Listen to Your Face: A Face Authentication Scheme based on Acoustic Signals

HUIMIN CHEN, Zhejiang University, China
CHAOJIE GU, Zhejiang University, China
LILIN XU, Zhejiang University, China
RUI TAN, Nanyang Technological University, Singapore
SHIBO HE, Zhejiang University, China
JIMING CHEN, Zhejiang University, China

Face authentication (FA) schemes are widely adopted in smart homes nowadays. However, existing FA systems for smart appliances are commonly camera-based and hence experience performance degradation in poor illumination conditions. Mainstream FA systems based on radio frequency require dedicated hardware that is inaccessible to many appliances. In this paper, we propose an acoustic signals-based FA scheme that extracts acoustic signal features associated with facial 3D geometries to achieve FA named *SoundFace*. This scheme can be widely deployed on most appliances in home environments. We propose a novel two-stage locating approach based on acoustic sensing to capture the signal variation of the user's face and separate the face region echoes from multipath interferences in the distance dimension. To obtain distinguishable facial features, we design a Convolutional Neural Network (CNN)-based feature extractor. In addition, the acoustic signal is highly susceptible to different changes in practical authentication. To overcome it, we utilize a transfer learning technique with little training overhead to enable SoundFace resilient to various authentication changes. Extensive evaluations demonstrate that SoundFace achieves an average true authentication rate of over 96.2% and an equal error rate of 4.2%, and it is robust to various real-world settings.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Acoustic Sensing, Face Authentication, Smart Home

1 INTRODUCTION

With the development of the Internet of Things (IoT), smart homes have widely spread in consumers' daily lives. According to the latest report [30], the smart household penetration rate worldwide has achieved 16.4% in 2023, and is expected to hit 33.2% by 2028. Within a smart home ecosystem, smart household appliances are experiencing a surge in demand due to their integration with various features that significantly enhance convenience, energy efficiency, and overall functionality. Nowadays, many smart household appliances store sensitive information, such as personal interests, hygiene habits, and health status, to facilitate customized services for the residents. However, the potential leakage of such information poses a risk of unauthorized access to personal data and lifestyles. Therefore, ensuring secure user authentication for smart home appliances is important.

In recent years, user authentication (UA) technology has been extensively applied in the smart home, including traditional knowledge-based methods (e.g., passwords, patterns [32]), and biometric-based approaches (e.g., fingerprint [22], face [29], voice [9] and gait [24]). These UA systems employ diverse sensing modalities, including vision, mmWave, and acoustic, to capture biometrics. Notably, acoustic-based UA systems have garnered increasing interest

[0]Part of this work was completed when Huimin Chen was visiting Nanyang Technological University.

Authors' Contact Information: Huimin Chen, bethanychm@zju.edu.cn, Zhejiang University, Hangzhou, China; Chaojie Gu, gucj@zju.edu.cn, Zhejiang University, Hangzhou, China; Lilin Xu, lilinxu@zju.edu.cn, Zhejiang University, Hangzhou, China; Rui Tan, tanrui@ntu.edu.sg, Nanyang Technological University, Singapore; Shibo He, s18he@zju.edu.cn, Zhejiang University, Hangzhou, China; Jiming Chen, cjm@zju.edu.cn, Zhejiang University, Hangzhou, China.

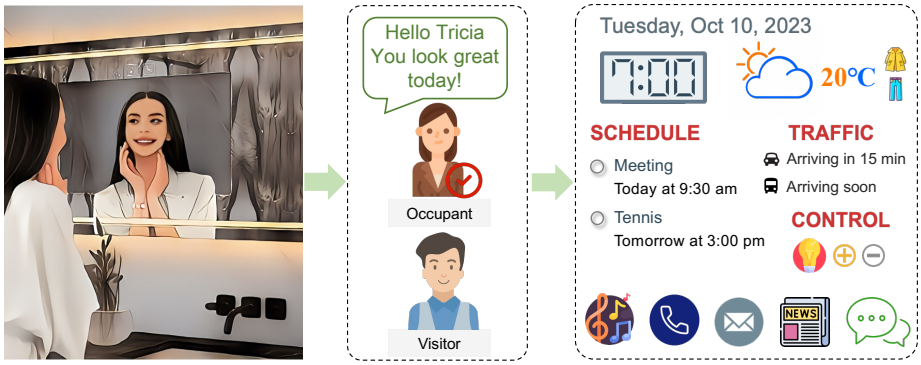


Fig. 1. An application scenario for SoundFace. In a home, the user can access various personalized services of the smart mirror after face authentication with the speaker and a pair of microphones built inside the mirror.[1]

due to their capability to operate without dedicated devices, while preserving most of the advantages associated with Radio Frequency (RF) based UA systems [39–41]. These advantages include insensitivity to lighting conditions [8] and sensitivity to materials encountered during signal propagation [34], which make the system resilient to complex lighting conditions and potential spoofing attacks, respectively. Moreover, with the development in voice recognition, acoustic components (i.e., speakers and microphones) are widely available in smart homes, such as smart speakers [2], smart TVs [28], and thermostats [1]. As a result, various acoustic-based UA approaches are emerging by extracting biometric features from the ear canal [7, 31, 37], dental occlusion [36, 38], face structure [4, 15, 16, 46].

However, it is worth noting that existing acoustic UA systems have not addressed a common application scenario. Since most smart home appliances/devices equipped with speakers and microphones are usually stationary, the challenge lies in enabling these appliances to reap the benefits of an acoustic UA system or leveraging acoustic technology to enhance multi-modal UA. To this end, in this paper, we propose *SoundFace*, an acoustic signals-based face authentication (FA) scheme that can be easily implemented on most smart home appliances.

Specifically, SoundFace uses a speaker to transmit acoustic signals towards the user's face and a pair of microphones to collect the acoustic signals reflected from the human face. SoundFace performs UA by extracting the features that depict the personalized differences in 3D facial geometry. Fig. 1 shows a promising application scenario for SoundFace. When a person approaches the mirror and looks up directly at it, the smart mirror authenticates the person's face by using acoustic signals, and then instructs the mirror to list personalized services such as message display, music playback, fitness and health monitoring, and augmented reality-based makeup.

However, it is not trivial to design SoundFace and several challenges need to be tackled: (1) *Combat multipath interferences.* When we apply acoustic signals for face authentication, apart from the reflection signals from the target face, there are other signals as interferences, including stronger signals reflected from surrounding objects and the most potent direct path signal traveling from the speaker to the microphone. Thus, it is important to separate the face-reflection signals from other interfering reflections to preserve the 3D feature information of the user's face. (2) *Distinguishable facial feature extraction.* Face-reflection acoustic signals are very weak due to the absorption and reflection by the skin, which makes it extremely difficult to find a representative feature to depict the differences in individuals' faces for authentication. As a result, it is critical to extract effective face features from the echo signal for

[1]The left image is generated from the original image [23] animated by AIGC.

downstream authentication tasks. (3) *Robust authentication.* In real-life application scenarios, due to the sensitivity of acoustic signals, changes in authentication environments, newly emerging individuals, authentication distance variation caused by users' unconscious position movements, and ambient noises may lead to undesirable degradation of authentication accuracy and robustness. Consequently, the system cannot achieve consistent robust authentication performance when came to these practical factors.

To address the first challenge, we propose a two-stage locating approach. In the first stage, we utilize the pilot signal variations contributed by the user's head-up at the beginning to conduct multi-granularity acoustic sensing for major face localization. In the second stage, given the major face localization result, we utilize cross-correlation on the pulse chirps to estimate face distances with higher accuracy, then further segment the face region echoes from other interferences.

For the second challenge, we explore a feature that can be used for distinguishing different users' faces. The segmented face echoes are composed of reflected signals from various face surfaces, resulting in unique combinations of echoes based on each individual's facial structure. We first apply a Short-Time Fourier Transform (STFT) to describe the weak segmented signals both in terms of the raw acoustic signal and the converted mixed signal after chirp mixing. We then utilize a Convolutional Neural Network (CNN)-based feature extractor to capture the high-level spatial characteristics of human faces and then classify them.

Regarding the various changes in authentication, such as the changes in the authentication environment, we adopt a transfer learning technique to fine-tune the trained feature extractor and the support vector machine (SVM) model when dealing with different authentication environments. To cope with the distance changes, we first estimate the user's distance from the mirror using the two-stage locating approach, then fine-tune the pre-trained models using a small amount of data, typically 3 s data per person, collected at this distance. When there are new individuals in the application scenario, if they are legitimate new users, we retrain the SVM model using their data with a low training overhead. Conversely, if they are not recognized as legitimate users, the trained One-Class Support Vector Machine (OC-SVM) model is used for authentication directly.

SoundFace is implemented on a research-purpose hardware platform Bela with a Commercial off-the-shelf (COTS) speaker and Micro-Electro-Mechanical Systems (MEMS) microphones. We conduct extensive experiments to evaluate SoundFace's performance. The results show that SoundFace can achieve over 96.2% true accept rate and around 4.2% equal error rate. Our contributions can be summarized as follows:

- We develop SoundFace, a novel acoustic-based FA system, that can be easily implemented on most appliances in home environments with low hardware overhead. It extracts the 3D facial geometry features of users' faces from acoustic signals to achieve FA.
- We propose a two-stage acoustic sensing pipeline with fine-grained face localization to capture face region echoes from severe multipath interferences. Besides, we apply data augmentation for the face echoes and transfer learning for the feature extractor model to enhance the robustness of the system.
- We build a prototype of SoundFace with COTS acoustic devices on a research platform. Extensive performance evaluations demonstrate the ability of SoundFace for robust FA authentication.

The remainder of this paper is organized as follows. §2 reviews the related work. §3 introduces the design details of SoundFace. Then, §4 presents the evaluation results in different scenarios and §5 discusses some related issues. Finally, §6 concludes the paper.

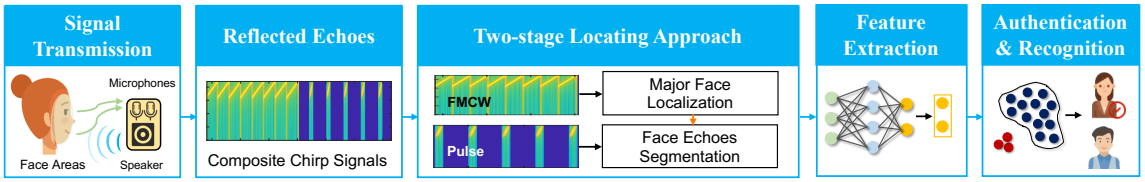
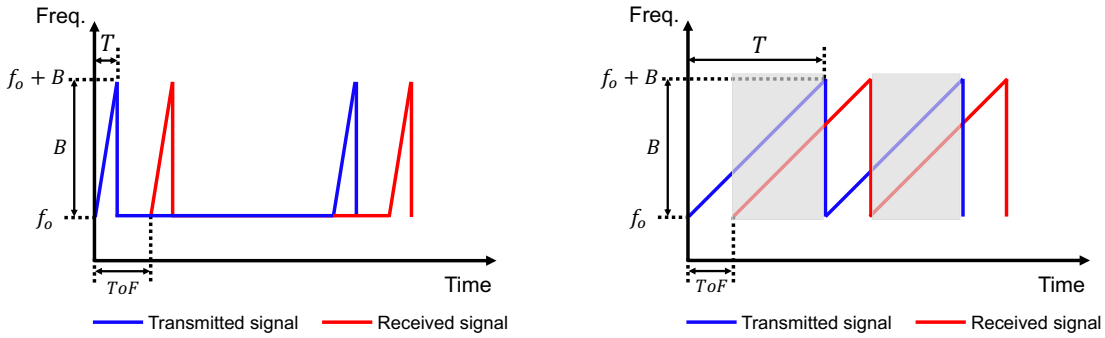


Fig. 2. System overview of *SoundFace*.

2 RELATED WORK

User Authentication for Smart Homes. Personal Identification Number (PIN) or a graphical password [32] is the earliest and still most widely used authentication methods for smart homes, but they are easily leaked to other people. With the development of smart household appliances, biometric-based authentication methods have been proposed, including fingerprint authentication [22] for smart doors, face recognition [29] for smart refrigerators, and voiceprint authentication [9] for smart speakers, etc. However, these methods are all vulnerable to replay attacks, which are easy to implement and require neither sophisticated equipment nor specific expertise. To mitigate the vulnerabilities associated with the existing authentication methods, recent works realize user authentication based on human movements for smart homes. FingerPass [17] and ThumbUp [44] leverage the channel state information (CSI) of WiFi signals and the signal collected from the Inertial Measurement Unit (IMU), respectively, to authenticate users through specific finger gestures. TouchAuth [43] performs authentication by having the user wearing a customized wristband touch an analog-to-digital (ADC) pin of the IoT device. P2Auth [20] builds a secure and intuitive authentication method that authenticates device users by comparing the simple operations sensed by devices and those captured by the user's wristband. These works require users either to wear pre-deployed skin-contact devices or perform a series of actions for user authentication, both of which are intrusive and lead to poor user experiences.

Facial Authentication. The most widely deployed face authentication (FA) systems in home environments rely on RGB cameras to collect facial information [11, 42], such as for verifying identities on smart TVs, smart locks, and smart appliances equipped with touchscreens. They are of high accuracy and decent robustness, but have several practical limitations. First, they are susceptible to lighting variations. 2D cameras that are not equipped with infrared lights for night vision may fail in poor light conditions [8]. Second, it is always worrying that using cameras has the risk of privacy information leakage. Third, such vision-based FA systems make it easy for spoofing attacks like photos or videos [6]. To overcome these limitations, wireless-based FA systems are designed to capture facial characteristics and achieve authentication. RFace [40] builds an anti-spoofing face authentication system based on passive RFID arrays to counteract 2D and 3D attacks. The mmFace [41] utilizes the mmWave signals to capture both the facial biometric features and structure features, facilitating the registration by only three photos for registration. However, the implementations of both FA systems require dedicated RFID array and costly mmWave radar, which hinders the wide deployment of these works in practical home scenarios. Compared with the RFID and mmWave signals-based FA systems, acoustic components have been widespread in smart home appliances. EchoPrint [46] combines acoustic signals with vision images to verify the user's face on a commodity smartphone. Such an acoustic-based FA system offers protection against 2D spoofing attacks. However, it has the strict requirement for face alignment given by the camera, which makes it still suffer from limitations such as poor lighting. In contrast, SoundFace aims to develop a user-friendly, robust and reliable FA scheme based only on acoustic signals.



(a) Pulse chirps for authentication.

(b) FMCW chirps for localization.

Fig. 3. The chirp signal representations for the acoustic signal.

3 SOUNDFACE DESIGN

Fig. 2 presents an overview of SoundFace, which consists of two primary phases: the registration phase and the authentication phase. Users can register and authenticate by simply lifting their heads and posing their faces in front of the acoustic devices for a few seconds. In the registration phase, the speaker actively emits designed composite chirp signals to the user's face, and the microphones receive reflected echoes. SoundFace first utilizes a two-stage locating approach to separate the face region echoes from multipath interferences. Then, SoundFace pre-trains a CNN-based acoustic feature extractor to obtain the facial features of different users. Finally, the extracted features are fed into a OC-SVM to train an authentication model and an SVM to train a recognition model. In the authentication phase, SoundFace commences the same process to segment the face region echoes of the user. The acoustic features are then extracted from the pre-trained acoustic feature extractor and fed into the trained OC-SVM model to determine if the user is authorized. If the user is authorized, their features are then fed into the pre-trained SVM model for recognition.

3.1 Design Consideration of Acoustic Signals

In this section, we present the design of the transmitted signals for sensing and the rationale behind the design.

3.1.1 ***Signal Design for Authentication.*** A unique 3D facial contour is composed of various reflecting surfaces (e.g., forehead, cheek), which can produce a distinctive mix of echoes. For user authentication, there are several considerations in the design of the transmitted signal. First, the signal should carry rich and distinct geometric information from the target face region. Second, it should be robust to severe multipath interferences from the surrounding objects. Lastly, the acoustic signal should be inaudible and little annoyance to the human ear to ensure a good user experience.

As a result, the transmitted signal is designed as a frequency-modulated continuous-wave (FMCW) chirp whose frequency increases linearly from 16 kHz to 22 kHz in a period of 1 ms and a 50 ms interval between adjacent chirps, as shown in Fig. 3(a). On one hand, the longer period of chirp can achieve a higher signal-to-noise ratio (SNR). However, if the chirp duration is too long, it could lead to the overlap of echoes from various distances. For example, the direct transmission from the speaker to the microphone will overlap with the echoes from the nearby face region. Meanwhile, a short interval between the continuous chirps may detrimentally affect the facial echoes by mixing them with the reflected signals from the far-away objects for the previous chirp. Assuming a comfortable distance from the human

nose to the transceivers is 25-50 cm, corresponding to ~1.4 -2.8 ms at the speed of sound. Thus setting the 1 ms chirp duration is moderate to isolate signals effectively in the time domain, and the 50 ms delay is sufficient to separate echoes from two consecutive chirps. On the other hand, a wider bandwidth can generate richer frequency-domain information and a higher sensing resolution. Therefore, we choose the 6 kHz bandwidth starting from 16 kHz, which may be slightly audible to some users. We finally apply a Hanning window to the chirps to increase its peak to sidelobe ratio (PSR), thereby increasing the low SNR and reducing the audible effects of the pulse chirp signal.

The pulse chirp is fed into the speaker and emitted towards the face, then the co-located microphones receive the echoes via multiple paths. It is straightforward to analyze the reflected signals by performing cross-correlation [35] between echoes and the transmitted signals, whereby the resulting correlation peaks reflect the distances of different objects. The ranging resolution is calculated as $\delta d = \frac{c}{2F_s} = \frac{343}{2\times44100} = 3.89$ mm. For extracting the echoes from the face area, the cross-correlation peaks within the typical face distance (e.g., 25-50 cm) need to be found. Since faces absorb and attenuate sound waves more than other reflectors, the comparatively weak facial echoes can easily be overwhelmed by other nearby stronger reflection signals. This makes the localization of target echoes based on cross-correlation unstable and prone to failure.

3.1.2 ***Signal Design for Face Localization.*** As cross-correlation-based distance estimation is vulnerable to strong reflections, it is highly likely to misjudge the target face area. To solve this problem, Echoprint uses the camera image from a smartphone to calibrate the outliers of the distance measurements from acoustic [46]. However, such a vision-aided technique is not available for our scheme. Although chirp signals have been considered effective in combating strong background noise or interference and enabling the separation of reflections from different distances, the pulse chirps designed for authentication in Fig. 3(a) are too short to accumulate enough SNR for basic chirp mixing.

Therefore, at the beginning of the transmitted pulse signals, we add some pilot FMCW chirps whose duration is 40 ms, sweeping from 16 kHz-22 kHz for finer face area localization as shown in Fig. 3(b). Such transmitted pilot chirp can be represented as

$$x(t) = \cos(2\pi(f_0 t + \frac{B}{2T}t^2)), \tag{1}$$

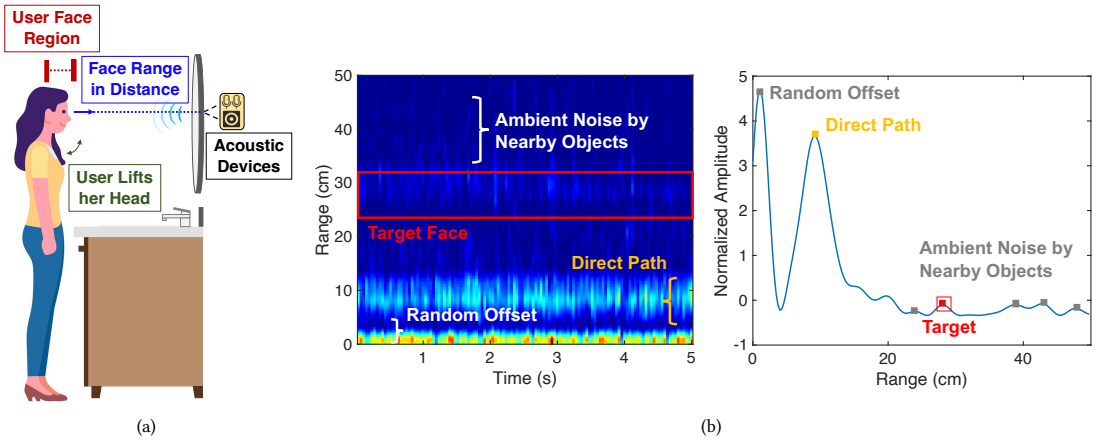
where f_0 is the starting frequency, B is the bandwidth, and T is the sweep time. The reflection signal is a time-delayed version of the transmitted signal that can be represented as

$$y(t) = \alpha\cos(2\pi(f_0(t-\tau) + \frac{B}{2T}(t-\tau)^2)), \tag{2}$$

where α is the signal amplitude attenuation factor and τ is the Time-of-Flight (ToF) of the signal in the air. The transmitted and received signal can be processed to generate the mixed signal as $m(t) = x(t)\cdot y(t)$, which contains the ToF information of the received signals reflected by the object. After applying a low-pass filter, the resulting mixed signal can be represented as

$$m(t) = \frac{1}{2}\alpha\cos(2\pi(\frac{B}{T}\tau t + f_0\tau - \frac{B}{2T}\tau^2)). \tag{3}$$

Clearly, $m(t)$ is a single tone signal with the beat frequency $f_d = \frac{B}{T}\tau$. By performing Fast Fourier Transform (FFT) on the mixed signal, we can calculate the absolute distance between the object and sensing device as $d = \frac{cf_dT}{2B}$, where c is the speed of sound in the air, and the factor '2' accounts for the round trip of the reflected signal. We can thus obtain the absolute range resolution $\delta d = \frac{c}{2B} = \frac{343}{2\times6000} = 2.86$ cm. This means if two reflectors are located within 2.86 cm apart with respect to the sensing device, they cannot be differentiated. Thereby, the weak facial reflected signals in the 25-50 cm distance range can be easily distinguished from other strong interfering signals originating outside this range.



(a) (b)

Fig. 4. (a) Face localization illustration; (b) Massive interferences of raw acoustic signals in range profile.

As for the strong reflection signals within this typical distance range that may disturb the target localization, we will further investigate the mixed signal and illustrate how to identify the face target by signal variance in the presence of nearby interference objects in Sec 3.2.

3.2 Two-stage Locating Approach

When we apply acoustic signals for face authentication, besides the reflection signals bounced off the face region, there is massive ambient noise due to surrounding objects, including multipath reflections by other appliances and walls, the direct path signal from the speaker to the microphone, and the signal offset caused by random system delay [45]. Such noise can interfere with the separation of face region echoes. To further understand the impact of these interferences, we conduct a preliminary experiment indoors using a speaker and a microphone on the Bela platform. We ask a user to stand 30 cm away and face directly to the devices for a while. Fig. 4(b) shows the range measurements generated from raw pilot FMCW chirps after mixing FFT, both in terms of heatmap over a period of time and an amplitude profile within one chirp. It clearly shows that acoustic signals contain massive interferences. Therefore, separating the face-reflection signals from other interfering reflections is critical. In this section, we introduce the two-stage locating approach, which employs a multi-granularity acoustic sensing technique to first locate the major face, and then locate the facial areas, thus separating the face region echoes effectively.

3.2.1 ***Major Face Localization.*** The face region of a user typically includes major surfaces, such as the forehead and cheeks, that produce strong echoes, as well as other surfaces, such as the nose and chin, that generate weaker echoes. In *Stage-1*, SoundFace captures the signal variation when users lift their heads at the beginning when the pilot FMCW chirps transmission, as shown in Fig. 4(a). For ease of observing the signal variations, in this section, the user is asked to perform several head-ups with long pilot chirps of 5 s. The movements of the user's head help locate the major surfaces, called *face major localization*.

• **Background Multipath Subtraction.** Before estimating the distance of the major face, we first remove the background multipath. To remove the delay caused by the operating system, we align the received signal with the

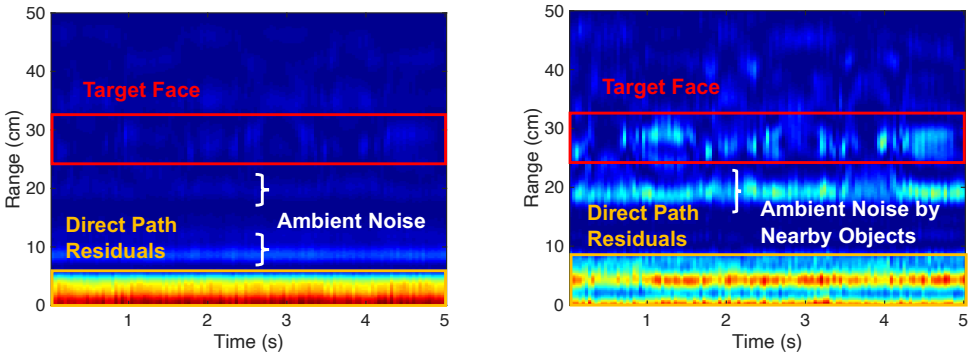


Fig. 5. Range profile heatmaps when the user performs head-ups: (a) Before background subtraction; (b) After background subtraction.

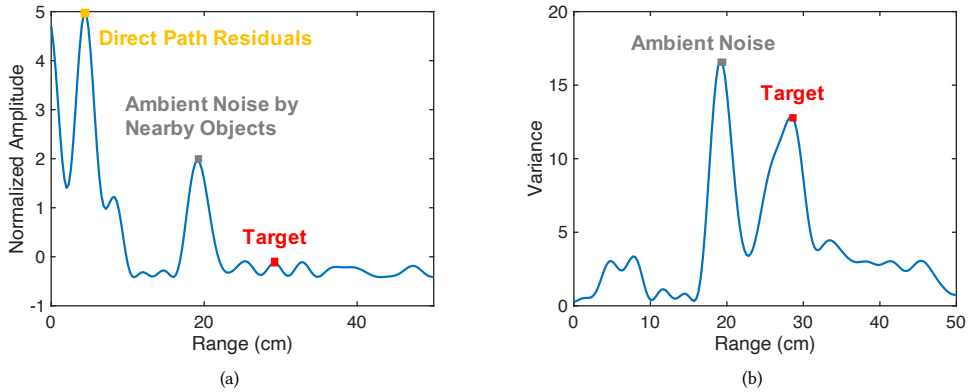
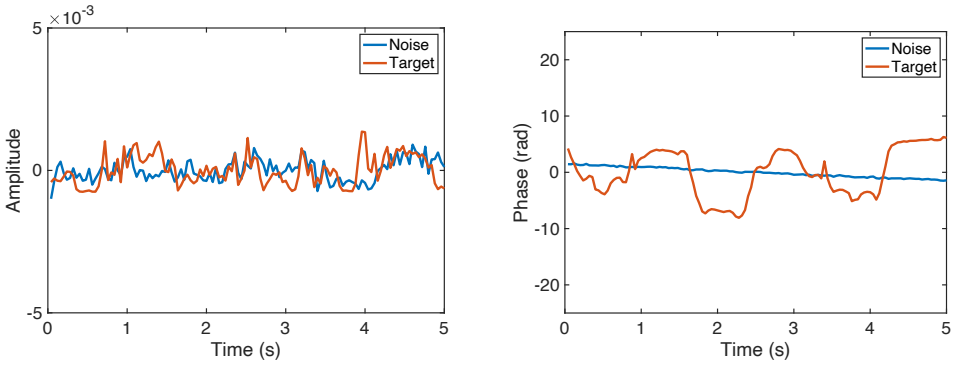


Fig. 6. (a) The FFT amplitude results for each range bin; (b) Variances of signal variations.

direct path signal [45]. To remove the interference of the direct path signal and reflections from static surroundings, we measure the background signal when there is no user standing and remove it later.

• **Multi-granularity Acoustic Sensing.** Although subtracting the background signals can remove most static interferences, there are still some residuals due to imperfect synchronization and scaling between received samples and pre-recorded signals. Fig. 5 shows the range profile heatmaps with and without background subtraction following signal alignment. Compared with the heatmap in Fig. 5(a), we can obtain a more clearly defined face region range profile after subtracting the background signals in Fig. 5(b), but at the same time, other interferences remain to exist. Inspired by the activity sensing using acoustic signals [18], we explore the signal variations when the user lifts his/her head, which facilitates differentiating face echoes from other interferences.

❶ Searching Target Candidates in Coarse-grained Distances.



(a) (b)

Fig. 7. (a) The signal amplitude variations; (b) The signal phase variations.

After performing background subtraction, the FFT results of the mixed signal reflect the frequency domain information for each range bin, as shown in Fig. 6(a). Among the range bins (e.g., 0-50 cm for typical face distance), there is one target bin that corresponds to the target's location and therefore contains the target's information. Generally, due to the user's head movement, there are much larger signal variations at the target bin than at other bins. To demonstrate this, we first compute the variance of the signal variations for all bins, then adopt an outlier detection algorithm to search the target candidate bins by finding the peaks using peak variance ratio (i.e., *pvr*) [18]. If the *pvr* value is larger than a pre-defined threshold (e.g., we empirically set it as 3), we consider it as a target candidate. As shown in Fig. 6(b), there are two potential target candidates whose signal variation variances are significantly higher than those of the other bins. Although the true target bin is successfully found, the bin with nearby noise is also mistakenly identified as a potential target candidate owing to its relatively large signal variations. As a result, we need to further identify the target bin among several candidates.

❷ Identifying Target Location with Fine-grained Signal Variations.

From Eq.3, the phase of the mixed signal is $\phi = 2\pi(f_d + f_0\tau - \frac{B}{2T}\tau^2)$. The first term $2\pi f_d$ corresponds to the coarse-grained target distance (i.e., range bin), and the second term $2\pi(f_0\tau - \frac{B}{2T}\tau^2) \approx 2\pi f_0\tau$ corresponds to the fine-grained distance change caused by the movement of the target within a range bin. To further identify the target from the nearby interference, we explore the signal variations for the potential candidate bins in Fig. 6(b). The FFT results of the mixed signal are complex values for each range bin, thus the signal variations contain both amplitude and phase changes, which are extracted in Fig. 7. We can obtain that the signal amplitude variations of two candidate bins are comparable (Fig. 7(a)), but the signal phase variation for the target movement is larger than that of the nearby object (Fig. 7(b)). Therefore, we identify the candidate bin with the largest phase change as the true target location.

3.2.2 ***Face Echoes Segmentation.*** In *Stage-2*, after the pilot FMCW chirps, the user just needs to look directly at the acoustic devices while the continuous short-time pulse chirps are transmitted. We utilize cross-correlation to the pulse chirps and find the peaks corresponding to the face localization result provided by the *Stage-1*, then we segment the reflected signals covering the whole face region as the target signals of interest.

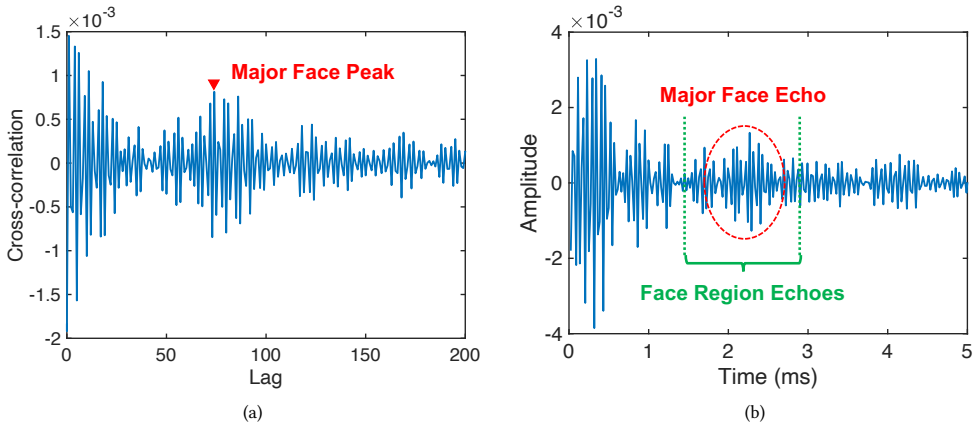


Fig. 8. (a) Cross-correlation result of pulse chirps; (b) Segmentation of face region echoes.

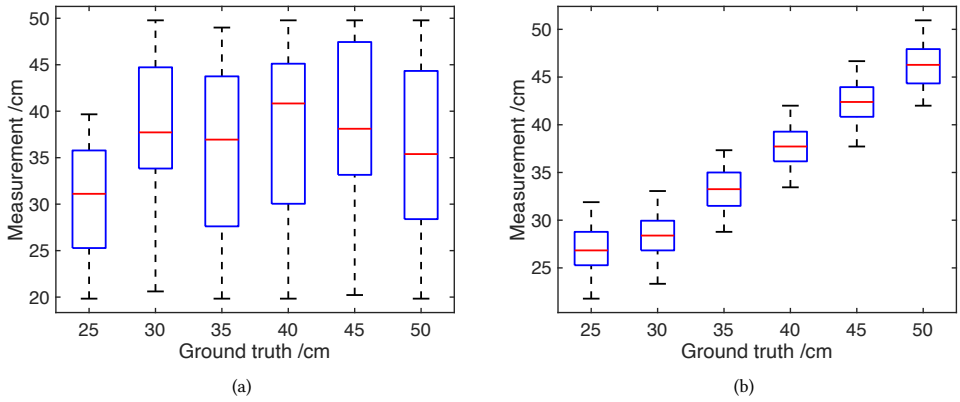


Fig. 9. Distance measurements of the major face: (a) Before narrowing down the searching range; (b) After narrowing down the searching range.

As mentioned in Sec. 3.1.1, the signal transmission for authentication is designed as the continuous short pulse chirps with gaps for better isolation of face region echoes from other multipath interferences in the time domain. However, the short period of chirps cannot achieve high SNR, and this, coupled with the fact that the face absorbs and weakens acoustic signals, makes it easy for the face echoes to be overwhelmed by other strong noises reflected from other objects. Fig. 9(a) shows the distance measurement results of the pulse chirps by directly finding cross-correlation peak location corresponding to major face distance range (i.e., from 25 cm to 50 cm with a step of 5 cm). We can observe that there are significant measurement errors when identifying the target peaks.

In *Stage-1*, we have successfully located the user's face at a distance with the assistance of their head-ups during the transmission of pilot FMCW chirps. Thereby, we narrow down the searching range for the cross-correlation peaks from 25-50 cm to near the given major face location. Fig. 8(a) shows the cross-correlation result of transmitted and received

pulse chirps after background signal subtraction. We can see that the peak corresponding to the major face can be successfully found. As a result, the distance measurements obtained are much more precise, as shown in Fig. 9(b).

After locating the major face pulse echo, we extend the segment by adding 10 samples before and after to cover the entire face region, which has a depth of 8 cm. As illustrated in Fig. 8(b), these segmented signals are used as the face region echoes for the following authentication.

3.3 Acoustic Feature Extraction

The echoes obtained from the face region are a combination of echoes from different facial surfaces such as the cheek, forehead, nose, chin, and others. These echoes capture the full information of the face. However, isolating the segmented signals into individual echoes in the time domain is difficult due to noises and frequency-selective fading, where the superposition of signals with different amplitudes and phases can be constructive at some frequencies while destructive at others. In this section, we first apply a time-frequency analysis to extract time-frequency spectrograms from the segmented face echoes both in terms of the raw acoustic signal and the converted mixed signal after chirp mixing. Then we design a CNN-based feature extractor to obtain the high-level acoustic features of users' faces.

3.3.1 ***Segmented Acoustic Signal Processing.*** In our system, the difference in each user's 3D facial profile lies in the segmented echoes from different facial surfaces that are at varying distances with different amplitudes from the acoustic device. On one hand, we convert the segmented face echoes into corresponding spectrograms that can directly manifest the distinctions of signal delays and attenuations in the time-frequency domain. On the other hand, we measure the delay of each echo in the segmented face echoes by the chirp mixing technique, where the resulting frequency shifts are proportional to the distances from face surfaces to the acoustic device [46]. We note that the designed pulse chirps are too short for basic chirp mixing, herein, we use a reference signal with a wider bandwidth (named extended reference signal) covering the entire segmented signals to achieve the chirp mixing [19]. Specifically, the extended reference signal is designed by adding an extra sweep time to the original transmitted pulse chirp and a corresponding increased bandwidth (e.g., the added extra sweep time from 1 ms to 1.5 ms increases the bandwidth from 6 kHz to 9 kHz). At the same time, to satisfy the Nyquist sampling theorem, the sampling rate of the extended signal is doubled from 44.1 kHz to 88.2 kHz accordingly. To obtain the extended mixed signal, we first upsample the received segmented face echoes by a factor of 2 using the cubic spline interpolation. Then we multiply it with the extended reference signal and its 90-degree phase-shifted version to derive the In-Phase (I) and Quadrature (Q) parts of the mixed signal respectively. Lastly, we downsample the I and Q components by a factor of 2 to form the complex extended mixed signal of the face region.

We use STFT to extract time-frequency spectrograms from the face region echoes and the extended mixed signals to analyze the frequency contents of finer time domain. To get a balanced trade-off between the time resolution and frequency resolution, carefully selecting the STFT parameters is of vital importance. First of all, we apply STFT on the two forms of the segmented signals for 66 samples, and later we infer an initial acoustic feature using the extracted raw as well as the mixed time-frequency spectrograms. To capture the time-frequency variation of reflected signals from facial surfaces more accurately, we apply a 64-point FFT with a hamming window to the 6-sample segment, and the step length between adjacent FFT windows is set at 1 sample, which allows us to achieve both higher frequency resolution and time resolution. The resulting time-frequency profiles are the 33×61 spectrogram from face region echoes and the 64×61 spectrogram from complex extended mixed signals.

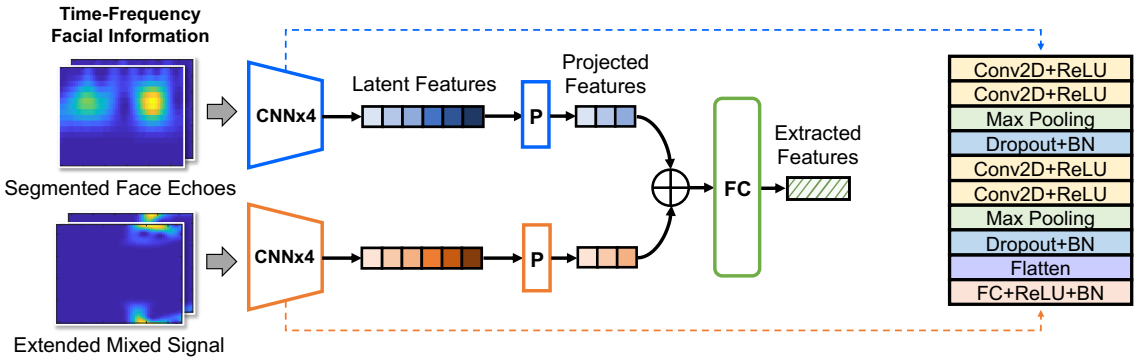


Fig. 10. The architecture of the CNN-based feature extractor.

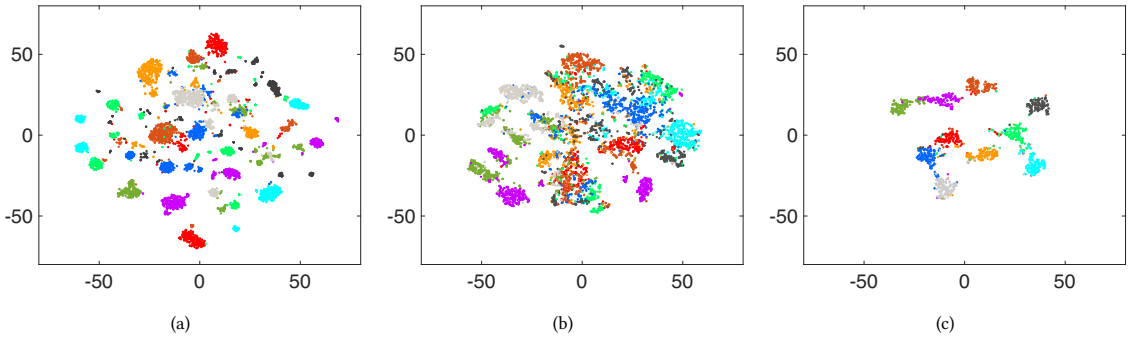


Fig. 11. (a) STFT feature from segmented face echoes of different subjects; (b) STFT feature from extended mixed signals of different subjects; (c) Features extracted from CNN-based feature extractor.

3.3.2 ***CNN-based Acoustic Feature Extractor.*** Although the derived two kinds of STFT profiles contain the information of user's facial structure, they cannot be directly used to distinguish different individuals. As shown in Fig. 11(a) and Fig. 11(b), we extract the two forms of STFTs from the 10 subjects and use the t-SNE scheme [33] to project them into a two-dimensional space, where different colored dots represent different subjects. We can see that the STFT features in raw and mixed do not show clear different clusters among subjects. Previous works has shown that Convolutional Neural Network (CNN) has good performance in extracting features from image-like data [10, 12, 21]. To further extract more distinguishable features for different subjects, we design a deep learning model based on CNN, whose designed architecture is shown in Fig. 10.

Since our system has one speaker and two microphones placed at different locations (i.e., two microphone-speaker links), we can measure the face region echoes from different observation angles. Here we take the dual-link STFTs of segmented face echoes and extended mixed signals as inputs of two parallel 4-layer CNN encoders with the same structure to obtain the latent intra-sensor features. The encoder network consists of multiple blocks. The first two blocks are cascaded convolutional stacks, with each stack containing two 2D convolutional layers following two Rectified Linear Units (ReLU) layers to enhance the extracted results, a max pooling layer to perform subsampling, a dropout

layer to prevent overfitting, and a batch normalization layer [13] to further improve the model performance and stability in sequential order. Specifically, the chained convolutional layers in two stacks use 32 and 64 kernels respectively with all 3×3 kernels, and the max pooling layers are all set to 2×2. Following the two stacks, the flatten layer combines the learned information and the encoded features are mapped into the 128 dimensional latent features by a ReLU-activated fully connected layer with the batch normalization. Then, we fuse the 2-channel latent features by concatenating the projected features after the projectors and generate the final 64 dimensional feature by fully connected layers. Generally, a basic feature extractor can be pre-trained on a multi-user data set by adding a softmax activation function as the output layer at the end to classify the extracted features into different classes. We use the cross-entropy as the loss function and apply the Adam algorithm [14] to minimize it.

We also visualize the extracted features from the CNN-based extractor for different 10 subjects in Fig. 11(c). We can see that the extracted features form distinguishable clusters for different subjects.

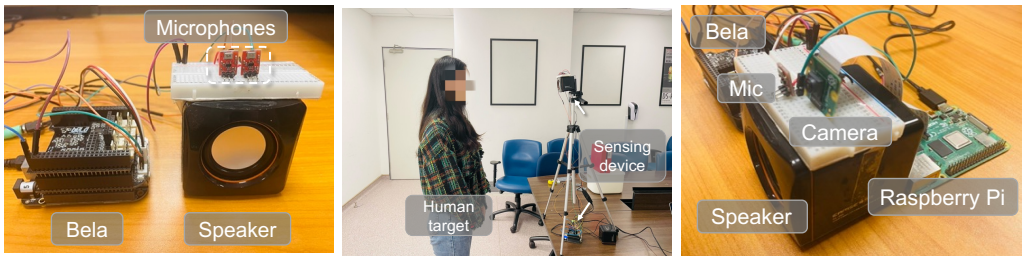
3.4 User Authentication

In SoundFace, we utilize a one-class SVM [5] and an SVM to authenticate and recognize users successively. Specifically, each legitimate user needs to provide a batch of signal samples to train the both SVM models for SoundFace registration. In the authentication stage, once receiving a login request, SoundFace first feeds the extracted acoustic facial features into the pre-trained one-class SVM model. Then, it outputs the authentication result by comparing the similarity with the registered users in the database. Since the extracted features of each person are distinct, unregistered users will fail the authentication owing to low similarity. After authenticating an authorized user, we then feed the extracted features into the pre-trained recognition SVM to further classify which legitimate user the person is.

Due to the single scenario and limited data used for registration, various factors such as newly emerging individuals, changes in the authentication environments, and authentication distance variation caused by users' unconscious position movements may affect the robustness of SoundFace in practice. Accordingly, we first apply the time warping and the scaling methods to adjust the amplitude and phase changes of acoustic signals to augment training data, which can cover most potential variables when performing authentication and augment our dataset. In this way, SoundFace will be less prone to the overfitting problem.

Then we employ transfer learning technique to further enhance the overall robustness of SoundFace. Although our two-stage locating approach can effectively segment the face region echoes, it is nearly impossible to entirely remove noisy echo signals from surrounding obstacles (e.g., the objects at a similar distance to the mirror) using signal processing methods. For smart mirror authentication scenarios, when the authentication environment is different from that of registration, the extracted acoustic features are more or less disturbed by the changed environmental interferences, thus affecting the authentication performance. In SoundFace, we train a base feature extractor model with the collected data in the registration phase and adapt the base model to different environments using transfer learning in the authentication phase. To perform transfer learning, we first obtain a small number of labeled samples of users in the authentication environment. Then, we freeze the parameters of the pre-trained CNN-based feature extractor except for the last layer (the Softmax layer) and retrain the last layer using the obtained data. In this way, the fine-tuned model can be adapted to extract acoustic facial features in the different authentication environments. Finally, we also fine-tune the SVM models by adding the data from the authentication environment, which costs low training overhead [40, 41, 46].

Although the smart mirror scenario ensures that the user's distance for each authentication is as fixed as possible, such as the washstand in Fig. 4(a) helps limit the distance, the imperceptible movement of the user may inevitably



(a) (b) (c)

Fig. 12. (a) Components of the Bela platform; (b) The experiment setup; (c) Implementation of the baseline.

cause the authentication distance to change. In SoundFace, we also employ fine-tuning to solve the impact of different authentication distances. In the registration phase, we collect a small amount of user data at several major distances. During the authentication phase, once we detect a change in the user's authentication distance utilizing the two-stage locating approach, we fine-tune the model by applying the pre-stored data under the corresponding distance, with a similar fine-tuning process as the one for authentication environment changes.

In the case that a new individual, whose data is not used to train the feature extractor, attempts to access the smart mirror, SoundFace uses a pre-trained feature extractor to obtain their acoustic facial features. If the individual is a legitimate new addition, we retrain the SVM models with their data for registration. If not, we feed their features directly into the trained OC-SVM model for authentication.

4 EVALUATION

4.1 System Setup

Fig. 12(a) and Fig. 12(b) give an overview of SoundFace system setup.

4.1.1 ***Implementation.*** We implement SoundFace on an off-the-shelf research-purpose hardware platform Bela [3]. It is flexible to transmit and receive acoustic signals for face authentication.

• **Bela platform**: Bela [3] is an open-source embedded computing platform for creating responsive, real-time interactive systems with audio and sensors. The Bela system, mainly based on the assembled Bela unit (Bela cape + BeagleBone Black), is widely used for research involving acoustic signals owing to its flexibility to support different numbers and locations of microphones and speakers. We implement a 2-array MEMS microphone and a general-purpose speaker system driven by BeagleBone Board and Bela board as shown in Fig. 12(a). With 8 analog inputs/outputs (16-bit), 16 digital I/O, 2 audio input/output channels, and 2 speaker amplifiers, the Bela board is known for audio signal processing with latency as low as 5 ms. We write C++ code on Beagle Bone to transmit acoustic signals and record data from two microphones at the same time.

• **Signal parameters**: As mentioned in Sec. 3, we adopt the frequency band from 16 kHz to 22 kHz which is only slightly audible to some people with a bandwidth of 6 kHz for finer sensing. The chirp duration of the pulse chirps for authentication is 1 ms with a 50 ms interval between two consecutive chirps. Meanwhile, we employ 40 ms as a chirp duration of the pilot FMCW chirps, which can be transmitted for 3 s. The sampling rate of the Bela platform is 44.1 kHz.

(a) Seminar room. (b) Meeting room. (c) Hall.

Fig. 13. Experimental environments.

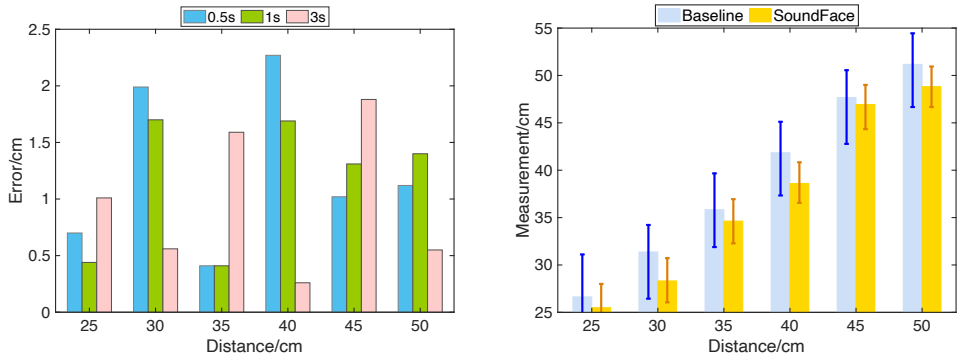
4.1.2 ***Experiment Setup.*** Fig. 12(b) shows the default deployment of SoundFace. We invite 20 volunteers (11 males and 9 females) in our experiments. Among the 20 volunteers, we choose 3 volunteers as legitimate users, 14 volunteers as unauthorized individuals, and another 3 volunteers as new users. During data collection, each participant is asked to look up and place their face around 30 cm in front of the sensing device, i.e., the acoustic transceiver that is mounted at a fixed height of 1.60 m with a tripod. In the registration phase, we collect groups of data at different times in three real environments (i.e., meeting room, seminar room, and hall, as shown in Fig. 13) under different ambient noises (i.e., quiet, music, and talking) for each user, which takes about 100 seconds for each environment. Then in the authentication phase, each user performs authentication for 400 times to evaluate the performance of SoundFace. Each default authentication attempt happens when the consecutive pulse chirps are transmitted every 50 ms.

• **Metrics**: We introduce six metrics to quantify the performance of SoundFace, including accuracy, True Accept Rate (TAR), False Accept Rate (FAR), False Reject Rate (FRR), Equal Error Rate (EER), and Receiver Operating Characteristic (ROC). In particular, TAR is the probability that a legitimate user is correctly authenticated by SoundFace. FAR represents the probability that SoundFace accepts an unauthorized individual as a legitimate user. FRR represents the probability that SoundFace rejects a legitimate user as an illegitimate one. EER describes the rate where FAR equals FRR. Additionally, ROC curve describes the relationship between the TAR and FAR under the various thresholds.

4.2 Two-stage locating Performance

We first evaluate the performance of the two-stage locating approach. We employ the distance between the cheek and the acoustic devices as the ground truth.

4.2.1 ***Major Face Localization.*** In the default real-world deployment, when the user heads up to the acoustic devices at the start, we set the first 3 s pilot 40 ms period FMCW chirps to fully capture the signal variation, thus locating the major face (Sec. 3.2.1). In this experiment, we evaluate the performance of the major face localization in the two-stage locating with different pilot signal lengths that relate to the user experience. We ask a subject to stand at six different viewing distances from the acoustic devices (i.e., the same experimental setup as in Fig. 12(b)) in turn and perform head-up to the devices under three different lengths of duration of pilot FMCW chirps. Fig. 14(a) demonstrates that despite the varying lengths of pilot chirps (0.5 s, 1 s and 3 s) used to capture signal variations, the errors on major face localization for different subject-to-device distances are all minor, e.g., the error drift is settled in the range from 0.26 cm at 40 cm to 2.27 cm at 40 cm. Additionally, even 0.5 s pilot chirps can capture sufficient signal variations from the user's head-up movement, implying that major face localization within a short period is user-friendly.



(a) (b)

Fig. 14. (a) Major face localization in distances with different pilot signal lengths; (b) Distance measurements.

4.2.2 ***Performance Comparison.*** We next evaluate the performance of two-stage locating with the vision-aided locating baseline method of Echoprint [46] for identifying the major face pulse echo in distance, which determines the center of the segmented echoes from the whole face region (Sec. 3.2.2). Since Echoprint uses the frontal camera of the smartphone to calibrate the acoustic distance measurements, we implement it on a Raspberry Pi 4 Model B [26], a Raspberry Pi Camera Module 2 [27], a Bela platform with a pair of speaker and microphone, as shown in Fig. 12(c). Fig. 14(b) shows the performance comparison on major face pulse echo localization for different subject-to-device distances (we set 1 s length pilot chirps here). We can see that our two-stage locating approach is comparable to the baseline in terms of distance error even though our method do not need an additional camera device. Moreover, due to the acoustic distance measurements d'_v estimated from the cross-correlation peaks are unstable, the vision calibrated distance measurement $d_v = \tau \cdot \frac{1}{d_p}$ drift larger than our motion-aided method in terms of the standard deviation of distance measurement (e.g., error bars in Fig. 14(b)). Thus, our two-stage locating can locate major face pulse echo accurately, which significantly improves the accuracy of segmenting the echoes of face region.

4.3 SoundFace Performance

We now evaluate the performance of SoundFace from different aspects.

4.3.1 ***Overall Performance.*** We first evaluate the overall performance of SoundFace in three different environments. We compare the performance of our acoustic facial feature extraction method with two baselines. **Baseline 1 (BS1)**: we utilize a single pair of speaker-microphone to extract the acoustic feature based on the STFT of mixed signal mentioned in Echoprint [46]. **Baseline 2 (BS2)**: We fuse the STFTs of mixed signal of 2-array microphones, and obtain the facial features from the acoustic feature extraction structure of Echoprint [46]. Fig. 15(a) shows the results of TARs in three different environments. We can see that the average TAR of SoundFace is 96.2%, while the average TARs of these two baselines are 83.9% and 88.3%, respectively. Meanwhile, as shown in Fig. 15(b), the average EER of SoundFace is 4.2% while the baselines' average EERs are 28.7% and 16.3%. The high TAR and low EER of SoundFace indicate its outstanding authentication performance in different environments. The comparison results with two baselines demonstrate that our acoustic feature extraction method, which takes into two STFT forms of facial echoes, can effectively distinguish

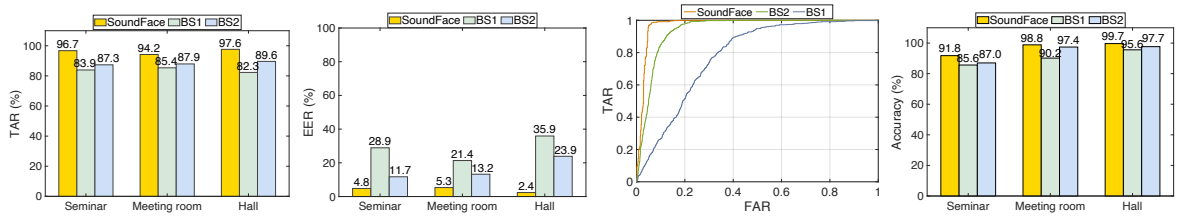


Fig. 15. (a) TAR under different environments; (b) EER under different environments; (c) ROC curves under the environment of seminar; (d) Recognition accuracy of legitimate users.

individuals' facial structure and improve the authentication accuracy. Fig. 15(c) depicts the ROC curves under the seminar room, where the corresponding area-under-curve (AUC) of SoundFace is larger than that of both Baseline 1 and Baseline 2. Additionally, Fig. 15(d) shows the recognition accuracy of the authenticated legitimate users, where SoundFace can achieve a high average accuracy of 96.8% and the two baselines can reach the average accuracy of 90.5% and 94.0%. These results demonstrate that SoundFace can authenticate users accurately and securely, and can further recognize legitimate users precisely, outperforming the other two baselines.

4.3.2 ***Continuous Authentication.*** We now evaluate the continuous authentication performance of SoundFace. One default authentication attempt mentioned in Sec. 4.1.2 happens when the consecutive pulse chirps are transmitted every 50 ms. Therefore, getting one authentication result from multiple attempts (corresponding to the continuous authentication duration of one cycle) is still user-friendly and does not affect the real-time experience. We pre-define an acceptable threshold for one continuous authentication. If the authentication accuracy within a cycle exceeds the threshold, the user will be accepted as a legitimate user. Fig. 16(a) shows the TAR results for three different environments with an acceptable threshold of 0.85 as the duration of continuous authentication varies. Although the shorter duration of continuous authentication will cause a slight decrease in TAR, the average TARs of three environments within 0.5 s, 1 s and 2 s are respectively 94.2%, 98.8% and 100%, which are still higher than 90%. The results indicate that SoundFace can achieve better performance if performing continuous authentication. Fig. 16(b) depicts the TAR and FAR results of 1 s continuous authentication with the variation of acceptable threshold in the seminar room. We find too high a threshold rapidly exacerbates the decline in TAR and the rise in FAR (e.g., the threshold is 0.95). We choose 0.85 as the empirically acceptable threshold.

4.3.3 ***Scalability.*** To further understand the impact of the user number and the stability of SoundFace, we explore how the performance of SoundFace changes as the number of legitimate users increases. To this end, we randomly choose 5 volunteers as unauthorized ones, then increase the number of legitimate users from 3 to 15 and calculate the corresponding accuracy. We repeat the random selection 5 times to ensure consistent results. As shown in Fig. 17(a), when the number of legitimate users reaches 15, TAR is still higher than 91% and EER is lower than 10%. Meanwhile, SoundFace also maintains a high recognition accuracy at 92.9%. It indicates that SoundFace has good scalability in the access control of multiple users.

4.3.4 ***Robustness.*** Robustness is a crucial issue for FA systems. Therefore, we evaluate the robustness of SoundFace in three aspects: new user authentication, cross-environment authentication and different ambient noises.

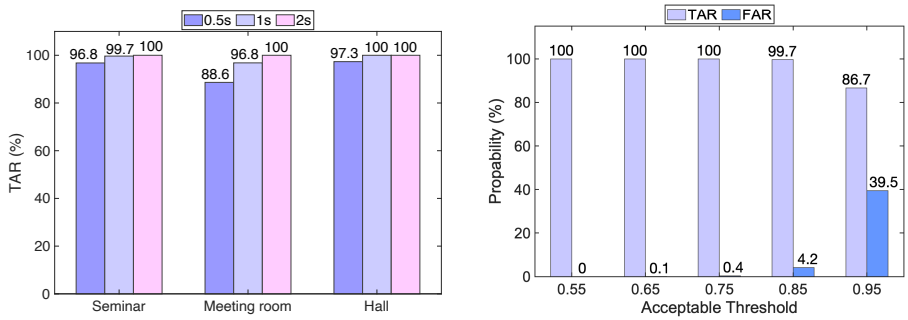


Fig. 16. (a) TAR under different durations of continuous authentication; (b) The impact of the acceptable threshold within a continuous authentication cycle.

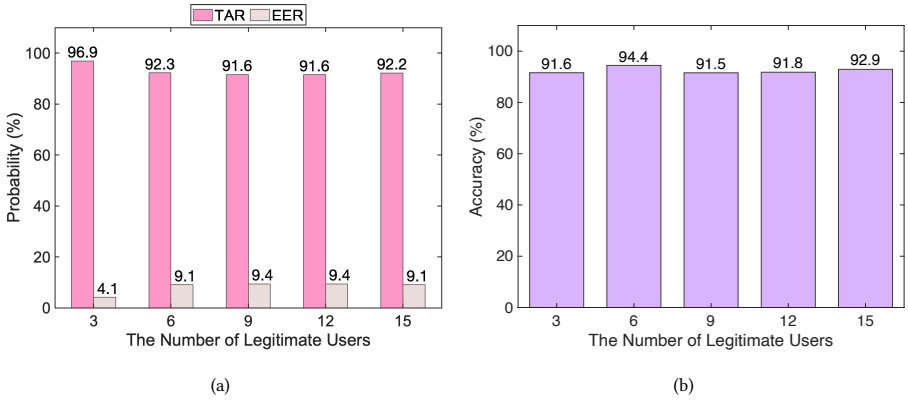


Fig. 17. The impact of legitimate user number on SoundFace performance in the seminar room. (a) TAR and EER under different legitimate users; (b) Recognition accuracy for different number of legitimate users.

• **Performance on new individuals:** To evaluate the performance of SoundFace for new individuals, we invite another 3 volunteers whose data are not used for feature extractor pre-training. Each volunteer collect 100 seconds data in three environments. The new individuals are regarded as unauthorized illegal users. Fig. 18(a) shows that the average TAR is 96.5% and the average EER is 3.2%, respectively. Additionally, Fig. 18(b) shows the recognition accuracy of the legitimated users. All these results demonstrate that SoundFace can achieve fairly good performance for new individuals.

• **Cross-environment authentication:** To assess the robustness of SoundFace in real-life application scenarios, instead of testing it in a fixed environment, we evaluate it in different authentication environments, i.e., cross-environment authentication settings. Fig. 19(a) shows the ROC curves for the meeting room and the hall, where only the SVM model is fine-tuned by adding data from these two authentication environments while the feature extractor is pre-trained from the seminar room and frozen in these two new environments. We can find that the performance of cross-environment authentication slumps since the background interferences of different places lead to unstable acoustic signals and the extracted facial features. Thus, our system uses transfer learning to fine-tune the base acoustic feature extractor,

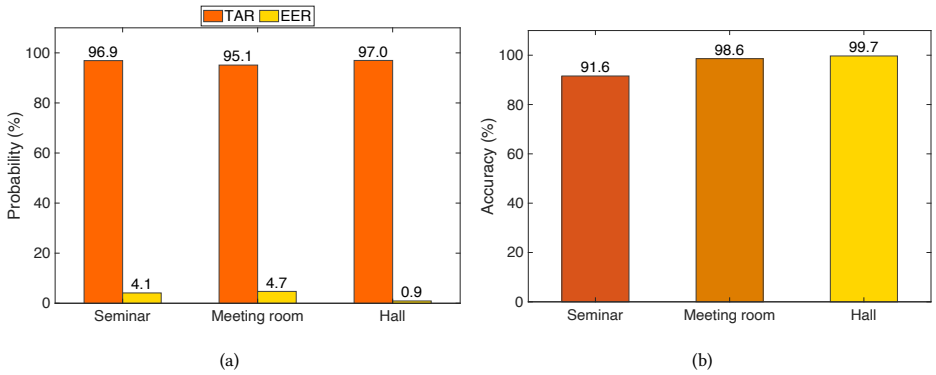


Fig. 18. (a) TAR and EER for the new individuals; (b) Recognition accuracy of legitimate users.

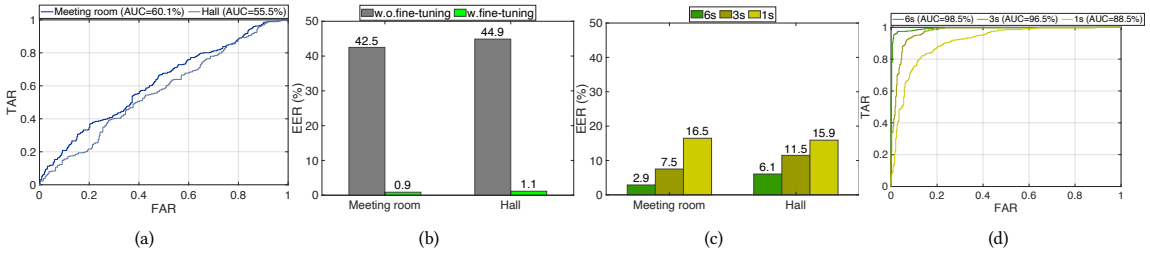


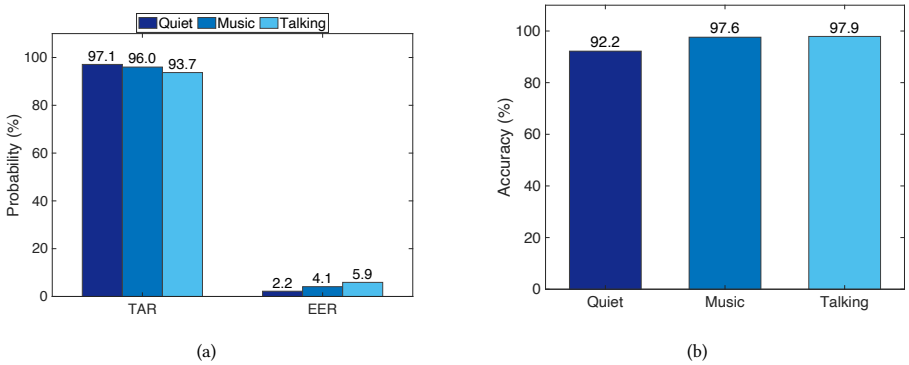
Fig. 19. (a) ROC curves of cross-environment authentication without model fine-tuning; (b) EER with and without model fine-tuning; (c) EER under different fine-tuning data lengths; (d) ROC curves varies in fine-tuning data length in the meeting room.

enabling it to adapt to various environments. Fig. 19(b) shows the comparative results of EER with and without the model fine-tuning for the feature extractor. It is clear that model fine-tuning can effectively reduce the error rate of cross-environment authentication. Then we explore how long of the target data used in fine-tuning can influence performance. We choose 1 s, 3 s and 6 s data of each user from the two authentication environments to fine-tune the feature extractor model, respectively. As shown in Fig. 19(c), generally, the EER decreases as more data are used in transfer learning. Fig. 19(d) depicts the ROC curves varies in data length for fine-tuning in the meeting room. To get a balanced trade-off between the authentication performance and the training overhead of model fine-tuning, we use 3 s data per user to fine-tune the base model for adapting the cross-environment authentication, which is also user-friendly in the real-life scenarios.

• **Impact of Authentication Distance:** We evaluate the impact of distance between the user's face and the acoustic devices in the seminar room. We guide volunteers to place their faces at two distances for multiple authentications. In the experiments, the data of 30 cm and 20 cm are used for registration in model training, respectively, and the authentication tests are conducted at the same or different distances. Table 1 shows the results of TAR, EER, and the recognition accuracy of the authenticated legitimate users for different distances. We find that when the authentication distance differentiates from the registration distance, the performance of SoundFace deteriorates significantly. However, after model fine-tuning with 3 s data per user, the impact of authentication distance can be mitigated and SoundFace is robust to distance variation.

Table 1. Impact of authentication distance.

	Train distance Test distance	30cm 30cm	30cm 20cm	20cm 20cm	20cm 30cm
Origin	**TAR**	0.9797	0.5972	0.9880	0.8364
	EER	0.0101	0.2309	0.0067	0.4050
	Accuracy	0.9595	0.6270	0.9920	0.5243
Finetuning	**TAR**		0.9808		0.9657
	EER	~	0.0064	~	0.0057
	Accuracy		0.9987		0.9829



(a) (b)

Fig. 20. Effect of different ambient sound noises. (a) TAR and EER in the seminar room; (b) Recognition accuracy of legitimate users.

• **Effect of Ambient Sound Noises:** We evaluate the performance of SoundFace under different ambient sound noises: quite, music, and talking. Fig. 20(a) shows the results of TAR and EER for three ambient noises in the seminar room. Fig. 20(b) shows the recognition accuracy of the authenticated legitimate users. We find that there is no major difference between 'quiet' and other two ambient noise conditions, which means the ambient sound noise has little impact on the performance of SoundFace.

5 DISCUSSION

Face deflection. SoundFace requires users to pose their faces directly to the acoustic devices, which makes it not difficult for them to find the right 90° angle. When the user significantly deviates from the devices, we can use our 2-array microphones to detect the face deflection angle and remind him to re-find the right facial angle.

User appearance changes. The current SoundFace is trained on limited data, far from exhaustive to be robust against users' appearance changes such as hats or glasses. To cope with such changes, retraining the SVM model with data of new appearances is an effective method, and such model updating will not introduce a large training overhead.

Time shift authentication. In a practical scenario, the performance of an FA system should be consistent across different days. The current version of SoundFace is implemented on the research-purpose platform Bela, which will inevitably undergo small variations when working for a long time. Such natural variation in devices may degrade

the performance over different days. To mitigate this problem, we can periodically fine-tune SoundFace with a small amount of the latest data to enhance the robustness across various days.

Location and height of acoustic transceiver. The location and height of speakers and microphones will have an impact on the performance of our scheme. Acoustic devices at different locations and heights collect different multipath reflections, and these new background interferences can introduce unstable acoustic signals and extracted facial features, thus affecting the performance of authentication. In SoundFace, the microphones are placed on the top of the speaker and mounted at a fixed height of 1.60 m with a tripod. We put them in fixed positions across three environments, and also ask users to face the acoustic transceiver for authentication. If the locations and heights of the speakers and microphones change, we can easily remind the users to realign their faces to the device and fine-tune SoundFace with a small amount of newly collected data to mitigate performance degradation.

Comparison with camera-based FA. We compare the camera-based FA with acoustic signals-based FA from two aspects. (1) Implementation and deployment: Most FA systems collect the user's facial features from RGB cameras, which are widely deployed in mobile products. Despite their widespread popularity, cameras have security problems like spoofing attacks and privacy leakage risks, which are also vulnerable to poor light conditions. The latest camera-based FA systems, such as Apple's Face ID, take defensive measures by installing a dot projector and an infrared depth sensor within a small area to perceive the 3D structure of the face. However, depth cameras and dot projectors are expensive and not widely adoptable in most devices. In contrast, acoustic components have also been widespread in smart devices, and the acoustic signals-based FA can effectively overcome the limitations of cameras. (2) Accuracy: The majority of FA techniques are designed atop vision-based image processing, e.g., FaceNet [29] and VGG-Face [25], which can achieve an accuracy of up to 99.63%. Comparatively, our acoustic signals-based FA of 96.2% accuracy is sufficient for daily authentication applications and can be an enhancement to multi-modal FA.

6 CONCLUSION

This paper presents a face authentication scheme named SoundFace based on acoustic signals for home appliances. SoundFace extracts facial geometry features from the acoustic signals reflected by human faces to achieve reliable face authentication, and recognition of legitimate users. Extensive evaluations demonstrate that SoundFace is resilient to various real-world settings with high accuracy and robustness.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. (62302439, U23A20296), Fundamental Research Funds for the Central Universities (226-2024-00004), Key Research and Development Program of Zhejiang Province (No. 2024C01065), and also in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 of RT14/22.

REFERENCES

[1] Amazon. [n. d.]. Smart Thermostat. https://www.amazon.com/ecobee-SmartThermostat-Voice-Control-Black/dp/B07NQT85FC?th=1.

[2] Apple. 2023. Apple HomePod. https://www.apple.com/shop/buy-homepod/homepod.

[3] Bela. 2024. Bela Platform. https://learn.bela.io.

[4] Huangxun Chen, Wei Wang, Jin Zhang, and Qian Zhang. 2019. Echoface: Acoustic sensor-based media attack detection for face authentication. *IEEE Internet of Things Journal* 7, 3 (2019), 2152–2159.

[5] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. 2001. One-class SVM for learning in image retrieval. In *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*, Vol. 1. IEEE, 34–37.

[6] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. 2016. Face recognition systems under spoofing attacks. *Face Recognition Across the Imaging Spectrum* (2016), 165–194.

[7] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 147–159.

[8] Habiba Farrukh, Reham Mohamed Aburas, Siyuan Cao, and He Wang. 2020. FaceRevelio: a face liveness detection system for smartphones with a single front camera. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[9] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[10] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. Sonicface: Tracking facial expressions using a commodity microphone array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.

[11] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. 2020. Learning meta face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6163–6172.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. pmlr, 448–456.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Chenqi Kong, Kexin Zheng, Yibing Liu, Shiqi Wang, Anderson Rocha, and Haoliang Li. 2024. M3FAS: An Accurate and Robust MultiModal Mobile Face Anti-Spoofing System. *IEEE Transactions on Dependable and Secure Computing* (2024).

[16] Chenqi Kong, Kexin Zheng, Shiqi Wang, Anderson Rocha, and Haoliang Li. 2022. Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. *IEEE Transactions on Information Forensics and Security* 17 (2022), 3238–3253.

[17] Hao Kong, Li Lu, Jiadi Yu, Yingying Chen, Linghe Kong, and Minglu Li. 2019. Fingerpass: Finger gesture-based continuous user authentication for smart homes using commodity wifi. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 201–210.

[18] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. Lasense: Pushing the limits of fine-grained activity sensing using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[19] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. Room-scale hand gesture recognition using smart speakers. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 462–475.

[20] Xiaopeng Li, Fengyao Yan, Fei Zuo, Qiang Zeng, and Lannan Luo. 2019. Touch well before use: Intuitive and secure authentication for iot devices. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[21] Chengwen Luo, Zhongru Yang, Xingyu Feng, Jin Zhang, Hong Jia, Jianqiang Li, Jiawei Wu, and Wen Hu. 2021. Rfaceid: Towards rfid-based facial recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–21.

[22] Davide Maltoni, Dario Maio, Anil K Jain, and Salil Prabhakar. 2009. *Handbook of fingerprint recognition*. Springer Science & Business Media.

[23] Mirlori. 2023. Gaze In, Find Your Beauty. https://mirlori.com/products/led-bathroom-vanity.

[24] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, and KJ Ray Liu. 2021. Gait-based people identification with millimeter-wave radio. In *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*. IEEE, 391–396.

[25] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*.

[26] Raspberry Pi. 2023. *Raspberry Pi 4 Model B*.

[27] Raspberry Pi Camera. 2023. *Raspberry Pi Camera Module 2*.

[28] Samsung. 2023. Smart TV | voice-assistants. https://www.samsung.com/us/tvs/smart-tv/voice-assistants/.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face face, and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[30] Statista. 2023. Smart Home | Worldwide | Statista Market Forecast. https://www.statista.com/outlook/dmo/smart-home/worldwide.

[31] Xue Sun, Jie Xiong, Chao Feng, Wenwen Deng, Xudong Wei, Dingyi Fang, and Xiaojiang Chen. 2023. Earmonitor: In-ear motion-resilient acoustic sensing using commodity earphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–22.

[32] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the security of graphical passwords: The case of android unlock patterns. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 161–172.

[33] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[34] Deepak Vasisht, Guo Zhang, Omid Abari, Hsiao-Ming Lu, Jacob Flanz, and Dina Katabi. 2018. In-body backscatter communication and localization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 132–146.

[35] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.

[36] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[37] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. An ear canal deformation based continuous user authentication using earables. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 819–821.

[38] Yadong Xie, Fan Li, Yue Wu, Huijie Chen, Zhiyuan Zhao, and Yu Wang. 2022. Teethpass: Dental occlusion-based user authentication via in-ear acoustic sensing. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1789–1798.

[39] Lilin Xu, Keyi Wang, Chaojie Gu, Xiuzhen Guo, Shibo He, and Jiming Chen. 2024. GesturePrint: Enabling User Identification for mmWave-based Gesture Recognition Systems. In *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1074–1085.

[40] Weiye Xu, Jianwei Liu, Shimin Zhang, Yuanqing Zheng, Feng Lin, Jinsong Han, Fu Xiao, and Kui Ren. 2021. RFace: anti-spoofing facial authentication using cots rfid. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[41] Weiye Xu, Wenfan Song, Jianwei Liu, Yajie Liu, Xin Cui, Yuanqing Zheng, Jinsong Han, Xinhuai Wang, and Kui Ren. 2022. Mask does not matter: Anti-spoofing face authentication using mmWave without on-site registration. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 310–323.

[42] Xiang Xu, Nikolaos Sarafianos, and Ioannis A Kakadiaris. 2020. On improving the generalization of face recognition in the presence of occlusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 798–799.

[43] Zhenyu Yan, Qun Song, Rui Tan, Yang Li, and Adams Wai Kin Kong. 2019. Towards touch-to-access device authentication using induced body electric potentials. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[44] Xiaojing Yu, Zhijun Zhou, Mingxue Xu, Xuanke You, and Xiang-Yang Li. 2020. Thumbup: Identification and authentication by smartwatch using simple hand gestures. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE Computer Society, 1–10.

[45] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can " Hear" Your Heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[46] Bing Zhou, Zongxing Xie, Yinuo Zhang, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2021. Robust human face authentication leveraging acoustic sensing on smartphones. *IEEE Transactions on Mobile Computing* 21, 8 (2021), 3009–3023.