2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON) | 978-1-6654-8643-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SECON55815.2022.9918548

VOGUE: Secure User Voice Authentication on Wearable Devices using Gyroscope

Shan Chang*, Xinggan Hu*, Hongzi Zhu†, Wei Liu‡ and Lei Yang§
*School of Computer Science and Technology, Donghua University, Shanghai, China
†Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
‡School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China
§Department of Computing, Hong Kong Polytechnic University, Hongkong, China
changshan@dhu.edu.cn, hxg@mail.dhu.edu.cn, hongzi@cs.sjtu.edu.cn, liuwei628@mail.dhu.edu.cn, tagsysx@gmail.com

***Abstract*—Voice assistants are popular to wearable devices with limited input and output capabilities, however vulnerable to voice attacks, which cheat a voice assistant by playing forged voice commands without user awareness. In this paper, we propose VOGUE, which captures unique yet stable pattern of speech movement sequences of speakers with embedded gyroscope in wearable devices, to distinguish between registered legal user and malicious attackers (human or machines). The design of VOGUE is based on two key observations. First, speech, as a type of highly complex motor task, inherently requires coordinated actions of many orofacial, laryngeal, pharyngeal, and respiratory muscles, and the collective movements of muscles propagate to distant body segments. Second, to generate a certain word, the speech movement sequence of a speaker is known to be distinctive, and can be captured by inertial sensors. We implement VOGUE on three kinds of COTS android devices including smart glasses, watches and phones, and conduct comprehensive evaluation on the performances. Experimental results show that VOGUE achieves a mean false-acceptance rate (FAR) and false-rejection rate (FRR) of 2.23% and 2.48%, respectively, even under sophisticated voice impersonation attacks.**

***Index Terms*—voice authentication, replay attack, speech movement sequence, gyroscope, liveness detection**

I. Introduction

Last decade has witnessed the wide prevalence of wearable devices, ranging from glasses, watches and headphones, to wrist-bands and jewelries. As most wearable devices are not equipped with proper input and output mechanisms (e.g., lacking large-size displays, and expressive input surfaces), traditional click-based (e.g., used in desktop scenarios) or tap-based (e.g., used in mobile scenarios) interactions are uncomfortable to use. In contrast, voice-based interaction (e.g., voice assistant apps such as Siri, Alexa, Cortana and Google Assistant) does not require users to type commands and allows users to make phone calls, play music and send messages, providing appealing user experience.

However, given the open sound channel, voice can easily be recorded without being noticed, making voice-based interaction inherently vulnerable to spoofing attacks. Due to the proliferation of high-quality, low-cost playback and recording devices, e.g., smartphones and digital recorders, it is easy to conduct replay attacks. Moreover, sophisticated attackers are able to fabricate voice passphrase of a victim by splicing a number of voice segments [1]. Existing studies demonstrate that, by learning only a very limited number of voice samples of the victim, an attacker is able to build a very close model representing the victim's voice. In addition, those counterfeit voice commands can be injected into the sound channel without the notice of human being [2] [3]. Such attacks often lead to severe consequences such as unlocking a device, tuning on a camera, and authorizing online payment, and thus have drawn much attention.

In the literature, to defend against voice spoofing attacks, liveness detection is applied to distinguish between legitimate voice samples from a human speaker and counterfeit ones from a machine. Existing liveness detection schemes fall into three categories. First, use a video camera to extract lip movements [4] or to detect a reliable face [5]. However, face recognition is restricted by illumination, and low-end wearables may not be equipped with a camera. Second, ultrasound-based verification leverages a built-in speaker of a smartphone to transmit a high frequency acoustic sound, and to listen the Doppler shift (due to articulatory gesture) of its reflections at the microphone when a user speaks a passphrase [6] [7]. The smartphone has to be placed very close to, and right in front of a speaker, such that ultrasound can be reflected by articulators, which makes the detection inconvenient for use on devices like glasses, necklaces, and headphones. Third, Feng *et al.* rely on a vibration sensor with very high sampling rate (11kHz), which is encapsulated as a hardware extension of wearables. It is placed close to the throat of speaker, in order the capture the vocal cord vibration and match it with the speech signal received by microphone [8].

In this paper, we propose VOGUE, a liveness detection scheme based on speech movements sequence. The core idea of VOGUE is to exploit the built-in gyroscope in wearable devices to capture user-specific speech movement sequence when a user speaks a passphrase, and verify the legitimacy the sequence. VOGUE is inspired by the following two key insights. First, speech is a kind of purposeful and task-related motor behavior, which inherently requires coordinated actions of many muscles and joints. During speaking, functional mandibular and head-neck movements are the results of activation of jaw as well as neck muscles, leading to simultaneous movements in the temporomandibular, atlantooccipital

and cervical spine joints. Those collective movements also propagate to distant body segments, since human skeleton is cross-linked, and muscle fibers adhere to each other. Second, there exist non-negligible differences in speech movement sequences (even for the same passphrase) among people due to individual diversity in the size, shape, strength of muscles, bones, and joints, and in the habitual way muscles exert forces. Such distinctive features could be utilized to detect an attacker who attempts to impersonate the speech movement sequence of a genuine user.

VOGUE contains an offline user enrollment process in which a user speaks a predetermined passphrase for several times, and the corresponding speech movement sequences are collected and stored as templates. During an online authentication process, the user speaks the registered passphrase for once, and the corresponding speech movement sequence is compared against those templates. The VOGUE design faces three main challenges as follows. First, how to effectively perceive micro-movements caused by speech is not straightforward. By analyzing sensory data from commonly used built-in sensors, we figure out that speech movement tends to have small accelerations however relatively large rotations on body segment, thus we utilize gyroscope to sample speech movement sequences.

Second, how to choose movement features that can characterize the speaker is not straightforward. We find that commonly used statistical features are unsatisfactory, because they cannot describe the whole trends of entire motions in fine grain, while feeding the entire motion sequence into the classifier not only introduces a high computational cost, but also implies the necessary of a large number of training samples, thus increasing registration burden of users. Therefore, we chose to characterize the similarity between registered templates and testing samples. A testing passphrase is compared with top-k templates to enhance the robustness of our similarity-based feature vectors.

Last but not least, when calculating the similarity between two motion signals, it is very difficult to eliminate the influence caused by different speaking rhythms. The problem is twofold. In the case of two passphrases spoken by the same speaker, due to different speaking rhythms, the movements related to same syllable occur on different locations in two gyroscope signals, which leads to a small similarity erroneously. We introduce *Voice Aware DTW (VA-DTW)*, in which we warp the gyroscope signals according to temporal aligned microphone signals. Then compare the similarity of the aligned gyroscope signals. In the case of two passphrases spoken by different speakers, it is required to avoid accidentally large similarity caused by matching the movements of different syllables mistakenly. We observe that a speech movement sequence from a genuine user should have small distortion when achieving the optimal matching to a template sequence, thus *Warping score* is defined to measure the similarity among speech movement sequences together with VA-DTW distance.

We implement VOGUE on three types of COTS android devices including a pair of smart glasses, a watch and a phone, and enroll 15 participants to conduct extensive real-world experiments. VOGUE is highly effective in distinguishing between legal users and malicious attackers (human or machines) of voice assistant systems, and can be used on wearables placed on various parts of human bodies, and does not require cumbersome operations or additional hardware. Experimental results demonstrate that VOGUE can achieve a low mean false-acceptance rate (FAR) and false-rejection rate (FRR) of 2.23% and 2.48%, respectively, even under sophisticated voice impersonation attacks.

II. Related Work

Existing work for detecting voice spoofing attacks can be divided into two categories: liveness detection and channel characteristic detection.

A. Liveness Detection

A liveness detection system distinguishes between the legitimate voice samples of human speakers and the replayed ones. Three methodologies of liveness detection have been proposed.

1) Conversation-face based Detection Methods: This kind of schemes identity speakers utilizing face and lips motion. By simultaneously recording voice and face signals, the similarity between them is calculated to defend against replay attacks [5] [9] [10] [11]. However, these detection schemes have two disadantages. First, the effectiveness of the schemes is subject to the lighting conditions. Second, they require access to the camera of devices, which introduces security risks.

2) Ultrasound-based Detection Methods: With the microphone, when the user speaks a passphrase, the built-in speaker emits ultrasonic waves, and the microphone records the Doppler shift in the reflected signal caused by multiple articulators of the sound system when the user utters the sound. The Doppler shift features are separated from the speech signal, and finally the similarity scores between a testing feature and the registered one is calculated. If the similarity score exceeds a predetermined threshold, a legitimate user is declared [6] [7]. The shortcoming of these methods is that the speaker needs to bring the smart device close to the mouth and to make a specific posture, in order to capture reflected signal clearly.

3) Vibration-based Detection Methods: The high-band speech signal is measured by using a specific acceleration sensor, and is compared with the voice signals recorded by the microphone. If the correlation between the two signals is lower than a threshold, the voice attack is detected [8]. The disadvantage of this method is that an additional sensor device is required and should be placed close to the throat of a speaker, in order to catch the vibration of vocal cords.

B. Channel Characteristic Detection

Channel characteristic-based schemes detect the difference between the data transcribed by a high-fidelity recording device and the original data on the sound channel. Based on the UBM (Universal Background Model), the channel is modeled

using the silent segment of the voice data to tell whether the channel to be authenticated is the same as the channel for training voices [12]. Nevertheless, the mute segment has a small amplitude, which is more susceptible to noise pollution than the speech segment, and based on the general background model of the speech segment, it is not always possible to train an accurate channel model.

Our VOGUE falls into the category of liveness detection. Comparing with existing solution, VOGUE does not make any restriction on the placement of wearables or require additional hardware or induce potential privacy risks.

III. Preliminaries

A. System and Attack Model

VOGUE is designed to secure task-dependent voice authentication (i.e., speaker recognition), in which a speaker is asked to speak a predetermined passphrase in enrollment, and the same phrase will be used for verification in future. VOGUE can be applied to different kinds of COTS wearable devices, e.g., necklaces, glasses, bracelets, helmets and mobile phones, as long as a device is equipped with a gyroscope, and has ability of computing. Furthermore, we require that the device should be carried or worn by a user, such that its inertial sensor can capture the micro-movement induced by speaking.

We consider an attacker who aims at fooling the victim's voice authentication system using fabricates voice command. The attacker can launch either a replay attack (i.e., recording the victim's voice in advance and replaying it), or a voice mimicry attack (i.e., injecting synthesized or imitated passphrase into audio channel). *Moreover, we assume that the attacker can obtain the victim's device, and know the detail design of VOGUE.* In particular, we consider the following two types of attack scenarios:

Scenario A: Zero-knowledge Impersonation Attack. When replaying (or broadcasting) a passphrase, the attacker wears the targeted device and mouths the passphrase simultaneously (but no sound must be uttered), to generate corresponding body movement. However, the attacker knows nothing about how the body of the victim moves when he or she pronounces the passphrase.

Scenario B: Snooping and Impersonation Attack. Snooping attack has the same process as zero-knowledge attack, except that the attacker first learns how its victim pronounces the passphrase, for example by taking a video of the victim, and then rehearses before launching the attack.

B. Speech Movement Sequence

Normally, speech is produced with pulmonary pressure from the lungs which generates sound by phonation through the glottis in the larynx, and then the sound is modified by the vocal tract into different vowels and consonants. During the production of certain syllable, relatively invariant combined muscles are involved, and moment-to-moment coordinated movements of several separate body parts distributed from the diaphragm to the lips are activated [13].

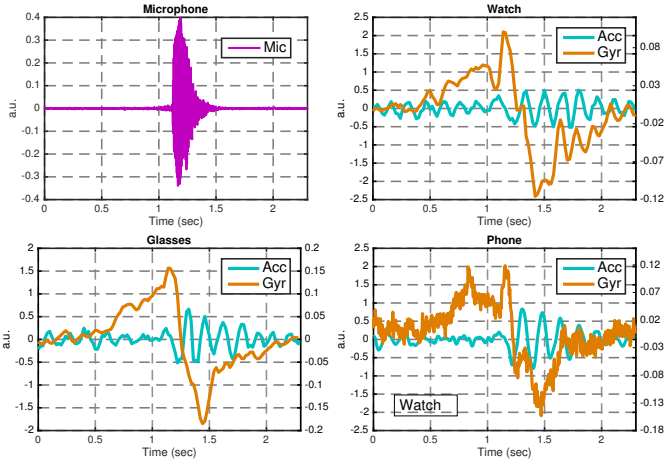


Fig. 1. The raw audio data sampled by microphone when a speaker pronounces "*home*", and the corresponding motion sequences collected from accelerometers and gyroscopes of smart watch, glasses, and phone

Definition 1. A **Speech Movement Sequence** refers to a series of complex body movements caused by speaking, which contains: 1) immediate synchronous movements over 100 muscles related to speech, including diaphragm, pectoralis major, mentalis, buccinator, platysma etc.; 2) ensuing traction-induced movements (displacements) of other distant muscles and body segments, the causes of which are muscle fibers adhere to each other, and human skeleton is cross-linked.

C. Capturing Diversity of Speech Movement Sequence

Considering that a speech movement sequence implies the contraction of muscles and displacements of body segments, we believe that it is possible to capture the speech movement sequence using motion sensors. Thus, we use both accelerometer and gyroscope, which are widely equipped on wearable devices, to measure the body movements of a speaker. We find out that, first, gyroscope is more sensitive than accelerometer, which is consistent with the intuition that speech movement tends to have small accelerations however relatively large rotations on body segments. Figure 1 shows the raw audio data sampled by microphone (left top) when a speaker pronounces "*home*", and the corresponding motion sequences collected from accelerometers and gyroscopes of smart watch (worn on the wrist), glasses (worn on the head), and phone (hold in the hand), respectively. It can be seen that speaking causes a much bigger fluctuation on gyroscope data than on accelerometer data, and, even a device located on distant body part, i.e., hand and wrist, can perceive micro-movements caused by speech.

Furthermore, because of the individual differences in the size, shape, strength of muscles, bones, and joints, as well as in the way muscles exert forces, we think that, for generating certain passphrase, the corresponding speech movement sequence of a speaker will show a distinctive and repeatable spatial-temporal characteristic. Figure 2 shows time series of gyroscope data of a smart watch sampled during pronouncing "*America*" twice by two volunteers, $\mathcal{A}$ (left) and $\mathcal{B}$ (right),

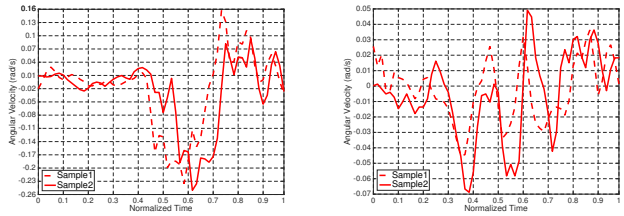


Fig. 2. Gyroscope signals of a smart watch sampled during volunteer $\mathcal{A}$ (left) and $\mathcal{B}$ (right) pronouncing "*America*" twice.

respectively. It can be seen that, two samples from certain speaker show high similarity, however, there exists obvious difference between samples from different speakers.

IV. Design of VOGUE

A. Overview

The key idea of VOGUE is to leverage built-in gyroscopes of COTS wearable devices to capture and recognize distinctive and repeatable spatiotemporal pattern of speech movement sequences of speakers. When a user speaks a passphrase to a device, its voice and body movements will be record by microphone and gyroscope, simultaneously. Once the microphone signal passes voice authentication, gyroscope data will be further verified by VOGUE.

VOGUE authentication has two phases: offline training and online authentication. In the training phase, a user speaks a passphrase for several times, in order to collect templates of voice and motion signals, which are employed to build a classifier. In the authentication phase, a testing input will be compared with those templates to determine whether it comes from a registered user or not.

Original templates are not suitable for training the classifier for two reasons. First, a speaker may pronounce the same passphrase with different speeds, which means those templates are of different lengths. Second, the amount of data points in a template is large, especially for a long passphrase. It implies a large number of training templates are necessary to guarantee high accuracy, which increases the registration burden of users. A reasonable choice is to extract effective features from original signals to form feature vectors, which is however non-trivial. The performance of normally used statistical features in time series, like maximum and minimum value, standard deviation, and their combinations, is unsatisfactory. Since the essential factor for classification is the trend of variation of a time series of gyroscope data rather than the distribution of it, which however cannot be represented using statistical features. In other words, the reason is that features should occur on the expected positions of a data sequence. For example, Figure 3 displays two segments of gyroscope data of glasses, which belong to different speakers pronouncing "*do*" . We can see that the two segments are of different shapes. However, the maximum, minimum values, and standard deviations of them are very similar.

TABLE I
Command List

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

We introduce a Voice Aware DTW (VA-DTW) algorithm to enable shape matching between two voice-aligned movement sequences. If two voice commands are spoken by the same person, the two corresponding movement sequences will have a small VA-DTW distance, as well as a short warping path. We conduct VA-DTW on speech movement sequence pairs to extract distance vectors. Both VA-DTW distance and warping score are utilized to training the classifier.

B. Data Collection

We develop an Android application for collecting both microphone and gyroscope signals simultaneously (sampling rates are 44100 and 50, respectively), and run it on Google Nexus 5X smartphone, MOTO 360 Android Wear smart watch and Storm Mirror Smart VR glasses. We recruit 15 volunteers, 4 women and 11 men, aged 21 to 27. During data collection, each volunteer holds the smartphone in dominant hand, puts the glasses on head, and wears the watch on the wrist of non-dominant hand. When issuing a voice command, a volunteer is asked to bend the forearm with smart device and place it in front of the chest naturally. We emphasize that, during trace collection, volunteers can rest or quit at any time.

As preparation, each volunteer predetermines 10 commands listed in Table I, which contains 30 voice commands divided into three groups according to length, satisfying that the number of commands selected in each group are roughly the same. Furthermore, we allow volunteers negotiate to ensure each command is selected by 4 to 6 times (expectation is 5). Then we collect the following two traces:

Trace A is collected for 30 days. In general, each volunteer speaks every selected command for *10* times a day in different speaking volumes and paces. To be specific, each command is spoken *six* times at a normal speaking volume with *two* times

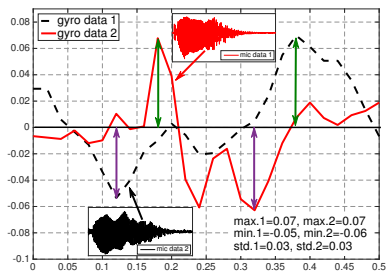


Fig. 3. Two segments of gyroscope data, belonging to different speakers, with different shapes, but similar *maximum*, *minimum*, and *standard deviation*.

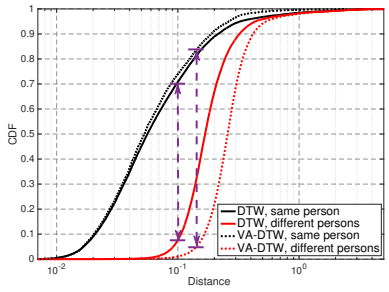


Fig. 4. CDF curves of the DTW and VA-DTW distances between gyroscope signals from same and different persons.

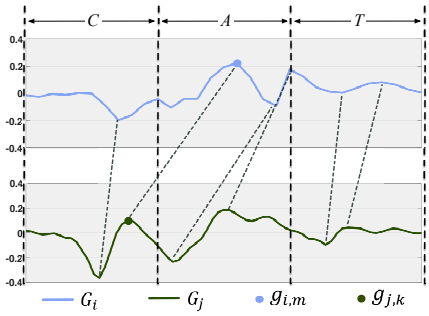


Fig. 5. An example of mismatching. Segments of gyroscope signals belonging to different phonemes are matched mistakenly.

each at regular, slow, and fast paces, and *two* times at a higher and a lower volumes, respectively, with regular paces. In order to investigate whether VOGUE is affected by body postures, we ask volunteers complete data collection while *standing, sitting straight* and *hump*, in the first, second and last *ten* days, respectively. In this way, for each command, we collect 300 pieces of microphone signals, as well as corresponding gyroscope signals, per volunteer. Additionally, We record the process of data collection using a video camera, which will be used in launching impersonation attacks.

Trace B is about impersonation attacks. We select *five* volunteers as victims, and the others act as attackers. Firstly, each attacker launches zero-knowledge impersonation attack (see Scenario A) against each command of victims *ten* times. After that, attackers watch the videos recording the way victims speaking their commands arbitrarily, and then speak each command of victims for *10* times, trying their best to mimic the way victims speaking (i.e., Scenario B). In this way, for both type of impersonation attacks, every command of a victim is attacked for 100 times.

C. Pre-processing

Assume that we collect a pair of time-ordered sequences from microphone and gyroscope, i.e., sampled voice command and speech movement synchronized in time. We perform the following two-step preprocessing.

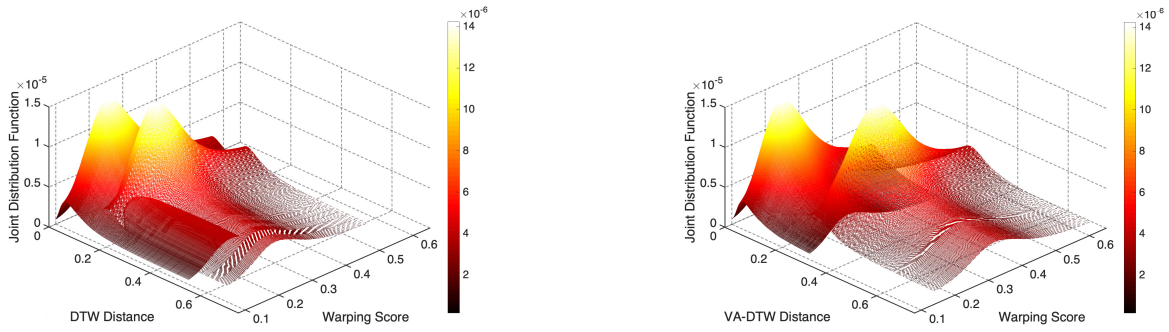
Removing unvoiced segments: since we only care about speech related movements, we use silence removal [14] to remove unvoiced segments in the audio signal. After that, a voice command V might be cut into several segments as well as the corresponding gyroscope signal G, obtaining several pairs of microphone and gyroscope data segments, each of which is denoted by $S_i = (V_i, G_i), i \geqslant 1$, where V_i and G_i represent the audio and motion segments, respectively.

Normalization: intuitively, the loudness of sounds closely correlates with the intensity of speech movements, and thus the amplitudes of gyroscope signals. To eliminate the influence of speaking volume on gyroscope signals, we normalize the amplitude of each gyroscope signal segment G_i.

D. Voice Aware DTW based Similarity

To measure the similarity between two segments of gyroscope signals with different length, a popular distance-computing algorithm is DTW. However, we DTW distance is not adequate for distinguishing whether two segments of gyroscope signals are from the same person or not. Figure 4 depicts CDF curves (solid line) of the DTW distances between two segments of gyroscope signals (same voice command) from same and different persons, respectively. It can be seen that gyroscope signals from different speakers could also have small DTW distances, e.g., 30% are less than 0.15. We figure out that warping gyroscope signals arbitrarily may lead to segments belonging to different phonemes being matched, which results in a small DTW distance (by coincidence) but a long warping path. Figure 5 gives an example of such mismatching. The data points $g_{i,m}$ and $g_{j,k}$ belonging to '*C*' and '*A*', respectively, are matched mistakenly. Thus, before comparing two gyroscope signals, it is necessary to align them according to their corresponding audio signals, and gyroscope signals from same speaker should have not only small warping distance but also small distortion on the optimal matching. We introduce VA-DTW distance to eliminate such mismatching when comparing gyroscope signals. In general, given S_i and S_j, we conduct DTW on their microphone signals, and record the corresponding warping function, and apply it to the corresponding gyroscope signals obtaining voice aligned gyroscope signals. Then, we calculate the DTW distance between them as their VA-DTW distance, as well as the corresponding warping score, to form a two-tuple of distance, which characterize the similarity between gyroscope signals and serve as inputs to the subsequent classifier. Specifically, the following procedures should be conducted.

1) Downsampling of Audio Signal: Since the sampling rate of microphone is much larger than that of gyroscope, which means the warping function between V_i and V_j cannot be applied directly on gyroscope signals, we conduct downsampling on microphone signals before doing DTW, which also reduces computational cost of DTW. A naive way of downsampling is to decimate an original microphone signal by an integer factor D; that is, keep only every D-*th* sam-



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

Fig. 6. The joint distribution functions of VA-DTW and DTW, and their corresponding warping scores, of two gyroscope data segments from the same and different persons

ple. However, the sampling rate of microphone is 882 (i.e., 44100/50) times that of gyroscope, such a sparse resampling is very likely to miss peak values in original signals, which leads to the tendency of a resampled signal dissimilar to that of the original one. Instead, we find the envelopes of V_i, then decimate the upper envelop with *D=882* to obtain a downsamped audio sequence. The envelopes are determined using spline interpolation over local maxima separated by at least σ (we empirically set σ as 1500) samples.

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 $I \times J$ 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. 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 of $G_A^{'}$ and $G_B^{'}$ equal to that of A and B, respectively.

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 7, 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., $(\mathbb{C}(G_i^{'}, G_j^{'}), \Phi(G_i^{'}, G_j^{'}))$.

E. Classification

During the offline training phase, all training templates (each of which is a pair of voice and motion signals) collected will be compared with each other. Suppose two pairs of microphone and gyroscope data signals S_i and S_j, need to be compared. Firstly, they will be segmented into two sequences of signals according to IV-C, i.e., $S_{i_1}, ..., S_{i_n}$, and $S_{j_1}, ..., S_{j_n}$. Then we construct a distance vector $D_{i,j}$ between them, i.e., $\mathbb{C}(G_{i_1}^{'}, G_{j_1}^{'}), \Phi(G_{i_1}^{'}, G_{j_1}^{'}), ..., \mathbb{C}(G_{i_n}^{'}, G_{j_n}^{'}), \Phi(G_{i_n}^{'}, G_{j_n}^{'})$ All distance vectors obtained by comparing templates will be fed into a one-class classifier, e.g., KNN or SVM.

Additionally, we introduce a step to identify top-*K* training templates. Given $\mathbb{N}$ training templates, we first identify the *K* templates that are closest to all the training samples. For each template, we calculate its VA-DTW distance to other templates, and then choose *K* templates which have the smallest average distances. These *K* templates, namely top-*K* templates, are an empirical estimation of the centroid of the template

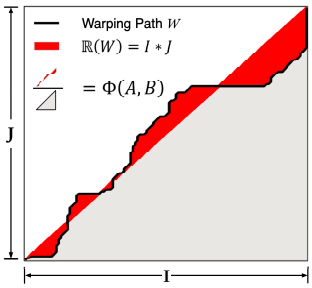


Fig. 7. An example of calculating the warping score of *A* to *B*

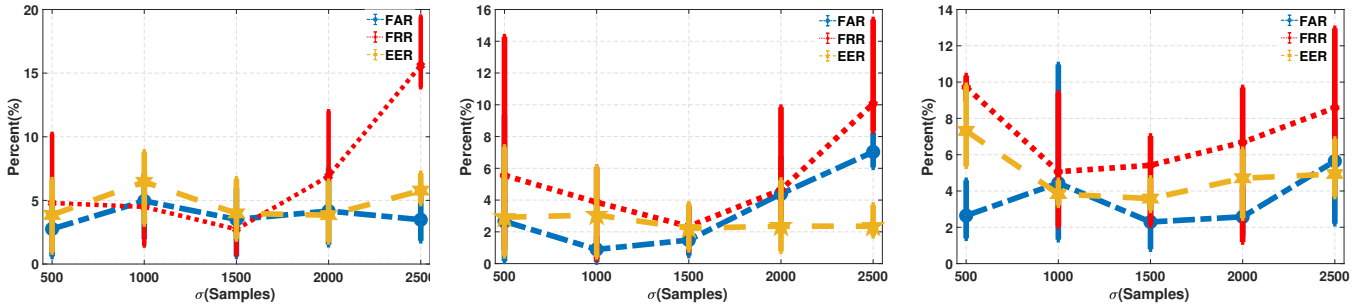


Fig. 8. Error rates under different σ (left: watch, center: glasses, right: phone)

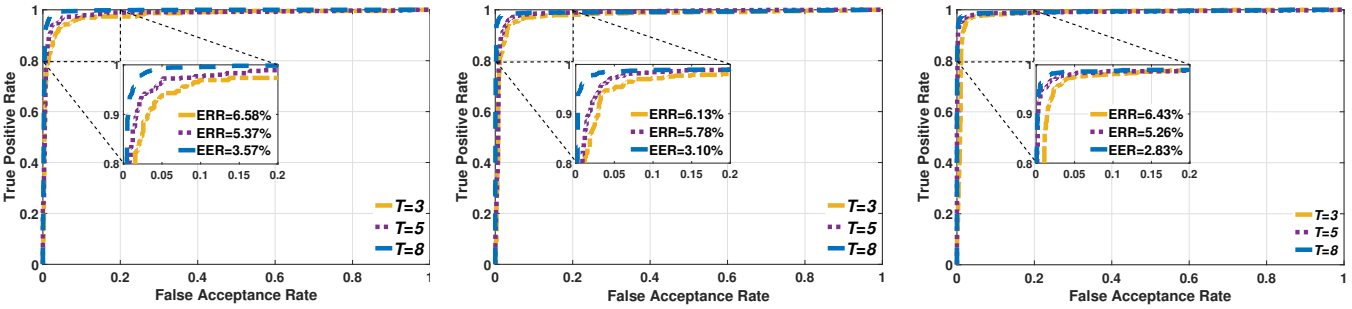


Fig. 9. ROC curve under different T (left: watch, center: glasses, right: phone)

space, and thus best represent the space among the collection of the training templates. In online authentication, focusing on the top-K templates rather than all templates brings the advantages of reducing computation overhead, and enhancing robustness against noises in training templates.

During online authentication, a testing sample S_i will be accepted or rejected by checking its distance vectors to top-K templates. Specifically, we compare S_i with top-K templates, and obtain K distance vectors, i.e., $D_{i,k}$ $(1 \leq k \leq K)$, then compute the center of these vectors, i.e.,$\overline{D}_i = \frac{1}{K}\sum_{k=1}^{K} D_{i,k}$. Finally, $\overline{D}_i$ is input to the classifier for judgment.

V. Evaluation

We conduct a comprehensive assessment of VOGUE based on traces described in Subsection IV-B. Both gyroscope and microphone data are transferred to a PC for offline analyzing. We employ the FAR and FRR as evaluation criteria. FAR is defined as the ratio between the number of falsely accepted illegitimate inputs and the number of all illegitimate testing inputs. FRR is defined as the ratio between the number of falsely rejected legitimate inputs and the number of all legitimate testing inputs. A Receiver Operating Characteristic (ROC) curve illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. We obtain the Equal Error Rate (EER) from the ROC curve where FAR and FRR are equal. We use **Trace A** in most experiments except impersonation attacks (in V-E), which utilize **Trace B**. To demonstrate the superiority of VA-DTW, we also compare VOGUE with two similar methods in which the distance vectors between two gyroscope signals are formed using traditional DTW distance, and the DTW distance plus the corresponding warping score, respectively.

A. Effectiveness of σ

Figure 8 plots the average, maximum and minimum of FRR, FAR, and EER over all volunteers as a function of σ, on different devices. We use three-word commands in this experiment, the value of σ is increased from 500 to 2500 at an interval of 500. We set K as 4 and use a one-class SVM classifier. It can be seen that, when σ =1500, all devices achieve best performances. Through analyzing the resulted envelops, we find that it is because a small σ makes the resulted envelopes contains to much unnecessary details, while a large σ will lead to big distortions on envelopes, both of which result in big differences between original and downsampled signals. Therefore, we set σ as 1500 in the following experiments.

B. Effectiveness of T

We examine VOGUE performance under different training data size, i.e., T = 3, 5, and 8. We set K as 4, and use a one-class SVM classifier and increase the discrimination threshold from 0.0 to 1.0 with an interval of 0.01, and plot ROC curves, in Figure 9. We observe that the average EERs drop gradually with the increase of T on all three devices. When T = 8, VOGUE delivers the best performance. It should be noticed

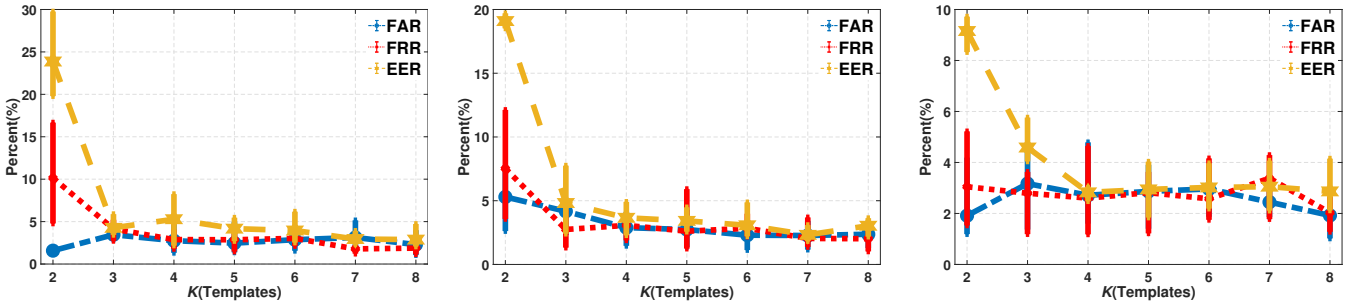


Fig. 10. Error rates under different K (left: watch, center: glasses, right: phone)

that the value of T has no effect on the computational overhead of the online authentication phase because the testing samples will only be compared with Top-K templates. Therefore, in the remainder of the study, we use *eight* training samples.

C. Effectiveness of K

Figure 10 plots the average, maximum and minimum of FRR, FAR, and EER over all volunteers as a function of K, on three devices, respectively. We use three-word commands in this experiment. The value of K is increased from 2 to 8 at an interval of 1. We have two observations. First, all three types of errors drop significantly when increasing K from 2 to 4, and then drop gradually as K being increased to 8 continually. Second, even a small K (e.g., 3), can achieve pretty good performances on all devices. During online authentication, a testing sample will be compared with Top-K templates. A large K implies high computational cost and long authentication delay. Thus, we set K as 3 in the remainder of the study, which achieves a best tradeoff between performance and cost.

D. Effect of Classifier

We compare the performance of two one-class classifiers, i.e., K-Nearest Neighbor [15] and SVM [16]. To select parameter κ of the KNN classifier, multiple tests with κ ranging from 1 to 10 are performed. The best parameter $\kappa = 3$ is selected. For the SVM classifier, we use the Radial Basis Function (RBF) as the kernel function. To obtain the appropriate parameters of c and g, we conduct a grid search over the same range of $[2^{-4},2^{4}]$ with cross validation on the training group. The experimental results are shown in Table II. It can be seen that both classifiers can achieve a satisfactory performance, and the SVM classifier outperforms KNN slightly. However, considering that, as the increasing of the training data size, the prediction efficiency of KNN will be decreased, we use the SVM classifier in the following experiments.

E. Defending against Impersonation Attacks

We examine the security of VOGUE according to the threat model presented in Section III-A by using Trace B.

TABLE II
COMPARISON OF TWO CLASSIFIERS

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

1) Zero-knowledge Impersonation Attack: Figure 11 illustrates the resilience of VOGUE to zero-knowledge attacks in general. It can be seen that, comparing with normal use scenarios, the FARs and FRRs of VOGUE only increase slightly (around 1%) for short commands (2 to 5 words), and almost remain unchanged for commands with 6 to 7 words, however both errors of DTW with or without warping score based schemes increase significantly in all cases.

2) Snooping and Impersonation Attack: Figure 12 demonstrates that, although further increased, both errors of VOGUE no more than 6% and 3%, for 2-3 and 4-7 word commands respectively, and it is still much more robust to impersonation attacks than the other two methods in all cases. For example, for commands with 4 to 7 words, the increases of average FARs of VOGUE are around 1%, while that of the other two methods exceed 7% and 5%, respectively.

VI. CONCLUSION

In this paper, we developed a spoofing detection system VOGUE for voice authentication, which leverages the built-in gyroscope of wearables to capture speech movement sequences of speakers. VOGUE is practical as no additional hardware and cumbersome operations are required, and it can be applied to wearables worn on different positions. Extensive experiments demonstrate the effectiveness and robustness of it. Nevertheless, VOGUE also has several limitations. For example, at the current stage, VOGUE can only be used when a speaker is stationary, otherwise, the FRR could be high. This also points out the direction of our future work. In the future, we will eliminate the impact of motion state of speakers by carrying out independent component analysis. In addition, we will evaluate VOGUE on more types of wearables, and large number of users, which are necessary for a mature solution.

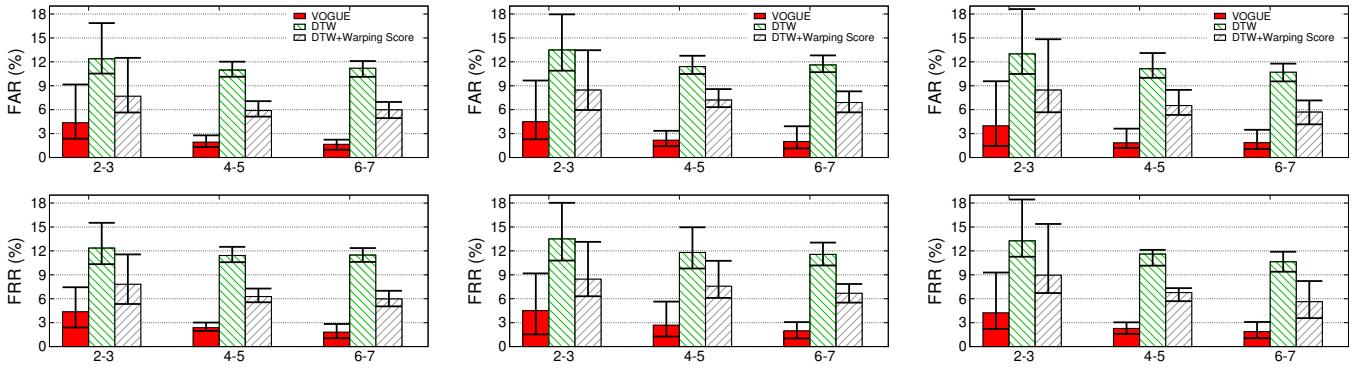


Fig. 11. FAR and FRR under zero-knowledge attacks (left: watch, center: glasses, right: phone)

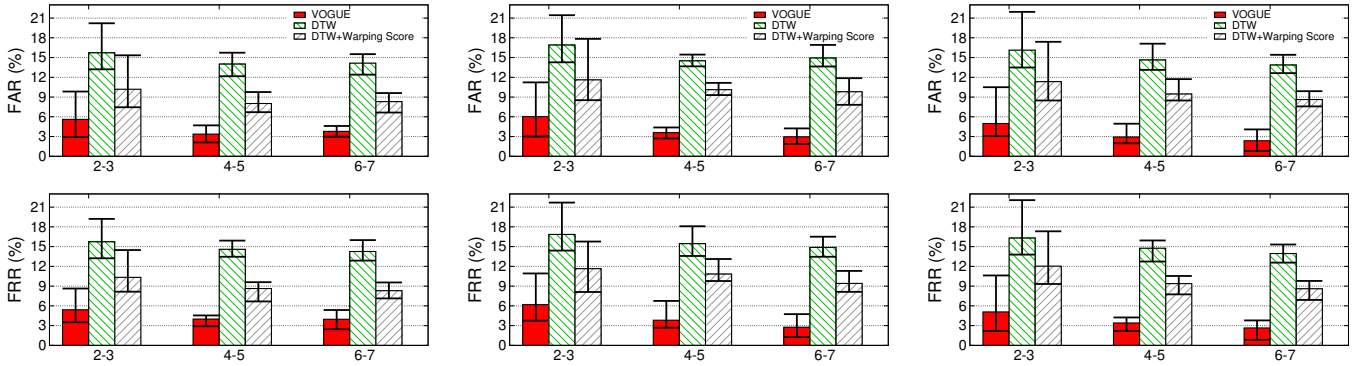


Fig. 12. FAR and FRR under impersonation attacks (left: watch, center: glasses, right: phone)

VII. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61972081), and the Natural Science Foundation OF Shanghai (Grant No. 22ZR1400200), the University Grant Committee of Hongkong (Grant No. 15204820).

References

[1] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *European Symposium on Research in Computer Security*, 2015, pp. 599–621.

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

[3] 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 *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 2014, pp. 63–74.

[4] 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.

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

[6] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1080–1091.

[7] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 57–71.

[8] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 343–355.

[9] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide, and H. Murase, "Detection of inconsistency between subject and speaker based on the co-occurrence of lip motion and voice towards speech scene extraction from news videos," in *2011 IEEE International Symposium on Multimedia*, 2011, pp. 311–318.

[10] H. Bredin and G. Chollet, "Making talking-face authentication robust to deliberate imposture," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1693–1696.

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

[12] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *2011 International conference on machine learning and cybernetics*, vol. 4, 2011, pp. 1708–1713.

[13] 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.

[14] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in matlab," *University of Athens, Athens*, vol. 2, 2009.

[15] P. Hall, B. U. Park, R. J. Samworth *et al.*, "Choice of neighbor order in nearest-neighbor classification," *The Annals of Statistics*, vol. 36, no. 5, pp. 2135–2152, 2008.

[16] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.