Fusing information from multiple biometric traits enhances authentication in mobile devices.

BY MIKHAIL I. GOFMAN AND SINJINI MITRA

Multimodal Biometrics for Enhanced Mobile Device Security

Millions of mobile devices are stolen every year, along with associated credit card numbers, passwords, and other secure and personal information stored therein. Over the years, criminals have learned to crack passwords and fabricate biometric traits and have conquered practically every kind of user-authentication mechanism designed to stop them from accessing device data. Stronger mobile authentication mechanisms are clearly needed.

Here, we show how multimodal biometrics promises untapped potential for protecting consumer mobile devices from unauthorized access, an authentication approach based on multiple physical and behavioral traits like face and voice. Although multimodal biometrics are deployed in homeland security, military, and law-enforcement applications, they are not yet widely integrated into consumer mobile devices. This can be attributed to implementation challenges and con-

key insights

- Multimodal biometrics, or identifying people based on multiple physical and behavioral traits, is the next logical step toward more secure and robust biometrics-based authentication in mobile devices.
- The face-and-voice-based biometric system covered here, as implemented on a Samsung Galaxy S5 phone, achieves greater authentication accuracy in uncontrolled conditions, even with poorly lit face images and voice samples, than single-modality face and voice systems.
- Multimodal biometrics on mobile devices can be made user friendly for everyday consumers.
cern that consumers may find the approach inconvenient.

We also show multimodal biometrics can be integrated with mobile devices in a user-friendly manner and significantly improve their security. In 2015, we thus implemented a multimodal biometric system called Proteus at California State University, Fullerton, based on face and voice on an Samsung Galaxy S5 phone, integrating new multimodal biometric authentication algorithms optimized for consumer-level mobile devices and an interface that allows users to readily record multiple biometric traits. Our experiments confirm it achieves considerably greater authentication accuracy than systems based solely on face or voice alone. The next step is to integrate other biometrics (such as fingerprints and iris scans) into the system. We hope our experience encourages researchers and mobile-device manufacturers to pursue the same line of innovation.

**Biometrics**

Biometrics-based authentication establishes identity based on physical and behavioral characteristics (such as face and voice), relieving users from having to create and remember secure passwords. At the same time, it challenges attackers to fabricate human traits that, though possible, is difficult in practice.21 These advantages continue to spur adoption of biometrics-based authentication in smartphones and tablet computers.

Despite the arguable success of biometric authentication in mobile devices, several critical issues remain, including, for example, techniques for defeating iPhone TouchID and Samsung Galaxy S5 fingerprint recognition systems.2,26 Further, consumers continue to complain that modern mobile biometric systems lack robustness and often fail to recognize authorized users.4 To see how multimodal biometrics can help address these issues, we first examine their underlying causes.

**The Mobile World**

One major problem of biometric authentication in mobile devices is sample quality. A good-quality biometric sample—whether a photograph of a face, a voice recording, or a fingerprint scan—is critical for accurate identification; for example, a low-resolution photograph of a face or
noisy voice recording can lead a biometric algorithm to incorrectly identify an impostor as a legitimate user, or “false acceptance.” Likewise, it can cause the algorithm to declare a legitimate user an impostor, or “false rejection.” Capturing high-quality samples in mobile devices is especially difficult for two main reasons. Mobile users capture biometric samples in a variety of environmental conditions; factors influencing these conditions include insufficient lighting, different poses, varying camera angles, and background noise. And biometric sensors in consumer mobile devices often trade sample quality for portability and lower cost; for example, the dimensions of an Apple iPhone’s TouchID fingerprint scanner prohibit it from capturing the entire finger, making it easier to circumvent.

Another challenge is training the biometric system to recognize the device user. The training process is based on extracting discriminative features from a set of user-supplied biometric samples. Increasing the number and variability of training samples increases identification accuracy. In practice, however, most consumers likely train their systems with few samples of limited variability for reasons of convenience. Multimodal biometrics is the key to addressing these challenges.

**Promise of Multimodal Biometrics**

Due to the presence of multiple pieces of highly independent identifying information (such as face and voice), multimodal systems can address the security and robustness challenges confronting today’s mobile unimodal systems that identify people based on a single biometric characteristic. Moreover, deploying multimodal biometrics on existing mobile devices is practical; many of them already support face, voice, and fingerprint recognition. What is needed is a robust user-friendly approach for consolidating these technologies. Multimodal biometrics in consumer mobile devices deliver multiple benefits.

**Increased mobile security.** Attackers can defeat unimodal biometric systems by spoofing a single biometric modality used by the system. Establishing identity based on multiple independent traits—a significantly tougher challenge.

**More robust mobile authentication.** When using multiple biometrics, one biometric modality can be used to compensate for variations and quality deficiencies in the others; for example, Proteus assesses face-image and voice-recording quality and lets the highest-quality sample have greater impact on the identification decision.

Likewise, multimodal biometrics can simplify the device-training process. Rather than provide many training samples from one modality (as they often must do in unimodal systems), users can provide fewer samples from multiple modalities. This identifying information can be consolidated to ensure sufficient training data for reliable identification.

**A market ripe with opportunities.** Despite the recent popularity of biometric authentication in consumer mobile devices, multimodal biometrics have had limited penetration in the mobile consumer market. This can be attributed to the concern users could find it inconvenient to record multiple biometrics. Multimodal systems can also be more difficult to design and implement than unimodal systems.

However, as we explain, these problems are solvable. Companies like Apple and Samsung have invested significantly in integrating biometric sensors (such as cameras and fingerprint readers) into their products. They can thus deploy multimodal biometrics without substantially increasing their production costs. In return, they profit from enhanced device sales due to increased security and robustness.

In the following sections we discuss how to achieve such profitable security.

**Fusing Face and Voice Biometrics**

To illustrate the benefits of multimodal biometrics in consumer mobile devices, we implemented Proteus based on face and voice biometrics, choosing these modalities because most mobile devices have cameras and microphones needed for capturing them. Here, we provide an overview of face- and voice-recognition techniques, followed by an exploration of the approaches we used to reconcile them.

**Face and voice recognition.** We used the face-recognition technique known as FisherFaces in Proteus, as it works well in situations where images are captured under varying conditions, as...
expected in the case of face images obtained through mobile devices. Fisher-Faces uses pixel intensities in the face images as identifying features. In the future, we plan to explore other face-recognition techniques, including Gabor wavelets and Histogram Oriented Gradients (HOG).

We used two approaches for voice recognition: Hidden Markov Models (HMM) based on the Mel-Frequency Cepstral Coefficients (MFCCs) as voice features, the basis of our score-level fusion scheme; and Linear Discriminant Analysis (LDA), the basis for our feature-level fusion scheme. Both approaches recognize a user’s voice independent of phrases spoken.

Assessing face and voice sample quality. Assessing biometric sample quality is important for ensuring the accuracy of any biometric-based authentication system, particularly for mobile devices, as discussed earlier. Proteus thus assesses facial image quality based on luminosity, sharpness, and contrast, while voice-recording quality is based on signal-to-noise ratio (SNR). These classic quality metrics are well documented in the biometrics research literature. We plan to explore other promising metrics, including face orientation, in the future.

Proteus computes the average luminosity, sharpness, and contrast of a face image based on the intensity of the constituent pixels using approaches described in Nasrolli and Moeslund. It then normalizes each quality measure using the min-max normalization method to lie between [0, 1], finally computing their average to obtain a single quality score for a face image. One interesting problem here is determining the impact each quality metric has on the final face-quality score; for example, if the face image is too dark, then poor luminosity would have the greatest impact, as the absence of light would be the most significant impediment to recognition. Likewise, in a well-lit image distorted due to motion blur, sharpness would have the greatest impact.

SNR is defined as a ratio of voice signal level to the level of background noise signals. To obtain a voice-quality score, Proteus adapts the probabilistic approach described in Vondrasek and Pollak to estimate the voice and noise signals, then normalizes the SNR value to the [0, 1] range using min-max normalization.

Multimodal biometric fusion. In multimodal biometric systems, information from different modalities can be consolidated, or fused, at the following levels:

Feature. Either the data or the feature sets originating from multiple sensors and/or sources are fused;

Match score. The match scores generated from multiple trait-matching algorithms pertaining to the different biometric modalities are combined, and

Decision. The final decisions of multiple matching algorithms are consolidated into a single decision through techniques like majority voting.

Biometric researchers believe integrating information at earlier stages of processing (such as at the feature level) is more effective than having integration take place at a later stage (such as at the score level).

Multimodal Mobile Biometrics Framework

Proteus fuses face and voice biometrics at either score or feature level. Since decision-level fusion typically produces only limited improvement, we did not pursue it when developing Proteus.

Proteus does its training and testing processes with videos of people holding a phone camera in front of their faces while speaking a certain phrase. From each video, the face is detected through the Viola-Jones algorithm and the system extracts the soundtrack. The system de-noises all sound frames to remove frequencies outside human voice range (85Hz–255Hz) and drops frames without voice activity. It then uses the results as inputs into our fusion schemes.

Score-level fusion scheme. Figure 1 outlines our score-level fusion approach, integrating face and voice biometrics. The contribution of each modality’s match score toward the final decision concerning a user’s authenticity is determined by the respective sample quality. Proteus works as outlined in the following paragraphs.

Let \( t_1 \) and \( t_2 \), respectively, denote the average face- and voice-quality scores of the training samples from the user of the device. Next, from a
Storing and processing biometric data on the mobile device itself, rather than offloading these tasks to a remote server, eliminates the challenges of securely transmitting the biometric data and authentication decisions across potentially insecure networks.
Phase 1 (face feature extraction). The Proteus algorithm applies Principal Component Analysis (PCA) to the face feature set to perform feature selection;

Phase 2 (voice feature extraction). It extracts a set of MFCCs from each preprocessed audio frame and represents them in a matrix form where each row is used for each frame and each column for each MFCC index. And to reduce the dimensionality of the MFCC matrix, it uses the column means of the matrix as its voice feature vector;

Phase 3 (fusion of face and voice features). Since the algorithm measures face and voice features using different units, it standardizes them individually through the z-score normalization method, as in score-level fusion. The algorithm then concatenates these normalized features to form one big feature vector;

Phase 4 (authentication). The algorithm uses Euclidean distance to determine the degree of similarity between the fused features sets from the training data and each test sample. If the distance value is less than or equal to a predetermined threshold, it accepts the test subject as a legitimate user. Otherwise, the subject is declared an impostor.

Implementation

We implemented our quality-based score-level and feature-level fusion approaches on a randomly selected Samsung Galaxy S5 phone. User friendliness and execution speed were our guiding principles.

User interface. Our first priority when designing the interface was to ensure users could seamlessly capture face and voice biometrics simultaneously. We thus adopted a solution that asks users to record a short video of their faces while speaking a simple phrase. The prototype of our graphical user interface (GUI) (see Figure 3) gives users real-time feedback on the quality metrics of their face and voice, guiding them to capture the best-quality samples possible; for example, if the luminosity in the video differs significantly from the average luminosity of images in the training database, the user may get a prompt saying, Suggestion: Increase lighting.

In addition to being user friendly, the video also facilitates integration of other security features (such as liveness checking) and correlation of lip movement with speech.

To ensure fast authentication, the Proteus face- and voice-feature extraction algorithms are executed in parallel on different processor cores; the Galaxy S5 has four cores. Proteus also uses similar parallel programming techniques to help ensure the GUI’s responsiveness.

Security of biometric data. The greatest risk from storing biometric data on a mobile device (Proteus stores data from multiple biometrics) is the possibility of attackers stealing and using it to impersonate a legitimate user. It is thus imperative that Proteus stores and processes the biometric data securely.

The current implementation stores only MFCCs and PCA coefficients in the device persistent memory, not raw biometric data, from which deriving useful biometric data is nontrivial. Proteus can enhance security significantly by using cancelable biometric templates and encrypting, storing, and processing biometric data in Trusted Execution Environment tamper-proof hardware highly isolated from the rest of
the device software and hardware; the Galaxy S5 uses this approach to protect fingerprint data.22

Storing and processing biometric data on the mobile device itself, rather than offloading these tasks to a remote server, eliminates the challenge of securely transmitting the biometric data and authentication decisions across potentially insecure networks. In addition, this approach alleviates consumers’ concern regarding the security, privacy, and misuse of their biometric data in transit to and on remote systems.

Performance Evaluation

We compared Proteus recognition accuracy to unimodal systems based on face and voice biometrics. We measured that accuracy using the standard equal error rate (EER) metric, or the value where the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. Mechanisms enabling secure storage and processing of biometric data must therefore be in place.

Database. For our experiments, we created a CSUF-SG5 homegrown multimodal database of face and voice samples collected from the University of California, Fullerton, students, employees, and individuals from outside the university using the Galaxy S5 (hence the name). To incorporate various types and levels of variations and distortions in the samples, we collected them in a variety of real-world settings. Given such a diverse database of multimodal biometrics is unavailable, we plan to make our own one publicly available. The database today includes video recordings of 54 people of different genders and ethnicities holding a phone camera in front of their faces while speaking a certain simple phrase.

The faces in these videos show the following types of variations:

Four expressions. Neutral, happy, sad, angry, and scared;

Three poses. Frontal and sideways (left and right); and

Two illumination conditions. Uniform and partial shadows.

The voice samples show different levels of background noise, from car traffic to music to people chatter, coupled with distortions in the voice itself (such as raspiness). We used 20 different popular phrases, including “Roses are red,” “Football,” and “13.”

Results. In our experiments, we trained the Proteus face, voice, and fusion algorithms using videos from half of the subjects in our database (27 subjects out of a total of 54), while we considered all subjects for testing. We collected most of the training videos in controlled conditions with good lighting and low background noise levels and with the camera held directly in front of the subject’s face. For these subjects, we also added a few face and voice samples from videos of less-than-ideal quality (to simulate the limited variation of training samples a typical consumer would be expected to provide) to increase the algorithm’s chances of correctly identifying the user in similar conditions. Overall, we used three face frames and five voice recordings per subject (extracted from video) as training samples. We performed the testing through a randomly selected face-and-voice sample from a subject we selected randomly from among the 54 subjects in the database, leaving out the training samples. Overall, our subjects created and used 480 training and test-set combinations, and we averaged their EERs and testing times. We undertook this statistical cross-validation approach to assess and validate the effectiveness of our proposed approaches based on the available database of 54 potential subjects.

Quality-based score-level fusion. Table 1 lists the average EERs and testing times from the unimodal and multimodal schemes. We explain the high EER of our HMM voice-recognition algorithm by the complex noise signals in many of our samples, including traffic, people chatter, and music, that were difficult to detect and eliminate. Our quality-score-level fusion scheme detected low SNR levels and compensated by adjusting weights in favor of the face images that were of substantially better quality. By adjusting weights in favor of face images, the face biometric thus had a greater impact on the final decision of whether or not a user is legitimate than the voice biometric.

For the contrasting scenario, where voice samples were relatively better quality than face samples, as in Table 1, the EERs were 21.25% and 20.83% for unimodal voice and score-level fusion, respectively.

These results are promising, as they show the quality of the different modalities can vary depending on the circumstances in which mobile users might find themselves. They also show Proteus adapts to different conditions by scaling the quality weights appropriately. With further refinements (such as more robust normalization techniques), the multimodal method can yield even better accuracy.

Feature-level fusion. Table 2 outlines our performance results from the feature-level fusion scheme, showing feature-level fusion yielded significantly greater accuracy in authentication compared to unimodal schemes.

Our experiments clearly reflect the potential of multimodal bio-

<table>
<thead>
<tr>
<th>Table 1. EER results from score-level fusion.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modality</strong></td>
</tr>
<tr>
<td>Face</td>
</tr>
<tr>
<td>Voice</td>
</tr>
<tr>
<td>Score-Level fusion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. EER results from feature-level fusion.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modality</strong></td>
</tr>
<tr>
<td>Face</td>
</tr>
<tr>
<td>Voice</td>
</tr>
<tr>
<td>Feature-Level fusion</td>
</tr>
</tbody>
</table>
metrics to enhance the accuracy of current unimodal biometrics-based authentication on mobile devices; moreover, according to how quickly the system is able to identify a legitimate user, the Proteus approach is scalable to consumer mobile devices. This is the first attempt at implementing two types of fusion schemes on a modern consumer mobile device while tackling the practical issues of user friendliness. It is also just the beginning. We are working on improving the performance and efficiency of both fusion schemes, and the road ahead promises endless opportunity.

Conclusion
Multimodal biometrics is the next logical step in biometric authentication for consumer-level mobile devices. The challenge remains in making multimodal biometrics usable for consumers of mainstream mobile devices, but little work has sought to add multimodal biometrics to them. Our work is the first step in that direction. Imagine a mobile device you can unlock through combinations of face, voice, fingerprints, ears, irises, and retinas. It reads all these biometrics in one step similar to the iPhone’s TouchID fingerprint system. This user-friendly interface utilizes an underlying robust fusion logic based on biometric sample quality, maximizing the device’s chance of correctly identifying its owner. Dirty fingers, poorly illuminated or loud settings, and damage to biometric sensors are no longer showstoppers; if one biometric fails, others function as backups. Hackers must now consider biometric schemes, and the road ahead promises endless opportunity.

References
23. Zarobedian, J. Samsung Galaxy S5 fingerprint reader hacked—It’s the iPhone 5S all over again! Naked Security (Apr. 17, 2014); https://makesecuritycool.com/2014/04/17/samsung-galaxy-s5-fingerprint-hacked-iphone-5s-all-over-again/

Mikhail I. Gofman (mgofman@fullerton.edu) is an assistant professor in the Department of Computer Science at California State University, Fullerton, and director of its Center for Cybersecurity.

Sinjini Mitr (smir@fullerton.edu) is an assistant professor of information systems and decision sciences at California State University, Fullerton.

Copyright held by authors.
Publication rights licensed to ACM. $15.00

 contributed articles

APRIL 2016 | VOL. 59 | NO. 4 | COMMUNICATIONS OF THE ACM 65