TeethPass: Dental Occlusion-based User Authentication via In-ear Acoustic Sensing

Yadong Xie* Fan Li* Yue Wu† Huijie Chen‡ Zhiyuan Zhao* Yu Wang§
* School of Computer Science, Beijing Institute of Technology, Beijing, China.
† School of Software, Tsinghua University, Beijing, China.
‡ School of Computer, Beijing University of Technology, Beijing, China.
§ Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA.
Email: {ydxie, fli, 3120205519}@bit.edu.cn, ywu92@mail.tsinghua.edu.cn, chenhuijie@bjut.edu.cn, wangyu@temple.edu

***Abstract*—With the rapid development of mobile devices and the fast increase of sensitive data, secure and convenient mobile authentication technologies are desired. Except for traditional passwords, many mobile devices have biometric-based authentication methods (e.g., fingerprint, voiceprint, and face recognition), but they are vulnerable to spoofing attacks. To solve this problem, we study new biometric features which are based on the dental occlusion and find that the bone-conducted sound of dental occlusion collected in binaural canals contains unique features of individual bones and teeth. Motivated by this, we propose a novel authentication system, TeethPass, which uses earbuds to collect occlusal sounds in binaural canals to achieve authentication. We design an event detection method based on spectrum variance and double thresholds to detect bone-conducted sounds. Then, we analyze the time-frequency domain of the sounds to filter out motion noises and extract unique features of users from three aspects: bone structure, occlusal location, and occlusal sound. Finally, we design an incremental learning-based Siamese network to construct the classifier. Through extensive experiments including 22 participants, the performance of TeethPass in different environments is verified. TeethPass achieves an accuracy of 96.8% and resists nearly 99% of spoofing attacks.**

I. INTRODUCTION

Nowadays, mobile devices are becoming powerful with a large storage capacity. They are often used to process sensitive information (e.g., private documents edit, health information record, and online payment). However, the leakage of user privacy data is increasingly serious. According to a survey from Cisco [1], 89% of users care about privacy data security, and 79% of them are willing to act to protect it. Another report from IBM [2] shows that personally identifiable information, such as login data, fingerprints, and voice, is the most frequently lost or stolen type of data. Thus, it is necessary to study a reliable and convenient authentication system.

To prevent the leakage of user privacy data, many authentication methods are adopted on mobile devices (e.g., PIN code, unlock pattern, and fingerprint). But the token used by these methods is susceptible to being inferred or stolen. Specifically, the PIN code and unlock pattern are the most popular authentication methods, but they are vulnerable to attacks [3] and require tedious input by users. Besides, many types of biometric features are studied for user authentication, such as fingerprint [4], [5], voiceprint [6], [7], and face recognition [9], [10], which are also adopted on commercial systems (e.g., Apple Touch ID [11], TD VoicePrint [12], and Amazon Rekognition [13]). However, these methods are vulnerable to replay attacks. For instance, an attacker can record the victim's face or voice, then replay the records to spoof the authentication system. Even fingerprints can be stolen through photos and made into fingerprint film for attacks. Recently, more types of behavioural [14] and biometric [15]–[17] features are leveraged to enhance the security of mobile authentication. SmileAuth [15] adopts the image of users' dental edge for authentication. LipPass [16] extracts features of users' speaking lips using audio devices on smartphones for authentication. However, these methods require the user to hold a phone towards the mouth, which is inconvenient and works in a limited scenario. EarEcho [17] uses unique features of human ear canal and assesses acoustic features of in-ear sound for authentication. But it is susceptible to interference from environments, such as position shift of device.

Motivated by the above limitations, we design a secure, convenient, and reliable user authentication method, TeethPass, based on the bone-conducted sound [18]–[20] of dental occlusion. Specifically, when a user occlude teeth, the occlusal sounds are absorbed, reflected, and dispersed by the skull and then transmitted to ear canals. Thus, the received sounds present individual differences due to the unique density and elasticity properties of his/her skull. With this characteristic, the occlusal sound (received in ears) can be used for authentication. Additionally, wireless earbuds are used widely in recent years. A survey [21] shows that the number of wireless earbuds in 2024 will reach 520 million, and many users report that they tend to wear earbuds all day. Most of all, commercial earbuds (e.g., Apple AirPods Pro, Sony WF-1000XM4, and Bose QuietComfort) have inward-facing microphones to collect the sound in ear canals for noise reduction. These principles inspire the basic idea of TeethPass: to use the inward-facing microphones of earbuds to capture bone-conducted sounds of dental occlusion, and then to extract the unique biometric

Fan Li is the corresponding author. The work of Fan Li is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No.62072040, 61772077 and Beijing Natural Science Foundation under Grant No. 4192051. The work of Huijie Chen is partially supported by China Postdocporal Science Foundation under Grant No. 2021M700302.

features from these sound for user authentication.

Despite its simple idea, three major challenges underlie the design of TeethPass. Firstly, although the inward-facing microphone is very close to the ear canal, it still can record air-conducted noises (e.g., speaking, environment noise). Besides, some daily activities also generate bone-conducted noises (i.e., eating, speaking, and walking). So, how to distinguish the bone-conducted occlusal sounds from the collected noisy acoustic signal is the first challenge. Secondly, TeethPass relies on unique bone-conducted occlusal sounds for user authentication. Thus, how to extract unique biometric features contained in the bone-conducted sounds to characterize the skull biometrics and occlusal location diversity is the second challenge. Finally, it is necessary for an authentication system to collect as little registration data as possible to achieve user-friendliness. Thus, we need to achieve an efficient authentication model with limited registration data.

To address the above challenges, we first analyze occlusion events and typical daily actions. We find that the duration of an occlusion event is much shorter than that of eating and speaking. In addition, the frequency of occlusal sounds ranges from $100Hz$ to $2.5kHz$, while the frequencies of bone-conducted sounds of walking and speaking are mainly below $300Hz$. Thus, occlusion events can be distinguished from daily actions in terms of duration and the ratio between Power Spectral Density (PSD) of different frequency bands. Then, to characterize a user's skull biometrics and occlusion location, we extract 3 biometric features, including the dispersion [22] related to the physical properties of bone and tissue, the acoustic delay related to the occlusal location, and Mel-Frequency Cepstral Coefficients (MFCC) of the bone-conducted sound. Finally, a Siamese network-based [23] authentication scheme is designed for registered users. It is worth mentioning that we apply data augmentation methods (i.e., time warping and time-frequency masking) to the limited training data for improving user experience. Combined with incremental learning [24], we can quickly update the parameters of the Siamese network to authenticate the newly registered user.

We implement TeethPass by using 3 kinds of earphones with inward-facing microphones. We recruit 22 participants (13 males and 9 females) and ask them to put on earphones for occlusion in diverse scenarios. We also simulate different attacks to test the anti-attack ability of our system. The results demonstrate that TeethPass is accurate in different environments, and can resist various spoofing attacks.

Our contributions are summarized as follows:

- We propose a novel authentication system, TeethPass, which uses earbuds to collect bone-conducted sounds of dental occlusion in binaural canals. To the best of our knowledge, we are the first to sense occlusal sounds by earbuds for authentication.
- We propose effective methods to filter out interferences of daily actions. We also design 3 unique biometric features for authentication, including physical features of bone and tissue, location features of occlusion, and integral features of occlusal sounds.
- We build an authentication scheme based on the Siamese network. And we combine the scheme with incremental learning, which can quickly update the parameters of the network for authenticating newly registered users.
- We evaluate TeethPass by using 3 prototypes in different application scenarios. The results show that TeethPass can authenticate users with an average accuracy of 96.8%, and resist 98.9% of spoofing attacks.

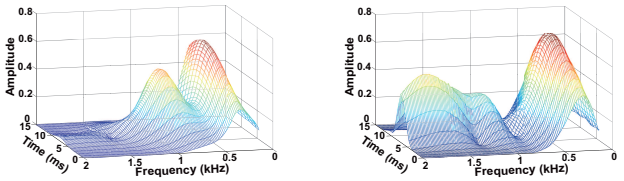
II. RELATED WORK

In this section, we review 3 kinds of biometrics-based authentication systems related to TeethPass.

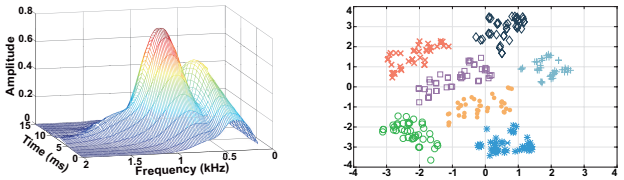
1) Voiceprint-based user authentication: Among various biometric-based authentication methods, voiceprint is one of the most commonly used biometrics for authentication. But traditional voiceprint-based methods [6]–[8] are vulnerable to replay attacks. To improve security, VoicePop [25] leverages pop noises that are produced when a user is breathing and can be hardly maintained in records to achieve authentication. LipPass [16] extracts unique features from users' speaking lips leveraging active acoustic sensing on smartphones. EarPrint [26] aims to extend voiceprint by building on body sounds that transmit from the throat to the ear for authentication. However, these systems require users to speak and are not suitable in some environments (e.g., library and conference room).

2) Teeth-based user authentication: Teeth biometrics, such as size, shape, and edge envelope, are intrinsically unique among individuals [27]. SmileAuth [15] extracts dental edge features by slightly moving the smartphone to capture a few images from different camera angles for authentication. An authentication approach [28] utilizes the contour information of teeth to extract coarse-grained features and further employs voice data to improve the accuracy. But these methods are sensitive to light and vulnerable to replay attacks due to a lack of liveness detection. Most recently, BiLock [29] extracts features from the sounds generated by a user's occlusion, which are recorded by the built-in microphone of a smartphone placed close to the user's lips. However, it only uses air-conducted occlusal sounds for authentication, so its principle is more similar to traditional voiceprint-based authentication, which makes it vulnerable to spoofing attacks and noises.

3) In-ear authentication: Recently, the development of smart earbuds provides a new way for user authentication. EarEcho [17] extracts the features by emitting sounds from the earphone. The sounds are reflected through the ear canal which can be recorded by inward-facing microphones. An authentication system [30] utilizes the microphone-integrated earphone to capture the static ear canal geometry. It extracts features of the reflected signals from a ear canal to distinguish different users. EarDynamic [31] makes earbuds emit an inaudible signal to probe the ear canal. Then the signal reflected from the ear canal are captured by the inward-facing microphone that can be further utilized to extract the deformation of the ear canal. However, most of these methods require earphones to emit ultrasound, and may impair the health of users if users are exposed to the ultrasound with high volume [32].

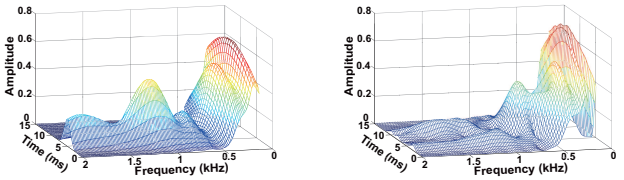


(a) The CWT of user #1. (b) The CWT of user #2.

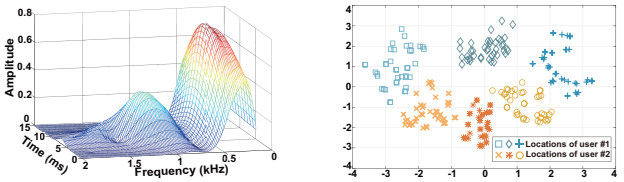


(c) The CWT of user #3. (d) t-SNE visualization.

Fig. 1: Inter-user study.



(a) The CWT of location #1. (b) The CWT of location #2.



(c) The CWT of location #3. (d) t-SNE visualization.

Fig. 2: Intra-user study.

Different from these works, TeethPass uses earbuds to collect bone-conducted sounds of dental occlusion in binaural canals to achieve authentication, which can effectively resist both replay attacks and mimic attacks. Furthermore, the sound of dental occlusion is more imperceptible and unobtrusive than voice, so it is more socially acceptable.

III. PRELIMINARY

In this section, we first introduce attack models and theoretical analysis of occlusal sound, then exploit the feasibility of using bone-conducted sounds of occlusion for authentication.

A. Attack Model

Voiceprint-based and teeth-based authentication systems often suffer from spoofing attacks. Here we list two main types of spoofing attacks for off-the-shelf authentication, i.e., mimic attack and replay attack. We also consider an extreme scenario in which the spoofers implement mimic attacks and replay attacks at the same time to achieve hybrid attacks.

1) Mimic attack: To attack a voiceprint-based authentication, spoofers first observe the way of speaking when legitimate users login, then practice to mimic the tone, speed, and pronunciation to perform the attack. If spoofers attempt to conduct mimic attacks on TeethPass, they first need to know which teeth and how much force users use to occlude when using TeethPass. Then, they can wear users' earbuds and mimic the dental occlusion to spoof TeethPass.

2) Replay attack: The voiceprint-based authentication requires users to make sounds, leading to a high probability that spoofers eavesdrop and record the voice of legitimate users. Then the spoofers can spoof the authentication system by playing back the recorded voice. For our system, spoofers may collect air-conducted sounds of dental occlusion at a location close to the users, and replay them to perform attacks.

3) Hybrid attack: We also consider an extreme situation, that is, spoofers can not only collect the air-conducted sounds of dental occlusion but also know the occlusal location and force of users during authentication. So they can mimic the occlusion of users while playing the recorded occlusal sounds by speakers in spoofers' mouths.

B. Theoretical Analysis of Occlusal Sound

Bone-conducted sounds of dental occlusion have unique biometric features. On the one hand, previous research [33] shows that dental structure is stable over time, and even a single tooth of a person is unique. The uniqueness is caused by the diversities of dental shape, size, and so on. Occlusion [34] refers to the action between the upper and lower teeth when they approach each other, so the occlusal sound has unique features for an individual. On the other hand, the occlusal sound passes through the maxilla, mandible, and zygoma, finally arrives at the auditory meatus. These bones have unique physical features [35], such as shape, bone-muscle ratio, and density, which lead to the unique dispersion, absorption, and refection of occlusal sounds. Given the theoretical analysis of bone-conducted sounds of occlusion, we conduct feasibility experiments to verify the uniqueness of occlusal sounds.

C. Feasibility Study

To verify the feasibility, we collect occlusal sounds of 7 users. Before collection, we explain the principle of TeethPass to them. Then, they are asked to find comfortable occlusal locations and practice occlusion a few times. We use a pair of inward-facing microphones to record bone-conducted sounds. The experiments are conducted in a quiet environment.

1) Inter-user study: Firstly, we ask 3 users to occlude with the same teeth. The Continuous Wavelet Transform (CWT) results of their bone-conducted sounds are shown in Fig. 1(a)(b)(c). The results show noticeable individual differences refer to duration time, frequency range, and energy distribution. Besides, the occlusal sounds collected from all users are visualized in Fig. 1(d) with the t-distributed Stochastic Neighbor Embedding (t-SNE) [36] method. We can find that the occlusal sound of each user shows a unique and consistent pattern, which presents that the individual difference in skull biometrics can be captured by the bone-conducted sounds.

2) Intra-user study: Then, we study the bone-conducted sounds of a user when occluding with different teeth. Fig. 2(a)(b)(c) show the results of CWT on the bone-conducted sounds of three locations (i.e., left, middle, and right teeth).

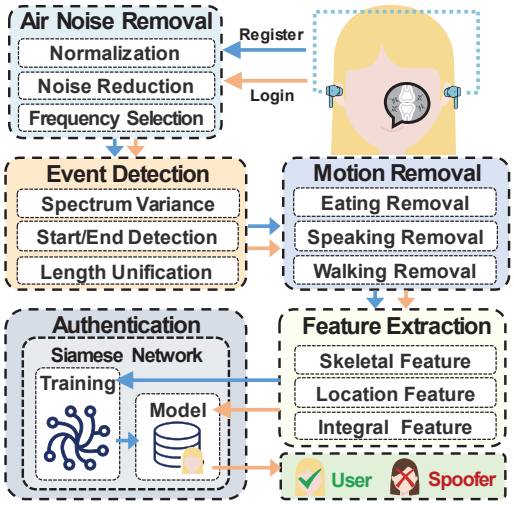


Fig. 3: System architecture of TeethPass.

We can see that when the same user occludes different teeth, the time-frequency domain features are different. Fig. 2(d) shows the bone-conducted sounds after visualization of 2 users at 3 locations. The result demonstrates that different occlusal locations also lead to different occlusal sounds.

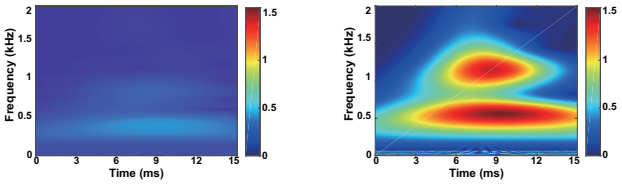
IV. SYSTEM DESIGN

In this section, we first present the system overview of TeethPass and then detail the behind techniques.

A. System Overview

Fig. 3 shows the architecture of TeethPass, which can be divided into two phases, i.e., register phase and login phase.

In the register phase, TeethPass aims to collect data from a user and train the classifier. Before collection, the user needs to find a comfortable occlusal location and practice occlusion a few times. During collection, the user occludes teeth several times, and occlusal sounds are recorded by a pair of inward-facing microphones in the ears. The recorded sounds first go through *Air Noise Removal*, including normalization, noise reduction, and frequency selection, to reduce air-conducted ambient noises. Then, TeethPass performs *Event Detection* to segment each bone-conducted sound event. Specifically, we calculate spectrum variance to capture the energy in different frequency bands of sound, then adopt a double-threshold method to detect the start and end points of each event. Since the sound may reach ears at different times, we unify the length of each event received by the two microphones. Some daily actions (i.e., eating, speaking, and walking) also produce bone-conducted sounds and can be captured by *Event Detection*, so the events that contain these daily actions are removed in *Motion Removal*. We analyze each event's duration and PSD to distinguish occlusion from the daily actions. For each occlusal event, three biometric features are extracted, including the dispersion related to physical properties of bone, the acoustic delay related to occlusal location, and the MFCC related to bone-conducted sound. Finally, the extracted features are used to train a Siamese network in *Authentication*.



(a) Air-conducted sound. (b) Bone-conducted sound.

Fig. 4: Occlusal sound spectrums of air and bone conduction.

In the login phase, TeethPass first records bone-conducted sounds by a pair of inward-facing microphones in the ears. Then, through *Air Noise Removal*, *Event Detection*, and *Motion Removal*, and *Feature Extraction*, the features are sent to the *Authentication* module to determine whether a user is a legitimate user or a spoofer.

B. Air Noise Removal

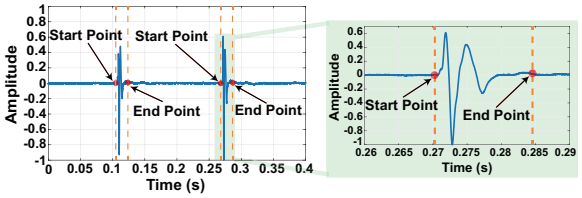
When a user puts on earbuds, TeethPass starts to monitor sounds in ear canals in real-time. Although the inward-facing microphone faces the ear canal, it may still record air-conducted ambient noises, such as human voice and road noise. Thus, the raw recorded sounds need to be processed to filter out these noises. To ensure real-time, we add a sliding window to the sounds. We find that the duration of the bone-conducted occlusal sound is usually between $10ms$ and $20ms$, so the length of the sliding window is $50ms$ and it slides $10ms$ each time. The sound in each window is a frame.

1) Normalization: There may be a slight difference in the tightness and angle of the earbuds each time the user puts on them, so the volume of sounds recorded by inward-facing microphones may be unstable. Before noise reduction, we normalize each frame. The most common method is peak normalization [37]–[39], which adjusts the sounds based on the highest volume level in each frame. However, it leads to the problem that the average volume is inconsistent across frames. So we use another normalization based on loudness, which adjusts the average volume of each frame to a desired volume. We set the desired volume to $-24dB$, which is the same as the standard loudness recommended by ATSC [40].

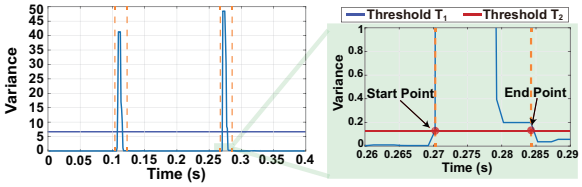
2) Noise reduction: Then, we reduce air-conducted ambient noises in each frame to improve the signal-to-noise ratio. Considering the limited computing capability of mobile devices, we adopt power spectral subtraction which has the advantages of small computation and high processing speed. Suppose the audio signal of a frame is $x(m)$, and $X(k)$ denotes the fast Fourier transform (FFT) results of $x(m)$. The amplitude $|\hat{X}(k)|$ after spectral subtraction can be calculated by

$$\left|\hat{X}\left(k\right)\right|^2 = \begin{cases} |X(k)|^2 - a \times D(k), & |X\left(k\right)|^2 \geqslant a \times D(k), \\ b \times D(k), & |X\left(k\right)|^2 < a \times D(k), \end{cases} \tag{1}$$

where a and b are constants, representing the over-subtraction factor and the spectral floor parameter, respectively. $D(k)$ denotes the amplitude of environmental noise recorded by outward-facing microphones. Through inverse FFT of $|\hat{X}(k)|$, we get the audio signal $\hat{x}(m)$ after noise reduction. It is



(a) The detection results of occlusal sounds.



(b) The corresponding spectrum variance.

Fig. 5: The process of event detection.

obvious that power spectral subtraction can reduce the impact of different noise caused by the different environments.

In addition, we study whether air-conducted sounds of occlusion affect noise reduction. Fig. 4(a) shows the air-conducted occlusal sound received by the outward-facing microphone. We can see that the energy of the sound conducted by air is much lower than that conducted by bone (Fig. 4(b)). Since the air-conducted sounds attenuate faster than bone-conducted sounds [41]. It proves that air-conducted occlusal sounds do not affect noise reduction.

3) Frequency selection: Finally, we use a band-pass filter to select the frequency band of occlusal sounds. Through our observations, we find that most bone-conducted sounds of dental occlusion range from $100Hz$ to $2.5kHz$. Therefore, we adopt a Butterworth band-pass filter ranging from $100Hz$ to $2.5kHz$ for each frame. After filtering, we can eliminate other out-band interferences and prepare for *Event Detection*.

C. Event Detection

After *Air Noise Removal*, the sound in each frame almost only contains bone-conducted sound. Then, we detect and segment each bone-conducted sound event caused by user's actions. A common event detection method is based on Short Time Energy (STE) [24], [42], [43], which is widely used in speech recognition and motion detection. But we find that the energy of bone-conducted sound of dental occlusion varies greatly with frequency, while most other bone-conducted sounds are more evenly distributed in the spectrum. Thus, we divide the spectrum into several bands and study an event detection method based on spectrum variance.

1) Spectrum variance calculation: Firstly, we apply a sliding window with a length of $2.5ms$ that slides $1ms$ each time on each frame. The audio signal in the i-th window is $x_i(m)$, we divide the amplitude $|X_i(m)|$ obtained by the FFT into q sub-bands evenly. And each sub-band is formed as

$$S_i(n) = \sum_{k=1+(n-1)p}^{1+(n-1)p+(p-1)} |X_i(k)|, n \in [1, q], \tag{2}$$

where p is the number of frequency points in each sub-band. Then, the spectrum variance D_i can be calculated by

$$D_i = \frac{1}{q-1} \sum_{k=1}^{q} \left[S_i(k) - \frac{1}{q} \sum_{s=1}^{q} S_i(s) \right]^2. \tag{3}$$

By analyzing Eq. 3, we find that the greater the fluctuation between the frequency bands, the greater the D_i. Fig. 5 shows the bone-conducted sounds of two occlusion and the corresponding spectrum variance. It is shown that spectrum variance can be used to capture the occlusal sounds well.

2) Start/End detection: Then, we adopt a double-threshold method [44] to detect the start and end points of each event. Specifically, we first set a threshold T_1 on spectrum variance, and the segment which is larger than T_1 can be considered to contain an event. Then, another threshold T_2 ($T_2 < T_1$) is set to find the start and end points of the event. We search from the beginning of the segment to the left and find the first point that intersects with T_2 as the start point of the event. In the same way, we search from the ending of the segment to the right to find the end point of the event. Fig. 5 also shows the detection results of start and end points on occlusal sounds and spectrum variance. We can see that the double-threshold method identifies start and end points precisely.

3) Length unification: We collect bone-conducted sounds using a pair of inward-facing microphones and process the sounds of the two microphones separately. Since the user can choose the occlusal location arbitrarily, the occlusal sound conducts to the two ear canals in different paths, which causes the start and end points of the occlusal sounds received by the two microphones to be different. To facilitate *Feature Extraction* later, we unify the length of two events received by the two microphones. Specifically, we choose the smaller one of the two start points as the new start point of the two events, and the larger one of the two end points as the new end point of the two events. After length unification, each occlusal sound produces two events with the same length.

D. Motion Removal

Although we filter out most of the air-conducted ambient noises in the *Air Noise Removal*, some actions also produce bone-conducted sounds (i.e., eating, speaking, and walking) and are extracted from *Event Detection*, so we need to remove these non-occlusion events from the detected events.

1) Eating removal: When users eat, they usually need to use their teeth to chew food. In the process of chewing, the collision and friction between teeth and food can produce bone-conducted sounds. And different foods may lead to different bone-conducted sounds. We experiment with different foods and find that the frequency range of eating is similar to that of occlusion, as shown in Fig. 6. However, the duration of an eating event is generally greater than $250ms$, while the duration of an occlusal event is usually between $10ms$ and $20ms$. So we can determine whether the event is eating or not by analyzing the duration time.

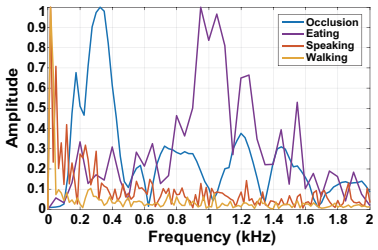


Fig. 6: Spectrum of 4 actions.

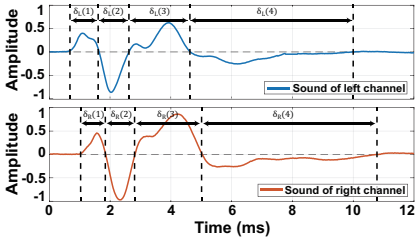


Fig. 7: Dispersion of 2 channels.

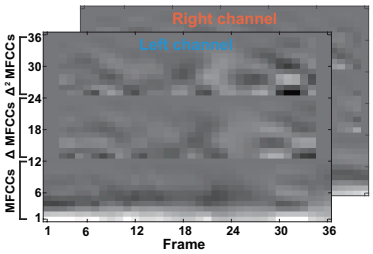


Fig. 8: MFCCs of 2 channels.

2) Speaking removal: The human voice is produced by the vibration of vocal cords, which can also be transmitted to the ear canal through bone conduction. Relevant research [45] and our experiment show that the base frequency of human voice is between $80Hz$ and $300Hz$, while the frequency of occlusion is between $100Hz$ and $2.5kHz$, as shown in Fig. 6. To detect whether the event is speaking, we calculate the energy ratio of the PSD of the $100Hz$-$300Hz$ to the PSD of the $100Hz$-$2.5kHz$. If the energy ratio is greater than a threshold, we consider that the event is speaking. It should be noted that users can not authenticate while eating or speaking, so we just discards these two events when they are detected.

3) Walking removal: Authentication while walking is a common scenario, and we find that the inward-facing microphone receives a noticeable sound as soon as the heel touches the ground. So we need to filter out bone-conducted sounds of walking from the recorded sounds. Fig. 6 shows the spectrum of the bone-conducted sound of walking, we can see that the frequency of the sound produced by heel landing is mainly concentrated below $100Hz$. The frequency difference between walking and occlusion is mainly caused by the different propagation paths. The sound produced by the heel landing needs to travel through the entire body to reach the ear canal, while the occlusal sound can reach the ear canal through a very short path. To remove the interference of walking, we adopt a band-pass filter ranging from $100Hz$ to $2.5kHz$ in *Air Noise Removal*. Thus, before the event detection, we can filter out bone-conducted sounds of walking.

E. Feature Extraction

To accurately authenticate users and resist spoofing attacks, it is necessary to extract reliable biometric features from bone-conducted sounds of dental occlusion. In this section, we present approaches for extracting three features from the skeletal structure, occlusal location, and occlusal sound.

1) Skeletal feature: When the upper and lower teeth collide with each other, surface acoustic waves (SAWs) are generated. SAWs can travel along the surface of teeth and bone. Relevant research [46] shows that bone is a dispersive medium, which means that the speed of SAWs is related to the frequency of SAWs and physical properties of the bone (e.g., density, elastic, and inertia properties). The speed of the high-frequency part of the SAWs is faster than that of the low-frequency part, which causes that the SAW spreads out and changes shape as it travels. Based on this, we try to extract the dispersion of occlusal sound as the skeletal feature.

Fig. 7 shows the events of a dental occlusion detected by two inward-facing microphones. We can see that the high-frequency sound first reaches the microphones, then the low-frequency sound, so we calculate the distance $\delta_L(i)$ (and $\delta_R(i)$) between two zero-crossing points. The zero-crossing sequence of left channel is defined as $ZS_L = [\delta_L(1), \delta_L(2), \cdots, \delta_L(n)]$. The zero-crossing sequence of right channel ZS_R is similar. We regard ZS_L and ZS_R as the skeletal features related to the physical properties of bones.

2) Location feature: Users can choose any location of teeth to register, so the paths and times of occlusal sounds from different locations to the two microphones are also different. We analyze the delay between the two occlusal sounds received by two microphones as the location feature. However, different paths may cause different dispersion of occlusal sounds, which makes the delay calculation based on cross-correlation not accurate enough. To solve this problem, we first divide the sound into 5 frequency bands and then compute cross-correlation $R_{l,r}(i)$ for each band. Finally, we get the cross-correlation sequence $R_{l,r} = [R_{l,r}(1), R_{l,r}(2), \cdots, R_{l,r}(5)]$ as the location feature.

3) Integral feature: Finally, we extract the integral feature contained in the occlusal sound. In recent years, MFCC [37] is commonly applied as the sound feature in speech recognition and sound classification. Thus, we calculate MFCC for each occlusal event. Before extracting MFCC, the occlusal sound is first cut into 36 frames with overlap. Then, we extract 12-dimensional MFCCs, 12-dimensional first-order derivatives (Δ MFCCs), and 12-dimensional second-order derivatives (Δ^2 MFCCs) from each frame. We combine the 36-dimensional features of 36 frames to form a 36×36 bicolor image. Due to the use of two microphones, we can get two bicolor images as shown in Fig. 8. Finally, we combine the two images into a two-channel image with size of $36 \times 36 \times 2$.

F. Authentication

After getting three biometric features, we design two methods for user authentication. First, we classify the integral features. Traditional classifiers (e.g., SVM, RF, and DNN) usually need a large number of positive and negative data to train classifiers, so they are not suitable for the single-user situation. To solve this problem, TeethPass leverages a Siamese network [47] as the classifier, which is especially suitable for

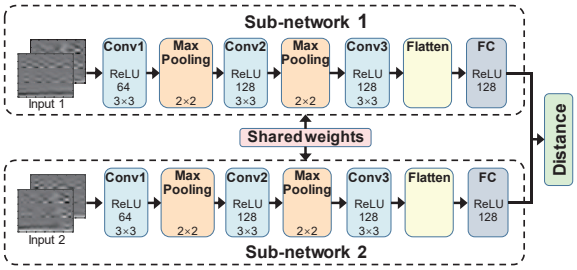


Fig. 9: Structure of the Siamese network.

(a) Prototype #1 (b) Prototype #2 (c) Prototype #3

Fig. 10: Prototype earphones.

solving the classification problem with unknown numbers of classes and few training data. The basic idea of the Siamese network is to use a pair of networks with the same structure and weights to compute a similarity for two inputs.

Fig. 9 shows the structure of the Siamese network, which has two identical sub-networks. Each sub-network mainly consists of three convolutional (Conv) layers, two max pooling layers, and a fully connected (FC) layer. Given a pair of integral features as the inputs, the Siamese network can extract user identity information through two identical sub-networks and compute the distance of user identity information as the similarity of the inputs. In the training phase, suppose the weights of the sub-network are W, then the loss function is

$$L(W) = \sum_{i=1}^{N} Y(D_W^i)^2 + (1-Y)\max(M - D_W^i, 0)^2, \quad (4)$$

where D_W^i denotes the Euclid distance of the i-th pair of input features. M is the margin that represents the decreased interval. If the input features are from the same user, then $Y = 1$, otherwise, $Y = 0$. The network is trained to minimize the loss $L(W)$. In other words, we try to minimize the distance between the features of the same user and maximize the distance between the features of different users.

We first pre-train the network. We ask 4 volunteers to collect bone-conducted sounds of occlusion. Then, any two integral features form a pair of inputs which are sent to the network for pre-training. After that, the network has the preliminary ability to distinguish integral features from different users. We deploy the network on a mobile device, when a new user registers the device, the user is required to perform occlusion several times to extract the integral features. To reduce the time of occlusion, we augment the integral features by using time-warping and frequency-time masking [26], [48]. Then, the user's integral features are combined with the integral features of himself/herself, the 4 volunteers, and other registered users (if any) to form new pairs of inputs. To reduce training cost, we borrow ideas from incremental learning [24] to make the network authenticate the new user. Specifically, instead of retraining the network completely, we continue to train on the existing network using the new pairs of inputs. In the login phase, the network compares the similarity between the received integral feature and each registered user's integral feature on the device. If the similarities are all less than a threshold, we consider there is a spoofer, otherwise, the identity is assigned to the registered user with the highest similarity.

In the registration phase, we also collect skeletal features of the new user and average zero-crossing sequences to get $\tilde{ZS}_L$ and $\tilde{ZS}_R$. In the login phase, we calculate zero-crossing difference between ZS_L of the received skeletal feature and $\tilde{ZS}_L$ of the registered user assigned by the network as $DS_L = \frac{1}{n}\sum_{k=1}^{n}[\delta_L(k) - \tilde{\delta}_L(k)]$. Then we get DS_R of the right channel in the same way. For the location feature, we use the same method to calculate cross-correlation difference $DR_{L,R}$. If any two of DS_L, DS_R, and $DR_{L,R}$ are less than corresponding thresholds, the user is considered legitimate.

V. IMPLEMENTATION AND EVALUATION

In this section, we introduce the implementation details and provide the evaluation results.

A. Experiment Setup

There are several commercial earbuds equipped with inward-facing microphones. But due to hardware limitations, we can not get sound data from inward-facing microphones. Thus, we implement TeethPass by attaching a microphone in front of the speaker in an earphone, which is similar to most commercial earbuds. We design 3 prototypes, as shown in Fig. 10. We recruit 22 participants (13 males and 9 females, aged from 18 to 52), 15 of them register TeethPass as legitimate users, and the rest 7 participants are spoofers. Each legitimate user chooses a comfortable prototype earphone and occlusal location. The occlusal sounds are recorded in 4 environments, including lab, park, car, and mall. Finally, we collect more than 2,000 occlusal sounds for legitimate users. Part of the data collected in the lab are used to train the network, and the remaining data are used as the test set. Spoofers perform mimic, replay, and hybrid attacks on TeethPass. All procedures are approved by the Institutional Review Board (IRB) at our institute.

B. Evaluation Methodology

We mainly evaluate TeethPass from the following aspects.

Confusion matrix. Each row and each column of the matrix represent the ground truth and the authentication result, respectively. Each entry represents the percentage of a user that is classified into each identity.

False reject rate (FRR). The probability that TeethPass authenticates a legitimate user as a spoofer.

False accept rate (FAR). The probability that TeethPass authenticates a spoofer as a legitimate user.

Fig. 11: Confusion matrix of TeethPass.

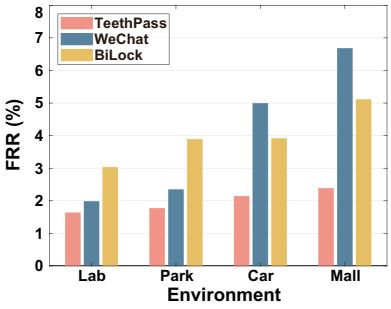


Fig. 12: FRR of TeethPass, WeChat, and BiLock under 4 environments.

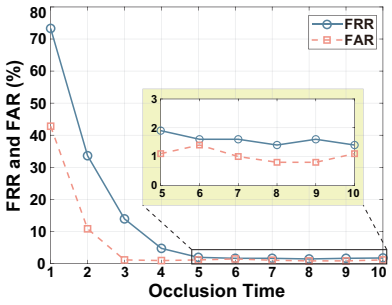


Fig. 13: FRR and FAR under different occlusion times for registration.

TABLE I: Response time during an authentication.

Phase	ANR	ED	MR	FE	Train	Auth	Total
Reg(s)	0.3	0.2	0.1	0.2	3	None	3.8
Login(ms)	58	45	28	54	None	92	277

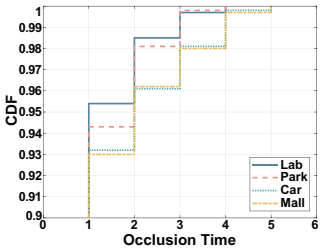


Fig. 14: CDF of occlusion times for login.

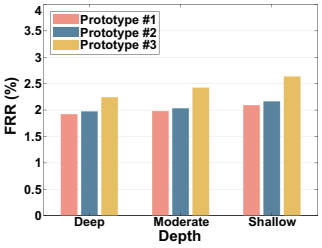


Fig. 15: FRR under different wearing depths.

C. Overall Performance

We first evaluate the overall performance of TeethPass. Fig. 11 shows the confusion matrix of 15 legitimate users (denoted as $U1$, $U2$, $\cdots$, $U15$) and 7 spoofers (denoted as SP). It shows that TeethPass achieves an average accuracy of 96.8% for users authentication and 98.9% for the detection of 3 types of attacks. Among the legitimate users, the lowest accuracy is 90.8% (user #5). The results indicate that TeethPass can accurately authenticate legitimate users and detect spoofers.

We compare the performance of TeethPass with that of WeChat voiceprint lock [49] and BiLock [29]. Fig. 12 shows the FRR of the 3 systems in 4 environments. We can see that the FRR of TeethPass is 1.6%, which is slightly better than 1.9% and 3.1% of WeChat and BiLock in the lab. But in cars and malls with loud ambient noises, the FRR of WeChat and BiLock increase over 5.0% and 3.9% respectively, while TeethPass keeps a stable FRR with a slight increase to 2.3%. With the help of noise reduction methods, TeethPass has stable performance in various environments. Another reason is that the shell of the earphone also helps to isolate ambient noises.

D. Performance on User Experience

1) Occlusion time for registration: In the register phase, more times of occlusion can improve the effect of training the network. But too many times of occlusion may lead to a poor user experience. Hence, we evaluate the FRR and FAR of TeethPass under different occlusion times for registration, and the results are shown in Fig. 13. It is obvious that with the occlusion time increases, the FRR and FAR of TeethPass decrease sharply at the beginning. And TeethPass only needs 5 occlusion to achieve 1.9% FRR and 1.1% FAR, which is mainly because we design a suitable Siamese network and adopt data augmentation. In order to balance performance and usability, we fix the occlusion times to 5 in all the evaluations.

2) Occlusion time for successful login: In the login phase, we evaluate the occlusion times required for successful authentication under 4 environments. Fig. 14 shows CDF of the occlusion times. We can see that more than 93% of login operations are successful with only one occlusion under each environment. And an average of 98.8% of login operations can be successfully authenticated within 4 occlusion, which is acceptable for users. In some special environments, such as a very loud sound source near the user, TeethPass requires 5 or more occlusion to successfully authenticate the user. Thus, when there occurs 5 consecutive times of unsuccessful authentication, the device is automatically locked for a while.

3) Authentication response time: Then we study the response time of TeethPass from receiving an occlusal sound to producing the authentication result. During the evaluation, all the registration and login data are transmitted to a PC with a $3.2GHz$ Intel i7 CPU and $16GB$ memory. Table I shows the average response times of *Air Noise Removal* (ANR), *Event Detection* (ED), *Motion Removal* (MR), *Feature Extraction* (FE), *Network Training* (Train), and *Authentication* (Auth). In the registration phase, the training of the network takes the most time, and the total response time for each new user to register TeethPass is about $3.8s$. But response time of registration has little impact on user experience. In the login phase, TeethPass can produce authentication results within $280ms$ after the user completes the occlusion, which indicates that TeethPass can achieve a satisfactory user experience.

4) Earphone wearing depth: The types and depths of earphones that each user is accustomed to wearing are different, so we study the impact of different earphone types and depths on system performance. The occlusal sounds are collected at different in-ear depths, including deep, moderate, and shallow positions. It can be seen from Fig. 15 that prototype #1 has the lowest FRR at all three depths since it has the most stable wearing way. The FRR of prototype #2 increases slightly at

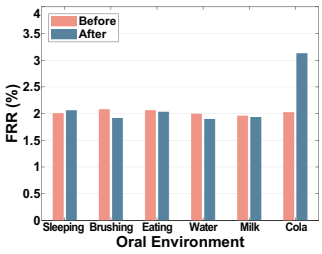


Fig. 16: FRR under different oral environments.

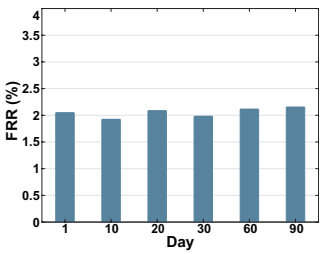


Fig. 17: FRR over different time of period.

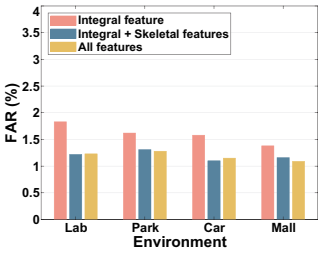


Fig. 18: FAR under mimic attacks.

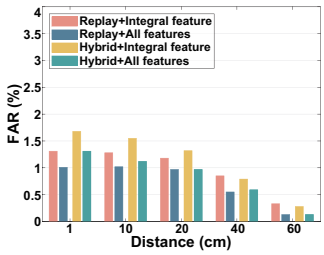


Fig. 19: FAR under replay and hybrid attacks.

shallow depth, but its in-ear structure can ensure that the FRR is less than 2.2%. Prototype #3 adopts a semi-in-ear structure, so its wearing stability and sound insulation are not good, causing FRR to rise to about 2.6% at shallow depth. In general, the type of earphone has a certain impact, while the wearing depth of the earphone has little effect on performance.

5) Oral environment: We also evaluate the impact of different oral environments on system performance. We ask users to use TeethPass before and after sleeping, brushing teeth, eating, drinking water, drinking milk, and drinking cola, respectively. Fig. 16 shows that the FRR after sleeping, eating, and drinking milk increases slightly, while brushing teeth and drinking water can reduce the FRR. In particular, we find that drinking cola has a great impact on system performance, the FRR after drinking cola increases to 3.2%. The reason may be that the ingredients with high viscosity (e.g., caramel color and syrup) in cola adhere to the surface of teeth, resulting in the change of biometric features of the occlusion.

In addition, the oral environment may change over time. To evaluate the robustness of TeethPass over time, we collect occlusal sounds of the users for more than 3 months. Fig. 17 shows the authentication FRR of 1, 10, 20, 30, 60, and 90 days after registration. We can observe that TeethPass can maintain high performance over a long time. Specifically, the FRR of TeethPass is still less than 2.3% after 90 days. It is worth noticing that two users extract wisdom teeth during the experiment, and the FRR after teeth extraction does not increase significantly.

E. Performance on Attack Resistance

In order to prove that TeethPass can resist attacks described in Section III-A, we conduct several experiments to verify the effectiveness of the three biometric features under the mimic attack, replay attack, and hybrid attack. Spoofers use the same prototype earphone as the legitimate user to attack TeethPass.

1) Mimic attack: To conduct mimic attacks, we assume that spoofers know which teeth and how much force the legitimate user uses to occlude. Fig. 18 shows the FAR of mimic attacks under different environments. It shows that the FAR is over 1.8% in the lab when using only integral features. But when two or all features are used, the average FAR is stable at about 1.2%, which indicates that TeethPass using all three features can resist mimic attacks well. The reason is that although the spoofer can mimic the location and force of occlusion of the legitimate user, the biometric features of the spoofer's teeth and bones are still different from those of the legitimate user.

2) Replay attack: To conduct replay attacks, we place a microphone at different distances to the user's mouth to eavesdrop on the air-conducted sound of dental occlusion and then replay it to the prototype earphone to attack. Fig. 19 depicts the results of replay attacks at different eavesdropping distances. It is obvious that when the distance is greater than $20cm$, the FAR is reduced to 1% by using only integral features, since the air-conducted sounds of occlusion are more close to impulse waves and decay fast. In addition, due to the fact that the air-conducted sound does not contain skeletal and location features, the FAR of TeethPass using all features is lower than that using only integral features.

3) Hybrid attack: Finally, we consider that the spoofer mimics the occlusion of the user while playing the eavesdropped occlusal sound by a speaker in the spoofer's mouth. The setting of eavesdropping is the same as that in replay attacks. Fig. 19 shows the results of hybrid attacks at different eavesdropping distances. When the distance is greater than $20cm$, the FAR is reduced to about 1%. But at smaller distances, the FAR of hybrid attacks is higher than that of replay attacks. In actual scenarios, it is difficult to eavesdrop on users in such a short distance. Generally, TeethPass can resist various attacks effectively in different environments.

VI. CONCLUSION AND FUTURE WORK

In this paper, we design and implement TeethPass, which uses inward-facing microphones in earbuds to collect bone-conducted sounds of dental occlusion in binaural canals to achieve authentication. We present effective methods to filter out interferences of ambient noises and daily actions. We extract biometric features, including physical features of bone and tissue, location features of occlusion, and integral features of occlusal sound, then adopt a Siamese network based on incremental learning as the classifier. The extensive experiments show that it achieves an average authentication accuracy of 96.8%, and resists 98.9% of spoofing attacks.

The experiments are conducted without other music being played by earphones. In the future, we will try to work around this limitation by analyzing the correlation between the played and received sounds. We will further integrate TeethPass with other existing authentication methods to provide users with all-round authentication services.

REFERENCES

[1] Cisco, "2020 Cisco consumer privacy survey," https://www.freevacy.com/news/cisco/2020-cisco-consumer-privacy-survey/894, 2020.

[2] IBM, "2020 Cost of a data breach report," https://securityintelligence.com/posts/whats-new-2020-cost-of-a-data-breach-report/, 2020.

[3] J. Yu, L. Lu, Y. Chen, Y. Zhu, and Linghe Kong, "An Indirect Eavesdropping Attack of Keystrokes on Touch Screen through Acoustic Sensing," *IEEE TMC*, vol. 20, no. 2, pp. 337–351, 2021.

[4] N. K. Ratha, V. D. Pandit, R. M. Bolle, and V. Vaish, "Robust fingerprint authentication using local structural similarity," in *IEEE WACV*, 2000, pp. 29–34.

[5] K. K. M. Shreyas, S. Rajeev, K. Panetta, and S. S. Agaian, "Fingerprint authentication using geometric features," in *IEEE THS*, 2017, pp. 1–7.

[6] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End text-dependent speaker verification," in *IEEE ICASSP*, 2016, pp. 5115–5119.

[7] R. G. M. M. Jayamaha, M. R. R. Senadheera, T. N. C. Gamage, K. D. P. B. Weerasekara, G. A. Dissanayaka, and G. N. Kodagoda, "VoizLock – human voice authentication system using hidden markov model," in *IEEE ICIAFS*, 2008, pp. 330–335.

[8] P. L. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE TASL*, vol. 20, no. 8, pp. 2280–2290, 2012.

[9] I. Song, H. J. Kim,and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *IEEE ICCE*, 2014, pp. 564–567.

[10] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *IEEE SIBGRAPI*, 2018, pp. 471–478.

[11] Apple, "Use touch ID on iPhone and iPad," https://support.apple.com/en-us/HT201371, 2020.

[12] TD Bank, "TD voicePrint," https://www.td.com/privacy-and-security/privacy-and-security/how-we-protect-you/weprotect.jsp, 2021.

[13] Amazon, "Amazon Rekognition," https://aws.amazon.com/cn/rekognition/, 2017.

[14] H. Kong, L. Lu, J. Yu, Y. Chen, and F. Tang, "Continuous authentication through finger gesture interaction for smart homes using WiFi," *IEEE TMC*, vol. 20, no. 11, pp. 3184–3162, 2021.

[15] H. Jiang, H. Cao, D. Liu, J. Xiong, and Z. Cao, "SmileAuth: Using dental edge biometrics for user authentication on smartphones," *ACM IMWUT*, vol. 4, no. 3, pp. 84:1–84:24, 2020.

[16] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM*, 2018, pp. 1466–1474.

[17] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "EarEcho: Using ear canal echo for wearable authentication," *ACM IMWUT*, vol. 3, no. 3, pp. 81:1–81:24, 2019.

[18] M. Li, I. Cohen, and S. Mousazadeh, "Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones," in *IEEE ChinaSIP*, 2014, pp. 1–5.

[19] B. Huang, Y. Gong, J. Sun, and Y. Shen, "A wearable bone-conducted speech enhancement system for strong background noises," in *IEEE ICEPT*, 2017, pp. 1682–1684.

[20] A. Shahina and B. Yegnanarayana, "Language identification in noisy environments using throat microphone signals," in *IEEE ICISIP*, 2005, pp. 400–403.

[21] Canalys, "Global smart accessories forecast 2021," https://www.canalys.com/newsroom/global-smart-accessories-market-2021-forecast, 2021.

[22] H. Kim, A. Byanjankar, Y. Liu, Y. Shu, and I. Shin, "UbiTap: Leveraging acoustic dispersion for ubiquitous touch interface on solid surfaces," in *ACM SenSys*, 2018, pp. 211–223.

[23] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *IEEE CVPR*, 2014, pp. 1701–1708.

[24] Y. Xie, F. Li, Y. Wu, and Y. Wang, "HearFit: Fitness monitoring on Smart speakers via active acoustic sensing," in *IEEE INFOCOM*, 2021, pp. 1–10.

[25] 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 *IEEE INFOCOM*, 2019, pp. 2062–2070.

[26] Y. Gao, Y. Jin, J. Chauhan, S. Choi, J. Li, and Z. Jin, "Voice In Ear: Spoofing-resistant and passphrase-independent body sound authentication," *ACM IMWUT*, vol. 5, no. 1, pp. 12:1–12:25, 2021.

[27] M. P. G. Salazara and J. R. Gasgaa, "Microhardness and chemical composition of human tooth," *Materials Research*, vol. 6, no. 3, pp. 367–373, 2003.

[28] D. S. Kim, K. W. Chung, and K. S. Hong, "Person authentication using face, teeth and voice modalities for mobile device security," *IEEE TCE*, vol. 56, no. 4, pp. 2678–2685, 2010.

[29] Y. Zou, M. Zhao, Z. Zhou, J. Lin, M. Li, and K. Wu, "BiLock: User authentication via dental occlusion biometrics," *ACM IMWUT*, vol. 2, no. 3, pp. 152:1–152:20, 2018.

[30] T. Arakawa, T. Koshinaka, S. Yano, H. Irisawa, R. Miyahara, and H. Imaoka, 'Fast and accurate personal authentication using ear acoustics," in *IEEE APSIPA*, 2016, pp. 1–4.

[31] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "EarDynamic: An ear canal deformation based continuous user authentication using in-ear wearables," *ACM IMWUT*, vol. 5, no. 1, pp. 39:1–39:27, 2021.

[32] T. G. Leighton, "Are some people suffering as a result of increasing mass exposure of the public to ultrasound in air?," *Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 472, no. 2185, pp. 1–57, 2016.

[33] K. Krishan, T. Kanchan, and A. K. Garg, "Dental evidence in forensic identification – an overview, methodology and present status," *Open Dentistry Journal*, vol. 9, no. 1, pp. 250–256, 2015.

[34] S. Davies and R. M. J. Gray, "What is occlusion?," *British dental journal*, vol. 191, no. 5, pp. 235–245, 2001.

[35] W. E. Siri, "The gross composition of the body," *Elsevier Advances in biological and medical physics*, vol. 4, pp. 239–280, 1956.

[36] Y. Xie, F. Li, Y. Wu, and Y. Wang, "HearFit$^+$: Personalized fitness monitoring via audio signals on smart speakers," *IEEE TMC*, pp. 1–1, 2021.

[37] Y. Wu, F. Li, Y. Xie, S. Yang, and Y. Wang, "HDSpeed: Hybrid detection of vehicle speed via acoustic sensing on smartphones," *IEEE TMC*, pp. 1–1, 2020.

[38] M. Chen, P. Yang, J. Xiong, M. Zhang, Y. Lee, C. Xiang, and C. Tian, "Your table can be an input panel: Acoustic-based device-free interaction recognition," *ACM IMWUT*, vol. 3, no. 1, pp. 3:1–3:21, 2019.

[39] X. Xu, J. Yu, Y. Chen, Q. Hua, Y. Zhu, Y. Chen, and M. Li, "TouchPass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *ACM MobiCom*, 2020, pp. 1–13.

[40] Advanced Television Systems Committee, "Techniques for establishing and maintaining audio loudness for digital television," https://www.atsc.org/atsc-documents/type/1-0-standards/, 2013.

[41] M. O. Culjat, D. Goldenberg, P. Tewari, and R. S. Singh, "A review of tissue substitutes for ultrasound imaging," *Ultrasound in medicine & biology*, vol. 36, no. 6, pp. 861–843, 2010.

[42] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *IEEE TAEECE*, 2013, pp. 208–212.

[43] X. Guo, J. Liu, and Y. Chen, "FitCoach: Virtual fitness coach empowered by wearable mobile devices," in *IEEE INFOCOM*, 2017, pp. 1–9.

[44] J. Vartiainen, J. J. Lehtomaki, and H. Saarnisaari, "Double-threshold based narrowband signal extraction," in *IEEE VETECS*, 2005, pp. 1–5.

[45] S. Lee, J. Kim, I. Yun, G. Y. Bae, D. Kim, S. Park, I. Yi, W. Moon, Y. Chung, and K. Cho, "An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition," *Nature Communications*, vol. 10, no. 1, pp. 1–11, 2019.

[46] A. M. Abduljabbar, M. E. Yavuz, F. Costen, R. Himeno, and H. Yokota, "Frequency dispersion compensation through variable window utilization in time-reversal techniques for electromagnetic waves," *IEEE TAP*, vol. 64, no. 8, pp. 3636–3639, 2016.

[47] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE VETECS*, 2005, pp. 1–8.

[48] Y. Xie, F. Li, Y. Wu, S. Yang and Y. Wang, "Real-time detection for drowsy driving via acoustic sensing on smartphones," *IEEE TMC*, vol. 20, no. 8, pp. 2671–2685, 2021.

[49] WeChat, "Voiceprint: The New WeChat Password," https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/, 2015.