EchoFace: Acoustic Sensor-Based Media Attack Detection for Face Authentication

Huangxun Chen, *Student Member, IEEE,* Wei Wang, *Member, IEEE,*
Jin Zhang, *Member, IEEE,* and Qian Zhang, *Fellow, IEEE.*

***Abstract*—Face authentication systems have gained widespread popularity because of their user-friendly usage and increasing recognition accuracy. Unfortunately, the boom in mobile social networks has bought with it media-based facial forgery; a critical threat where an adversary forges or replays the victim's photo/video to fool the system. In this paper, we propose EchoFace, an effective and robust liveness detection system to enhance face authentication in defending against media-based attacks, which works with today's smartphones/smartwatches without any hardware modification. EchoFace uses active acoustic sensing to differentiate the uneven stereostructure of the face and the flat forged media. Our proposed scheme effectively extracts the desired reflection profiles from the target. Moreover, we propose effective similarity measurements of reflection profiles to distinguish live users from forged media, which works robustly under various environmental conditions. EchoFace only requires low-cost and universally equipped acoustic sensors without human intervention for liveness detection, which can be easily deployed in a variety of application scenarios. We implement EchoFace on commercial smartphones, and experiment results show that EchoFace achieves an average detection accuracy higher than 96% and false alarm rate lower than 4% across various media attacks and different levels of background noise. This shows its great potential to enhance the security of widely-deployed face authentication systems in real scenarios.**

***Index Terms*—Acoustic sensing, Liveness detection, Face authentication, Media attack detection.**

I. Introduction

In recent years, biometric authentication has attracted considerable attention for its natural advantages over traditional credential-based authentication. Among various biometric methods, facial authentication has been widely adopted due to the user-friendly usage and increasing recognition accuracy. Nowadays, facial authentication systems are not only in mobile devices like iPhones, various Android smartphones, but also in wearables [1] and many other IoT devices in retail stores [2], hotels [3] and airports [4], [5].

However, some existing face authentication systems (True Key [6], FaceLock Pro [7] and Visidon [8]) are proved to be vulnerable against the vicious media attacks, where an adversary forges or replays the victim's photo/video to fool the system. Prior study [9] has shown that 53% of facial photos from the social media, such as Facebook and Google+, can easily be utilized to spoof face authentication systems, which raises great public concern for the security of these systems and calls for a universal solution to deal with the threat of these media attacks.

Despite extensive efforts in proposing various liveness detection methods to defend media attacks, there is still space for improving security levels, and the universality and robustness of liveness detection systems. Video-based liveness detection systems [10]–[12] rely on video or multiple image inputs and assume that images used in media forgery attacks move differently to a live face. Thus, they are vulnerable to the well-executed video forgery attacks. True Key [6] simply integrated another biometric, fingerprint to detect the live users. Unfortunately, the fingerprint biometric can also be forged [13]. FaceLive [14] and Chen et al. [15] both require a user to hold and move a mobile device over a short distance in front of her/his face for liveness verification. Obviously, such user-involved methods can not be applied to many face authentication systems in retail stores [2], hotels [3] or airports [4], [5], since it is impossible for the user to move bulky machines. FaceHeart [16] requires relatively good ambient illumination to extract subtle photoplethysmograms from the face videos for liveness detection, thus their performance may be affected in a relatively dim environment. Apple's FaceID achieves good performance at the extra cost of additional hardware (e.g., dot projector, flood illuminator and infrared camera), which may not be available in most low-end devices. In contrast, EchoFace aims to provide a solution suitable for a wide range of existing low-end IoT/wearables and smartphones.

It is observed that spatial structures are the biggest difference between the live user's face and her/his photo/video. As shown in Fig. 1, the live user's face is an uneven stereostructure, including complex curved surfaces, while the photo/video should be a flat plane. Thus, to defend media forgery attacks, the devices equipped with a speaker and two microphones could use active acoustic sensing to detect such spatial structures. Based on the above observation, we propose *EchoFace*, an acoustic-based liveness detection system to help face authentication defend itself from the media forgery attacks. EchoFace turns the devices equipped with speakers and microphones into an active sonar to detect the spatial structures of the target. In our system, the speaker emits well-designed signals, and two microphones with a little space separation collect reflection signals in the meantime.

Our system achieves the desired properties as follows: 1) *Effectiveness*: the system can effectively distinguish the live user and forged media (photos and videos). The uneven stereostructure of face results in quite different multipath effects on the signals reflected to two microphones, while the flat photo/video induces similar reflection effects. Therefore, EchoFace could distinguish between live users and the forged media by analyzing the similarity between reflected signals at two microphones. 2) *Passiveness*: our approach requires no explicit user involvement, which makes the solution universally suitable to enhance face authentication on both mobile devices

Huangxun Chen and Qian Zhang are with Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China (Email: hchenay@connect.ust.hk, qianzh@cse.ust.hk)

Wei Wang is with School of Electrical Information and Communication, Huazhong University of Science and Technology, Wuhan, 430074, China (Email: weiwangw@hust.edu.cn).

Jin Zhang is with Department of Electrical and Electronic, South University of Science and Technology of China, Shenzhen, China (Email: zhang.j4@sustc.edu.cn).

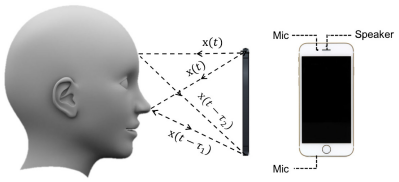


Fig. 1: Working principle of EchoFace.

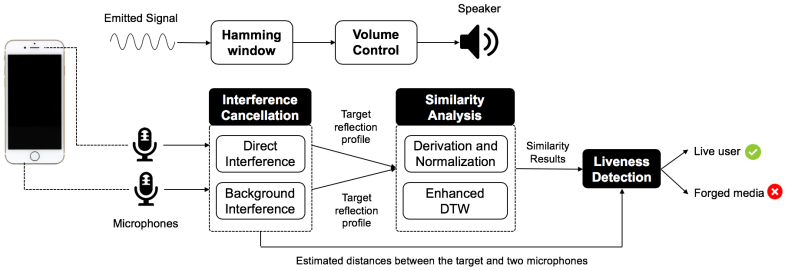


Fig. 2: EchoFace Overview.

and stationary machines [2]–[5], considering speakers and microphones are not only commonly equipped on mobile devices but also gradually adopted on many machines to provide voice interaction to the users. 3) *Robustness*: the system performance is insensitive to different environmental factors. EchoFace would not be affected by ambient illumination, and would also be resistant to ambient noise by well-designed emitted signals. Compared with the most recent effort [17] which mainly utilizes a trained Convolutional Neural Network(CNN) to extract reliable acoustic features, EchoFace requires no training process. Thus, the liveness detection function provided by EchoFace can work in a plug-and-play manner to enhance any existing face authentication method.

However, there exists two main challenges to achieve the above acoustic-based liveness detection system. Firstly, most off-the-shelf speakers and microphones are omni-directional. Thus, acoustic signals received by the microphones include not only desired reflection from the detection targets (the live user or the impostor), but also direct transmission from the nearby speaker and background reflection from other objects. Besides, the magnitude of direct transmission is generally two to three orders larger than that of target reflection. Secondly, EchoFace requires effective similarity measurements to work robustly under various unknown settings, including different target-device distances, hardware imperfection and ambient noise level. To deal with the above challenges, we develop a distance estimation scheme to extract the desired target reflection, which effectively eliminates the interference from direct transmission and background reflection. Then, we leverage two microphones with a little separation to collect the reflection signals. Since the two microphones work simultaneously and the target is closer to the device than other acoustic sources, the similarity characteristics between signals collected from two microphones are relatively stable under different settings.

Our main contributions are summarized as follows: (i) We propose EchoFace, an acoustic-based liveness detection system to help face authentication effectively defend the media(photo/video) forgery attacks, which only requires commonly available acoustic sensors without explicit user involvement. (ii) We propose a distance estimation scheme to collect the desired reflection profile from the target without interference from the speaker and background objects. Moreover, we propose effective similarity measurements of reflection profiles to distinguish live users and the forged media, which achieves good performance on liveness detection under various environmental conditions. (iii) We implement EchoFace on the commercial smartphones and conduct extensive experiments to evaluate its performance on liveness detection. Experiment results show that EchoFace achieves 96% average detection accuracy and 3.57% false alarm rate, which shows the potential of EchoFace to enhance the security of widely-deployed face authentication systems on mobile devices and in other real life scenarios [2]–[5].

II. System Overview

This section first illustrates how EchoFace leverages active acoustic sensing to differentiate a live user and the forged media (photo/video), then introduces the threat model, and then finally gives an overview of the EchoFace system.

A. EchoFace's Working Principle

EchoFace turns devices equipped with one speaker and two microphones into an active sonar to detect the spatial structures of the target. The speaker emits the well-designed signals, and meanwhile the two microphones collect reflection signals. As shown in Fig. 1, supposing the emitted signal is $x(t)$ and two reflected signals from the target surface arrive at a microphone with time delay τ_1 and τ_2 respectively, the received signal can be represented as $y(t) = Ax(t-\tau_1) + Ax(t-\tau_2)$, where A is the amplitude of the reflected signal. Considering the target is quite close to the device, the amplitude distortions from two transmission paths are regarded as the same. Thus, the Fourier amplitude spectrum of $y(t)$ can be expressed as follows:

$$\begin{aligned} Y(f) &= Ae^{-j2\pi f\tau_1}X(f) + Ae^{-j2\pi f\tau_2}X(f) \\ &= 2Ae^{-j2\pi f(\tau_1+\tau_2)}\cos\pi f(\tau_1-\tau_2)X(f), \end{aligned} \quad (1)$$

where $X(f)$ is the Fourier amplitude spectrum of $x(t)$.

This equation makes it clear that the received signals have uneven attenuation at different frequencies when the emitted signal has a flat amplitude spectrum. The reason is that different reflections with delayed phases may be constructive at some frequencies while destructive at other frequencies.

We regard the above Fourier amplitude spectrum of reflected signals as the multipath profile. Since two microphones are located at different places, they may obtain different multipath profiles if their reflected signals go through different reflection multipaths. The typical wavelengths of acoustic signals are around 2-3 centimeters. Thus, when the sensing target is a live user's face, the uneven stereostructure of face results in quite different multipath effects on the reflected signals to two microphones. However, for the flat forged photo/video, the reflected signals go through similar multipaths before arriving at the two microphones, thus, their multipath profiles will show more similarity. Even a printed photo can be bent to simulate a facial structure, it is very difficult to simulate the detailed fluctuation of the face without losing essential image information. Therefore, *EchoFace* could distinguish between live users and forged media by analyzing the similarity between signals collected from two microphones.

The attack scenarios for our system are defined as follows:

- *Simple Media Attack*: The attacker replays a victim's photo/video to fool the system.
- *Advanced Media Attack*: The attacker bends a printed photo of the victim to simulate the unevenness of the facial structure.

B. Overview

The system overview of EchoFace is shown in Fig. 2. We adopt the earpiece speaker instead of the main speaker to emit the chirp signals. The reasons are two-fold. Firstly, the earpiece speaker is on the frontal panel and faces the human face, which facilitates most acoustic signal reflected off the face for collection and analysis. Secondly, although the main speaker at the bottom is generally more powerful, its signal may be serious blocked by users' hand when the users hold the phone for liveness detection. Besides, the main speaker faces the bottom, the signal reflected off the users' body may also affect the system performance. The predefined acoustic signals (elaborated in Section III-A) would be multiplied with a Hamming window function and then go through a volume control system to reduce annoyance. Next, the signals are emitted by the speaker, and in the meantime two microphones collect reflection signals.

The collected signals include not only the desired reflection from the detection target but also a direct transmission from the nearby speaker and background reflection from other objects. Thus, both reflected signals at two microphones would be processed in the interference cancellation module to eliminate two interference described above. In the meantime, the system will estimate distance between the target and two microphones. Now we obtain the two multipath profiles induced by the spatial structure of the target from different angles of view. Next, the system conducts derivation and normalization operations on two profiles and applied Dynamic Time Warping(DTW) to measure their similarity. Finally, the liveness detection algorithm can distinguish between the live user and the forged photo/video based on previous similarity results, and also the estimated distances between the target and the two microphones.

III. System Design

This section elaborates on the detailed design of EchoFace, including the emitted signal design, the interference cancellation, the similarity analysis and the liveness detection algorithm.

A. Emitted Signal Design

The emitted signal should be carefully designed to satisfy the properties as follows: 1) *Diverse*: The emitted signals with diverse frequency components could facilitate EchoFace to capture more spatial structural features in the target multipath profile. 2) Less-annoying: The disturbance to the user would be limited as much as possible. 3) Robust: The emitted signal would be resistant to ambient acoustic sources so that the system performance would remain stable. In the following, we illustrate our considerations on signal frequency and duration, signal diversification and annoyance reduction.

1) Signal Frequency and Duration: It is well-known that most smartphones support a 44.1kHz sampling rate of their microphones [18], so the highest sensed frequency is 22kHz. Thus, EchoFace adopts chirp signal sweeping from 12kHz to 21kHz to sense the target. We discard the frequency below 12kHz because most background noise from human activities fall into such range [19]. The acoustic signals from 20kHz to 22kHz are inaudible to avoid annoyance [20]. However, most off-the-shelf speakers/microphones' frequency responses above 20kHz on commodity smartphones are very poor. Some microphones of commodity smartphones even experience 30dB signal degradation at 22kHz [21]. The reason is that they are originally built for playing and recording sounds falling in audible range (far below 20kHz). Therefore, when the system emits chirp signals from 20kHz-22kHz, on one side, the emitted signal from the speaker is weak in terms of energy due to the speaker's poor frequency response above 20kHz; on the other side, the signal that reflected off the face and then collected by the microphone is even weaker due to the microphone's poor frequency response above 20kHz. This situation makes it difficult to collect valid face features for liveness detection. The proposed method adopts chirp signals from 12kHz to 21kHz to overcome the signal degradation issue. Besides, as shown in Equation (1) in Section II, the uneven stereo structure of face results in uneven attenuation at different frequencies of the received signals. Thus, a wider bandwidth of the emitted chirps helps the system capture richer and more accurate face features for liveness detection. Although the emitted signals of EchoFace are audible, we apply two effective approaches to reduce the annoyance, which is illustrated in Section. III-A3.

Besides frequency range, the duration of the chirp signals requires careful selection since it would affect the signal-to-noise-ratio (SNR) of the received signals. On the one hand, the long chirp signal can enable more energy to be collected on each frequency for more accurate sensing. On the other hand, a long chirp signal may overlap with the reflections from the nearby target, which affects the system performance. In daily usage of facial authentication, the users generally keep their faces at around 25cm to 40cm away from the camera without any blockage so that the camera could capture sufficient and valid facial features. Based on the above analysis, we choose 50-sample emitted signals for sensing, and collect only 50 samples of the received signals within the target range(as shown in Fig. 3) to eliminate the interference from outside the target range.

2) Signal Diversification: As mentioned before, the emitted signals with a diverse frequency component could help facilitate EchoFace in capturing more spatially structural features in the target multipath profile. To further enhance the features in multipath profiles, we adopts two strategies, signal repetition and piecewise frequency sweeping.

Signal Repetition: Echoface emits 16 chirp signals consecutively to sense the targets, where the duration of each 50-sample chirp signal is $50/44.1kHz \cong 0.0011$ second. Here, it is important to select the time interval between two chirps, which is related to both the sensing speed and the detection accuracy of EchoFace. On the one hand, the larger the time interval, the longer the time EchoFace needs for target sensing. On the other hand, an excessively short waiting time may affect sensing results, since the received signals might include the remote reflections of the previously emitted chirp. For example, if EchoFace senses a target 30cm away from a device using two chirp signal with a 50-sample interval, the reflection signal from objects at a distance of about 68cm[1] will be added as noise to the received target reflection profiles of the next chirp. Based on the above analysis, the time interval

[1]The sampling rate of EchoFace is 44.1kHz and sound travels at a speed of 338m/s in air, so we have $(50+50)/44100 \times 338/2 \times 100 + 30 \cong 68$cm

between two chirps in EchoFace is set to 3000 samples as in [22]. Thus, the total duration of sensing time is equal to $16 \times (50 + 3000)/44100 \cong 1.12$ second.

Piecewise Frequency Sweeping: Inspired by [22], we divide all sixteen chirp signals into four groups, each of which contains four 50-sample signals. Four groups will sweep at the piecewise frequency range of 12-15kHz, 14-17kHz, 16-19kHz and 18-21kHz respectively. Compared with one 50-sample segment sweeping the entire 12–21kHz, the piecewise frequency sweeping keeps enough energy at each frequency, which would facilitate EchoFace to capture reliable spatial structural features in the target multipath profile. A 250-sample pilot is added before the frequency sweep for synchronization between speakers and microphones, since the operating system delays are not consistent in commodity phones. We adopt a similar synchronization process as in [23] and 11.025kHz pilot tone as in [22]. Besides that, another 8000 samples follow the pilot before the chirp signals are played. Different to [22], four chirps with an incremental frequency range are played consecutively before the next frequency sweeping, which increases the independence between two neighboring chirps to help capture more spatial structural features in the target multipath profile.

3) Annoyance Reduction: As mentioned before, the disturbance of the emitted signals to the user should be limited as much as possible. EchoFace adopts two strategies to minimize the annoyance of audible components in the emitted signals.

Multiplying Hamming Window: To effectively lower annoyance, EchoFace adds the Hamming window function on the emitted signals. Abrupt cutoff in the rectangular window function broadens the frequency spectrum of emitted signal and produces more audible components. The side lobe of the Hamming window function degrades more than the rectangular window function, which avoids such a truncation effect. Thus, we choose the Hamming window instead of a rectangular window on the emitted signal. Our experiments shows that the volunteers feel the sound strength is obviously lower after we apply the Hamming window function to the emitted signals.

Volume Control: To further reduce annoyance, we adjust the sound volume without degrading the detection accuracy. In our preliminary experiments, when the volume of emitted signal is full (i.e., 100%), the detection accuracy is instead reduced because the received signals are saturated by the direct transmission (i.e., the sound emitted directly from the speaker). Moreover, emitting sound at full volume makes the sensing process more annoying to the users. When the volume is only 6%, the reflections are too weak to be picked up by two microphones. Through experiment, we found that it is optimal to set the speaker volume at 13% for Galaxy C7. Although the optimal volume setting varies among different devices, it only requires a one-time calibration to figure out the optimal volume. Given only 0.018 second (50/44100*16) occupied by the emitted chirps and above two methods applied, the volunteers hardly notice the emitted sound in a laboratory.

B. Interference Cancellation

After the speaker emits the designed signals, two microphones harvest the reflection signals, which includes multiple reflections from the desired targets and the surrounding objects. EchoFace needs to eliminate the interferences to extract the targeted reflection signals. There are two types of interference: (i) direct transmission from the nearby speaker, and (ii) background reflection from other objects. The magnitude of the direct transmission is 2-3 orders larger than that of the target reflection. Moreover, these interferences overlap with the target reflection in time when the difference between their propagation delay is smaller than the chirp duration (e.g., about 1.1ms in EchoFace). To minimize the impact of the interference, EchoFace adopts an interference cancellation scheme, which utilizes the property of emitted chirp (frequency sweeping) signals to produce the reflection profile(as shown in Fig. 3) and eliminates the interference from outside the target range.

Firstly, we accomplish the time synchronization between speaker and microphones by multiplying the received signals with the pilot. Then we pass the received signals to a matched filter to estimate the propagation delay of the target reflection.

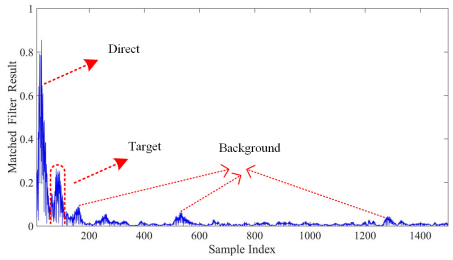


Fig. 3: An example of reflection profile.

The unit impulse response of the matched filter is as follows.

$$h[n] = \begin{cases} s[n], & \text{if } n = 1,2,...,N_s \\ 0, & \text{if } n \text{ is other} \end{cases}, \tag{2}$$

where $s[n]$ is the emitted signal and N_s is length of the emitted signal. if the received signal is $r[n]$, then the signal after match filtering $r'[n]$ can be written as $(n = 1, 2, ..., N_r)$:

$$r'[n] = \sum_{i=1}^{N_s} h[i]r[n + i - 1], n = 1, 2, ...N_r \tag{3}$$

where N_r is the length of the received signals/$N_r \geq N_s$).

As shown in Fig. 3, the first and largest peak represents the direct transmission. After obtaining the propagation delay of direct path, we can subtract the interference of direct transmission to obtain $\hat{r}[n], n = 1, 2, ...N_r$.

Then, we pass the residual received signals by an adaptive matched filter, and figure out the propagation delay of the target reflection indicated by the second peak in Fig. 3. The detailed algorithm is illustrated in Algorithm 1. Once we obtain the propagation delay of the target reflection, we can extract the reflection profiles as shown in Fig. 3 only with the reflection from the target range, without direct transmission and background interference. We denote the reflection profiles of the main microphone as P_α and denote those of the vice microphone as P_β for the following similarity analysis.

C. Similarity Analysis of Multipath Profiles

With two multipath profiles P_α and P_β, EchoFace estimates their similarity to distinguish the live user face and forged media for liveness detection as explained in Section II.

The intuitive method is to adopt the simple correlation expressed as follows.

$$r(P_\alpha, P_\beta) = \frac{\sum_k (P_\alpha(k) - \bar{P_\alpha})(P_\beta(k) - \bar{P_\beta})}{\sqrt{\sum_k (P_\alpha(k) - \bar{P_\alpha})^2} \sqrt{\sum_k (P_\beta(k) - \bar{P_\beta})^2}} \tag{4}$$

Algorithm 1 Adaptive Propagation Delay Estimation

Input:
The residual received signals from two microphones: $\{\hat{r}_i^{m_1}[n], i = 1, 2, ..., 16\}$ and $\{\hat{r}_i^{m_2}[n], i = 1, 2, ..., 16\}$.
Output:
The estimated propagation delay of reflection signals from the target τ_{m_1} and τ_{m_2}.

```
Initialization: idx_l ← 40, idx_r ← 110, idx_thres ← 15
reference standard deviation Δ ← MAXINT, λ ← 1/2
for pt ← idx_l to idx_r − 50 do
    for i ← 1 to 16 do
        Feed r̂_i^{m1}[pt : pt + 50] and r̂_i^{m2}[pt : pt + 50] to the
        matched filter h[n], and figure out the sample index
        of the first and largest peak: I_i^{m1} and I_i^{m2}
    end for
    Compute the mean of I_i^{m1} and I_i^{m2}(i = 1, 2, ..., 16) as
    μ_I^{m1}, μ_I^{m2}
    if λδ_I^{m1} + (1−λ)δ_I^{m2} < Δ and |μ_I^{m1} − μ_I^{m2}| < idx_thres
    then
        Δ ← λδ_I^{m1} + (1 − λ)δ_I^{m2}
        τ_{m1} ← μ_I^{m1}
        τ_{m2} ← μ_I^{m2}
    end if
end for
```

where $r(P_\alpha, P_\beta)$ is the correlation coefficient, $\bar{P_\alpha} = \frac{1}{N}\sum_{k=1}^{N} P_\alpha(k)$ and $\bar{P_\beta} = \frac{1}{N}\sum_{k=1}^{N} P_\beta(k)$, $k = 1, 2, ..., 50$. However, there exists an inevitable error in the propagation delay estimation, which causes the shift in frequency features for chirp signals. The experimental results show that simply correlating the profiles can only achieve the detection accuracy as high as 82.3%, which cannot satisfy the requirement of real-life scenarios.

Then we try to utilize Dynamic Time Warping (DTW) [24] to characterize the overall similarity of two profiles, which are widely used in speech recognition to deal with the word matching problem under various speech speeds [25]. Similarity analysis of multipath profiles in EchoFace has a similar nature to speech pattern matching. Given two sequences, and a cost matrix, DTW searches for an alignment that maps each point in the first sequence to one or more points in the second sequence, such that the mapping cost summed over all point pairs is minimized. In the context of EchoFace, we have two multipath profiles P_α and P_β harvested by the main microphone and the vice microphone respectively. For any pair of points in the two profiles, $P_\alpha(i)$ and $P_\beta(j)$, we define the cost of mapping these two points to each other as the Euclidean distance between the two power values:

$$C(i, j) = |P_\alpha(i) - P_\beta(j)|. \tag{5}$$

For such an input, DTW looks for the best alignment of the two profiles that minimizes the total cost, using standard dynamic programming [24]. In EchoFace, the two multipath profiles are of the same length N. We refer to the candidate alignment set as $\{M = m_1, m_2, ..., m_l, ..., m_L\}$, where $m_l = (i_l, j_l)$ indicates the mapping of point i_l in sequence P_α with point j_l in sequence P_β. What is more, these candidate alignments need to satisfy the following requirements:

(i) $i_l \in [1, N]$ and $j_l \in [1, N]$;
(ii) $m_1 = (1, 1)$, $m_L = (N, N)$, and $N \leq L$;
(iii) $\forall l \in [1, L-1]$, $i_{l+1} \geq i_l$, $j_{l+1} \geq j_l$;
(iv) $\forall l \in [1, L]$, $|i_l - j_l| < N$.

We refer to the minimum total cost associated with the best alignment as the distance between two profiles, $D(P_\alpha, P_\beta)$, then

$$D(P_\alpha, P_\beta) = \min_M \sum_{l=1}^{L} C(i_l, j_l), \tag{6}$$

where $C(i_l, j_l)$ refers to the mapping cost between point $P_\alpha(i_l)$ and point $P_\beta(j_l)$. EchoFace leverages the distance between two multipath profiles to represent the similarity of them. The value of distance and the similarity are in inverse proportion, i.e., a larger distance indicates a smaller similarity and vice versa.

Microphone Hardware Difference: A practical challenge arises when we directly apply the above DTW algorithm to compare the multipath profiles: scaling of feature values (e.g., peak heights, valley depths in the multipath profiles). Feature values obtained by the main microphone are 2-5 times larger than that of the vice microphone due to the differences in microphone hardware. It is common that the sensitivity of microphones on the same device differ significantly with each other. The intuitive method to eliminate the effect of such a variation is to normalize each microphone's multipath profile by its maximum value. However, normalization alone is not enough. Note the multipath profiles are generally sensitive to the relative distance and orientation between the target and device, which cases peaks and valleys to be scaled differently across different multipath profiles, independent of the microphones' hardware.

To address these potential variations in feature values, we leverage a variant of DTW [26]. Instead of performing DTW directly on the two multipath profiles P_α and P_β, EchoFace firstly computes their derivatives: P'_α and P'_β. Next, each derivative sequence is normalized by its standard deviation. Then EchoFace applies the DTW algorithm to align the two normalized derivative sequences. The cost of this alignment is recorded as the distance between the two multipath profiles. It has been shown in [26] that such a design allows DTW to focus on the high level features of "shape", rather than being bogged down by the absolute values of the sequences, which meets the needs of EchoFace. Our preliminary experiments demonstrate that derivative DTW provides a robust metric to evaluate the similarity between multipath profiles collected by two microphones.

D. Liveness Detection Algorithm

In EchoFace, the speaker emits 16 chirp signals, covering the frequency ranges of 12-15kHz, 14-17kHz, 16-19kHz, and 18-21kHz respectively. After reflection signals are collected by the two microphones, EchoFace firstly applies the target reflection extraction algorithm described in III-B to obtain 16 target reflection signals from each microphone. The extracted 16 pairs of target reflection profiles are expressed as $\left(P_i^{m_1}[n], P_i^{m_2}[n]\right)$, $i = 1, 2, ..., 16$, where the superscript m_1 and m_2 indicate that the signals are harvested by the main microphone and the vice microphone respectively. Then, we process the extracted signal pairs with the derivative DTW, and figure out the distance between $P_i^{m_1}[n]$ and $P_i^{m_2}[n]$ as $D(P_i^{m_1}, P_i^{m_2})$, $i = 1, 2, ..., 16$. Next, we develop a liveness detection algorithm to distinguish forged media from a live user's face, which takes the estimated profile distances $D(P_i^{m_1}, P_i^{m_2}), i = 1, 2, ..., 16$ and the estimated distances between the target and the two microphones d_{m_1}, d_{m_2} as inputs. ($d_{m_1} = \tau_{m_1} \times c, d_{m_2} = \tau_{m_2} \times c, c = 340m/s$). The

liveness detection algorithm basically utilizes profile distances to distinguish forged media from a live user's face. When most of the distances of profile pairs are larger than the empirical thresholds, the system regards the target as a live user's face; otherwise, the target is regarded as a forged media. Here, we adopted adaptive thresholds T_d^f based on the estimated distances between the target and the two microphones and also the frequency range of emitted signals.

IV. Implementation and Evaluation

We implement the EchoFace on commercial smartphones and conduct extensive experiments to evaluate the system performance. In this section, we first introduce the implementation and our experimental setting. Then, we evaluate the performance of EchoFace under different conditions to validate the effectiveness and robustness of our system.

A. System Implementation and Experimental setting

The acoustic signal emission and reflection collection is implemented on a Sumsang Galaxy C7(Android 7.0). Collected signal analysis including interference cancellation, adaptive propagation delay estimation (Algorithm 1) and liveness detection (Algorithm **??**) are implemented in MATLAB R2015a. We conduct our experiments in the laboratory setting.

It is worth mentioning that the goal of proposed method is to enhance the existing face authentication methods with liveness detection capability. Face authentication itself requires the user to hold the phone at the position where the front camera can capture valid face features. Thus, the ranges from cameras should be neither too close (partial captured face) nor too far (small or blurry captured face). To satisfy this requirement, the volunteers were asked to take some selfies before the system evaluation. The sefies are required to contain whole face features to pass photo-based face authentication. It is found that even though different volunteers have different arm lengths, the smartphone-to-face distances generally fall into the range of 23cm-41cm to capture their faces for authentication. Thus, we conduct evaluation within this range to validate the system performance.

In the following evaluation, there are two settings for data collection: 1) the volunteer holds the smartphone with different smartphone-to-face distances for evaluation; 2) the smartphone is set up on a office table for evaluation. The second one is the particular setting for evaluating system performance across various smartphone-to-face distances, where the distances can be controlled and measured precisely between different volunteers' trials for comparison.

In the normal usage mode, a volunteer's face will be in front of the smartphone. In the media attack mode, a victims's photo (flat or bent) will be in front of the smartphone. When we collect the experiment data, other people in the room are free to do daily activities including talking and walking around. Some people pass by frequently since the table is located near the door. Moreover, there exists environmental noise including the noise of the air-conditioner and keystrokes. We recruit six volunteers(1 female and 5 males) and amass about 540 trials collected at different periods of time (9:00-12:00am, 14:00-17:00pm, and 19:00-22:00pm). This is done over three days to evaluate EchoFace's performance under different conditions, which includes different smartphone-to-face distances, different threat models, and different ambient noise levels.

We mainly use two metrics to demonstrate the performance of the liveness detection: detection accuracy and false alarm rate. TP (true positive) denotes the system correctly regards a live user as a live user. TN (true negative) denotes the system correctly regards a forged photo as a forged photo. FP (false positive) denotes the system wrongly regards a forged photo as a live user. FN (false negative) denotes the system wrongly regards a live user as a forged photo. Detection accuracy is calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, and false alarm rate is calculated as $\frac{FP}{FP+TN}$. A good liveness detection system is expected to have a high detection accuracy and a low false alarm rate.

B. Performance across Various Distances

As analyzed before, the collected multipath profiles are significantly affected by the distance between the user's face and the smartphone. To validate the effectiveness of our method on handling the distance factor, we conduct the following experiment to show the performance of our system under different smartphone-to-face distances. We collect the reflected signals when the volunteers' faces are 23cm-41cm away from the smartphone. At each location, we also collect the reflected signal of the volunteers's photos. We put the reflected signals of all volunteers' faces and their photos at the same distance into one group. Within each distance group, we calculate the TP, TN, FP, FN to derive detection accuracy and false alarm rate. As shown in Fig. 4, EchoFace can achieve a stable performance regardless of smartphone-to-face distances. The average accuracy is 96.02%, and the average false alarm rate is 3.97%. Since the face authentication requires an appropriate distance (typically from 25cm to 40cm) between the target face and the smartphone to capture valid image information, the above evaluation results shows the proposed liveness detection system has an adequate effective working range to co-work with face authentication in real scenarios.

C. Performance across Various Users

The stereostructure of the face varies across different people. To validate the effectiveness of our system on different users, we conduct the following experiment to show the detection accuracy of our system under different users. We collected the reflected signals from six volunteers. The volunteers were asked to hold the smartphone at slightly different distances (29cm, 32cm and 35cm; 20 trials of each distance) from their faces. The distances may not be very precise due to the small tremble of their hands. And at each location, we also collect the reflected signal of the volunteers's photos. We put the reflected signals of one specific volunteer' faces and his/her photos at different distances into one group. Within the group of each volunteer, we calculate the TP, TN, FP, FN to derive detection accuracy and false alarm rate. As shown in Fig. 5, the system achieves a stable performance regardless of the variety of users' face and small tremble of users' hand. The average detection accuracy is 96.2%, and the average false alarm rate is 3.57%.

D. Performance under Advanced Media Attack

In the previous experiment, the system can distinguish the 2D media from the face of a live user, relying on the fact that the live user's face has more a complex stereostructure while the 2D media only has a flat surface. In this experiment, we test the system performance under a more advanced attack, i.e.,

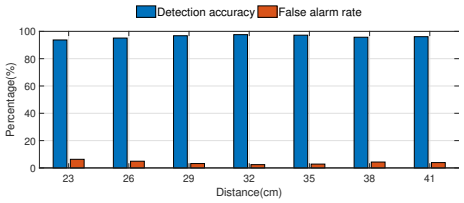


Fig. 4: Performance across various distances.

Fig. 5: Performance across various users.

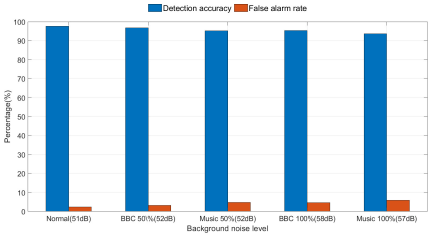
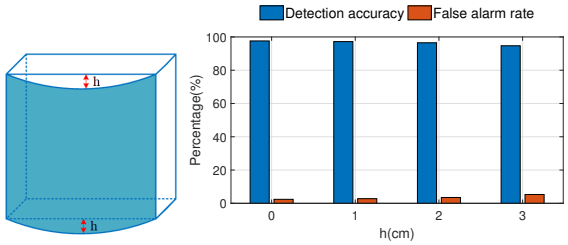


Fig. 6: Noise Resistance Experiments



(a) Curved photo. (b) Performance across different bending degrees.

Fig. 7: Performance under Advanced Media Attack.

we try to use the bent photos to fool the system. We bent the printed photo by attaching it to the curved surface of a plastic box. As shown in Fig. 7(a), the curvature can be described by the height h. We set the height at 1cm, 2cm and 3cm to imitate different bending degrees of the printed photo. The experiment setting is the same as previous experiments except we collect the reflected signal of the volunteers's bent photo for evaluation instead of flat photos. The results in Fig. 7(b) shows that EchoFace achieves a satisfying performance when the printed photo is slightly bent. The average detection accuracy is 96.5%, and the average false alarm rate is 3.5%. We did not test the cases with higher h because when the photo is over-bent, the facial features on the photo are too distorted to pass the original face authentication, where liveness detection may not be required. Due to the same reason, it is meaningless to try to fool the detection system with other objects like a cup or a toy doll because they do not have the correct facial features to pass the original face authentication.

E. Noise Resistance

In this experiment, we validate the noise resistance capability of the system. In real-life scenarios, EchoFace is supposed to work well in the presence of noise from TV, people chatting, and air conditioning fans. To test the robustness against noises, we collect reflection signals in a laboratory environment when another smartphone is placed 30cm away on the same surface. That phone is set to either replay a clip of BBC news or a song ("Because of you-Kelly Clarkson") at different volumes (50% and 100%). As shown in Fig. 6, EchoFace can work well under wide background noise levels including chatting and music even at a 100% volume. Relying on carefully-designed emitted signals and active acoustic sensing, EchoFace is robust to most background noises in real scenarios.

V. Related Works

Researchers have proposed various liveness detection methods to help face authentication defend against media attacks. Early works focused on different features between replayed photos/videos and live users. Li et al. [10] exploited the dynamic characteristics of captured videos for liveness detection. Tan et al. [11] applied the Lambertian model to extract different surface properties of a live human face or a photograph, while Bao et al. [12] leveraged the differences in optical flow fields generated by movements of two dimensional planes and three dimensional objects. These liveness detection systems [10]–[12], [27] rely on video input or multiple images and assume that playback attacks will have different motions to that of a live face. Thus, they are still vulnerable to well-executed video playback attacks. Therefore, the researchers attempted to integrate additional sensors to achieve more secure liveness detection. True Key [6] simply integrated another biometric, fingerprints to detect live users. However, it impairs the convenience of face authentication, and the fingerprint biometric can also be forged [13]. FaceLive [14] and Chen et al. [15] both leveraged the correlation between readings of inertial sensors and facial videos from front-facing camera to achieve liveness detection. To perform the liveness verification in their system, a user needs to hold and move a mobile device over a short distance in front of his/her face. These user-involved methods can not be applied to many face authentication machines in retail stores [2], hotels [3] and airports [4], [5], since it is impossible for the user to move the bulky machines. FaceHeart [16] achieved liveness detection by comparing the two photoplethysmograms independently extracted from the face videos taken with the front camera and fingertip videos taken with the rear camera on smartphones, which requires relatively good ambient illumination to extract subtle photoplethysmograms from face videos. Apple's FaceID achieves a good performance at the extra cost of additional hardware (e.g., dot projector, flood illuminator and infrared camera), which may not be available in most low-end devices. Komeili et al. [28] fused ECG and fingerprint for liveness detection, which also requires specialized hardware. By contrast, EchoFace aims to provide a solution suitable for a wide range of existing low-end IoT/wearables and smartphones. Active acoustic sensing is an active research field [29]–[31], and the previous works [32], [33] have utilized it to enhance voice authentication. In this paper, EchoFace leverages active acoustic sensing to help face authentication defend against media forgery attacks.

VI. Conclusion

This paper proposes EchoFace, an effective and robust liveness detection system to enhance face authentication in defending against media-based attacks. EchoFace uses active acoustic sensing to differentiate the uneven stereostructure of the face and the flat forged media. EchoFace only requires the low-cost and universally equipped acoustic sensors, without explicit user involvement for liveness detection, which can be easily deployed in a variety of application scenarios. Experiment results show that EchoFace achieves an average accuracy

higher than 96% and false alarm rate lower than 4% across various media attacks and different levels of background noise. This shows its great potential for enhancing the security of widely-deployed face authentication systems in real scenarios.

Acknowledgement: This work was supported in part by the RGC under Contract CERG 16203719, 16204418 and in part by the Guangdong Natural Science Foundation No. 2017A030312008.

REFERENCES

[1] "Orcam myme: wearable camera for facial recognition." [Online]. Available: https://myme.orcam.com/
[2] "Alipay's 'Smile to Pay' facial recognition system at KFC outlet." [Online]. Available: https://www.scmp.com/tech/start-ups/article/2109321/alipay-rolls-out-worlds-first-smile-pay-facial-recognition-system-kfc
[3] "NEC's hotel face recognition solution." [Online]. Available: https://www.nec.com/en/global/solutions/hospitality/security_face/
[4] "Hong Kong international airport's face recognition departure system." [Online]. Available: https://www.futuretravelexperience.com/2017/10/smart-departure-system-goes-live-at-hong-kong-international-airport/
[5] "CBP's biometric exit technology at miami international airport." [Online]. Available: https://www.cbp.gov/newsroom/local-media-release/cbp-deploys-biometric-exit-technology-miami-international-airport
[6] "True key — mcafee." [Online]. Available: https://www.truekey.com/en
[7] "Facelock." [Online]. Available: http://www.facelock.mobi
[8] "Visidon." [Online]. Available: https://www.visidon.fi
[9] Y. Li, K. Xu, Q. Yan, Y. Li, and R. H. Deng, "Understanding osn-based facial disclosure against face authentication systems," in *Proceedings of the 9th ACM symposium on Information, computer and communications security*. ACM, 2014, pp. 413–424.
[10] J. Bai, T.-T. Ng, X. Gao, and Y.-Q. Shi, "Is physics-based liveness detection truly possible with a single image?" in *Circuits and systems (ISCAS), Proceedings of 2010 IEEE international symposium on*. IEEE, 2010, pp. 3425–3428.
[11] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *European Conference on Computer Vision*. Springer, 2010, pp. 504–517.
[12] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*. IEEE, 2009, pp. 233–236.
[13] "Fake fingerprint fools iPhone 6 TouchID." [Online]. Available: https://arstechnica.com/information-technology/2014/09/fake-fingerprint-fools-iphone-6-touch-id-why-its-not-so-serious/
[14] S. Chen, A. Pande, and P. Mohapatra, "Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 2014, pp. 109–122.
[15] Y. Li, Y. Li, Q. Yan, H. Kong, and R. H. Deng, "Seeing your face is not enough: An inertial sensor-based liveness detection for face authentication," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1558–1569.
[16] Y. Chen, J. Sun, X. Jin, T. Li, R. Zhang, and Y. Zhang, "Your face your heart: Secure mobile face authentication with photoplethysmograms," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*. IEEE, 2017, pp. 1–9.
[17] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 2018, pp. 321–336.
[18] "Guideline of android platforms." [Online]. Available: http://developer.android.com/reference/android/media/AudioRecord.html
[19] S. P. Tarzia, P. A. Dinda, R. P. Dick, and G. Memik, "Indoor localization without infrastructure using the acoustic background spectrum," in *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '11. ACM, 2011, pp. 155–168.
[20] C. J. Plack, *The sense of hearing*. Routledge, 2018.
[21] P. Lazik and A. Rowe, "Indoor pseudo-ranging of mobile devices using ultrasonic chirps," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '12. ACM, 2012, pp. 391–392.
[22] Y.-C. Tung and K. G. Shin, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. ACM, 2015, pp. 525–536.
[23] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: A high accuracy acoustic ranging system using cots mobile devices," in *Proceedings of the 5th International Conference on Embedded Networked Sensor Systems*, ser. SenSys '07. ACM, 2007, pp. 1–14.
[24] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
[25] J. Wang and D. Katabi, "Dude, where's my card?: Rfid positioning that works with multipath and non-line of sight," in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, ser. SIGCOMM '13. ACM, 2013, pp. 51–62.
[26] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining*. SIAM, 2001, pp. 1–11.
[27] W. Kim, S. Suh, and J.-J. Han, "Face liveness detection from a single image via diffusion speed model," *IEEE transactions on Image processing*, vol. 24, no. 8, pp. 2456–2465, 2015.
[28] M. Komeili, N. Armanfard, and D. Hatzinakos, "Liveness detection and automatic template updating using fusion of ecg and fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1810–1822, 2018.
[29] H. Yin, A. Zhou, L. Liu, N. Wang, and H. Ma, "Ubiquitous writer: Robust text input for small mobile devices via acoustic sensing," *IEEE Internet of Things Journal*, 2019.
[30] Z. Yu, H. Du, D. Xiao, Z. Wang, Q. Han, and B. Guo, "Recognition of human computer operations based on keystroke sensing by smartphone microphone," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1156–1168, 2018.
[31] T. Wang, D. Zhang, L. Wang, Y. Zheng, T. Gu, B. Dorizzi, and X. Zhou, "Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2959–2973, 2018.
[32] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1080–1091.
[33] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 57–71.

Huangxun Chen received the B.S. degree in computer science from Shanghai Jiao Tong University, Shanghai, China in 2015. She is currently pursuing the Ph.D. degree at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong.

Wei Wang (S'10-M'16) received the Ph.D. degree from the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology. He is currently a Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology. His research interests include PHY/MAC design and mobile computing in wireless systems. He served on TPC of INFOCOM and GBLOBECOM. He served as Editors for IJCS, China Communications, and Guest Editors for Wireless Communications and Mobile Computing and the IEEE COMSOC MMTC COMMUNICATIONS.

Jin Zhang (S'06-M'09) is currently an assistant professor in Electrical and Electronic Department, South University of Science and Technology of China. She graduated from Department of Electronic Engineering at Tsinghua University in 2004 with a bachelor's degree and in 2006 with a master's degree. She received the Ph.D. degree from Department of Computer Science and Engineering, Hong Kong University of Science and Technology. Her research interests are mainly in next-generation wireless networks, network economics, mobile computing in healthcare, cooperative communication and networks.

Qian Zhang (M'00-SM'04-F'12) joined Hong Kong University of Science and Technology in Sept. 2005 where she is a full Professor in the Department of Computer Science and Engineering. Before that, she was in Microsoft Research Asia, Beijing, from July 1999, where she was the research manager of the Wireless and Networking Group. She is a Fellow of IEEE for "contribution to the mobility and spectrum management of wireless networks and mobile communications". Dr. Zhang received the B.S., M.S., and Ph.D. degrees from Wuhan University, China, in 1994, 1996, and 1999, respectively, all in computer science.