Multi-Modal Face Authentication using Deep Visual and Acoustic Features

Bing Zhou, Zongxing Xie, Fan Ye
ECE Department, Stony Brook University
Stony Brook, NY 11790
Email: {bing.zhou, zongxing.xie, fan.ye}@stonybrook.edu

***Abstract*—User authentication on smartphones is the key to many applications, which must satisfy both security and convenience. We propose a multi-modal face authentication system, which pushes the limit of state-of-the-art image based face recognition solutions by incorporating a new dimension of sensing modality – acoustics. It actively emits almost inaudible acoustic signals from the earpiece speaker to "illuminate" the user's face and extracts features from the echoes using a customized convolutional neural network, which are fused with sophisticated visual features extracted from state-of-the-art face recognition models, for secure face authentication. Because the echo features depend on 3D facial geometries and material, our multi-modal design is not easily spoofed by images or videos like image based face recognition systems. It does not require any special sensors thus eliminating the extra costs in solutions like FaceID. Experiments show that our design achieves comparable face recognition performance to the state-of-the-art image based face authentication, while able to block image/video spoofing.**

I. Introduction

With ubiquitous access to the Internet via mobile devices, user authentication on smartphones has drawn much attention due to the plethora of daily Apps, such as social networks, shopping and banking [1, 2]. Traditional PIN number requires the user to remember/manage the corresponding PIN number/password for each account, which is inconvenient. Biometric based solutions are preferred due to their uniqueness and persistence for human subjects, while they suffer from security issues. For instance, face recognition based authentication can be easily spoofed by images or videos of the user [3]. Authentication of iris [4] and fingerprint [5], and the latest effort, Apple's FaceID [6] using an infrared depth sensor to sense the 3D shape of the face, achieve high security. But they all require extra special sensors and have constraints on deployment due to the limited screen space on smartphones. We seek to develop an alternative solution using existing sensors for user authentication with secure and convenient user experience.

In this paper, we propose a multi-modal user authentication system, which leverages both acoustic and visual features for secure and convenient face authentication, without the need of any special sensors. As shown in Figure 1, it combines user's facial features of both acoustic and vision extracted from pre-trained Convolutional Neural Networks (CNNs), and jointly trains a classification model that describes the user's face. Similar to FaceID, our facial features depend on 3D facial geometries, thus it is resilient to images/videos spoofing.

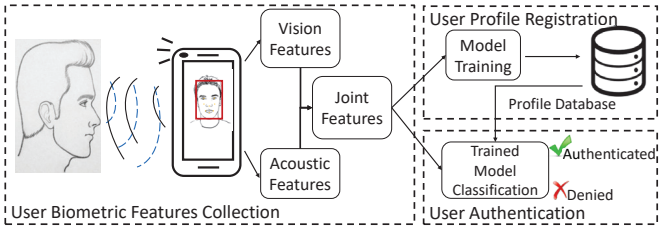

Fig. 1. The smartphone actively emits sound signal towards the user's face, and collects image and echo data for authentication. Sophisticated visual features from face recognition models and acoustic features extracted from a customized CNN are jointly used for the final classification.

It saves efforts of direct touch or managing passwords, thus avoiding the usability issues such as wet fingers that pose difficulties to fingerprint sensors and management overhead to PIN numbers.

To achieve resilient, secure and easy-to-use authentication using acoustic and vision, we must address the following two major challenges: i) echo signals are highly sensitive to the relative position between the user's face and the device (i.e., pose), which makes it extremely hard to extract reliable pose-invariant features for robust authentication; ii) sophisticated visual features are extracted using state-of-the-art face recognition model, which is not suitable for mobile devices.

We make the following contributions in this work:

- We propose a novel face authentication approach, which pushes the limit of existing image based face authentication solutions by incorporating both sophisticated visual features with acoustic sensing on smartphones.
- We design an end-to-end, distributed machine learning pipeline, which extracts reliable acoustic features on the mobile device using neural networks and offloads vision feature extraction to a server machine, thus making real-time recognition possible given the limited computational resources on mobile devices.
- We build a prototype, conduct extensive experiments and find that our solution inherits the advantages of face recognition capability as the state-of-the-art image based solutions, while it is resilient to images/video spoofing attacks. It achieves 99.96% precision and 88.84% recall in the tests of 10 participants.

To the best of our knowledge, this work is the first attempt leveraging two *sophisticated* features both extracted from deep

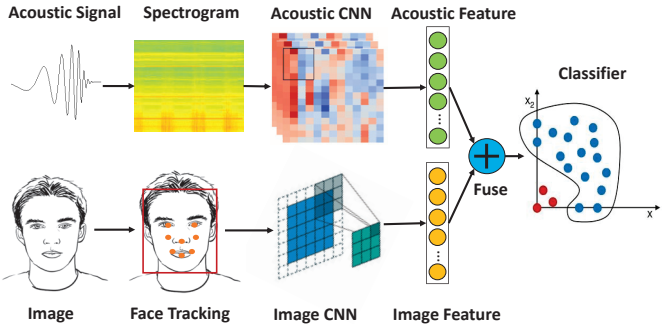

Fig. 2. The system takes both image and acoustic echoes as input, and extracts features using two pre-trained neural networks, which are fused for classification using SVM.

neural networks for smartphone user authentication, demonstrating robust performance without requiring any additional special sensor.

II. Overview

Considering the limitations of existing solutions, we aim to build a secure, convenient, and resilient multi-modal authentication system that is available to most existing smartphones without requiring any special sensors. By integrating the acoustic features extracted from our customized CNN with the sophisticated visual features, we believe such "free" acoustic-aided authentication will play an important role in mobile authentication developments. Figure 2 shows the overall design of our approach, which takes a joint biometric representation for user authentication, combining acoustic sensing and facial feature extraction. For authentication, the user just needs to hold the smartphone in front of the face for facial feature detection and acoustic sensing, and thereby the extracted representations are fed into the trained SVM classifier for final authentication.

III. System Design

Our system design has three major components: acoustic feature extraction, visual feature extraction, and multi-modal authentication.

A. Acoustic Feature Extraction

1) Sensing Hardware Selection: The earpiece speaker and the top microphone are selected as the acoustic sensing hardware combination, since it is a highly standard design across most smartphones. Besides, they are co-located with the frontal camera, thus causing less alignment problem with visual sensing.

2) Acoustic Signal Design: We choose a linear increasing frequency chirp (FMCW) as our base signal design due to its capability of distance measurement sensing. The linear increasing frequencies from 16 - 22KHz is selected based on our survey and experiments. To achieve higher SNR of echoes, a short signal length is set to avoid self-interference from the speaker, and a Hanning window [7] is applied to reshape the pulse envelop. The designed signal is almost inaudible to most users.

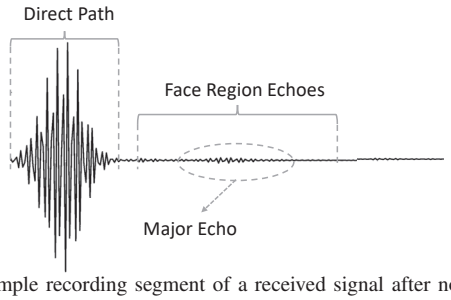

Fig. 3. Sample recording segment of a received signal after noise removal.

3) Acoustic Signal Processing: Before we can feed acoustic signal into the CNN for feature extraction, we need to remove background noise and segment out the echoes from the face area.

Background Noise Removal. The received raw signal goes through a 16 - 22KHz Butterworth band-pass filter to remove background noises, such that weak echoes from human faces will not be overwhelmed by the noise. A sample recording segment of a received signal after noise removal is shown in Figure 3. The *direct path* segment is the emitting signal traveling from speaker to the microphone directly, which ideally should be a copy of the emitting signal and has the highest amplitude. The *major echo* corresponds to the mix of echoes from the major surfaces (e.g., cheek, forehead) of the face. The *face region echoes* include all these echoes, capturing the full 3D geometry of the face.

Signal Segmentation. We take two steps to extract the face region segment. Firstly, the direct path segment in raw recordings is located based on the peak detection method. Then we leverage cross-correlation to locate the major echo thus face region segment after the direct path segment. We follow the similar approach used in our previous work EchoPrint [8], and segment the echoes from face regions. From our experiments, occasional offsets of direct path signal still happen after cross-correlation, due to ambiguities from comparable peak values in the cross-correlation result. Due to the hardware (speaker/microphone) imperfection, the received sound signal is usually slightly different from the designed emitting signal. To get an accurate "template" signal for cross-correlation, we perform emitting and recording in a quiet environment only once when the user registers, so that the direct path signal can be reliably detected and saved as a calibrated template for future cross-correlation in the authentication process.

Next, we use the similar approach for locating the major echo. However, human face echoes can be so weak that clutters nearby can have even stronger amplitudes. This makes the estimation unstable and leads to occasional location "jumping". We leverage the distance measured from vision and narrow down the search region of the major face echo, as the actual distance between phone and face can be estimated roughly but robustly from vision. For example, if the phone is closer to the face, then landmarks such as two outer corners of eyes are getting more apart on the image. By leveraging this vision-aided acoustic echo finding, we are able to reliably segment the major echo from the face. Since the depth of human face

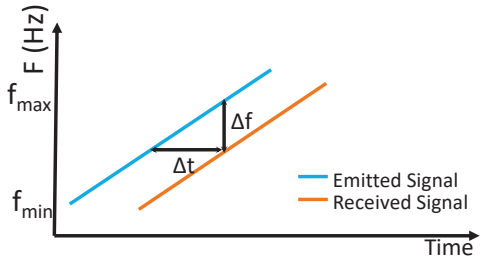

Fig. 4. Illustration of FMCW.

is limited, we extend 10 sample points before and after the major echo segment to cover the whole face region (allowing a depth range of $\sim 7cm$), which are later used as inputs for authentication.

FMCW Transformation. Since the face region echoes are a spatial and temporal combination of individual echoes, it is challenging to isolate individual echoes in time domain due to self-interference and noise. To measure the propagating distance of each echo, we adopt the Frequency-Modulated Continuous Wave (FMCW) technique [9] used in radars. Traditionally, the transmitter emits continuous chirp signals with linearly increasing frequency, from f_{min} to f_{max}. As Figure 4 shows, the frequency shift Δf between the received signal and the emitted signal is proportional to the elapsed time Δt, thus the relative distance given the sound wave propagation speed. The frequency shift Δf can be estimated by comparing the frequency of the echo signal to that of a reference signal using a technique called signal mixing. Therefore, finding Δf gives the distance (i.e., Δf multiplying a constant coefficient).

4) CNN Feature Extraction: The spectrogram of the segmented face region echoes after FMCW signal mixing is used as input for CNN training. We leverage the same CNN architecture as used in our previous work [8]. To use the pre-trained CNN as feature extractor, the last layer is removed so that we get a 128-dimensional vector as acoustic features.

B. Visual Feature Extraction

Getting a low-dimensional representation is crucial for efficient classification on mobile devices where the resource is limited. As the intrapersonal image variations such as angles, distances and even facial expressions can cause difficulty in classification, we adjust and normalize the face before the actual feature extraction.

Figure 5, shows the four stages to pre-process the image input for training the face representation neural network. On the mobile device, we get the aligned face image input once the App detects there appears the face in the red box while the red box stays in between the two green boxes as shown in Figure 6. We can drastically reduce the computation effort for further image processing with the alignment by the first stage. In the second stage, we aim to locate where the eyes, nose and lips are. To mitigate the constraints from illumination conditions, we leverage a pre-trained detector based on Histogram of Oriented Gradients [10] to layout the face landmarks. Next, we must consider the case that the

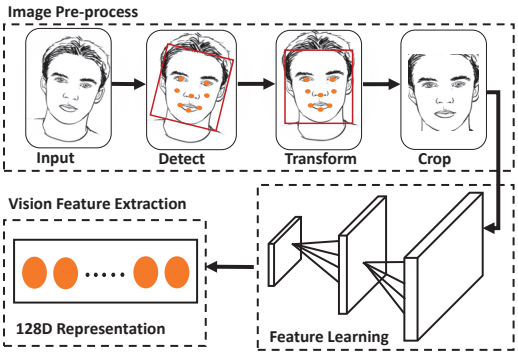

Fig. 5. Image pre-processing and visual feature extraction.

relative angle of the user's face may differ for each sample. To make it easier for facial recognition, thus authentication, we project all the face landmarks to our predefined positions using affine transformations, which have the expression shown the Equation 1:

$$T = A_{2\times2} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + B_{2\times1} = M_{2\times3} \cdot \begin{bmatrix} x, y, 1 \end{bmatrix}^T \quad (1)$$

$$M_{2\times3} = \begin{bmatrix} A_{2\times2} & B_{2\times1} \end{bmatrix}$$

where M is obtained based on the predefined landmark positions and the positions of detected landmarks in raw image inputs. The affine transformation provides a derived affine map for every pixel, such that no matter in what angle the raw image is taken, we can achieve a well adjusted and normalized image input for training. Obtaining the image with landmarks of known positions, we can crop the picture to have compact image thus further reducing the complexity in training.

Now that the reduced size of the normalized input space is obtained, we can have the deep convolutional network with less parameters to be trained to achieve a desirable low-dimensional representation, which can generalize well to faces that are new to the neural network. We achieve this goal by taking advantage of OpenFace's neural network[11], which is a reduced version of *nn4* proposed by Google's FaceNet[12]. The network is trained by using a combination of classification and *triplet loss*, which minimizes the distance between faces of the same identity and enforces a margin between different identities. After training, we leverage the pre-trained model as a feature extractor to map the face image input to a 128-dimensional vector, in which faces from the same identity should be close and form well separated clusters, such that they can be easily recognized/classified. And the extracted vision feature later will be combined with acoustic features together to form a joint embedding, which is a generic representation for anybody's face, for classification (i.e., final authentication). Instead of continuously training the deep neural network during the whole life of the face authentication application, we only need to update the classifiers (e.g., One-class SVM in our design) for the final authentication, and this is more cost efficient and practical for computing on mobile devices.

C. Authentication Model

We realize multi-modal authentication, fuse the acoustic and visual features as a joint description of a particular

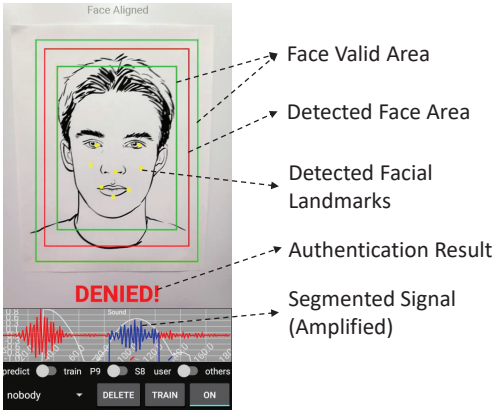

Fig. 6. Prototype user interface.

user, and train a one-class SVM, which is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set, based on the training data.

IV. Implementation

In the current stage, we offload the image feature extraction to a server to lower the computation complexity on the mobile. Apart from this, the face tracking, facial landmark detection, acoustic sensing and model inferences are running on the mobile. Figure 6 shows the user interface of our prototype on Android, where the acoustic data would be collected only when the user face is aligned within the valid area, which implicitly mitigates the impact of the smartphone's location variations on the acoustic signal. It is the same as the prototype we developed in [8], however the underlying implementation is different: this work offloads image data to a server for more sophisticated visual feature recognition, and combines with acoustic features locally extracted for final authentication. In an alternative implementation of our design, the sophisticated visual feature extraction is still portable to mobile devices as the computation power of mobile devices nowadays grows drastically and some light facial feature encoders are also available. Therefore, issues of network connection or privacy would be minimized.

V. Evaluation

A. Data Collection

For acoustic signal feature extraction, we leverage the pre-trained CNN model in our previous work EchoPrint [8] removing the last layer. We use OpenFace's model as the vision feature extractor, and it was trained based on *triplet loss* [11], which directly serves the goal of clustering, thus recognition and verification. Hereby, both acoustic and vision representations are 128-dimensional vectors. We invited 10 volunteers whose data are not used for the acoustic training. For each one of them, a $\sim 2mins$ acoustic data and 20 image samples are recorded for evaluation. These data are collected in uncontrolled environments where noise and lighting conditions may vary. The total acoustic samples count is 13806. Since the image data capturing is slower, we populate the image samples by interpolating images using the neighbor image to match the number of the acoustic samples as the visual features are relatively stable over a short time period. We also have 5 non-human classes: printed/displayed human faces on different materials such as paper, desktop monitor, photo on paper box, wall and a marble sculpture.

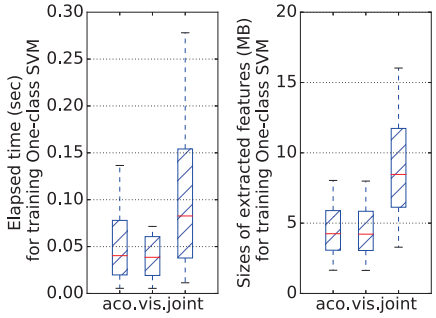

Fig. 7. Elapsed time and size of extracted features for training one-class SVM classifiers using acoustic, vision and joint features.

B. User Authentication Accuracy

To better evaluate the performance, we introduce precision, recall, F-score and balanced accuracy (BAC) as metrics. Precision is denoted as $P = \frac{TP}{TP+FP}$, and a high precision means the unauthorized user is seldom passed. Recall is denoted as $R = \frac{TP}{TP+FN}$, and a high recall means the authorized user is seldom denied. However, precision and recall alone can be misleading when the class distribution is imbalanced, whereby F-score and balanced accuracy (BAC) are introduced as both are insensitive to the class distribution. $F\text{-}score = 2\frac{P \cdot R}{P+R}$, is the harmonic mean of precision and recall with a best value of 1 and worst value of 0. $BAC = \frac{1}{2} \cdot (TPR + TNR)$, is the average of true positive rate ($TPR = \frac{TP}{TP+FN}$) and true negative rate ($TNR = \frac{TN}{TN+FP}$). A BAC of 1 means no false positive (i.e., successful attack) or false negative (i.e., denied access of legitimate users). For better convenience, we want to achieve lower false negative rate ($FNR = \frac{FN}{FN+TP}$), thus authenticated users can easily pass the verification; meanwhile, we want to achieve lower false positive rate ($FPR = \frac{FP}{TN+FP}$), thus unauthenticated users have little chance to pass the verification. Higher BAC hereby is desired.

To train the authentication model and evaluate the performance, we shuffle and split the collected data into two parts, 80% for training and 20% for testing. From Figure 7, we note that the elapsed time for training SVM with vision features is lower than that with acoustic features, even with the same amount of features. This is because the larger margin between different classes, the shorter time for training SVM with the same regularization parameter C, and the visual feature extractor is trained based on *triplet loss* [12], which encourages clustering representations of different identities.

To analyze the impact of acoustic, visual and the combination of both on the overall authentication performance, we compare the above performance metrics using individual features and the joint features. For each user, we use the

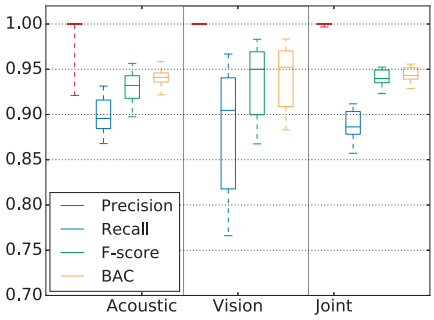

Fig. 8. The precision, recall, f-score and BAC of one-class SVM model using acoustic, vision and joint features.

TABLE I
MEAN/MEDIAN AUTHENTICATION ACCURACY OF NEW USERS WITH VISION, ACOUSTIC AND JOINT FEATURES.

	Acoustic	Vision	Joint
Precision (%)	98.62 / 100.0	100.0 / 100.0	99.96 / 100.0
Recall (%)	89.83 / 89.56	87.98 / 90.46	88.84 / 88.63
F-score (%)	93.15 / 93.21	93.46 / 94.99	94.07 / 93.97
BAC (%)	94.15 / 94.10	93.99 / 95.23	94.42 / 94.32

positive samples in its test set as positive testing samples and use all the data from other users as negative samples, trying to attack the model. Figure 8 shows the results. As we can see, leveraging acoustic features only, the precision is above 90% with large variances, which is inferior to using visual feature or the joint feature. However, the recall and F-score are significantly better compared to visual features, and slightly better than the joint features. This is because 1) when registering new users, the number of vision samples is limited; 2) when verifying the authentication, joint features examine both modalities in order to produce a positive prediction, which brings down the recall slightly. Table I shows the authentication results of new users. Thanks to the sophisticated visual features, the precision of joint features is better than pure acoustic features, demonstrating higher security. Note that, the results are based on the data collected from real human subjects, however the performance of the vision based method would deteriorate in the presence of image spoofing attacks, where the joint feature method can outperform the others. In our extensive experiments, we even pass a twin test, i.e., able to distinguish twin sisters from each other using the joint features. Since we only test one pair of twins as a preliminary evaluation, more experiments and analysis will be conducted in our future work. Next, we evaluate the capability of leveraging different features for image/video anti-spoofing.

C. Image/Video Spoofing Attacks.

Anti-spoofing capability, especially image/video spoofing, is of great importance for face authentication systems. We take photos of the participants, and leverages photos (printed on paper, or displayed on electronic devices such as smartphones or tablets) to attack the authentication model. As expected, these attacks pass pure image based face recognition solutions [11] easily. More advanced vision solutions incorporate liveness detection, for example, requiring eye blinks when the user is doing authentication. These methods require active interaction from the users and can also be spoofed by recorded videos. Leveraging both acoustic and visual features, our design shows the capability of anti-spoofing as no successful attacks are observed. This is not surprising: such objects create significantly different acoustic features compared to that of human faces. However, we admit that large scale experiments are needed to verify the robustness of our system against spoofing attacks in future work.

TABLE II
MEAN/MAX RESOURCE CONSUMPTION.

Device	Memory (MB)	CPU (ms)
$S8$	27.0 / 51.5	6.54 / 31.25
$P9$	32.5 / 61.3	8.37 / 27.63

D. Impact of Background Noise.

It is necessary to investigate the impact of background noise on the performance when acoustic sensing is used. We collect data samples under background noise in multiple environments (noisy laboratory and crowded classroom with people talking and walking by). No obvious performance degradation is observed, which demonstrates the robustness against background noise in our daily scenarios.

E. Resource Consumption.

We evaluate memory, CPU usage using the Android Studio IDE Profiler tool, and power consumption using Qualcomm's Trepn Profiler tool [13] on our test device. Table II shows the resource consumption on two testing devices, Samsung Galaxy S9 and Huawei P9 smartphones. The memory consumption has an average $\sim 30MB$ and maximum $\sim 50-60MB$, which appears when CNN feature extraction using tensorflow inference is running. The average amount of time for the CPU to complete all the machine learning inferences is low on all phones ($6.5 \sim 8.5ms$). The maximum CPU time is around $\sim 30ms$, still very low. Compared with resource consumption evaluated in EchoPrint [8], our latest design leveraging sophisticated visual features only requires slightly more computation resource. This is because of the heaviest computation task – sophisticated visual feature extraction – is offloaded to the server. The image data is $\sim 1.4MB$ for each test. Depending on the network quality, the delay time varies. However, unless the wireless networking is highly congested, the delay should be acceptable for most use cases.

VI. Related Work

User Authentication. Personal Identification Number (PIN) is widely used due to simplicity, however it is easily exposed to someone close by or forgotten by the user. Speech recognition and vision based face recognition both suffer the replay attack where the voice/image/video is recorded. Fingerprint sensors have achieved great security and convenience. However the sensor takes a lot of precious space, and it is proven susceptible to attacks [5]. Apple's FaceID [6] uses special TrueDepth sensors, bringing extra hardware costs and requiring significant

design changes. A similar approach, EchoPrint [8] combines acoustic features and coordinates of a small set of facial landmarks for authentication using existing sensors. It does not utilize sophisticated 2D visual features, which should be incorporated for better performance. Unlike all the above solutions, our multi-modal solution is the first to leverage sophisticated acoustic-visual joint embeddings for user authentication.

Acoustic-based Face Recognition. Acoustics has been used for face recognition in some prior work [14–16]. I. E. Dror *et al.* [15] recognize five human faces with an accuracy over 96% using special ultrasonic sensors. K. Kalgaonkar *et al.* [16] propose a sensing mechanism based on the Doppler effect to recognize talking faces using ultrasound. K.K. Yoong *et al.* [14] classify up to 10 still faces with an accuracy of 99.73% using hand-crafted features from ultrasound echo signals. Compared to all the above work using special ultrasonic sensors which are not available in consumer electronics, our solution uses commodity smartphone speakers and microphones, and combine them with sophisticated visual features from camera for authentication.

Acoustic Sensing on Smartphones. Acoustic sensing is widely used on mobile platforms for localization, tracking, vita signal monitoring, etc. EchoTag [17] recognizes different locations leveraging unique echo frequency responses, a series of work [18–20] builds indoor floor plans using echo signals, and BatTracker [21] enables high-precision infrastructure-free mobile device tracking. ApenaApp [22] monitors the minute chest and abdomen breathing movements using FMCW [23], and SonarBeat [24] monitors breathing beat using signal phase shifts. Compared to them, our solution leverages acoustic features from deep neural networks and combined them with sophisticated visual features for user authentication.

VII. Conclusion

In this paper, we propose a multi-modal user authentication solution leveraging both sophisticated acoustic and visual features from deep neural networks for smartphones, without requiring any special sensors. Experiment results show that our multi-modal design has comparable face recognition performance as state-of-the-art 2D image based solutions, while it is resilient to image/video attacks which is a well known drawback for 2D image based solutions.

Acknowledgement

This work is supported in part by US NSF grants 1513719, 1730291.

References

[1] Google. (2014) Android pay. https://www.android.com/pay/.
[2] Apple. (2018) Apple pay. https://www.apple.com/apple-pay/.
[3] N. M. Duc and B. Q. Minh, "Your face is not your password face authentication bypassing lenovo–asus–toshiba," *Black Hat Briefings*, vol. 4, p. 158, 2009.
[4] J. G. Daugman, "Biometric personal identification system based on iris analysis," Mar. 1 1994, uS Patent 5,291,560.
[5] (12/3/2007) How to fool a fingerprint security system as easy as abc. http://www.instructables.com/id/How-To-Fool-a-Fingerprint-Security-System-As-Easy-/.
[6] "About face id advanced technology," https://support.apple.com/en-us/HT208108, 6/8/2018.
[7] E. C. Ifeachor and B. W. Jervis, *Digital signal processing: a practical approach*. Pearson Education, 2002.
[8] 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.
[9] K. G. Derpanis, "Overview of the ransac algorithm," *Image Rochester NY*, vol. 4, no. 1, pp. 2–3, 2010.
[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *international Conference on computer vision & Pattern Recognition (CVPR'05)*, vol. 1. IEEE Computer Society, 2005, pp. 886–893.
[11] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016.
[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
[13] "Qualcomm trepn power profiler," https://developer.qualcomm.com/ software/trepn-power-profiler, 2018.
[14] P. McKerrow and K. K. Yoong, "Classifying still faces with ultrasonic sensing," *Robotics and Autonomous Systems*, vol. 55, no. 9, pp. 702–710, 2007.
[15] I. E. Dror, F. L. Florer, D. Rios, and M. Zagaeski, "Using artificial bat sonar neural networks for complex pattern recognition: Recognizing faces and the speed of a moving target," *Biological Cybernetics*, vol. 74, no. 4, pp. 331–338, 1996.
[16] K. Kalgaonkar and B. Raj, "Recognizing talking faces from acoustic doppler reflections," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.
[17] K. G. S. Yu-Chih Tung, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," *MobiCom '15 Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pp. 525–536, 2015.
[18] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2017, pp. 42–55.
[19] ——, "Demo: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 2017, pp. 519–521.
[20] ——, "Towards scalable indoor map construction and refinement using acoustics on smartphones," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2019.
[21] ——, "Battracker: High precision infrastructure-free mobile device tracking in indoor environments," in *Proceedings of the 15th ACM Conference on Embedded Networked Sensor Systems (SenSys 2017)*. ACM, 2017.
[22] R. Nandakumar, S. Gollakota, and N. Watson, "Contactless sleep apnea detection on smartphones," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2015, pp. 45–57.
[23] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F-Radar and Signal Processing*, vol. 139, no. 5. IET, 1992, pp. 343–350.
[24] X. Wang, R. Huang, and S. Mao, "Sonarbeat: Sonar phase for breathing beat monitoring with smartphones," in *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*. IEEE, 2017, pp. 1–8.