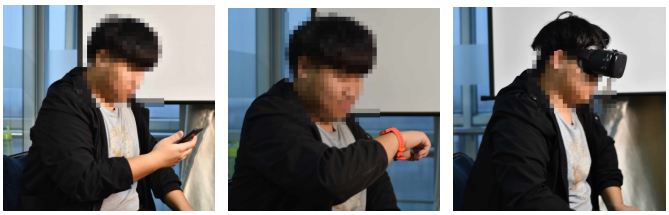


Fig. 5: Data Collection Experimental Setup for GyroTalk. smartphone (left), smartwatch (middle) and VR glasses (right).

4.2 Data Collection

We develop an Android application that can simultaneously collect microphone and gyroscope signals with sampling rates of 44.1kHz and 50Hz, respectively. The experimental setup during the data collection is shown in Figure 5. We deploy and test the application on three devices: a smartphone (Google Nexus 5X), a smart watch (MOTO 360), and a pair of VR glasses (Storm Mirror). 15 volunteers are recruited for the study, including 4 women and 11 men aged between 21 and 27 years old. During the data collection, each participant wears the watch on their non-dominant wrist, holds the smartphone in their dominant hand, and wears the glasses on their head. When giving a voice command, we instruct the volunteers to naturally bend their forearm with the smart device and hold it in front of their chest. It's noteworthy that the participants are allowed to take a break or quit at any time.

As a preparation, each volunteer preselects *10* commands from Table 1, which categorizes 30 voice commands into three groups based on their length, ensuring an approximately equal number of commands chosen from each group. Besides, volunteers are allowed to negotiate, aiming to have each command selected 4 to 6 times, with an expected average of 5. Then, we collect data in three traces as follows:

- *Trace-A: One-month Data Collection.* During this period, each volunteer speaks *10* times per day for each selected command in various speaking volumes and paces. Specifically, every command is uttered *10* times in a standard speaking tone with *2* times each at regular, slow, and fast pace. In addition, each command is spoken at a higher and a lower volume *2* while retaining the standard speaking pace. To examine the potential impact of body postures on GyroTalk, we instruct volunteers to conduct data collection while assuming three different postures (i.e., *standing, sitting straight, and hump*) over the course of the initial, middle, and final *10* days, respectively. For every command, we gather 300 microphone signals and corresponding gyroscope data per volunteer. Additionally, we record the data collection process with a video camera for potential use in impersonation attacks (i.e, spoofing attacks with priori knowledge).
- *Trace-B: Spoofing Attacks. 5* participants are selected to act as victims and the remaining *10* as attackers. Firstly, each attacker launches spoofing attack without priori knowledge (see Scenario-A) by replaying each command of victims *10* times. In the next scenario (Scenario-B), attackers watch the videos showing how victims perform during speaking voice commands. They then proceed to repeat each victim's command *10* times, making an effort to mimic their speaking styles. Therefore, for both types of spoofing attacks, each victim's command is subjected to 100 attack attempts.
- *Trace-C: One-week Data Collection after six Months of Trace-A.* Each volunteer utters every selected command *10* times daily, maintaining a regular speaking volume and pace.

TABLE 1: Command List

Command		
2-3 words (*nine* commands)		
1. Take a photo.	2. Turn off camera.	3. Call Emily.
4. Open WeChat.	5. Play some music.	6. Lock phone.
7. Turn off Wi-Fi.	8. Check my voicemail.	9. Next song.
4-5 words (*ten* commands)		
10. Navigate to my home.	11. Show me my first message.	
12. Turn off all alarms.	13. Show me the nearest mall.	
14. Send an email to John.	15. Go to google scholar.	
16. What's my next appointment?	17. Do you speak Morse code?	
18. What is the weather like?	19. What's 299 divided by 8?	
6-7 words (*eleven* commands)		
20. What is the meaning of life?	21. What song am I listening to?	
22. What is my schedule for today?	23. Set an alarm for 6 pm.	
24. Did the Golden State Warriors win today?	25. What are some attractions in City?	
26. What time is it now in Shanghai?	27. Show me pictures of Mount Rushmore.	
28. Show me restaurants near my campus.	29. Who is the Producer of Star Wars?	
30. Do I need an umbrella today?		

4.3 Pre-processing

Assume we collect a pair of time-synchronized sequences from a microphone and gyroscope. We conduct a two-step preprocessing as follows:

- *Step-1: Removing Unvoiced Segments.* Since we only need to analyze the movements with human speech, we utilize *silence removal* [32] to remove the unvoiced segments in the audio signal. Specifically, given an audio signal V, it can be cut into multiple segments according to the voice commands, along with the corresponding gyroscope signal G, resulting in several pairs of microphone and gyroscope data segments. Each of these segments is represented as $S_i = (V_i, G_i), i \geqslant 1$, where V_i and G_i denote the audio and movement segments, respectively.
- *Step-2: Normalization.* When users speak, the volume of their voice and the intensity of their motion while speaking affect the amplitude of the gyroscope signal. We normalize the amplitude of each G_i segment of the gyroscope signal to remove the impact of speech volume on the signal, which aims to ensure the amplitude of the gyroscope signal will not be affected by changes in the intensity or volume of the speech movements.

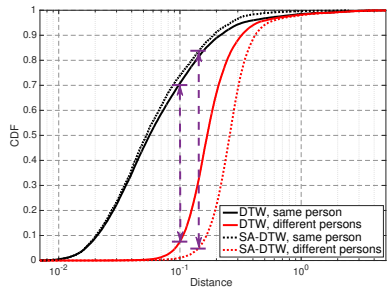


Fig. 6: DTW and SA-DTW distances between gyroscope signals coming from different and same persons shown in CDF curves.

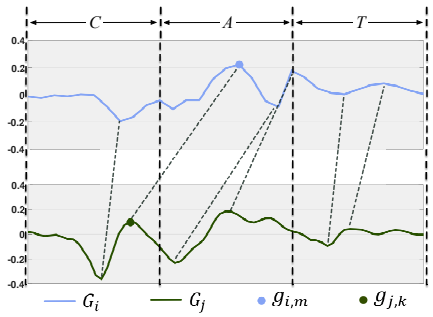


Fig. 7: An example of a mismatch. Different segments of the gyroscope signals are mistakenly matched.

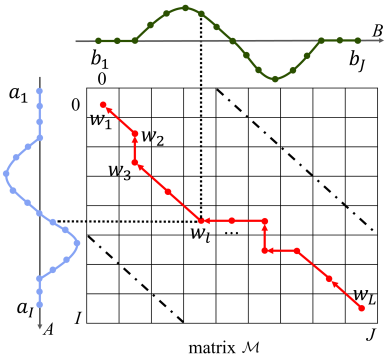


Fig. 8: An example of the resulting accumulated cost matrix $\mathcal{M}$ and its corresponding warping path $\mathcal{W}$.

4.4 Speech Aligned DTW based Similarity

The Dynamic Time Warping (DTW) algorithm is commonly used to compute the distance between two segments of gyroscope signals with different lengths. However, the way of using only the DTW distance to determine whether two gyroscope signal segments belong to the same person might lack sufficient accuracy. In Figure 6, the solid line depicts the Cumulative Distribution Function (CDF) curves of the DTW distance between two gyroscope signal segments corresponding to the same voice command from the same speaker and different speakers shows the narrow DTW distances can also occur between gyroscope signals belonging to different speakers, e.g., 30% of the signals have distances less than 0.15. In the case where gyroscope signals are blindly warped, they could lead to segments matching up that are part of different phonemes, which would result in a short DTW distance, but a longer warping path by chance. Such mismatching is illustrated in Figure 7, where the data point $g_{i,m}$ belonging to 'C' is incorrectly matched to $g_{j,k}$ belonging to 'A'. Hence, it is crucial to align gyroscope signals to their corresponding audio signals before comparison. In addition, it is essential for gyroscope signals from the same speaker to exhibit minimal distortion on optimal matching, as well as a tiny warping distance.

To eliminate such mismatching during signal comparison, we introduce the SA-DTW distance. We perform DTW on the microphone signals of two segments S_i and S_j, and record the corresponding warping function. To acquire the voice-aligned gyroscope signals, we apply this warping function to the corresponding gyroscope signals. Next, we determine the SA-DTW distance by calculating the DTW distance between them, along with the associated warping score. This information is then combined into a two-tuple distance representation, capturing the similarity between the gyroscope signals. Figure 9(a) and Figure 9(b) display the joint distribution functions of SA-DTW and the corresponding warping score, and DTW and the corresponding warping score, of two gyroscope data segments from the same and different persons, and respectively. It can be seen that the two-tuple of distance is more effective for distinguishing speakers. These two-tuple distances are utilized as inputs for the subsequent classifier. Specifically, the following steps need to be executed.

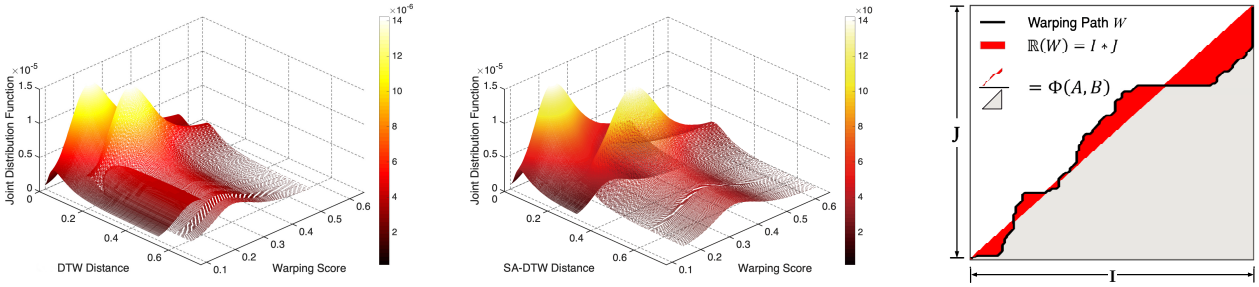
4.4.1 *Downsampling Audio Signal*

After analyzing the signal sampling rates, we discover that the gyroscope signals cannot be directly subjected to the warping function between V_i and V_j due to the considerably smaller sampling rate in comparison to the microphone signals. To address this issue, we downsample the microphone signals before conducting DTW. Downsampling is commonly accomplished by decimating the microphone signal with an integer factor D such that only the *D-th* sample is kept. However, since the sampling rate of microphone is 882 (i.e., 44100/50) times higher than that of gyroscope, this significant disparity in sampling rates may result in the omission of peak values from the original signals, rendering the resampled signal notably dissimilar to the original. To avoid this, we choose to extract the envelope of V_i and then extract the upper part of this envelope at a sampling rate of *D=882*, resulting in a downsampled audio sequence. The determination of the envelope is performed by spline interpolation between local maxima, and the sample interval between these local maxima is at least σ (σ is set as 1500 empirically).

4.4.2 *Warping Gyroscope Signal*

Given the downsampled signals of V_i and V_j, denoted by $A = \{a_1, \cdots, a_I\}$ and $B = \{b_1, \cdots, b_J\}$, we conduct DTW on them. To align A and B with DTW, we construct an $I \times J$ accumulated cost matrix $\mathcal{M}$, where the (i^{th}, j^{th}) element in $\mathcal{M}$ is defined as $d(i,j) = (a_i - b_j)^2$, implying the alignment of a_i and b_j. During calculation of the DTW distance between A and B, we obtain a corresponding warping path $\mathcal{W} = w_1, \cdots, w_l, \cdots, w_L$ simultaneously, (where $w_l = (i,j)_l$, and $\max(I,J) \leq L \leq I + J - 1$), on which $\frac{1}{L}\sqrt{\sum_{l=1}^{L} d(w_l)}$ reaches its minimum. It implies that the shortest warping path coincides with the diagonal line $j = i$ of the matrix when A equals to B, and long path means severe warping. Figure 8 shows an example of resulting matrix $\mathcal{M}$ and the purple line indicates the corresponding warping path. Then, we stretch $G_A = \{g_1^{(A)}, \cdots, g_I^{(A)}\}$ and $G_B = \{g_1^{(B)}, \cdots, g_J^{(B)}\}$ according to $\mathcal{W}$, and obtain the stretched gyroscope signals $G'_A = \{g_{(i_1)}^{(A)}, \cdots, g_{(i_l)}^{(A)}, \cdots, g_{(i_L)}^{(A)}\}$ and $G'_B = \{g_{(j_1)}^{(B)}, \cdots, g_{(j_l)}^{(B)}, \cdots, g_{(j_L)}^{(B)}\}$, where i_l and j_l refer to the value of i and j in w_l. Notice that the lengths

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3418908



(a) DTW+Warping Score (b) SA-DTW+Warping Score

Fig. 9: The joint distribution functions of SA-DTW and DTW of two separate gyroscope data segments from the same and different persons, along with their corresponding warping scores.

Fig. 10: An example of calculating the warping score of A to B

of $G_A^{'}$ and $G_B^{'}$ equal to that of A and B, respectively. For example, suppose the optimal warping path $\mathcal{W} = \{(0,0),(1,1),(2,1),(3,2),(4,3)\}$, and given the original gyroscope signals

$$G_i = \{g_1^{(i)}, g_2^{(i)}, g_3^{(i)}, g_4^{(i)}, g_5^{(i)}\}$$
$$G_j = \{g_1^{(j)}, g_2^{(j)}, g_3^{(j)}, g_4^{(j)}\},$$

the stretched gyroscope signals $G_i^{'}$ and $G_j^{'}$ are

$$G_i^{'} = \{g_1^{(i)}, g_2^{(i)}, g_3^{(i)}, g_4^{(i)}, g_5^{(i)}\}$$
$$G_j^{'} = \{g_1^{(j)}, g_2^{(j)}, g_2^{(j)}, g_3^{(j)}, g_4^{(j)}\}$$

4.4.3 Measuring VA-DTW Distance and Warping Score

After getting $G_i^{'}$ and $G_j^{'}$, we calculate the DTW distance between them, denoted by $\mathbb{C}(G_i^{'}, G_j{'})$, which is actually the VA-DTW distance of G_i and G_j .

Definition 2. Given two sequences A and B with same length, i.e., $A = a_1, a_2, ..., a_i, ..., a_I$, $B = b_1, b_2, ..., b_j, ...t, b_J$, where $I = J$. The area enclosed with the diagonal line $j = i$ of the matrix $\mathcal{M}$ and the warping path $\mathcal{W}$, denoted by $\mathbb{R}(\mathcal{W})$, can be calculated as

$$\mathbb{R}(\mathcal{W}) = \sum_{l=1}^{L} \left| \frac{(i_{l+1}-i_l)[j_l+j_{l+1}-\frac{J}{I}(i_l+i_{l+1}-2)-2]}{2} \right|.$$

Then the *Warping Score* of A to B is

$$\Phi(A,B) = \frac{2\mathbb{R}\mathcal{W}}{I^2}.$$

An example is illustrated in Figure 10, and the value of $\Phi(A,B)$ is between 0 to 1. Now, we compute the warping score of $G_i^{'}$ to $G_j^{'}$, i.e., $\Phi(G_i^{'}, G_j^{'})$, and form a tuple of distance, i.e., $(\mathbb{C}(G_i^{'}, G_j^{'}), \Phi(G_i^{'}, G_j^{'}))$.

4.5 Classification

In the offline training stage, each collected training template, comprising a pair of voice and movement signals, is subjected to mutual comparison. Specifically, we compare two sets of microphone and gyroscope data signals to each other, denoted as S_i and S_j. Firstly, we segment them into two signal sequences, $S_{i_1}, ..., S_{i_n}$ and $S_{j_1}, ..., S_{j_n}$, according to 4.3. Then, we construct a distance vector $D_{i,j}$ between them, which consists of elements such as $\mathbb{D}(G_{i_1}^{'}, G_{j_1}^{'}), \Psi(G_{i_1}^{'}, G_{j_1}^{'}), ..., \mathbb{D}(G_{i_n}^{'}, G_{j_n}^{'}), \Psi(G_{i_n}^{'}, G_{j_n}^{'})$. We input all distance vectors, obtained by comparing training samples, to a one-class classifier, such as KNN or SVM.

Furthermore, to enhance the efficiency and robustness of online authentication, we introduce a step to identify the top-K training samples from a given set of $\mathbb{N}$ samples. These top-K samples represent the centroid of the sample space and are best suited to represent the training samples. To identify them, we first calculate the SA-DTW distance between each sample and all training samples. Then, we choose the K samples that have the smallest average distances to other samples. By focusing on these top-K samples instead of all samples during online authentication, we can reduce calculation burden and increase robustness against noise in training samples. Overall, this approach provides a qualitative assessment of the sample space and improves the effectiveness of online authentication.

Online authentication relies on the evaluation of distance vectors to determine whether a test sample S_i should be accepted or rejected. This is achieved by comparing S_i with the top-K samples, which results in K distance vectors $D_{i,k}$ $(1 \leq k \leq K)$. The center of these vectors, denoted as $\overline{D_i}$ is then computed by taking their average, i.e., $\overline{D_i} = \frac{1}{K}\sum_{k=1}^{K} D_{i,k}$. Finally, the classifier evaluates $\overline{D_i}$ to make a judgment on whether to accept or reject S_i.

5 Evaluation

We carry out a thorough evaluation of GyroTalk using the traces described in Subsection 4.2. The data from both the gyroscope and microphone are sent to a PC for later analysis.

For the majority of experiments, we use *Trace-A*, except for spoofing attacks and long-term performance tests (as described in Subsection 5.3 and 5.2), where we utilize *Trace-B* and *Trace-C* correspondingly. To prove the efficacy of SA-DTW, we perform a comparative analysis with two comparable methods. Specifically, these methods involve generating distance vectors between the two gyroscope signals using traditional DTW distance and the DTW distance plus the corresponding warping score, respectively. We also examine the corresponding warping score for each method in comparison to GyroTalk. In order to ensure the robustness and applicability of our study, the postures we choose are

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3418908

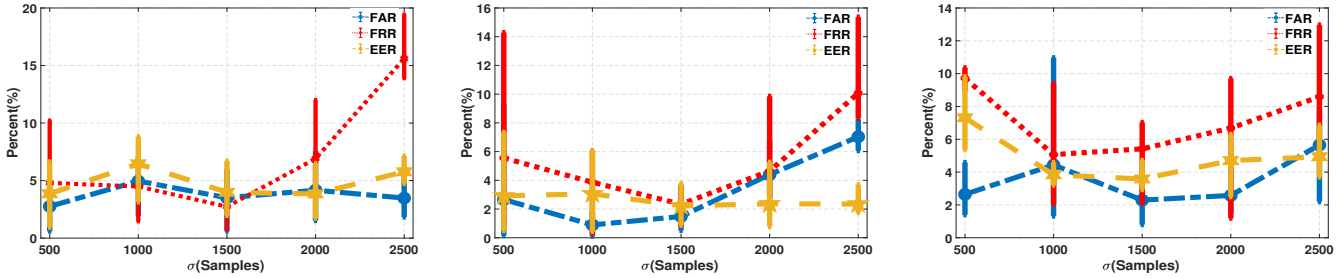


Fig. 11: Error rates with different σ (left: watch, center: glasses, right: phone)

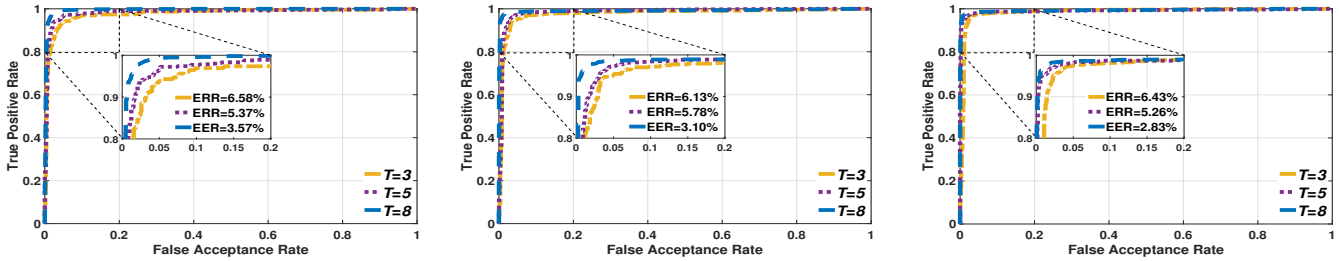


Fig. 12: ROC curve with different T (left: watch, center: glasses, right: phone)

the most common ones for using smartphones and wearable devices in our daily life. Moreover, collecting data of multiple poses for training can bring diversity, which helps to analyze the influence of different postures (i.e., body mechanical structures) on the performance of GyroTalk .

Metrics. We use the FAR and FRR to evaluate GyroTalk. The FAR measures the proportion of illegitimate testing inputs that are falsely accepted, while the FRR calculates the percentage of valid testing inputs that are incorrectly rejected. To assess the diagnostic performance of a binary classifier, we vary its discrimination threshold using the Receiver Operating Characteristic (ROC) curve. The Equal Error Rate (EER) is calculated at the intersection of the FAR and FRR curves on the ROC curve.

5.1 Micro-benchmarks

Within this subsection, we evaluate how GyroTalk performs when subjected to various real-world factors.

Efficacy of σ. In Figure 11, the mean, maximum, and minimum values of FAR, FRR, and EER across all participants are depicted as a function of σ on various devices. This experiment involves three-word commands, with σ ranging from 500 to 2500 at intervals of 500. A one-class SVM classifier is used, with K set to 4. The corresponding analysis of the resulting envelope shows that too small a value of σ causes the envelope to contain unnecessary detail, while too large a value of σ distorts the envelope, leading to significant deviations between the downsampled signal and the original signal. Hence, σ is set to 1500 for the subsequent experiments.

Efficacy of T. We evaluate the performance of GyroTalk across various training data size, specifically at T = 3, 5, and 8, with K fixed at 4. We use a SVM classifier (one-class), which discrimination threshold is increased from 0.0 to 1.0 at a 0.01 interval and its corresponding ROC curves are shown in Figure 12. Our analysis reveals that increasing T results in a gradual drop in the average EERs across all three devices studied. Notably, GyroTalk achieves the best performance when T = 8. It is essential to note that the computational burden of the online authentication stage remains unaffected by the value of T, as only the Top-K samples are compared against the testing samples. Thus, we utilize a training sample size of *eight* for the remainder of this study.

Efficacy of K. In Figure 13, the mean, maximum, and minimum FAR, FRR, and EER for all volunteers are plotted against K on three different devices. The experiment involves commands with three words, with K varying from 2 to 8 at an interval of 1. The results show that all three error rates decrease significantly when K is increased from 2 to 4 and gradually as K is further increased to 8. Additionally, even with a small value of K, such as 3, high performance is achieved across all terminals. In the process of online authentication, the top-K samples are used as reference to compare with the testing sample. Large K value can result in significant computational overhead and long authentication latency. Therefore, for the rest of the study, K is set to 3 as it delivers the optimal trade-off between computation cost and performance.

Influence of Classifier. In this study, we evaluate the performance of two one-class classifiers: K-Nearest Neighbor (KNN) [33] and Support Vector Machine (SVM) [34]. We conduct multiple tests with KNN, varying the parameter κ from 1 to 10, and determine that the optimal value is 3. For SVM, we adopt Radial Basis Function (RBF) as the kernel function, and perform a grid search with cross-validation on the training set to determine the values of c and g, which are chosen from the range of $[2^{-4}, 2^{4}]$. The results are presented in Table 2, indicating that both classifiers can achieve satisfactory performance, and SVM slightly outperforms KNN. Nonetheless, given the decrease in prediction efficiency of KNN as the training data size increases, we decide to use the SVM classifier for subsequent experiments.

Influence of Speaking Volume and Pace. We divide *Trace-A* into two parts to investigate the impact of speaking volume and pace. Specifically, we utilize data from different devices,

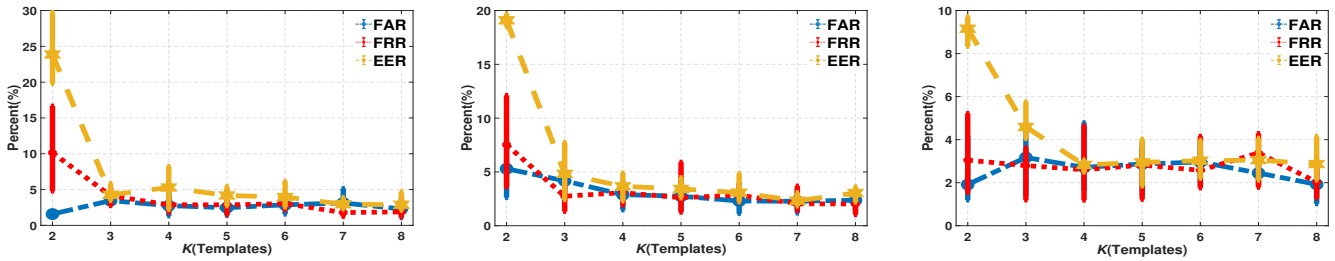


Fig. 13: Error rates with different K (left: watch, center: glasses, right: phone)

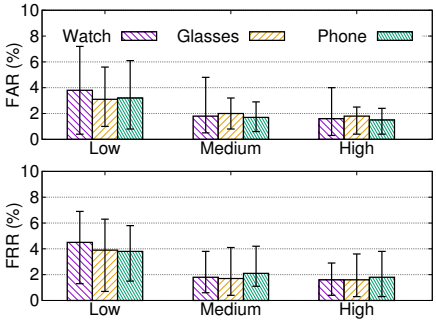


Fig. 14: FAR and FRR with different speaking volume

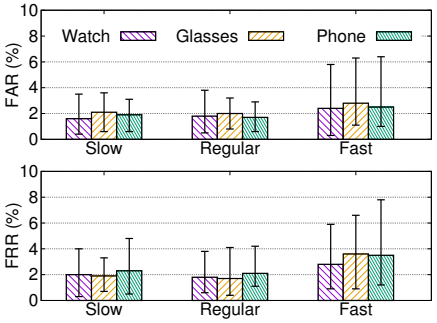


Fig. 15: FAR and FRR with different speaking pace

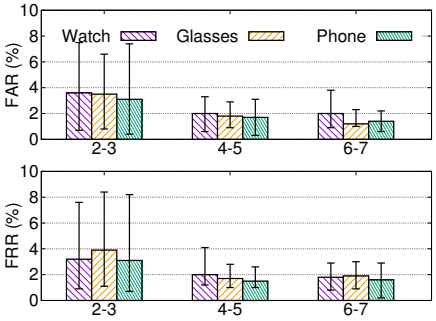


Fig. 16: Long term performance

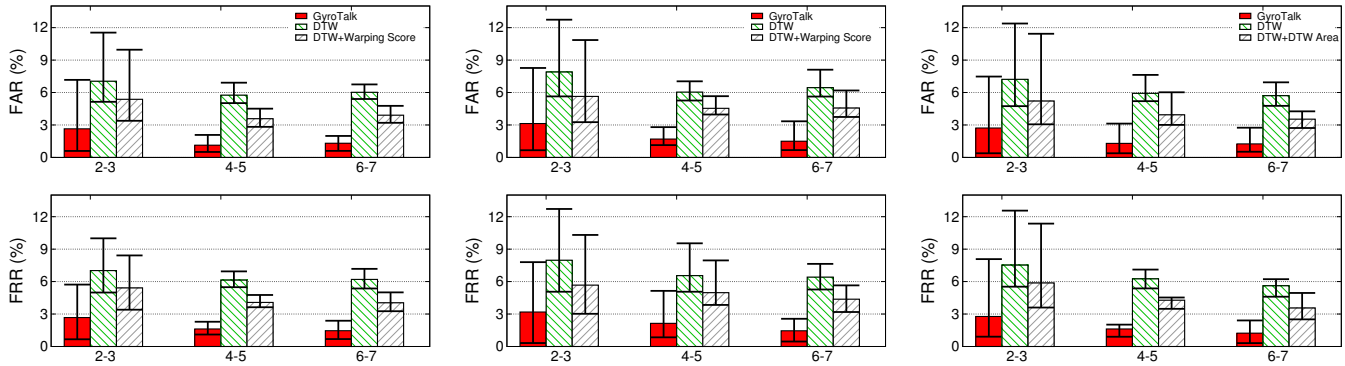


Fig. 17: FAR and FRR vs. Length of voice command (left: watch, center: glasses, right: phone)

TABLE 2: Comparative Analysis of Two Classifiers

Classifier	Watch		Smartphone		Glasses	
	FAR	FRR	FAR	FRR	FAR	FRR
KNN	4.09%	5.85%	5.03%	5.17%	3.82%	4.79%
SVM	3.45%	4.13%	3.24%	3.53%	3.01%	3.35%

consisting of microphone and corresponding gyroscope signals, where the microphone data is collected at normal volume but with different paces, and at normal pace but with different volumes, to test the performance of GyroTalk. Experimental results are depicted in Figure 14 and Figure 15, which indicate that when speaking at a fast pace or low volume, the associated FARs and FRRs are approximately 3%-4%, which is 1%-2% higher than in other situations. Additionally, our findings suggest that the FRRs are more easily influenced by changes in speaking pace and volume compared to the FARs.

Influence of Voice Command Length. Figure 17 demonstrates that, as the length of voice commands increases, both the FAR and FRR decrease. Specifically, the findings of the experiments show that when the length of voice command rises from 2–3 words to 4–5 words, the average FAR and FRR of GyroTalk on a smartphone decrease from 2.82% and 2.88% to 1.42% and 1.86%, respectively. These findings support our conjecture that longer voice commands provide stronger protection and decrease false rejections. Notably, GyroTalk outperforms the comparative techniques in all cases, e.g., for 6-7 words command, both types of errors for GyroTalk are only around 1.3%, whereas those of DTW with and without warping score still exceed 6% and 3%, respectively. Moreover, longer voice command (e.g. 10-11 words) can further improve GyroTalk's performance. However, excessively lengthy voice commands for the voice assistant's wake-up words will inevitably reduce user convenience. Typically, the optimal length of a voice wake-up command should contain 2-3 words.

5.2 Behavioral Variability

For behavioral biometric systems, behavioral variability is a significant problem. In this experiment, we examine the influence of long-term behavioral unpredictability on GyroTalk. To this end, we use *Trace-C* (collected *six* months after *Trace-A*) to test GyroTalk trained and configured using *Trace-A*. As shown in Figure 16, the FARs and FRRs do not vary

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3418908

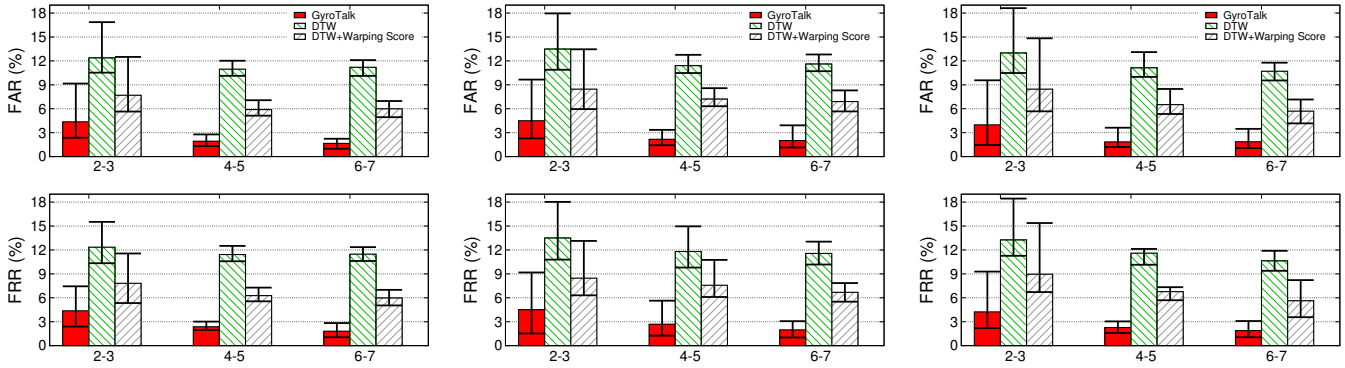


Fig. 18: FAR and FRR under spoofing attacks without priori knowledge(left: watch, center: glasses, right: phone)

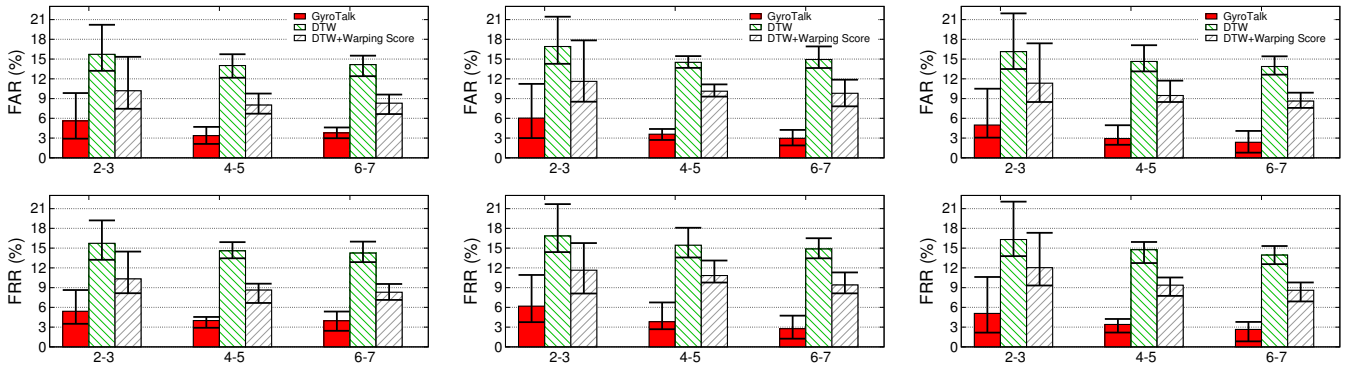


Fig. 19: FAR and FRR under spoofing attacks with priori knowledge(left: watch, center: glasses, right: phone)

extremely greatly over time in all cases. For example, for 2-3 word commands, we observe FARs are slightly increased from 3% (see Figure 17) to around 3.5%, and for 4-7 word commands, both FARs and FRRs are no more than 2%, which means GyroTalk is resistant to long-term behavioral variability.

5.3 Combating Spoofing Attacks

We use *Trace-B* to analyze the security of GyroTalk based on the threat model discussed in Subsection 3.1.

5.3.1 Spoofing Attack without Priori Knowledge

Figure 18 demonstrates the resilience of GyroTalk to attacks without priori knowledge. In comparison to normal usage scenarios, the FARs and FRRs of GyroTalk experience only slight increases (around 1%) for commands consisting of 2-5 words, and display nearly no change for commands consisting of 6-7 words. In contrast, both DTWs, with and without warping score, exhibit significant errors in all scenarios. The results indicate that, when subjected to attacks without priori knowledge, the performance of traditional voice recognition methods deteriorates significantly. This is likely due to the fact that such methods rely solely on the spectral characteristics of the user's voice, making them more vulnerable to attacks that introduce subtle changes to the acoustic properties of the voice signal. Overall, these findings highlight the effectiveness of GyroTalk's approach to voice recognition, which combines both audio and gyroscopic sensor data to improve accuracy and mitigate the impact of spoofing attacks. The results suggest that the security and dependability of voice-based authentication systems can be improved by incorporating additional sensor data, particularly in scenarios where attacks without priori knowledge are a concern.

5.3.2 Spoofing Attack with Priori Knowledge

The results presented in Figure 19 indicate that GyroTalk demonstrates stable performance even when subjected to increasingly complex spoofing attacks. Specifically, for commands consisting of 2-3 words, the error rate of GyroTalk increases by no more than 6%, while for commands consisting of 4-7 words, the increase is limited to no more than 3%. Furthermore, it is worth noting that in all circumstances, GyroTalk performs better than the other two studied approaches when it comes to resilience against spoofing assaults. More specifically, considering commands of 4-7 words, GyroTalk experiences an average FAR increase of approximately 1%, which is significantly lower than the corresponding FAR increases of over 7% and 5% observed for the other two methods. These findings suggest that GyroTalk's approach to voice recognition is effective in mitigating the impact of various types of spoofing attacks, allowing it to maintain elevated accuracy rates even in the face of sophisticated attempts to deceive the system. Overall, the results demonstrate GyroTalk's superior robustness compared to alternative approaches and highlight its potential as a reliable solution for secure voice-based authentication.

5.4 GyroTalk Implementation

We implement GyroTalk as an app on a MOTO 360 Android Wear watch. Upon voice authentication by the user, the app operates surreptitiously in the background to capture gyroscope and microphone data, which are subsequently

TABLE 3: Measured Processing Delay of GyroTalk

Command Length	Processing Latency(s)	Time Breakdown (%)		
		Pre-processing	SA-DTW	Classification
7	1.357	0.75	99.22	0.03
5	0.882	0.93	99.03	0.04
3	0.521	1.22	98.71	0.07

processed. Once the processing is finished, the app outputs a decision (accept or reject). We employ Fast DTW [35] to reduce the computational load from (n^2) to (n) while maintaining system performance. Table 3 displays the average processing delays measured for GyroTalk app concerning voice commands of 7, 5, and 3 words, which are 1.357, 0.882 and 0.521 seconds, correspondingly. The lightweight nature of the app makes it feasible for deployment on low-end wearables.

6 Discussion and Limitation

GyroTalk provides a novel approach to defend against replay attacks on advanced voice interfaces. It demonstrates robust defense without the need of additional hardware for wearable devices. However, GyroTalk remains some limitations. For instance, presently, it is only functional when the speaker remains still, otherwise, the FRR may be high. We speculate that this is an effect caused by the difference in sensor performance on different devices. Hence, this study identifies a future direction for research to eliminate the impact of the speaker's motion state by employing Independent Component Analysis (ICA) [36]. ICA can separate mixed signals that are simultaneously input and reconstruct them into signals based on independent components. In this case, the voice signal and the user motion noise signal are mixed based on independent components, so they can be separated by ICA. However, ICA also has the problem of how to select the independent components, so further research and improvement are needed in practice. Besides, GyroTalk is strongly dependent on the built-in gyroscope. When the wearable device cannot move properly (i.e., gyroscope movement is limited), the system cannot obtain effective gyroscope data, thus, the performance of GyroTalk will be greatly influenced. Finally, in order to create a mature solution, we will assess GyroTalk on a variety of wearables across a sizable user base.

7 Conclusion

In this study, we create a voice authentication spoofing detection system called GyroTalk that uses the embedded gyroscope of wearables to record SMSs from speakers. The system captures the SMSs of speakers without the need for any additional hardware or cumbersome operations. The practicality of GyroTalk makes it applicable to wearables that can be worn in different positions. Extensive experiments are conducted, and the outcomes illustrate GyroTalk's robustness and effectiveness. Hence, this study identifies a future direction for research to eliminate the impact of the speaker's motion state by employing ICA. In addition, in order to develop a mature solution, we will evaluate GyroTalk across a wide range of wearable products and a large user base.

References

[1] X. Wang, H. Zhu, S. Chang, and X. Wang, "Sevi: Boosting secure voice interactions with smart devices," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2020, pp. 2204–2212.

[2] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *Proc. IEEE Symp. Secur. Privacy*, 2021, pp. 694–711.

[3] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *Proc. Eur. Symp. Res. Comput. Secur.*, 2015, pp. 599–621.

[4] C. Kasmi and J. L. Esteves, "Iemi threats for information security: Remote command injection on modern smartphones," *IEEE Trans. Elect. Compat.*, vol. 57, no. 6, pp. 1752–1755, 2015.

[5] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2020.

[6] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proc. 4th ACM Workshop Secur. Privacy*, 2014, pp. 63–74.

[7] T. Sugawara, B. Cyr, S. Rampazzi, D. Genkin, and K. Fu, "Light commands: laser-based audio injection attacks on voice-controllable systems," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 2631–2648.

[8] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 343–355.

[9] G. Chetty and M. Wagner, "Automated lip feature extraction for liveness verification in audio-video authentication," *Proc. Image and Vision Computing*, pp. 17–22, 2004.

[10] M.-I. Faraj and J. Bigun, "Synergy of lip-motion and acoustic features in biometric speech and speaker recognition," *IEEE Trans. on Comput.*, vol. 56, no. 9, pp. 1169–1175, 2007.

[11] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1080–1091.

[12] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 57–71.

[13] Y. Meng, J. Li, M. Pillari, A. Deopujari, L. Brennan, H. Shamsie, H. Zhu, and Y. Tian, "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 1077–1094.

[14] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2019, pp. 2062–2070.

[15] T. Liu, F. Lin, C. Wang, C. Xu, X. Zhang, Z. Li, W. Xu, M.-C. Huang, and K. Ren, "Wavoid: Robust and secure multi-modal user identification via mmwave-voice mechanism," in *Proc. 36th Annu. ACM Symp. on User Interface Softw. and Technol.*, 2023, pp. 1–15.

[16] F. Han, P. Yang, H. Du, and X.-Y. Li, "Accuth $^{+}$: Accelerometer-based anti-spoofing voice authentication on wrist-worn wearables," *IEEE Transac. on Mobile Comput.*, 2023.

[17] Y. Jiang, H. Zhu, S. Chang, and B. Li, "Mauth: Continuous user authentication based on subtle intrinsic muscular tremors," *IEEE Transac. on Mobile Comput.*, 2023.

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

[19] H. Bredin and G. Chollet, "Making talking-face authentication robust to deliberate imposture," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, 2008, pp. 1693–1696.

[20] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 321–336.

[21] L. Huang and C. Wang, "Pcr-auth: Solving authentication puzzle challenge with encoded palm contact response," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 1034–1048.

[22] W. Xu, W. Song, J. Liu, Y. Liu, X. Cui, Y. Zheng, J. Han, X. Wang, and K. Ren, "Mask does not matter: Anti-spoofing face authentication using mmwave without on-site registration," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 310–323.

[23] J. Wang, L. Lu, Z. Ba, F. Lin, and K. Ren, "Shift to your device: Data augmentation for device-independent speaker verification anti-spoofing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, 2023, pp. 1–5.
[24] Y. Lee, Y. Zhao, J. Zeng, K. Lee, N. Zhang, F. H. Shezan, Y. Tian, K. Chen, and X. Wang, "Using sonar for liveness detection to protect smart speakers against remote attackers," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–28, 2020.
[25] P. Jiang, Q. Wang, X. Lin, M. Zhou, W. Ding, C. Wang, C. Shen, and Q. Li, "Securing liveness detection for voice authentication via pop noises," *IEEE Transac. Depen. Secur. Comput.*, vol. 20, no. 2, pp. 1702–1718, 2022.
[26] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, vol. 4, 2011, pp. 1708–1713.
[27] Y. Meng, J. Li, M. Pillari, A. Deopujari, L. Brennan, H. Shamsie, H. Zhu, and Y. Tian, "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *Proc. 31th USENIX Secur. Symp.*, 2022, pp. 1077–1094.
[28] Y. Meng, J. Li, H. Zhu, Y. Tian, and J. Chen, "Privacy-preserving liveness detection for securing smart voice interfaces," *IEEE Transac. Depen. Secur. Comput.*, 2023.
[29] Q. Yang, K. Cui, and Y. Zheng, "Room-scale voice liveness detection for smart devices," *IEEE Transac. Depen. Secur. Comput.*, no. 01, pp. 1–14, 2024.
[30] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, 2019.
[31] A. Smith, L. Goffman, H. N. Zelaznik, G. Ying, and C. McGillem, "Spatiotemporal stability and patterning of speech movement sequences," *Experimental Brain Research*, vol. 104, no. 3, pp. 493–501, 1995.
[32] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in matlab," *University of Athens, Athens*, vol. 2, 2009.
[33] P. Hall, B. U. Park, R. J. Samworth *et al.*, "Choice of neighbor order in nearest-neighbor classification," *Ann. Statist.*, vol. 36, no. 5, pp. 2135–2152, 2008.
[34] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Math. Program.*, vol. 127, no. 1, pp. 3–30, 2011.
[35] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, 2007.
[36] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.

Shan Chang (Member, IEEE) received the Ph.D. degree in computer software and theory from Xi'an Jiaotong University, Xi'an, China, in 2012. From 2009 to 2010, she was a Visiting Scholar with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. She was also a Visiting Scholar with BBCR Research Lab, University of Waterloo, Waterloo, ON, Canada, from 2010 to 2011. She is currently a professor at the Department of Computer Science and Technology, Donghua University, Shanghai, China. Her research interests include security and privacy in mobile networks and sensor networks. Dr. Chang is a member of IEEE Computer Society, Communication Society, and Vehicular Technology Society.

Luo Zhou received the M.S. degree from the Department of Electronics and Communication Engineering, Jiangsu University of Science and Technology, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Donghua University, Shanghai, China.

His research interests include ubiquitous and pervasive computing, mobile computing/sensing and edge computing.

Wei Liu received the Ph.D. degree in computer science from the Donghua University in 2021. He is now an associate professor with the Department of Computer Science & Technology, Anhui University of Finance & Economics, Bengbu, China.

His research interests include ubiquitous and pervasive computing, mobile computing and wireless sensing.

Hongzi Zhu received the PhD degree in computer science from Shanghai Jiao Tong University, in 2009. He was a post-doctoral fellow with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, and the Department of Electrical and Computer Engineering, University of Waterloo, in 2009 and 2010, respectively. He is now a professor at the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include mobile sensing and computing, and Internet of Things. He received the Best Paper Award from IEEE Globecom 2016. He was a leading guest editor for IEEE Network Magazine. He is an associate editor for the IEEE Transactions on Vehicular Technology. He is a member of the IEEE, IEEE Computer Society, IEEE Communication Society, and IEEE Vehicular Technology Society. For more information, please visit http://lion.sjtu.edu.cn.

Xinggang Hu received the B.S.degree from Anhui Polytechnic University in 2017. He received the M.S. degree from the Department of Computer Science and Technology, Donghua University, Shanghai, China, in 2020.

His research interests include Internet of Things, mobile sensing and wearable computing. He is a member of ACM.

Lei Yang (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Computer Science and Technology, School of Software, Xi'an Jiaotong University. He is currently working as an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University. Previously, he was a Post-Doctoral Fellow at the School of Software, Tsinghua University. His research interests include the Internet of Things, RFID and backscatters, and wireless and mobile computing.