Face Recognition In Harsh Conditions: An Acoustic Based Approach

Yanbo Zhang
Nanyang Technological University (NTU)
Singapore
yanbo001@e.ntu.edu.sg

Panrong Tong
Alibaba Group
Hangzhou, China
panrong.tpr@alibaba-inc.com

Songfan Li
Hong Kong University of Science and Technology (HKUST)
Hong Kong, China
lisongfan@ust.hk

Yaxiong Xie
University at Buffalo SUNY
New York, United States
yaxiongx@buffalo.edu

Mo Li
HKUST and NTU
Hong Kong, China
lim@cse.ust.hk

ABSTRACT

The accuracy of vision-based face recognition suffers in challenging scenarios, such as foggy or smoky weather, poor lighting, and blockage by objects like facial masks. This paper proposes an acoustic-based facial recognition system based on acoustic facial spectrum – a novel acoustic representation of human faces in 3D space. Specifically, we divide the 3D space into cubes and profile the distribution of the acoustic signal reflected by the human face inside each cube. Generating such a per-cube acoustic profile is challenging in relating each reflected signal path back to the physical location of its reflecting cube. To address the challenge, we propose a novel multipath resolving algorithm that is capable of distinguishing signal reflection happened within different cube. Based on the facial spectrum, we propose a discriminator-recognizer network that can robustly recognize human faces under varying face-microphone distances or even in presence of facial mask blockage. Extensive experimental results demonstrate that the proposed system achieves over 95% average recognition accuracy for cases with and without mask blockage. The research artifacts accompanying this paper are available via DOI: 10.5281/zenodo.11094213.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Security and privacy → Security services**.

KEYWORDS

Acoustic sensing, Face recognition, Discriminator network

ACM Reference Format:
Yanbo Zhang, Panrong Tong, Songfan Li, Yaxiong Xie, and Mo Li. 2024. Face Recognition In Harsh Conditions: An Acoustic Based Approach. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3643832.3661855

1 INTRODUCTION

With the advent of digital era, intelligent identification techniques plays an important role in supporting identity sensitive authorizations such as safety entrance, secure payment, and authenticated device access. Conventional human identification schemes are designed based on explicit human biometrics such as fingerprint [16], iris [14], and voiceprint [26, 45], or implicit physical characters like bone density [39] and vocal tract structure [20]. The individual unique habitual behaviors such as lips movement [19, 31] or breath dynamics [7] are recently explored for improved accuracy, among which facial biological information [28, 44] is particularly favored due to its high identification accuracy and long-term stability.

Vision-based techniques use photosensitive elements (*e.g.*, CMOS or CCD) to record facial reflected visible light, which carries the fine-grained spatial (*e.g.*, face contour, mouth shape) and frequency (*e.g.*, skin color) information due to its small wavelength (nanometer scale) and high bandwidth (terahertz bandwidth), respectively. On the other side of the coin, the nature of visible light also makes it extremely sensitive to challenging environmental conditions, such as foggy or smoky weather, poor lighting or blockage by objects like facial masks. Vision-based face recognition also raises privacy concerns [12] regarding the leakage of sensitive personal information including gender, age, skin color, which may be leveraged by malicious third parties [43].

Wireless sensing technique, when applied for face recognition, naturally alleviates the two problems of the vision-based approach, since wireless signals (e.g., infrared signal, millimeter wave or acoustic signals) exhibit remarkable resistance to environmental or lighting conditions and their intrinsic physical properties prevent them from carrying any sensitive identity information. Apple Face ID projects thousands of infrared dots on human face which are then captured by a infrared camera for facial depth map construction [1]. This method however requires expensive cameras and chips, making it unsuitable for most scenarios that require low-cost deployment (e.g., smart access control). mmFace [38] implements a facial authentication system by leveraging a SAR-based millimeter-wave radar for facial imaging. EchoPrint [44] combines acoustic signal

Mo Li is the corresponding author.

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06
https://doi.org/10.1145/3643832.3661855

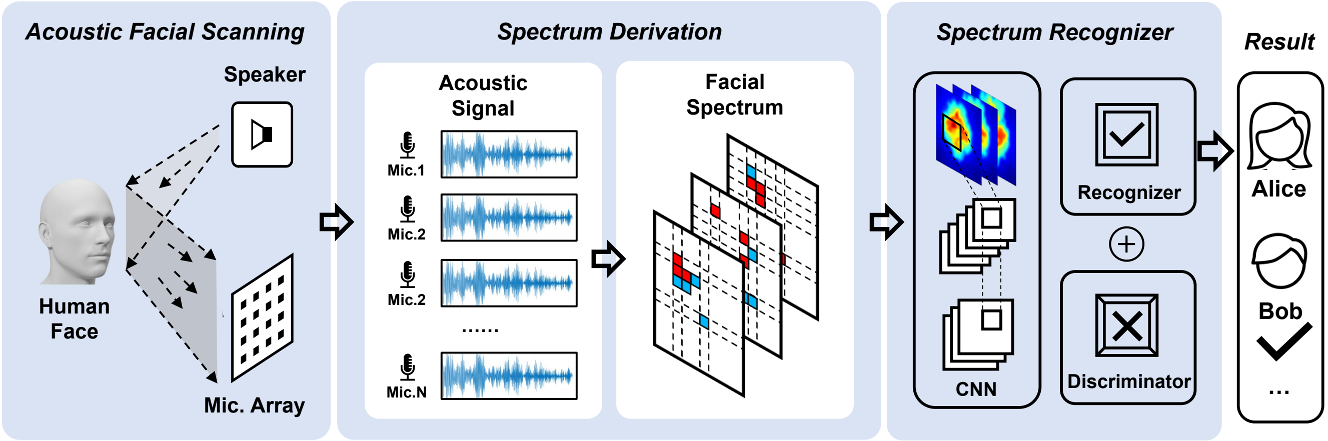


Figure 1: AcFace derives facial spectrum to fingerprint the spatial characteristics of human face and utilizes a recognizer-discriminator network to achieve accurate and robust spectrum recognition.

processing with vision-based image processing to achieve facial authentication. RFID signal [21, 37] is also used to distinguish different faces and combat spoofing attacks. These systems, however, either require expensive dedicated hardware support (e.g., mm-Face [38]) or inherent vulnerability of vision-based techniques (e.g., EchoPrint [44]).

In this paper, we propose AcFace, the first facial recognition system that purely exploits acoustic signals to achieve accurate recognition, even with facial features that are partially obstructed by objects such as facial masks. When compared with RF signals, acoustic signals have much lower frequencies (kHz-level) and propagate by means of particle displacement, exhibiting better obstacle penetration capability at near-field [36]. Due to the low propagation speed, acoustic signal can achieve millimeter-level ranging resolution with low-cost ADC (kHz-level sampling rate would be enough) and thus has the potential of characterizing detailed facial features, e.g., the height of the nose bridge or the depth of the eye socket, by using inexpensive acoustic hardware.

To recognize the human face, we propose the *acoustic facial spectrum*: a novel acoustic representation of the human face that preserves rich facial features. Specifically, we divide the free 3D space into cubes and profile the distribution of the signal reflected by the human face inside each cube (an empty cube containing no human face has no reflections). Depending on the shape of the human face inside each spatial cube, the acoustic signal of one incident angle will be reflected towards various reflection angles. Therefore, the spatial distribution of the signal reflection that happens within one cube characterizes the 3D surface of the human face inside that cube. Accordingly, the combination of all the cubes in the 3D space fingerprints the entire human face.

Obtaining the signal distribution of distinct cubes requires categorizing all the reflected paths into separate cubes. This is challenging because the signal reflected from different cubes does not present any distinguishable signal features. To address this challenge, we utilize a microphone array to observe signals from a certain reflection cube from multiple spatial locations. Consequently, the reflected signals of a certain cube acquired distinctive features based on their measurements across different microphone positions, including signal strength and path delay. Leveraging these features, we developed a cube multipath resolving technique to identify the reflections inside each cube and further combine the spatial characteristics of all the cubes to derive the facial spectrum.

To further alleviate the impact of mask blockage and varying face-microphone distances, we design a deep neural network to extract transferable features from various facial spectrums collected across different domains (i.e., different distances with and without masks) by playing a minimax game to minimize the impact of the two factors while maximizing the weight of the transferable facial features. The proposed model is able to achieve accurate and robust face recognition without prior domain knowledge.

We build a complete end-to-end system that prototypes the above design. The system is implemented using purely commercial low-cost acoustic hardware and machine learning model hosted on a general PC with mainstream NVIDIA GPUs. We experimentally evaluate the performance of the proposed system with comparison to state-of-the-art vision based techniques. The results demonstrate that our system generally provides comparable accuracy with vision based solutions, and performs even better in challenging scenarios (e.g., users with mask, low lighting condition).

The contribution of this paper is summarized as follows:

- An acoustic facial spectrum is proposed and utilized to profile the human faces, which is accurate and more robust than vision based solutions in harsh application conditions.
- A novel signal processing technique is devised to resolve the multipath reflected by different facial areas and produce fine-grained facial spectrum.
- A deep neural network model with domain adaptivity is designed, which is able to accurately recognize facial spectrums collected with different use conditions, even those with facial mask blockage.

The rest of this paper is structured as follows. Section 2 provides an overview of the system design. Section 3 details the feature that we adopt for facial recognition and defines the facial spectrum.

Section 4 and 5 presents our core designs for the derivation and recognition of the facial spectrum. Section 6 describes the prototype implementation. Section 7 details experimental evaluation results. Section 8 discusses limitations and possible solutions. Section 9 summarizes the related work and Section 10 concludes this paper.

2 SYSTEM OVERVIEW

The AcFace system consists of two main blocks. The first block identifies multipath reflections from different facial areas and recombines them to derive the full spectrum. The second block utilizes the recognizer-discriminator network to achieve accurate and robust spectrum recognition. Figure 1 illustrates the collaboration between the functioning blocks.

In the following sections, we first provide the specific definition of facial spectrum in Section 3, and then detail the technical designs that support the operation of each functioning block in Section 4 to 5.

3 FACIAL SPECTRUM

Energy distribution of scattered signals. The human head is capable of reflecting the acoustic signal. To quantitatively capture the signal reflections, we grid the 3D space where a human head locates into $M \times N \times Q$ cubes. Depending on the 3D surface of the human face inside each cube (an empty cube has no reflections), the acoustic signal from one incident direction could be scattered into various departure directions, as illustrated in Figure 2. The energy distribution of the scattered acoustic inside each cube characterizes the facial area inside that cube. We, therefore, leverage the energy distribution of all cubes inside the whole 3D space to fingerprint the human face.

Facial spectrum. We sample the signal scattered from one cube at multiple spatial points with a microphone array, coherently combine the signal collected from all the microphones, and calculate the strength of the combined signal, as shown in Figure 3. We repeat the above process to derive the *facial spectrum*: a $M \times N \times Q$ matrix, where each element represents the strength of the combined signal of one cube. The facial spectrum preserves rich facial features and thus is efficient at differentiating the human faces.

Challenge. We face a major challenge when generating the highly accurate facial spectrum. The scattered signals from all the $M{\times}N{\times}Q$ cubes superimpose at the microphone array. We must resolve the signal from each cube in order to derive the facial spectrum. This is however difficult because the signal reflected from different cubes does not present any distinguishable signal features.

4 RESOLVING CUBIC MULTIPATH

In this section, we detail the design of resolving the multipath signal that is scattered from different cubes. The processing is composed of two steps – multipath decomposition and multipath re-combining, as illustrated in Figure 4. The first step decomposes the delayed echoes at each microphone from time domain with FMCW-based multipath separation. The second step identifies a set of echoes that are reflected from a certain facial area and combines them in a coherent way. We detail each step separately in the following subsections.

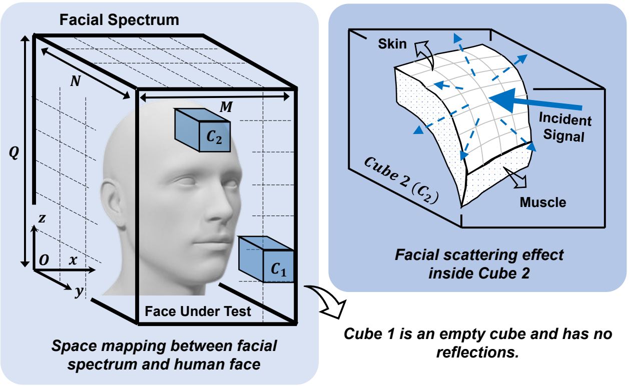

Facial Spectrum
N
M
Q
C_2
C_1
z
O
x
y
Face Under Test
Space mapping between facial spectrum and human face
Skin
Incident Signal
Cube 2 (C_2)
Muscle
Facial scattering effect inside Cube 2
Cube 1 is an empty cube and has no reflections.

Figure 2: Illustration of the 3D space mapping from human face to facial spectrum (the left part) and the facial scattering effect inside a cube (the right part).

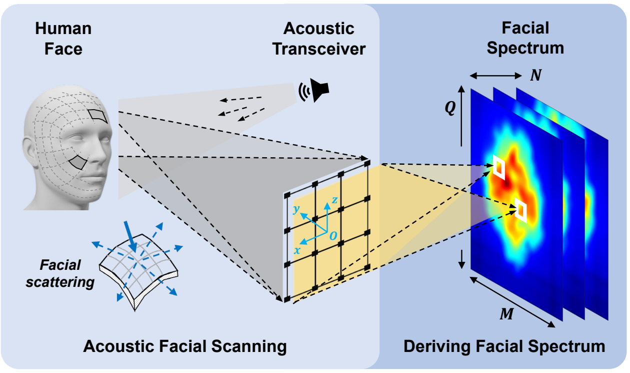


Figure 3: The scattering effect of different facial areas superimpose at the single microphone array. We need to disentangle the multipath superposition to derive the facial spectrum.

4.1 Multipath Decomposition

We transmit consecutive FMCW symbols for facial scanning. The signal will be reflected by different facial areas which further generates numerous multipath components with varying travelling distances towards different directions. The multipath signals finally superimpose at each microphone. We perform FMCW-based multipath separation on the raw receiving of each microphone to separate the superimposed multipath components from time domain. By following standard FMCW ranging processing [32], we can obtain the beat signal, which embodies the propagation delays of different multipath components. The beat signal $y_{beat}(t)$ can be formulated as below (refer [32] for the detailed derivation),

$$y_{beat}(t) = \sum_{i=1}^{N} A_i e^{j(2\pi s\tau_i t - \pi s\tau_i^2 + 2\pi f_{init}\tau_i)} \tag{1}$$

where A_i and τ_i denote the amplitude and propagation delay of the i-th component, respectively. We use s and f_{init} to denote the slope and starting frequency of an FMCW chirp.

As we can see from Equation 1, the derived beat signal is the weighted summation of multiple single tones oscillating at the

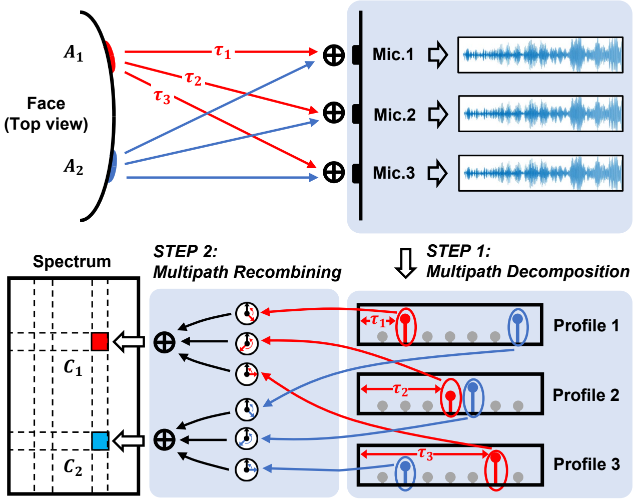


Figure 4: The derivation of facial spectrum reverses the effect of facial scattering and multipath superposition by devising the two-step processing.

frequencies of $s\tau_i$. With performing FFT on the beat signal, we are able to obtain the beat spectrum where the signal with a certain time delay (e.g., τ_i) is identified as the corresponding beat frequency (i.e., $s\tau_i$). The beat spectrum is further scaled to derive the *multipath profile* where each component represents one reflected path of amplitude A_i and delay τ_i. Specifically, the multipath profile can be formulated as below,

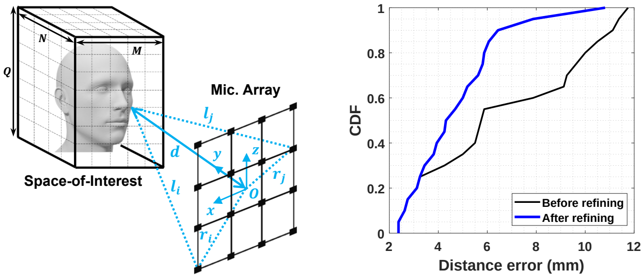
$$h(t) = \sum_{i=1}^{N} p_i \delta(t - \tau_i) \tag{2}$$

where $p_i = A_i e^{j(-\pi s \tau_i^2 + 2\pi f_{init} \tau_i)}$ represents the phasor induced by the i-th path and $\delta(t - \tau_i)$ denotes the delayed impulse function. Basically, the profile is characterized by a series of facial scattered paths with different amplitude, delay and phase rotation.

4.2 Multipath Re-combining

We construct the facial spectrum via multipath re-combining. Specifically, the re-combining process contains the following two steps. First, the *multipath identification* where we identify all the scattered multipath signals inside the multipath profiles whose reflection points are inside the same cube of the facial spectrum, as shown in Figure 2. Second, the *coherent combining* where we align the phases of all the identified multipath components and add them up to derive the facial spectrum.

Multipath identification. We identify the multipath signals from the same cube according to the *time-of-flight* of the signal. Figure 4 illustrates the idea of multipath identification with an example on a 2D plane. Given the location of the speaker and the microphones, we are able to calculate the propagation time T_{c_i,m_j} the signal takes to travel from the speaker, gets reflected by the human face inside i-th cube, and at last reaches j-th microphone inside the array. Then, we match the calculated T_{c_i,m_j} with the estimated time-of-flight of the signals inside the multipath profile to identify the signal that gets reflected within i-th cube. By repeating the above process for every microphone, we obtain a sampled distribution of the scattered signal from i-th cube:



(a)

(b)

Figure 5: The idea (a) and result (b) of locating Space-of-Interest (SoI).

$$S_{c_i} = [s_{c_i,m_1}, s_{c_i,m_2}, \cdots, s_{c_i,m_G}] \tag{3}$$

where s_{c_i,m_j} represents the signal that gets reflected within i-th cube and finally reaches j-th microphone. We have in total G microphones inside the array. We note that the signal s_{c_i,m_j} is empty if we cannot identify any reflections from the profile.

Theoretically, the signal strength $P_{s_{c_i,m_j}}$ of every signal s_{c_i,m_j} characterizes the energy distribution of the scattered signal and thus can be used for facial recognition. Accordingly, our facial spectrum becomes a tensor with the size of $M \times N \times Q \times G$. We note that processing such a four-dimensional tensor results in significant computational overhead, so we propose to coherently combine the identified G signals received and leverage the energy of the combined signal as our feature.

Coherent combining. To combine the selected path coherently, we reverse the phase rotation caused by the propagation delay of each path (as illustrated in Eq. 1). The processing can be formulated as below,

$$P(x_{F_i}, y_{F_i}, z_{F_i}) = \sum_{j=1}^{G} s_{c_i,m_j} e^{j(\pi s T_{c_i,m_j}^2 - 2\pi f_{init} T_{c_i,m_j})} \tag{4}$$

where $P(x_{F_i}, y_{F_i}, z_{F_i})$ represents the reflected energy inside the i-th cube. The processing is illustrated in Figure 4.

4.3 Locating Space-of-Interest

We define the space where we generate the facial spectrum as Space-of-Interest (SoI). Accurately locating the SoI is crucial for deriving an effective facial spectrum. If the SoI is shifted away from the actual position of the test face, the derived spectrum would contain a lot of empty cubes which contribute less feature. To locate the SoI, it is essential to measure the distance between the test face to the microphone array. We design a coarse-to-fine algorithm for this purpose.

We first estimate the distance by averaging the distances measured at all microphones, i.e., $d_r = \frac{1}{M}\sum_{i=1}^{M} d_i$. As shown in Figure 5a, for each microphone, the distance is calculated based on the Pythagorean equation of the space geometry. Specifically, $d_i = \sqrt{l_i^2 - r_i^2}$ where r_i denotes the Euclidean distance from i-th microphone to the center of the array (i.e., the coordinate origin), and l_i is approximated by taking half of the path length which exhibits the

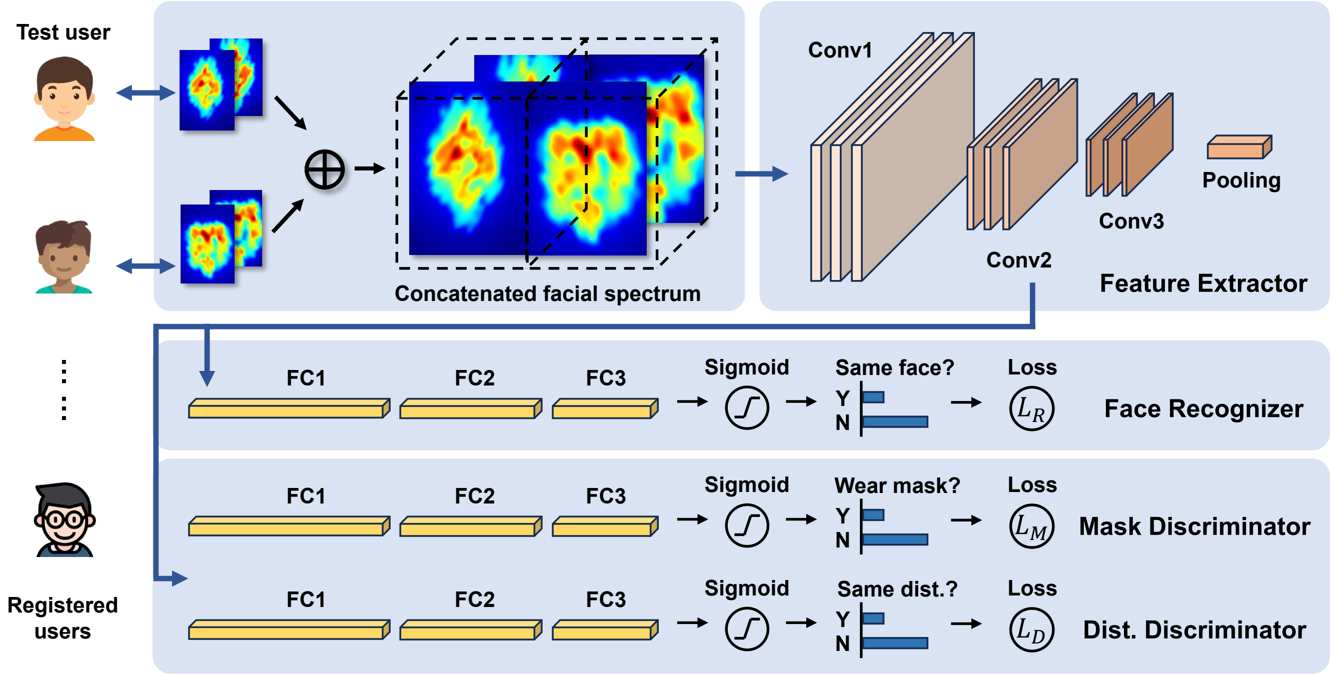


Figure 6: Structure of the proposed RD-Net.

highest power, i.e., $l_i \approx c\tau_{peak}/2$ as the speaker is closely located with the array.

The approximated result is further refined by leveraging the fact that the SoI contains more non-empty cubes if it is located more accurately. With such idea, we search over N cubes within the SoI located at each candidate distance $\hat{d}$ where $\hat{d} \in \left[d_r - \frac{\eta}{2}, d_r + \frac{\eta}{2}\right]$ (η defines the window length), and count the number of non-empty cubes for each distance with an empirical threshold ϵ representing the lowest power to be counted as non-empty. Finally, we adopt the distance that yields the maximal non-empty cubes.

Figure 5b benchmarks the accuracy of such algorithm against the ground-truth distances obtained from 15cm to 40cm at the step of 1cm. After refining, the distance error is less than 5mm in median which demonstrates the super resolution of the proposed algorithm. Even without refining, the coarse estimation still provides sub-centimeter accuracy and thus we may omit the second step for faster processing.

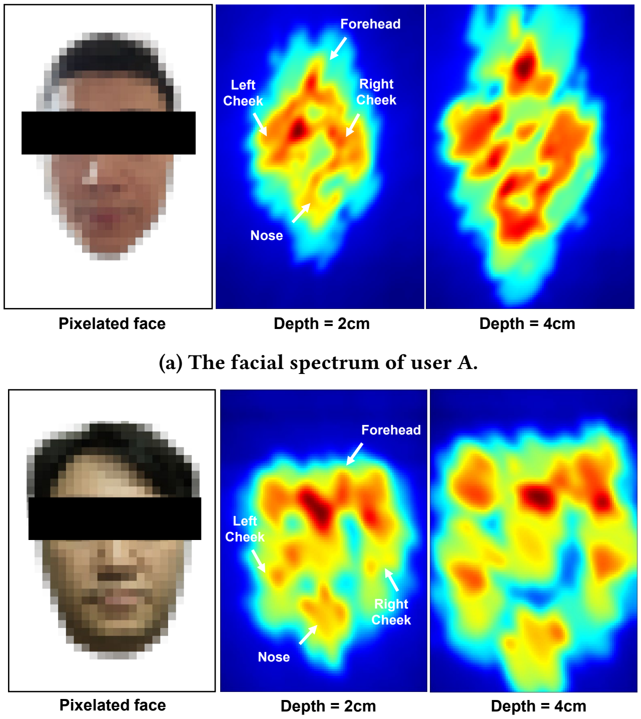
5 ROBUST FACE RECOGNITION

With the above processing, we derive facial spectrum which presents diverse spatial characteristics of human face. In figure 7, we showcase two example facial spectrums from different users (we provide anonymized facial images as ground-truth reference). For each spectrum, we plot the 2D energy distribution at different depth (i.e., different N) of 2cm and 4cm. From the spectrum of user A, we can identify several key facial landmarks such as the forehead, cheeks and nose. We also observe that the power disperses over larger area as the depth increases, which may be resulted from the increased facial area. The spectrum of user B also exhibits identifiable facial landmarks and manifests the dispersed energy distribution. When comparing the spectrum of the two users, we can clearly identify the different facial features including the facial contour, the locations of facial landmarks, power scattering at the landmarks, etc. The various features together characterize the appearance of a certain face.

In this section, we design a deep neural network to identify different users with their corresponding facial spectrums. In addition to the basic function of feature extraction and recognition, we adopt the discriminator network design [13, 17] to further alleviate the impact of facial mask blockage and varying face-microphone distance.

Impact of mask blockage and distance variation. To investigate the impact of facial mask, we process the raw acoustic samples collected with the same two users when they are wearing a surgical facial mask and derive their corresponding facial spectrums with mask blockage. The result is illustrated in Figure 8. When compared with normal scenario of bare face, for both users, we observe enhanced energy intensity over the region that the facial mask covers, which may be resulted from the combining between the signal paths scattered by facial mask and by the facial landmarks underneath. Fortunately, we observe that for both users, the essential facial features (e.g., facial contour and the upper facial landmarks) persist with similar energy distribution.

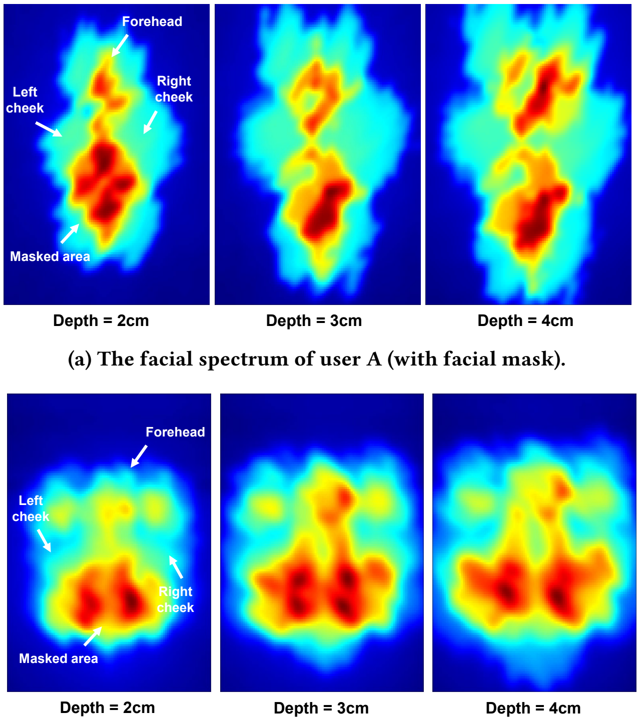
Additionally, distance variation also affects the facial spectrum in its energy distribution. Specifically, longer distances result in a reduced overall size of the facial spectrum, manifesting a "zoom out" effect. This reduction is particularly evident in the diminished signal strength at the facial spectrum's edges, which may due to a



(a) The facial spectrum of user A.

(b) The facial spectrum of user B.

Figure 7: Illustration of example facial spectrums derived with two different users. The 3D spectrum is presented as 2D energy distribution at different depths (i.e., different N in Figure 2). The anonymized facial images of the two users are provided as ground-truth.



(a) The facial spectrum of user A (with facial mask).

(b) The facial spectrum of user B (with facial mask).

Figure 8: Facial spectrum derived for user A and user B when they wear facial mask. Although the mask creates strong reflections that bury the lower partial facial landmarks, the facial contour and the upper partial landmarks still persist and can be leveraged to enhance the robustness to mask blockage.

decreased capacity of the microphone array to capture signals from the facial periphery at longer distances.

Based on the observations above, we believe that a signature matching network can be utilized to compare two facial spectrums and determine their similarity, thereby identifying facial identity information. Furthermore, since the primary features atop the spectrums remain recognizable despite wearing masks and changes in distance, it is feasible to eliminate the impact of these two factors during facial recognition. We will achieve this through a recognizer-discriminator based network. In the following content of this section, we will provide a detailed explanation of the network model.

Network design. We design a recognizer-discriminator network (RD-Net) to extract the persistent features and avoid the impact of the two factors. The network compares the facial spectrum of a test user with those of all registered users. The spectrum is first fed into a CNN, which we utilize to exploit the spatial diversity and extract the 3D feature shaped by a human face. A recognizer network is then used to judge the similarity between the derived feature maps, and at the same time a discriminator network is adopted to alleviate the impact from the two impact factors.

Figure 6 illustrates the structure of the proposed model. The network contains a CNN based feature extractor, a deep NN based recognizer and two discriminators. The CNN model contains three convolution layers, each using 3D kernels as filters, and the filtered channels are normalized with a batch norm layer. We avoid using down sampling process (e.g., dropout or pooling) to maintain complete feature space. The extracted features are flattened and fed into the face recognizer, which uses five fully connected layers to investigate the similarity between the feature maps derived from two facial spectrums.

The result, as we earlier mentioned, may be impacted by facial mask reflections and varying distance. Two discriminator networks are used to alleviate such impact. The discriminators use duplicated network structure from the recognizer but handle different tasks. Specifically, the mask discriminator aims at identifying whether the two facial spectrums are with the same mask condition (i.e., with or without mask), and the distance discriminator determines whether the two spectrums are captured at the same distance.

Loss design. The feature extractor, recognizer and discriminators are jointly optimized with a carefully designed loss function. To achieve robust face recognition with high accuracy, we need to enhance the recognizer's ability to identify the similarity of two facial spectrums, and at the same time to reduce the discriminator's sensitivity to different mask conditions and distances. To achieve so, the training process aims at minimizing the loss of face recognizer (L_R), but maximizing the losses of the two discriminators (L_M

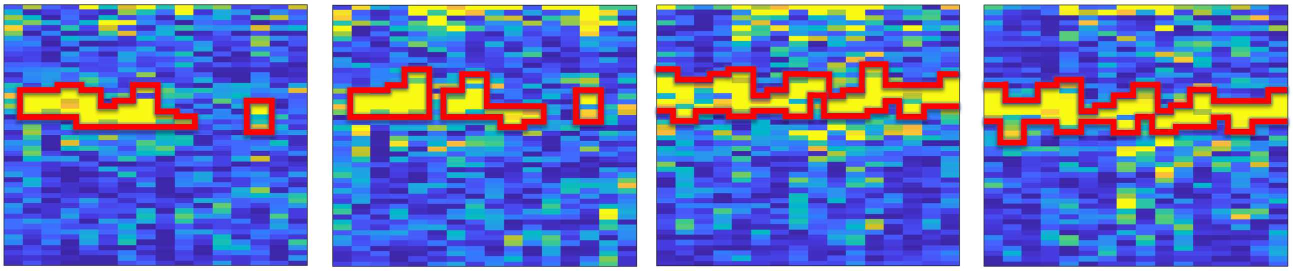


Figure 9: From left to right: the lightweight facial spectrum of user A at the environment of (a) meeting room and (b) office, as well as that of user B at the (c) meeting room and (d) office. The facial reflections of the same user create similar area of high intensity (enclosed by the red lines) between the different environments.

and L_D). We define the loss function of the integrated network as follows,

$$L = L_R - \frac{\alpha L_M + \beta L_D}{2}, 0 \leq \alpha, \beta \leq 1 \quad (5)$$

where α and β are two coefficients that balance the significance of L_M and L_D respectively. With the above definition, the overall loss function (L) is oppositely related to the loss of two discriminators. By optimizing the network for minimized L, we therefore strengthen the accuracy of the face recognizer in various use conditions.

Network scalability. The proposed network is designed for improved user scalability, primarily by leveraging the signature matching approach. With this approach, the network compares a target user's facial spectrum with all registered spectrums of different users and derives a probability which quantifies the similarity in their identity. Once trained, the obtained signature matching capability can facilitate a seamless inclusion of new users without necessitating re-training, which avoids the limitation of conventional multi-class classification models – they require extensive retraining for every new user inclusion. The scalability is validated with our experimental results in Section 7.2.3.

6 IMPLEMENTATION

6.1 A lightweight method

Through the aforementioned designs, we can implement an end-to-end facial recognition system. However, the current system operates with a high computational cost, primarily due to the requirement for iterative calculations of the intensity of each pixel at the facial spectrum generation stage. In order to alleviate the computational cost of this system, we present a lightweight solution in the following.

The solution reduces computational costs by simplifying the signal processing of generating facial spectrum. Specifically, this approach directly concatenates the multipath profiles obtained from multiple channels to generate a simplified version of the facial spectrum (named lightweight facial spectrum in the following). Figure 9 illustrates the lightweight facial spectrums obtained from measurements of two users in different environments. Each column in the spectrum represents the temporal distribution of reflected signals measured on a specific channel, while each row represents the distribution of reflected signals across different channels at a specific temporal tap. The areas enclosed by the red lines highlight regions with relatively strong facial reflected signals.

By comparing the spectrum of the same user (Figure 9 (a) and (b), Figure 9 (c) and (d)), we observe similar pattern of the area of high intensity, which demonstrates that the spectrogram is able to capture consistent facial features of an individual, regardless of the environmental changes. By comparing the spectrum of different user (Figure 9 (a) and (c), Figure 9 (b) and (d)), we see that the area of high intensity shows distinguishable pattern and structure, which verifies the sensitivity of the facial spectrum to varying spatial features of human face.

The lightweight solution essentially circumvents the requirement for iterative computation on each pixel value during facial spectrum generation, resulting in a significant saving in computational costs. Section 7.2 provides a detailed comparison between the lightweight solution and the fully-implemented system on their recognition accuracy and computational expenses.

Feature selection. The lightweight facial spectrum may be affected by environmental reflections because this method does not have a particular design for identifying facial reflections. In order to eliminate the impact from environmental dynamics, we conduct a software based feature selection by trimming the multipath profile obtained from each microphone channel with a specific range limit (R_l). Specifically, given R_l, the length of each facial spectrum to be saved equals to $\lceil R_l / r \rceil$, where r is the ranging resolution. For our implementation, we use R_l = 50 cm, which defines 71 bins from each multipath profile.

6.2 Hardware and Software

Hardware prototype. The system is prototyped with an acoustic transceiver which is primarily comprised by an omni-directional speaker [29] and a four-by-four rectangular microphone array [23] as depicted in Figure 10. The array adopts 16 SPH1668LM4H MEMS microphones which are synchronized with an internal clock for acoustic signal reception, and is stacked with a MCHStreamer USB interface which achieves real-time streaming of the raw samples from the microphone array to a PC for post-processing. The distance between adjacent microphone element is 44mm and the array spans the area of 132mm×132mm. All microphones and the speaker sample the analog signal with 48 kHz sampling rate which provides 24 kHz frequency bandwidth.

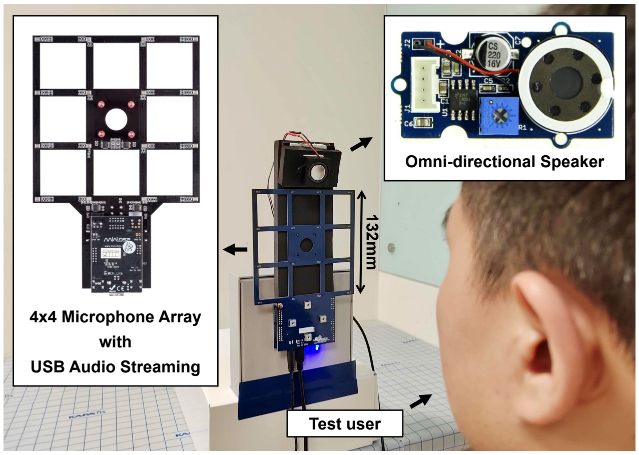


Figure 10: The hardware prototype of AcFace and experiment scenario.

Model training. The proposed recognizer-discriminator network model is implemented with PyTorch API and trained with Google Colab Pro+ GPU based compute engine where the allocated computational resource includes 16 GB GPU RAM (NVIDIA A100 Tensor Core GPU), 64 GB system RAM and 256 GB disk. The learning rate and batch size are set to 0.001 and 128 in respective for model training. The same learning rate and batch size are also adopted for training other models in our comparative evaluation (detailed in Section 7.2.1). The loss control coefficients α and β (defined in Equation 5) are empirically set to 0.03 and 0.02 to optimize the performance (the impact of α and β on recognition accuracy is detailed in Section 7.2.4).

Multipath profile calibration. The multipath profile obtained with the multipath decomposition process characterises the power and delay of each facial scattered path and plays an important role in deriving facial spectrum. However, the profile is shifted by an unknown time delay offset – In contrast to the front-end of mmwave FMCW radar [10] that incorporates a loop-back channel for providing a reference timing signal, the low-cost acoustic transceiver does not have built-in loop-back channel and thus cannot timing the arrival of signal path with absolute delay measurement. To solve this issue, we leverage the direct path leakage from the speaker to each microphone to conduct a calibration. Specifically, with the derived multipath profile, we align the first peak (the direct path always arrives as the first peak due to the shortest travel distance) to the actual time-of-arrival of the direct path which is pre-measured with a digital laser measure. The calibration is purely software based and is a one-shot effort as long as the speaker has a fixed location.

Facial spectrum sizing. The dimension of the facial spectrum space is determined by parameters M, N, and Q, as defined in Section 3. These parameters are empirically determined to ensure that the spatial dimensions comprehensively cover the typical size range of human faces. Additionally, the cube size within this space crucially influences the granularity of the facial spectrum, with smaller cubes theoretically enhancing granularity and providing more detailed information. However, it is important to note that despite the potential for increased granularity, the resolution of the facial spectrum is ultimately determined by the hardware bandwidth. For our current implementation, the dimension of facial spectrum is set to 30cm-by-25cm-by-5cm, and the cube size is set to 2mm.

7 EVALUATION

In this section, we evaluate the performance of AcFace with the prototype system. We begin by validating the derived facial spectrum by benchmarking its capability on capturing facial features. Following this, we conduct a series of end-to-end experiments to demonstrate the system's performance across various aspects and examine how different factors impact the performance.

Data collection. We conducted data collection with 15 volunteers. Each participant underwent a scanning procedure to capture their facial features using consecutive FMCW symbols. This process was performed twice for each individual: once with a mask and once without. During the scans, participants were instructed to gradually move their faces away from the acoustic transceiver, increasing the distance from 15cm to 35cm. Each scanning session lasted 90 seconds, enabling us to capture facial spectrums at various distances (The 90s facial scanning is required only for new user registration, and is not necessary for online testing). To guarantee comprehensive capture of all facial features, we maintained a perpendicular alignment between the center of the participant's face and the center of the microphone array throughout the scanning. The collected data, encompassing different users, distances, and mask conditions, were then paired to compile the final dataset. We randomized the dataset, allocating 80% for training and reserving 20% for testing. The data collection methodology received approval from the Institutional Review Board (IRB) of our university.

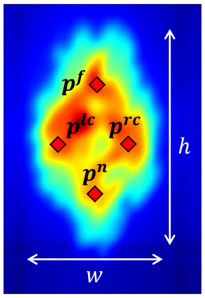
7.1 Facial Spectrum Validation

An effective spectrum should be a well representation of the essential facial features. In this section, we first define essential facial features, and then evaluate the spectrum's accuracy in representing these features, as well as its uniqueness among different users.

Essential facial features. We design a data structure $\mathbf{F} = \{\mathbf{s}, \mathbf{p}\}$, named *facialSpec*, to characterize essential facial feature obtained from 2D front face. Specifically, $\mathbf{s} = (\mathrm{w}, \mathrm{h})$ measures the width (w) and height (h) of a certain face, and $\mathbf{p}$ is a vector consisting of four 2D coordinates, i.e., $\mathbf{p} = \left[\mathbf{p}^{\mathrm{f}}, \mathbf{p}^{\mathrm{lc}}, \mathbf{p}^{\mathrm{rc}}, \mathbf{p}^{\mathrm{n}}\right]$ which represent the positions of four essential landmarks including the forehead, left cheek, right cheek and nose in respective. Figure 11 illustrates these parameters that are adopted for facial spectrum validation.

Metrics. Based on *facialSpec*, we introduce two metrics: *self-distance* (denoted as $\boldsymbol{\delta}$) and *cross-distance* ($\boldsymbol{\Delta}$). Self-distance measures the 2D distances between the positions identified from the spectrum and the pseudo-truth positions[1] which are obtained from an RGB facial image of the same user, and cross-distance measures the 2D distances between the positions estimated from different user's

[1] The positions obtained from facial images provide a benchmarking reference which however may not reflect the ground-truth positions of the facial landmarks, so we name the reference positions as pseudo-truth.



Facial contour	$s = (w, h)$
Position of essential landmarks	Center of forehead $p^f = (x^f, y^f)$
	Center of left cheek $p^{lc} = (x^{lc}, y^{lc})$
	Center of right cheek $p^{rc} = (x^{rc}, y^{rc})$
	Center of noise tip $p^n = (x^n, y^n)$
FacialSpec	$F = \{s, p\}$

Figure 11: Illustration of parameters included in defining essential facial features.

spectrums. Basically, self-distance evaluates the spectrum's capability in representing the facial features of a certain user, and cross-distance indicates the spectrum's capability in distinguishing the facial features among different users.

Specifically, the self-distance of the *i*-th user can be formulated as below,

$$\delta_{\mathrm{i}} = \|(\mathrm{d^s}_{\mathrm{i}}, \overline{\mathrm{d^p}}_{\mathrm{i}})\|_2 \tag{6}$$

where $\|*\|_2$ denotes the L2-norm of a certain vector, $\mathrm{d^s}_{\mathrm{i}} = \|\mathbf{s}_{\mathrm{i}}^{\mathrm{sp}} - \mathbf{s}_{\mathrm{i}}^{\mathrm{im}}\|_2$ represents the Euclidean distance of the *facial size feature* $\mathbf{s}_{\mathrm{i}}$ that is obtained from the spectrum ($\mathbf{s}_{\mathrm{i}}^{\mathrm{sp}}$) and the image ($\mathbf{s}_{\mathrm{i}}^{\mathrm{im}}$), and $\overline{\mathrm{d^p}}_{\mathrm{i}} = \overline{\|\mathbf{p}_{\mathrm{i}}^{\mathrm{sp}} - \mathbf{p}_{\mathrm{i}}^{\mathrm{im}}\|_2}$ represents the average of the Euclidean distances of the *landmark position features* $\mathbf{p}_{\mathrm{i}}$ that are extracted from the spectrum and image.

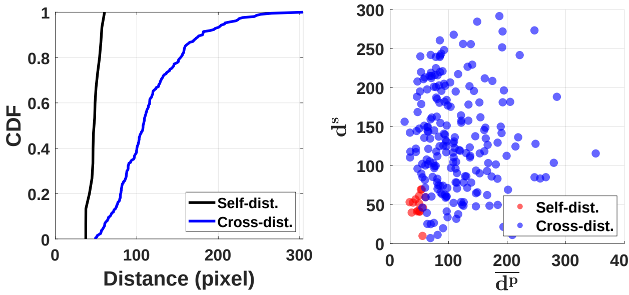
Similarly, the cross-distance between two different users (with index i and j, i ≠ j) is formulated as below,

$$\Delta_{\mathrm{ij}} = \|(\mathrm{d^s}_{\mathrm{ij}}, \overline{\mathrm{d^p}}_{\mathrm{ij}})\|_2 \tag{7}$$

where $\mathrm{d^s}_{\mathrm{ij}}$ and $\overline{\mathrm{d^p}}_{\mathrm{ij}}$ represents the same Euclidean distances as defined in Eq. 6 but are calculated between different users' facial spectrums.

Feature extraction. To extract the essential facial features and construct *facialSpec* of each user, we apply an edge detection algorithm [5] and a local maxima detection algorithm [11] to the grayscaled spectrum, based on which we further obtain the size of the face ($\mathbf{s}^{\mathrm{sp}}$) and the positions of essential landmarks ($\mathbf{p}^{\mathrm{sp}}$). To obtain the pseudo-truth for each user, we first re-size the RGB facial image to make it align with the dimension and the scale of the facial spectrum, and then process facial images with an open source landmark detection model [8] to obtain $\mathbf{F}^{\mathrm{im}}$. The size and positions are calculated from the row and column index of corresponding pixels.

Results. We demonstrate the efficacy of the spectrum by comparing the distribution of self-distance and cross-distance. Figure 12a depicts the CDF of the two distances where the median of self-distance is only 50 pixels whereas the median of cross-distance is over 100 pixels. The results indicate that the features extracted from the spectrum are closely distributed with that obtained from the facial image of the same user, and can be easily distinguished from the spectrum of a different user. Figure 12b further explains the result by demonstrating the distribution of the two distances. As we can see, the self-distance measurements locate at around 50-pixel distance occupying a narrow space, whereas the cross distances scattering over a large space which reveals the high deviation of the distribution.



(a) (b)

Figure 12: (a) The CDF of the self-distance and cross-distance and (b) the distribution of the distance measured over two different types of features (The feature of face size $\mathrm{d^s}$ and essential landmark positions $\overline{\mathrm{d^p}}$).

7.2 End-to-end Evaluation

We examine the end-to-end performance of AcFace in this section. We first demonstrate the face recognition accuracy of AcFace (with and without mask blockage), in comparison with the other three facial recognition approaches (one of them is also capable of recognizing masked faces). After the comparative evaluation, we further validate the robustness and scalability of the system. Specifically, we evaluate 1) the system's robustness to environment changes (with moving dynamics and background noise considered) and 2) the system's scalability to increased number of users. In addition, we investigate the efficacy of the discriminator-recognizer network design through an ablation study. The computational efficiency is evaluated at last.

Metrics. We evaluate the end-to-end performance with four metrics - *Precision*, *Recall*, *F1-score* and *Averaged Accuracy (AA)* according to their definitions for multi-category classification [30]. A higher precision indicates that the prediction is more accurate in its recognized face input, and a higher recall indicates that the system works more stable in not missing correct face input. F1-score gives an overall indication of the precision and recall. Averaged accuracy quantifies the overall accuracy of the prediction in all categories.

7.2.1 Comparative evaluation. We evaluate AcFace performance in comparison with three state-of-the-art vision based approaches, namely VGG-Face [6], FaceNet [28], and SRT [4]. Among the three comparative approaches, SRT uses a self-restrained triplet loss for the original ResNet-50 [15], and is capable of conducting face recognition with mask blockage. For fair comparison, we re-train the models with the same data size, learning rate and batch size. The dataset used for re-training is obtained from open face dataset VGGFace2 [6]. We use a software tool named MaskTheFace [3] to generate masked facial images.

The evaluation results are detailed in Table 1. While FaceNet provides the best performance (i.e., 98.86% accuracy and 98.79% precision), the performance sharply drops by near 15% when facial masks are applied, and by nearly 20% when used in dim environment. VGG-Face experiences similar performance drops during

Test setting	VGG-Face	FaceNet	SRT	AcFace
Without mask	98.05/97.73	**98.86 / 98.79**	98.81/97.06	95.88/96.12
With mask	83.16/83.25	85.63/86.66	**95.61 / 95.82**	**95.77 / 96.07**
With mask (dim)	77.67/77.32	78.11/79.67	81.57/83.79	**95.71 / 96.19**

Table 1: The averaged accuracy and precision of different face recognition approaches under three different test settings. AcFace outperforms the other three approaches in its robustness to mask blockage and poor lighting condition.

Environment	Precision (%)	Recall (%)	F1-score (%)	AA (%)
Meeting room	95.66/96.53	95.51/95.49	95.76/96.33	95.88/−
Lab	96.79/95.99	95.67/95.87	95.82/95.87	95.81/−
Office	95.29/95.90	94.82/95.83	94.66/95.86	95.45/−

Table 2: The performance of AcFace measure with different environments (stated with the average and median of the four metrics). The lab environment is with a lot ambient movements from people, and the open office is the most challenging due to its noisy and dynamic environment.

the test. SRT and our AcFace are able to provide comparable performance (i.e., 98.81/97.06% for SRT and 95.88/96.12% for AcFace) when tested without masks. Both can tolerate the facial mask blockage and provide comparable performance when the masks are applied (95.61/95.82% for SRT and 95.77/96.07% for AcFace). When further applied to dim environment where we reduce the ambient light AcFace stands out with reliable performance (i.e., 95.71/96.19% accuracy and precision) while SRT suffers over 10% loss and only achieves 81.57/83.79% accuracy and precision. The results suggest the high performance of AcFace when compared with vision based approaches, and in particular its comparative advantage when applied in harsh application conditions.

7.2.2 Different environments. We evaluate AcFace in practical indoor environment including a meeting room, a lab, and an open office. The meeting room is kept quite and clear without many people moving around. The lab environment is relatively quiet but with a lot ambient movements from people. The open office is the most challenging due to its noisy and dynamic environment. The measured noise levels in the three experimental environments are 33dB in a meeting room, 42dB in a lab, and 51dB in an open office. The test datasets for both the meeting room and the office are newly collected and are not included during model training.

Table 2 provides detailed results. The averaged accuracy is over 95% across experiments with the different environments. The Precision, Recall and F1-score also suggest high performance of AcFace even with the noisy and crowded office environment, which demonstrates the robustness of AcFace when applied to unfavoured scenarios. Although part of the facial scanning signal spans over the audible band, our system still achieves robust performance under noisy environment which primarily because of the noise-robust feature of FMCW chirp signal. The multipath identification process helps avoid the negative impact of human movement dynamics from background environment.

7.2.3 Scalability. We validate the scalability of AcFace by increasing the number of users of the test dataset from 10 to 15 and comparing the recognition accuracy. As summarized in Table 3, the accuracy remains stable (higher than 95%) with the increasing number of users. The reliable accuracy, even with unseen users, is resulted from two aspects: First, the proposed signal processing technique is carefully designed to construct facial spectrum that captures rich facial features, such as facial contours and important landmarks (e.g., forehead, cheeks and nose), as visualized in Figure 7 and Figure 8. Second, the design of RD-Net adopts the signature matching approach, which is inherently scalable to handle an increased number of users once the model has been adequately trained for signature recognition (unlike conventional multi-class classification model which requires re-training every time a new user is included).

Number of users	10	11	12	13	14	15
Ave. Accuracy (%)	95.88	95.33	95.97	95.01	96.13	95.67
Sign. matching delay (ms)	31.36	32.21	34.80	37.79	39.96	43.39

Table 3: The scalability of the proposed network model is validated by using a test dataset with increasing number of users. The averaged accuracy of AcFace remains stable for all the tests.

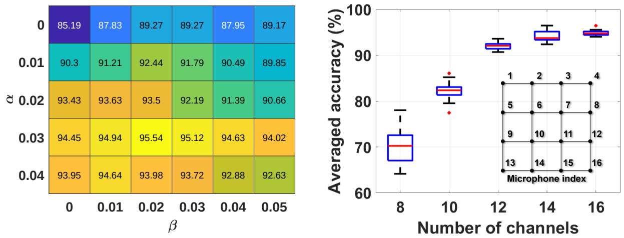


Figure 13: The impact of two internal parameters: (a) varying α and β – the RD-Net degrades to a normal CNN when α and β are set to zero. (b) different number of enabled audio channels.

The table also details the signature matching delay of the proposed RD-Net. As we see from the results, the delay increases from 31.36ms to 43.39ms when the number of users increases from 10 to 15, suggesting a linear increasing trending when user size becomes larger. This is due to the sequential execution of the signature matching process when it deals with different test samples. Nevertheless, the total delay of signature matching is only 43ms when there are 15 users. We may project to less than 5 second even when the user base increases to over 100. When applying parallel computing with clustered machine inference, the system can easily scale up to handle hundreds of thousands of users at second level.

7.2.4 Internal impact factors. We consider two impact factors that are internal to our system – 1) the selection of α and β and 2) the number of audio channels. We detail our evaluation as below.

The selection of α and β. The two parameters are used to weight the impact of mask blockage and distance variation, and need to be adjusted to balance their significance. We train and test the model over multiple rounds with using different α and β. Figure 13a shows the results. When α and β are set to zero, the model degrades into a normal CNN network (basically an ablation study where the discriminators are removed from the network). In such a case, the averaged accuracy is about 85.19% because the model can hardly recognize masked samples. With different settings, we obtain the optimal performance when $\alpha = 0.03$ and $\beta = 0.02$, which provides

Implementation	Spec. Gen. Time	Size	Train Time	Infer Time	Acc.
Full system	4.68s	85.14MB	5.06ms	1.45ms	95.64%
Lightweight	3.07ms	46.98MB	4.29ms	1.31ms	91.51%

Table 4: The computational effciency and accuracy of the fully implemented system and the lightweight solution.

over 95% accuracy. We adopt such a setting for our prototype system implementation.

The number of audio channels. AcFace uses a four-by-four array of microphones (i.e., audio channels) to collect facial scattered signal at different measurement locations. We evaluate how the number of audio channels impacts the recognition accuracy. Specifically, we adopt the acoustic samples of 8 channels (microphone 5 - 12, as indexed in Figure 13b), 10 channels (microphone 5 - 12 plus 2 and 3), 12 channels (microphone 5 - 12 plus 2, 3, 14 and 15), 14 channels (microphone 1 - 12 plus 14 and 15) and 16 channels. As Figure 13b suggests, using more channels improves the overall performance by providing higher accuracy (the median of averaged accuracy is higher) and better reliability (the 25th and 75th percentiles of averaged accuracy are closer).

7.2.5 *Computational cost.*

The system computational cost is summarized in Table 4. The two implementations are compared in their spectrum generation time and model training efficiency (detailed with the total number of parameters, model size, training time and inference time). The training time represents the training duration of one mini-batch (128) of the total samples and the inference time is the time required by executing the signature matching between the facial spectrum samples of two users. As we see from the table, the lightweight implementation reduces computational cost significantly (spectrum generation time from near 5s reduced to around 3ms) while at a cost of around 4% accuracy drop. The varied performance can be attributed to its approach of feeding the multipath estimates directly into the neural network. In this way it bypasses the computational steps involved in generating the facial spectrum, but on the other hand relies heavily on the neural network's ability to infer such features from the data.

8 DISCUSSION

Impact of face pose. The impact of different face pose (i.e., different θ as illustrated in Figure 14a) is resolvable. During data collection, we can let the user move her head and obtain spectrum samples at various directions – Figure 14b and 14c shows two examples at the direction of 45 and 90 degree. When trained with these samples, the RD-Net should be able to integrate the features extracted from different viewing angles and finally provides similar accuracy for different facing direction.

We conduct a preliminary validation of this hypothesis with an experiment. The experiment uses the facial spectrum samples collected when two users change their facing directions from 0 degree (front view) to 90 degree to train the RD-Net, and then test the recognition accuracy using the spectrum of different direction. The results is presented in Table 5, which demonstrates over 93% averaged accuracy for both users when their facing direction varies. The accuracy can be further improved by incorporating data fusion techniques to better integrate the features obtained from the different angles, which we leave as future work.

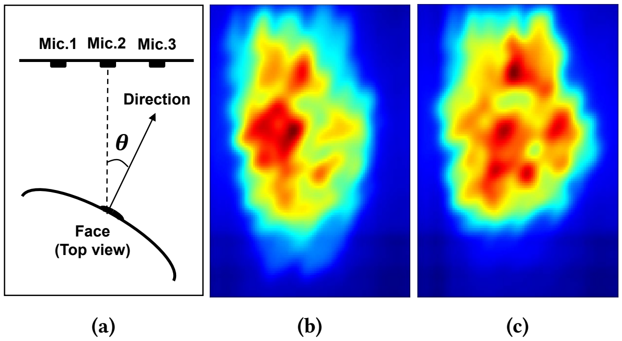


Figure 14: Facial spectrums derived when the facing direction θ equals (b) 45 degree and (c) 90 degree.

Facing direction (degree)	0	15	30	45	60	75	90
Ave. Accuracy - User 1 (%)	95.44	95.13	93.25	94.01	93.13	93.29	93.61
Ave. Accuracy - User 2 (%)	96.25	93.41	91.94	93.67	93.71	93.67	94.22

Table 5: The averaged accuracy when two users are in different facing directions.

Impact of appearance changes. Appearance changes such as hairdos and hats have a minor impact on the recognition results of our system, but facial reflectors (such as glasses) can prevent the system from identifying the current user. The reason is that different hairstyles and hats have a smaller effect on the reflection of acoustic signals in the main facial area (the part not higher than the forehead), whereas objects like glasses can cause significant changes in the reflection. To further adapt to the impact of facial reflectors, we may need to improve the design of the network structure (for example, by introducing a glass discriminator on top of the existing model) and re-train specifically for these scenarios.

Working with mobile devices. As illustrated in Figure 13b, the performance of AcFace may get impaired due to the reduced spatial diversity. To work well with mobile devices which are normally equipped with less number of microphones, we can let users hold the device and move it in front of his face in order to collect acoustic samples scattered by different facial areas. We leave this design as future work.

9 RELATED WORK

9.1 Face Recognition

Face recognition has been widely adopted for user identification due to the uniqueness and long-term stability of facial characteristics. The majority of face recognition techniques are designed atop vision based image processing [15, 25, 28]. For example, FaceNet [28] achieves high accuracy by using deep CNN to learn a mapping function that converts the similarity of face images to the distance defined on Euclidean space. VGG-Face [25] investigates the performance of numerous variants of existing CNN models and adopts a much simpler but effective network that achieves high accuracy. These techniques, however, do not work when the facial landmarks are blocked by obstacles (e.g., facial masks). Several techniques are proposed in recent years to achieve face recognition under masks.

Commercial solutions such as Apple Face ID [2] focuses on identifying features around the eyes when users wear mask which might compromise their accuracy. Furthermore, the facial recognition technology employed by the commercial devices (e.g., smartphones, tablets) necessitates sophisticated and expensive hardware, including a dot projector, flood illuminator, and infrared camera. Such requirements render it less practical for cost-sensitive applications, including secure entry systems or attendance monitoring. SRT [4] proposes a self-restrained triplet loss design, based on which the authors leverages the Embedding Unmasking Model (EUM) and achieves reasonable accuracy even with mask blockage. MaskTheFace [3] designs an open-source tool to generate masked face images from existing face datasets. The authors re-train the Facenet model with the masked dataset and achieve better performance. These techniques adopt the common idea of making the model converge to the reduced feature space extracted from the explicit landmarks (e.g., eye, forehead) when the user is with mask, but usually takes heavy overlead of data collection/generation and model training. Recently, wireless signal based face recognition technique emerge which explore the features from facial reflected acoustic/RF signal. EchoPrint [44] leverages the acoustic hardware (i.e., speaker, microphone) and camera to build a two-factor authentication system that leverages both acoustic reflections and vision landmarks. However, the system inherits the vulnerability to facial blockage since it relies on vision algorithms for landmark detection. RFace [37] exploits the facial features from RFID signal to distinguish different users and combat spoofing attacks. mmFace [38] proposes a millimeter wave based facial authentication system that works for masked faces. The system adopts the idea of Synthetic Aperture Radar to increase the field-of-view and sensing resolution of commercial millimeter wave devices. However, it requires a 2D slide trail to move the transceiver for facial scanning, which imposes extra overhead on system implementation for practical usage.

9.2 Human Identification

Intelligent user identification has become a cornerstone technology underpinning a variety of applications, driven by the growing demand for secure and convenient authentication methods. Over recent years, a plethora of research efforts have been dedicated to exploring the use of diverse biometric markers for user authentication, significantly enhancing both security and user experience across different platforms. Among these innovations, EchoFace [9] utilizes acoustic signal for detecting and defending against photograph/video-based attacks on facial recognition systems. It is essentially a vision-based facial recognition system but utilizing acoustic signals to discern between genuine human faces with uneven stereo-structure and 2D images or videos. VocalLock [20] utilizes the unique characteristics of an individual's vocal tract to offer a user authentication solution that is notably resistant to replay attacks, further optimized for ease of use by its passphrase-independence. Similarly, TouchPass [39] harnesses the distinctive physical attributes manifested through the act of touching, employing active vibration signals to discern the unique patterns of screen interaction attributable to different users. Regarding wearable technology, Bilock [45] introduces an innovative approach to authentication by capturing the unique biometric signature produced by human dental occlusion, thereby offering a novel and secure method for user verification. Building on this rich literature of biometric-based authentication solutions, AcFace distinguishes itself by leveraging the acoustic characteristics of 3D facial features. This method is specifically designed for improved robustness even in the face of adverse environmental conditions, marking a significant advancement in the field of intelligent user identification.

9.3 Acoustic Sensing

Acoustic signal has been exploited for diverse sensing applications [18, 27, 33–35, 41, 42] for its high ranging resolution due to its nature of low propagation speed. An acoustic based contactless respiration detection scheme is proposed in [33], which achieves high ranging resolution with using C-FMCW – a time domain correlation based method for ToF estimation. RobuCIR [35] conducts CIR estimation with least square channel estimation, which is based on time domain signal correlation, and thus inherits its instability when working with noisy channels. In [41], the authors extract tiny heartbeat motion accurately by compensating the random buffer delay at the acoustic front-end. The virtual transmission signal based two-phase mixing approach guarantees the correct relation between the first two paths, but assumes that the power of the first path must be stronger than that of the second path, which does not hold when the speaker is directional or LOS path is blocked. CAT [22] proposes to use distributed FMCW to combat the hardware induced delay between the unsynchronized transmitter and receiver, but it requires calibration to obtain the reference position. Strata [40] and FingerIO [24] enable fine-grained device free tracking by monitoring the continuous phase rotation over time, which is not applicable to in our system where there exists huge amount of multipath reflections. To the best of our knowledge none of existing acoustic sensing solutions can be applied to face recognition in harsh conditions as considered in this paper.

10 CONCLUSION

This paper studies an acoustic based face recognition approach that complements existing vision based solutions, in particular when applied in harsh environment conditions. The design of AcFace entails novel acoustic signal processing techniques as well as a special neural network design to alleviate the impact of facial mask blockage. One future work is to adopt AcFace in a mobile setting where the challenge is achieving comparable performance with reduced number of microphones.

ACKNOWLEDGMENTS

We sincerely thank our shepherd and the anonymous reviewers for their helpful feedback in improving this paper. We also thank all the volunteers for their support for the experiment implementation and dataset construction. This work was supported in parts by National Key R&D Program of China (2022ZD0160804), HKUST internal fund, and Singapore MOE AcRF Tier 2 MOE-T2EP20220-0004. This work was also supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI).

A ARTIFACT APPENDIX

The research artifacts accompanying this paper are available via DOI: 10.5281/zenodo.11094213.

REFERENCES

[1] 2023. About Face ID advanced technology. https://support.apple.com/en-us/102381#:~:text=The%20technology%20that%20enables%20Face,infrared%20image%20of%20your%20face.

[2] 2023. Face ID under mask recognition. https://support.apple.com/en-sg/guide/iphone/iph6d162927a/ios#:~:text=Use%20Face%20ID%20while%20wearing%20a%20face%20mask&text=Note%3A%20Face%20ID%20is%20most,then%20follow%20the%20onscreen%20instructions.

[3] Aqeel Anwar and Arijit Raychowdhury. 2020. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104* (2020).

[4] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2022. Self-restrained triplet loss for accurate masked face recognition. *Pattern Recognition* 124 (2022), 108473.

[5] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.

[7] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 278–291.

[8] Cunjian Chen. 2021. PyTorch Face Landmark: A Fast and Accurate Facial Landmark Detector. https://github.com/cunjian/pytorch_face_landmark Open-source software available.

[9] 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.

[10] Christian Wolff. [n. d.]. Frequency-Modulated Continuous-Wave Radar (FMCW Radar). https://www.radartutorial.eu/02.basics/Frequency%20Modulated%20Continuous%20Wave%20Radar.en.html

[11] Erik Demaine. [n. d.]. Introduction to Algorithms. http://courses.csail.mit.edu/6.006/spring11/lectures/lec02.pdf Peak Finding.

[12] EU's privacy concern on malicious camera usage. 2020. https://sciencebusiness.net/news/eu-makes-move-ban-use-facial-recognition-systems

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[14] Jutta Hämmerle-Uhl, Karl Raab, and Andreas Uhl. 2011. Robust watermarking in iris recognition: application scenarios and impact on recognition performance. *ACM SIGAPP Applied Computing Review* 11, 3 (2011), 6–18.

[15] 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.

[16] A.K. Jain, S. Prabhakar, L. Hong, and S. Pankanti. 2000. Filterbank-based fingerprint matching. *IEEE Transactions on Image Processing* 9, 5 (2000), 846–859. https://doi.org/10.1109/83.841531

[17] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 289–304.

[18] Jialin Liu, Dong Li, Lei Wang, and Jie Xiong. 2021. BlinkListener: "Listen" to Your Eye Blink Using Your Smartphone. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 73 (June 2021), 27 pages. https://doi.org/10.1145/3463521

[19] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1466–1474.

[20] Li Lu, Jiadi Yu, Yingying Chen, and Yan Wang. 2020. Vocallock: Sensing vocal tract for passphrase-independent user authentication leveraging acoustic signals on smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–24.

[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] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: High-Precision Acoustic Motion Tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (New York City, New York) *(MobiCom '16)*. Association for Computing Machinery, New York, NY, USA, 69–81. https://doi.org/10.1145/2973750.2973755

[23] MiniDSP. [n. d.]. UMA-16v2 USB mic array. https://www.minidsp.com/products/usb-audio-interface/uma-16-microphone-array

[24] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[25] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).

[26] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. 2018. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 82–94.

[27] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 474–485.

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

[29] Seeed studio. [n. d.]. Grove Speaker. https://wiki.seeedstudio.com/Grove-Speaker/

[30] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45, 4 (2009), 427–437.

[31] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. 2018. SilentKey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–18.

[32] TI. 2017. Intro to mmWave Sensing: FMCW radars. https://training.ti.com/intro-mmwave-sensing-fmcw-radars-module-1-range-estimation?context=1128486-1139153-1128542

[33] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 170 (Jan. 2018), 20 pages. https://doi.org/10.1145/3161188

[34] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 82–94.

[35] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 566–575. https://doi.org/10.1109/INFOCOM41043.2020.9155402

[36] Wikipedia contributors. 2021. Acoustic wave — Wikipedia, The Free Encyclopedia. [Online; accessed 25-September-2021].

[37] 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*. 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488737

[38] 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.

[39] Xiangyu Xu, Jiadi Yu, Yingying Chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[40] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 15–28.

[41] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can "Hear" Your Heartbeat! *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 161 (Dec. 2020), 24 pages. https://doi.org/10.1145/3432237

[42] Huanle Zhang, Wan Du, Pengfei Zhou, Mo Li, and Prasant Mohapatra. 2016. DopEnc: Acoustic-based encounter profiling using smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 294–307.

[43] Wei Zhang, Sen-Ching S Cheung, and Minghua Chen. 2005. Hiding privacy information in video surveillance system. In *IEEE International Conference on Image Processing 2005*, Vol. 3. IEEE, II–868.

[44] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-factor authentication using acoustics and vision on smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 321–336.

[45] Yongpan Zou, Meng Zhao, Zimu Zhou, Jiawei Lin, Mo Li, and Kaishun Wu. 2018. BiLock: User authentication via dental occlusion biometrics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–20.