Biometric Authentication on a Mobile Device: A Study of User Effort, Error and Task Disruption

Shari Trewin[1], Cal Swart[1], Larry Koved[1], Jacquelyn Martino[1], Kapil Singh[1], Shay Ben-David[2]
[1]IBM T.J. Watson Research Center
{trewin, cals, koved, jmartino, kapil}@us.ibm.com
[2]IBM Research Haifa
bendavid@il.ibm.com

ABSTRACT

We examine three biometric authentication modalities – voice, face and gesture – as well as password entry, on a mobile device, to explore the relative demands on user time, effort, error and task disruption. Our laboratory study provided observations of user actions, strategies, and reactions to the authentication methods. Face and voice biometrics conditions were faster than password entry. Speaking a PIN was the fastest for biometric sample entry, but short-term memory recall was better in the face verification condition. None of the authentication conditions were considered very usable. In conditions that combined two biometric entry methods, the time to acquire the biometric samples was shorter than if acquired separately but they were very unpopular and had high memory task error rates. These quantitative results demonstrate cognitive and motor differences between biometric authentication modalities, and inform policy decisions in selecting authentication methods.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Authentication; H.5.2 [**User Interfaces**]: Interaction styles

General Terms

Security, Human Factors

Keywords

Authentication, mobile, biometric, usability

1. INTRODUCTION

Mobile devices are rapidly becoming a key computing platform, transforming how people access business and personal information. Access to business data from mobile devices requires secure authentication, but traditional password schemes based on a mix of alphanumerics and symbols are cumbersome and unpopular, leading users to avoid accessing business data on their personal devices altogether [7].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACSAC '12 Dec. 3-7, 2012, Orlando, Florida USA
Copyright 2012 ACM 978-1-4503-1312-4/12/12 ...$15.00.

The rich set of input sensors on mobile devices, including cameras, microphones, touch screens, and GPS, enable sophisticated multi-media interactions. Biometric authentication methods using these sensors could offer a natural alternative to password schemes, since the sensors are familiar and already used for a variety of mobile tasks.

User frustration with password-based authentication on mobile devices demonstrates that a high level of usability must be achieved for a mobile authentication technique to be accepted. As biometric recognition algorithms continue to improve, the user experience will be an increasingly critical factor in the success of such techniques.

In this paper, we explore authentication techniques on mobile devices from the users' point of view. We study three biometric authentication modalities - voice, face and gesture, and combinations of voice with face and gesture. A typical 8-character password condition is included as a baseline.

This study is the first to measure user action times for authentication using different biometrics on a mobile device. It provides insight into user performance when using these techniques under favorable conditions.

The study examined:

1. The time taken to provide an authentication sample (password, biometric, or two biometrics);
2. Error rates in providing a sample of suitable quality for analysis by verification algorithms;
3. The impact of the user actions required for authentication on performance in a memory recall task; and
4. User reactions to the authentication methods.

To allow for comparison between authentication methods, the voice and gesture conditions use the same 8-digit authentication token. We find that speaking was the fastest biometric authentication method, but taking a photograph supported better performance in the memory recall task. Speaker verification was considered less usable than password, face and gesture (writing an 8-digit PIN). Combination conditions – simultaneously entering two biometric samples – were very unpopular. Failure rates were not significantly different among single conditions, but combining methods led to high error rates.

2. BACKGROUND AND RELATED WORK

2.1 The Mobile Context and Authentication

The proliferation of smartphones, such as those based on Apple, Android, Microsoft and Blackberry technologies, is rapidly changing the nature of interactive computing. Much

of this is driven by the multitude of digital sensors embedded within these devices, including GPS, touch screens, cameras and microphones. As a result, peoples' expectations around ease of use of mobile devices are changing.

Simple gestures (e.g. Android screen lock pattern), graphical passwords [11], and biometric authentication [22] are beginning to emerge as alternative mobile authentication mechanisms, but passwords and PINs remain the most common methods used today. Corporate use of mobile devices is frequently dictating the use of password strength policies, derived from desktop password policies, for device screen unlock. A typical company password policy requires a mix of alphabetic and numeric or symbol characters [7].

Bao et al [7] measured the time to type an 8-character, mixed-case alphanumeric password on desktops and mobile phones. On mobile devices with soft keyboards, entry of compliant passwords often requires the user to switch between different keyboard layouts. They found that while participants typed the password at 17wpm on a desktop computer, they only achieved a mean of 6wpm on their own phones. Mobile device users are acutely aware of this additional effort. Their participants found password typing on a mobile phone so onerous that they avoided business data access on their phones because it would have required a corporate-compliant device unlock password.

Even in desktop environments, users often select poor quality passwords [12][13]. The perceived effort of entering passwords on mobile devices will encourage further password simplification, for example placing non-alphabetic characters only at the beginning or end of the password. Recall aids such as writing down passwords and physically attaching them to devices [31] pose additional security risks for password authentication in a mobile context.

Interaction with mobile devices tends to be brief and interruption driven [24][25]. As a result, mobile devices have been caching the security credentials in the device to make it easier for users to authenticate. The result is that mobile devices have effectively become authentication tokens (e.g., [1][17]). Given that mobile devices are often borrowed [18], and perceived to be more frequently lost or stolen [23], users' personal and business resources are at greater risk of being lost or compromised.

2.2 Mobile Biometric Authentication

Biometric authentication is a well-studied area of research. Physical biometrics, such as face, voice and signature, are the most commonly used forms. Biometrics authentication systems have been evaluated against a rich set of metrics that incorporate both performance and usability aspects [10]. User attitudes have been explored [14][19][30], but relatively little attention has been paid to empirical comparison of the usability of biometric authentication methods. Toledano et. al's usability evaluation of multimodal (non-mobile) biometric authentication systems [32] is a notable exception. It proposes a testing framework for biometric usability analysis that uses ISO usability factors (i.e., effectiveness, efficiency and satisfaction) for evaluation.

We believe that the era of using biometric authentication for mobile devices is imminent. People are now accustomed to talking into small mobile devices, and seeing themselves through the device camera. As the quality of sensors and processing power of mobile devices improves, mobile biometric authentication has become a realistic proposition. Diverse usage environments, including poor lighting, motion/vibration, and ambient noise, pose significant challenges to biometric recognition algorithms. Research has explored algorithms suitable for use on mobile devices [16][21], and for processing face and voice data gathered in noisy mobile environments [2], or with low resolution cameras [29]. Researchers have also investigated fusion of multiple biometrics to compensate for loss of quality in one modality [3][8][34]. For example, Hazen et. al [15] explored the combination of face and voice recognition on an iPAQ device, finding significant improvements in recognition accuracy compared to either biometric alone. Krawczyk and Jain [20] explored signature and voice modalities on a tablet device. All of these studies focused on recognition performance. Combining biometrics also supports 'liveness testing' – the ability to differentiate a live user from a spoof. Efforts in this space [28] have focused both on biometric analysis and custom user challenges.

We are not aware of any existing comparison of user experience in password and biometric authentication on mobile devices, prior to this study. Little is known about the usability of these methods in comparison to each other, and to passwords. Further, little is known about the ease with which users can simultaneously provide two biometric samples, to support efficient multi-factor authentication.

2.3 Working Memory

When accessing information on mobile devices, authentication is an interruption in the user's primary task flow, and a disruption to working memory. The greater the demands on working memory from the authentication process, the greater the risk of forgetting aspects of the task at hand.

Tasks performed on mobile devices, and in particular those performed in the context of a business activity, involve multi-step procedures. In light of the brief nature of the tasks performed on these mobile devices [25], in this study we raise the question of how much of an impact authentication challenges have on users' working memory and thus on reliable task completion. Prior studies indicate that there is an impact, particularly just before task completion (e.g. [33]). Part of the present study is to assess the recall impact due to authentication modality, or combination of modalities, on a memory recall task in the absence of recall cues (e.g., [4]).

Working memory is the mental process by which information is temporarily stored and manipulated in the performance of complex cognitive tasks. The capacity of working memory is limited, and varies between individuals. Models of working memory describing a multi-component system including a phonological loop and visuo-spatial scratch pad were introduced in the early 1970s [5] and have decades of empirical support. The 'phonological loop' stores and rehearses verbal and other auditory information, while the 'visuo-spatial scratch pad' manipulates visual images. Information stored in working memory fades, or 'decays' over time. Subvocal (or even vocal) articulation is a commonly used memory strategy, in which an individual repeatedly subvocally verbalizes and hears an item in order to rehearse it and maintain its activation in working memory. Verbal authentication methods could interfere with this process.

3. USABILITY STUDY

Three different forms of user action for biometric authentication, password entry, and two combinations were exam-

ined in six experimental conditions described below. All voice and gesture conditions used the same authentication phrase, '35793579', providing a memorable consistent value across both modalities, and an audio sample long enough to be acceptable for an automated speaker verification technology. A repeated 4-digit sequence was used to increase memorability while still using a variety of gestures and speech sounds. Password entry was included as a reference point.

This paper uses the terms 'user action' and 'taking action' to refer to the actions taken by the user in providing an authentication sample (biometric or password). As authentication algorithms improve, these user actions will be an important determinant of technology acceptance. This study assumes a zero false rejection rate (FRR), the ideal scenario for a legitimate user.

The six experimental conditions were as follows:

1. Password: Enter an alphanumeric password using the built-in on-screen keyboard. In the spirit of typical corporate password policies, the easy to remember 8-character password *securit3* was used.
2. Voice: The user must speak the password phrase "three five seven nine three five seven nine".
3. Face: The user must take a photograph of their face using the front-facing camera.
4. Gesture: The user must write '35793579' on the screen with their finger.
5. Face+Voice: The user must say "three five seven nine three five seven nine" while simultaneously lining up their face and taking a photograph.
6. Gesture+Voice: The user must say "three five seven nine three five seven nine" while simultaneously writing the digits '35793579' on the screen with their finger.

3.1 Participants

Participants were 30 employees (13 women) of a large technology corporation, unconnected to the project, having 1.5 to 45 years with the company. They were recruited through email lists and personal contacts, and were given a small compensation. Twenty-nine have experience using a smartphone. Six use multiple smartphones. Twenty-one have used a tablet device with the iPad being the most common device and one month to two years of experience. Five used a smartphone and three used a tablet device to access protected company information, where policy required a mobile device screen lock password of at least 8 characters, including both alphabetic and numeric or symbol characters.

All participants had experience with password and PIN as an authentication method. Five occasionally used on-screen signature, four regularly used other types of gesture id and one occasionally did. Six occasionally used face id (3) or voice id (3). Ten occasionally used fingerprint while one regularly did. Some participants' work had at some time involved taking or analyzing facial images for verification (4), recording or analyzing speech samples for voice or speaker verification (7), or collecting or analyzing gestures (3).

3.2 Apparatus and Materials

3.2.1 Hardware

Participants used a Motorola Xoom touch screen tablet with 1GHz Dual Core processor, 1GB RAM, 32GB memory, and 10.1in HD widescreen 1280x800 resolution display. The tablet was running Android version 3.2.1 (Honeycomb).

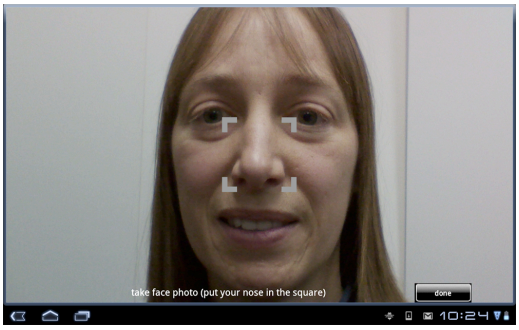

Figure 1: Face Authentication Screen

It measured 249.1mm x 167.8mm x 12.9mm (HxWxD) and weighed 708g. We used the built-in 2MP front-facing camera with automatic focus, located in the top center of the long side of the tablet, making landscape the natural device orientation for taking a photograph. The microphone was centered on the lower long edge.

3.2.2 Client Software

An Android app was developed in HTML, CSS and JavaScript, using PhoneGap v1.0.0rc2 [2] with custom-built audio, camera and gesture capture extensions. The app recorded photographs, gestures, audio recordings, and a time-stamped log of user and system actions.

Each condition presented a different authentication screen. Figure 1 shows the Face authentication screen. The gesture screen presented a plain white writing area with the instruction "write PIN". The Voice authentication screen showed a glowing microphone with the text "Say the PIN", and a counter showing the recording time.

In each condition, three practice trials were given. In Face and Face+Voice conditions, the software also instructed users to lower the device between attempts, so as to practice the full process of positioning the device.

After the practice trials, the software presented a series of memory task trials. This simulates the situation where a user performing a task must authenticate before they can complete the task. The memory task presented a randomly generated three-digit number and a two-character measurement unit randomly selected from 10 options, for example 'The value is 512mg'. Tapping an 'Authenticate' button activated the authentication screen for the current condition. After taking action, participants were asked "What is the value?", and entered their response using the on-screen keyboard. Buttons for 'Done' and 'Forgotten' were available. No feedback on response accuracy was provided.

In all conditions, users could start to take action as soon as the authentication screen was displayed. Specifically, the on-screen keyboard was automatically displayed, voice recording was on, the camera was active, or the gesture capture was active, as appropriate. Users pressed a button to complete their authentication action. Placement of these buttons was influenced by the expected user action. For example, the button on the face authentication screen was placed in the lower right, for convenient thumb activation while holding up the tablet with two hands (see Figure 1).

Each sample resulting from a user action (password or biometric) was immediately checked by the server. This simulates a likely usage scenario where an organization policy is

to control access to its information rather than authenticating the local device.

If the sample quality was not acceptable (as defined below), an error message was displayed, and the user was returned to the authentication screen. After three failed attempts, the software moved on to the next trial.

3.2.3 Acceptance Criteria

No automated verification was performed. Instead, a server on the local network assessed password, voice, face and gesture input quality. Voice input samples were quality checked by a remote server. Acceptance of the sample depended on passing the following simple quality checks:

1. Password: The password (*securit3*) was typed correctly. The error message provided for incorrect passwords was "Authentication failed, please try again"
2. Voice: The user provided a sample containing at least 1.5s of speech content with a speech level > 1000 (32767 indicates full dynamic range) and a signal-to-noise ratio $>= 20$dB. The error message provided for failed voice samples was "voice sample too short, too noisy, or no voice found, please try again"
3. Face: The photograph was accepted when it contained a face, as determined by the VeriLook SDK. This ensured that pictures of the ceiling, fuzzy images, and partially hidden faces would not be accepted. The error message provided for failed face samples was "no face found, please try again"
4. Gesture: A gesture is comprised of one or more strokes, each made up of line segments connecting recorded finger positions on the screen. The gesture was accepted when it contained at least 20 line segments. The error message provided for failed gesture samples was "gesture too short, please try again"
5. Face + Voice: The image and voice sample both met the quality criteria as above.
6. Gesture + Voice: The gesture and voice sample both met the quality criteria as above.

This approach establishes a best case scenario for the user, in which their biometric is always recognized so long as they provide a usable sample (FRR=0). The laboratory environment, tightly-specified task and presence of a researcher combined to ensure that participants performed the authentication correctly, minimizing false acceptances. Samples were manually examined for conformance.

3.2.4 Other Materials

The 10-question System Usability Scale (SUS) assessment tool [9] was used to gather subjective impressions of the usability of each authentication action. The word 'system' in the standard questionnaire was replaced with the word 'method'. After pilot testing, questions 5 and 6 were appended with further explanation shown in italics below:

5. I found the various functions in this method were well integrated (*I could remember the values in the task easily after authenticating*)
6. I thought there was too much inconsistency in this method (*I got different results for the same authentication input*)

Responses to each question are given on a five-point scale ranging from 'Strongly disagree' to 'Strongly agree'. An overall SUS score is a value between 0 and 100, where a higher value indicates a more usable method. An average SUS Score is 68 [27]. Sauro [27] analyzed over 500 studies using the SUS, allowing a raw SUS score to be transformed into a percentile, while Bangor et al [6] proposed an A-F grading scale, allowing for easy interpretation. Raw scores, percentiles and grades are all reported here.

An 11th question, using the same response scale, was added: "This method was tiring to use."

Participants were also asked "What did you like or dislike about this method?" A 10-question demographic questionnaire elicited background information including experience authenticating on mobile devices.

3.2.5 Location

Study sessions were conducted in three different interior rooms with overhead fluorescent office lighting; one small office, one larger office, and one 10-person conference room.

3.3 Procedure

After providing informed consent, participants used six different forms of authentication action, presented in random order, and then filled in the demographic questionnaire.

We chose to use a standing position. This makes interaction more challenging because the user must hold the device while operating it, and enabled participants to explore different lighting positions easily. All were advised that they could lean on a desk or a wall, move freely around the room as they wished, and rest at any time.

For each condition, a researcher showed a printed image of the authentication screen and described the user action to be taken. On-screen instructions were also provided. The instructions for taking a photograph were "Authenticate by taking a well-lit photo of your face. Put your nose in the box and use a neutral expression. Press 'done' when you are ready to take the photo." When Face was combined with Voice, participants were instructed to "Authenticate by saying the PIN AND taking a well-lit photo of your face. You can speak while lining up your face, or speak first and then take the photo. Put your nose in the box and use a neutral expression. Press 'done' when you are finished speaking AND are ready to take the photo." In the Gesture+Voice condition, the instructions were: "Authenticate by saying the PIN AND writing it on the screen with your finger. You can write and speak at the same time, or in any order you choose. Press 'done' when you have finished both writing and speaking".

Participants executed 3 practice trials then went on to a set of 8 memory task trials. They were not told that the system was not performing automated verification of their face/voice or gesture. A researcher observed participant actions, comments, position, and method of holding the tablet device. In voice conditions, participants were corrected by the researcher if they did not say the correct phrase. It was not possible to see their gestures during the sessions.

After completing each condition, participants sat down to fill in the usability questionnaire. This provided an opportunity to rest. The instruction given for the usability evaluation questionnaire was:

"Where these questions ask about "the method" we mean the authentication method you just used, within the context of the scenarios where you are trying to remember a number and unit. This includes the experience of sometimes having to repeat your actions to get a good sample, or correct

Table 1: Biometric performance summary

Condition	Failure to Enroll (FTE) % of participants	Failure to Acquire (FTA) % of attempts	User action time per error-free attempt (median sec)
Password	0.0	4.2	7.46
Voice	3.4	0.5	5.15
Face	6.9	3.1	5.55
Gesture	0.0	0.0	8.10
Face+Voice	10.3	21.3	7.63
Gesture+Voice	3.4	13.6	9.91

an error. For example, 'learning to use the method' means learning how to use it accurately, to avoid the need to repeat."

3.4 Data Available

Two participants ran out of time and attempted only 5 of the 6 conditions. A further 16 trials are missing due to technical problems. Three participants did not complete all conditions because they were unable to provide either face or voice samples that passed the acceptance test (see below for further details). Finally, one participant abandoned the Gesture+Voice condition after 2 scenarios due to frustration with that method.

Data from one participant, whose comments indicated that he was testing the authentication mechanisms rather than performing the requested tasks, were discarded.

Authentication attempts were coded as follows:

1. Success: The participant performed authentication correctly and was successful. (1229 samples)
2. Minor error: The participant performed well enough to succeed but may have included additional speech or corrected errors. (43 samples)
3. Error: The user attempted to provide the correct authentication but failed, for example a password with errors, a fuzzy picture, or a speech sample that did not meet the quality check. (100 samples)
4. Noncompliance: The user did not perform authentication correctly, for example speaking the value to be memorized ('529mg') instead of the PIN, saying nothing, or writing a squiggle. (35 samples)
5. Technical error: The sample was unusable due to technical problems. (14 samples, all empty or clipped speech files)

Technical errors and noncompliant attempts were excluded from the analysis.

4. RESULTS

4.1 Failure to Enroll (FTE)

The 'Failure to Enroll' metric (FTE) used in biometric usability research [10] is intended to identify the proportion of individuals who would never be able to use a biometric system. Table 1 summarizes the failure to enroll (FTE) rates for each condition.

Two of the 29 participants found that the Face condition did not work for them – they were not able to take a picture in which the face verification engine could locate their face. These participants contributed no data for the Face or Face+Voice conditions. One of these participants always wears dark, light blocking glasses.

One participant was not successful with the Voice condition – their voice samples did not meet the threshold for signal-to-noise ratio. They contributed no data for the Voice, Face+Voice and Gesture+Voice conditions.

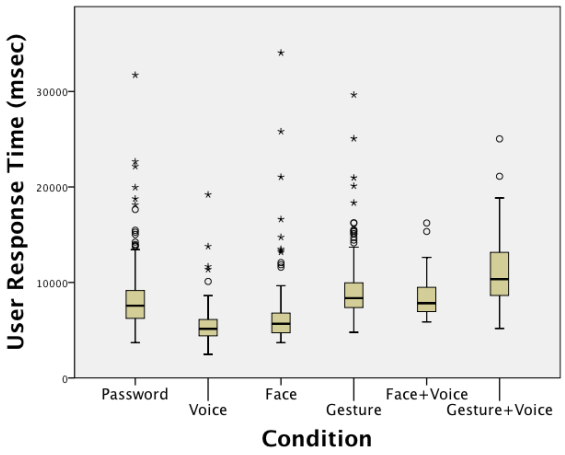

Figure 2: User response time by authentication condition

4.2 Failure to Acquire (FTA)

The 'Failure to Acquire' (FTA) metric [10] is used in biometric usability research to measure failure to provide a sample of sufficient quality. In this study it captures failures where a participant provides a sample that does not meet the predefined quality criteria. For biometric samples, such samples do not contain good enough data on which verification algorithms can operate.

1372 user actions were analyzed, of which 92.7% were successful. Table 1 summarizes the percentage of these attempts that were unsuccessful, in each condition. Face+Voice had the highest FTA rate, at 21.3%. A one-way ANOVA indicated a significant effect of condition on success ($F(5,1366) = 27.249$, $p<0.001$), with post-hoc pairwise comparisons using Bonferroni corrections indicating that FTA values for Face+Voice and Gesture+Voice are significantly different from each other ($p=0.013$) and all other conditions ($p<= 0.001$). The differences between the remaining conditions are not statistically significant.

One participant abandoned the Gesture+Voice condition after 2 scenarios, in which he succeeded only once out of 6 attempts, despite having success in the practice. If he had completed all 8 scenarios with the same low success rate, the overall FTA rate for Gesture+Voice would have been 18.7%.

4.3 User Action Time

User action time is time spent by the user taking action to provide the sample for authentication. It does not include processing time spent verifying the sample quality, performing authentication, or server response delays.

This measure was calculated for the 1229 successful trials (coded as 'Success'), with 184-221 samples per condition. Figure 2 illustrates the distribution of user response times in each condition. Voice authentication was both fast and consistent, with few outlier values. As shown in Table 1, the voice sample was fastest with a median of 5.15 seconds

Table 2: Memory task performance summary

Condition	Memory task preparation time (median sec)	Memory task (% success)
Password	4.3	73
Voice	5.4	76
Face	3.9	85
Gesture	4.2	72
Face+Voice	5.3	71
Gesture+Voice	5.7	65

and taking a photo took 5.55 seconds. The other conditions all took 7.46 seconds or more, with Gesture+Voice being the slowest at 9.9 seconds. The data are not normally distributed, so the Friedman test was used as a non-parametric alternative to a one-way ANOVA with repeated measures. There was a statistically significant difference in user action time depending on the authentication method ($\chi^2(5) = 430.339$, $P<0.001$). Post-hoc analysis with Wilcoxon Signed Rank tests was conducted. Applying Bonferroni correction, the significance level was set at $P<0.003$. All pairwise comparisons were statistically significant ($P<0.001$) with the exception of Password and Face+Voice (Z=-1.128, P=0.259).

4.4 Memory Task

The memory task required participants to enter a three-digit value and two-digit measurement unit they had been shown prior to the authentication action, using the on-screen keyboard. Trials containing technical errors or noncompliant attempts are excluded (N=21), leaving 1277 trials for analysis.

Table 2 shows the median memory task preparation time, defined as the time participants spent viewing the screen that showed the value before proceeding to the authentication screen. This gives an indication of time spent actively memorizing the value. Face had the least time with a median of 3.9s. Using the Friedman test as a non-parametric alternative to a one-way ANOVA with repeated measures, there was a statistically significant difference in preparation time depending on the authentication method ($\chi^2(5) = 81.334$, $P<0.001$). Post-hoc analysis with Wilcoxon Signed Rank tests was conducted with Bonferroni correction applied, resulting in a significance level set at $P<0.003$. There was a statistically significant difference between Face and all other conditions except Gesture (Password: Z=-3.121, P=0.002, Voice: Z=-4.297, $P<0.001$, Gesture: Z=-1.602, P=0.109, Face+Voice: Z=-3.340, P=0.001, Gesture+Voice: Z=-7.447, $P<0.001$). There was also a statistically significant difference in preparation time between Voice and Gesture (Z=-4.064, $P<0.001$), with participants spending approximately one second longer in the Voice condition. All other pairwise comparisons were not statistically significant.

In the 1277 memory task trials, the participants entered the correct response 74% of the time. The success rate for the 1204 trials where user action was successful at the first attempt was 75%, while the success rate for the remaining 64 trials was 56%. These memory task failures include typing errors as well as cases where the user pressed the 'Forgot' button, or omitted all or part of the response. Table 2 shows the percentage of correctly completed memory tasks for each condition (Memory task % success). There was an overall statistically significant difference in success depending on the authentication method ($\chi^2(5) = 28.261$, $P<0.001$). The combined Face+Voice condition was associated with significantly poorer performance than Face or Voice alone (Wilcoxon Signed-Ranks test with Bonferroni correction, significance level $P<0.003$, Voice: Z=-3.094, P=0.002, Face: Z=-5.000, $P<0.001$), and the combined Gesture+Voice condition was poorer than Face (Z=-3.299, P=0.001). Other pairwise comparisons were not statistically significant.

Table 3: System Usability Scale summary

Condition	SUS score	SUS response percentile (approx.)	SUS grade	Fatigue
Password	78%	80^{th}	C	2.5
Voice	66%	40^{th}	D	3.0
Face	75%	76^{th}	C	2.2
Gesture	77%	78^{th}	C	2.4
Face+Voice	46%	8^{th}	F	3.7
Gesture+Voice	50%	13^{th}	F	3.8

4.5 Usability Responses

Table 3 summarizes the overall score, percentile and grade for the System Usability Scale (SUS) for each condition, and level of agreement with the question "This method was tiring to use". These interpretations illustrate that none of the user actions were well liked in the context of the memory task, with grades ranging from C to F. Password, Face and Gesture were rated above the average SUS response value, while the combination conditions lagged behind, with ratings in the 10th percentile of typical responses. The combination conditions were also considered the most tiring to use, while Password, Face and Gesture were not tiring.

In Table 3, ratings from the three participants who experienced failure to enroll (FTE) are included. Excluding all ratings from these participants increases the scores for Face, Voice and Face+Voice by 1-2 points and does not impact the other scores, leading to the same overall assessment.

Participant responses also take into account the processing time used to communicate the sample to the server, assess the quality, and provide a response. Variable, and sometimes long, network delays were observed, and likely influenced these usability results. Median server response times were: Password=0.06s; Voice=2.04s; Face=1.49s; Gesture=0.13s; Face+Voice=4.28s; and Gesture+Voice=3.82s.

4.6 Participant Comments

Participants provided comments both while using the tablet, and in written form after each condition in response to the question "What did you like or dislike about this method?" Conditions were ordered randomly, so participants' first impressions of a biometric may have been in a single or combination condition.

4.6.1 Password

Participants liked the familiarity of password entry, commenting that there was "no need to learn new tricks", it was "comfortable, easy and familiar", "seemed to be the fastest method and easiest to remember the measurements" and "familiar = easy = like!"

However, they did not like that "the input requires many

steps (including switching back and forth between alphabet and number input)". One person commented that "1. Having to switch keyboards affected my memory terribly, 2. As well as having to have a number in it" (the password). Another observed "Keyboards that do not display letters AND numbers simultaneously can be irritating in this scenario." One person found that "Standing and keying in letters/digits is a bit of a challenge, balancing the pad on one hand."

4.6.2 Voice

Only three participants made positive comments, that speaker verification using a spoken number was "natural", "faster than other modes that required an additional biometric", or "easier to use than typing".

Most comments were negative. Nine participants commented that they experienced "Interference between the content of the authentication method and what I needed to remember" or it was "impossibly difficult to remember things after speaking".

Participants also expressed concern about the security aspects of this approach. Five participants commented that speaking a phrase out loud "doesn't feel secure". Participants felt that voice would not be a practical method in real contexts, saying "In real life there would be noise, and interference leading to huge frustration". One participant commented on the volume level required for speech "I learned from the last speech based system to speak more loudly. That helped. I still didn't like it."

The Voice recording user interface also received some criticism, that the timer indicator was "distracting and led to some confusion over how fast I should say the passphrase" and it was "confusing with recording on and off message - not sure if I tapped properly to start voice authentication".

4.6.3 Face

Eleven participants made positive comments that "it was easier to remember the numbers", or "I was able to mentally 'repeat' the value, even as I was taking a picture."

Four found it "easy" or "simple" to take the picture, but nine others complained that positioning the camera was "somewhat annoying", "a bit hard because of the reflection of myself I was getting" or "cumbersome to position the face". Participants commented on the lack of feedback when their face was positioned properly: "I didn't know when it worked well", or "not sure how accurately I need to place my nose in the box on the screen."

Participants took action to get better pictures: "I had to find a solid background and then it worked", or "I found a better lit spot in the room". Several participants felt uncomfortable taking a picture of themselves: "I have to suspend the fact that I might not like the picture", "felt too much like I was taking a vanity photo."

4.6.4 Gesture

Some participants found the gesture condition "fun", "fast", "easy to use", "fairly automatic", and "an intuitive way of entering passwords". One participant observed that "I could easily see what the system was getting from me (vs. audio where I don't hear the recording)".

However, in the context of the memory task, it was "mechanically easy to use but cognitively difficult", and "still easy to forget the value". Eight participants commented that it was difficult to remember the memory task value while writing the phrase, but four considered it easier than other conditions, for example "the writing of numbers is like a pattern which makes remembering the other number easier", and "I could use muttering to remember the codes". One suggested a shorter password, while another observed that it would have been easier if the phrase was a word.

4.6.5 Face+Voice

Only two positive comments were made about the Face+ Voice condition, that it had "simplicity" and provided a "double degree of security".

Seven participants commented on difficulty with the memory recall task, for example "I had to invent memory aids to remember the number and units to key after authenticating."

Eleven participants commented on the physical difficulty of the required actions. For example it was "cumbersome", "requires too much coordination", was "very annoying trying to get the camera at the right angle to get a photo", and "felt like a lot of work". Other comments included "Positioning nose in square on screen is not easy; once nose is in position scanning the screen for 'done' button resulted in moving my face", "I disliked having to center my nose in the target area – I seem to move the tablet about quite a bit without thinking about it and had to make an effort", "My arms get tired holding the tablet up and aligning it for a face shot", and "tilting the screen (both horizontally and vertically) seemed counter-intuitive – my first inclination to tilt it up or left was consistently wrong (moving my nose further away). Over time I overcame this with practice."

A further five felt that the method was not working correctly. Saying the voice performance was "erratic" or "didn't work well", or "too slow", and "Had a few failures when I moved around possibly because of lighting".

As with the Face condition, participants also mentioned a dislike of looking at their own images: "didn't like seeing myself at such close-up!" and "it makes me self-conscious".

4.6.6 Gesture+Voice

No positive comments were made about the combined Gesture+Voice condition. Eight participants commented on difficulty with the memory recall task. Seven participants commented that the performance "seemed slow", "the numbers I wrote appeared distorted", and it "did not seem to track the movement of my finger with good resolution".

Participants chose to speak as they wrote, but three commented on the awkwardness of slowing down their natural speech rate to match their writing speed: "Unlike the first experience w/ writing (alone) this seemed too slow – I guess because the voice channel is so much faster than the gesture feedback", "I can speak much faster than I can write so having to do both was off putting (because I was very aware of the 'slowness') whereas when I was just writing it 'felt' just right."

Some participants considered this condition "horrible", with "WAY too much distraction".

4.7 Researcher Observations

As participants performed the study, they often moved around the room. Some participants paced as they worked, while most stood or leaned against a wall or desk. Those who paced, stopped pacing to take a photograph, but continued pacing while entering a password, writing or speaking.

The tablet was normally held at chest or belly height. Participants were observed to switch positions as they became tired.

The method of holding the tablet was also strongly influenced by the experimental condition. When using the camera, 23 participants held it with two hands, one at each side, and held it up in front of their face, lowering it again afterwards. When tapping in a password, participants often held the device with one hand spread underneath, whereas the most common position for gesture was to hold the device with the left hand at the left side. When speaking, participants did not move the tablet, and 22 held it in their left hand.

While practicing with the camera, participants moved around the room and experimented with different tablet angles and positions, then used a single location and position throughout the remainder of the study. Taking a face picture was made more difficult by the distraction of seeing their reflection in the shiny screen, under the strong overhead lighting.

Even when the voice was clearly audible to a person in the room, the signal-to-noise ratio was sometimes low. Some participants needed to speak more loudly than was comfortable in order to reduce error rates. Those who experienced problems with the voice condition reacted first by speaking more loudly or slowly. Only two looked for or asked about the microphone location, and two moved the tablet closer to their mouth.

When voice was combined with face or gesture, participants appeared to speak with lower volume and have a tendency for their voice to trail off. This reduced the signal-to-noise ratio, causing voice quality failures.

The participants were highly motivated to perform well on the memory task, and employed techniques to help them remember the value and unit, including speaking the value aloud, or thinking of a mnemonic to help them remember. These techniques were used more often in conditions involving speech.

5. DISCUSSION

These data provide an understanding of the relative user effort required by the different authentication mechanisms under quiet, well-lit, stable conditions and may be representative of environments such as an office or home location. Work is ongoing on robust authentication algorithms that are effective in a broad range of environments that are noisy, low lighting, or involve movement (e.g., walking, public and private transportation), etc. and multi-factor biometric authentication. Privacy considerations may be addressed by cancellable biometrics [26].

The interfaces for biometric and password acquisition used here were simple. With the exception of a screen orientation to facilitate self portrait photos (landscape), we did not attempt to compensate for any perceived shortcomings of the device (e.g., reflections on the display surface, alternative keyboard layouts to minimize changing between alphabetic and numeric/symbol layouts). Our participants were novice users, and performance improvements with practice could be expected. Further field studies in natural environments with more experienced users are needed to provide a more complete understanding, including learning effects.

5.1 Time to provide an authentication sample

Clearly the Face and Voice conditions were faster than the Password and Gesture conditions. The Gesture entry was significantly slower than any of the other conditions, although that may be related to the substantial software lag time in responding to drawing on the touch screen. On average, the Face and Voice conditions had a 2.0-2.5 sec. lower user action time than the 7.5 sec. in the password condition. Participants were able to provide dual biometrics in less time than sequential entry of the same two biometrics, but with higher acquisition error rates. The error-free Face+Voice condition time was comparable to error-free password typing. Where there is a failure to provide an acceptable biometric sample, the overall time would quickly rise, underscoring the importance of an authentication interface that minimizes user error through appropriate feedback to the user, and recognition algorithms that can operate on real-world samples with minimal error. For the Face conditions, once participants found a place with good lighting, they tended to stay in that position. In outdoor or highly populated environments such as public transport, additional actions, and time, would be required to find a suitable location, and biometrics will sometimes not be appropriate.

5.2 Ability to provide a quality sample

With minimal instruction and very little practice, 90% of participants were able to use all of the biometric methods well enough to provide a sample that met the quality criteria. However, there were three participants who could not use one of the biometric modalities. In two cases, the reasons for these failures are not clear, and will be explored in further work. This failure rate underscores the importance of having multiple modalities for authenticating, with a reliable fallback method to support critical access scenarios.

The dual conditions had error rates much higher than the sum of the individual error rates. High error rates negate the benefit of dual conditions by increasing the overall time to acquire beyond the time that would be required for single biometrics in sequence. There are multiple possible explanations for the higher error rates. Given the low error rate in the Gesture condition, but high lag time for displaying the gesture, the high error rates for Gesture+Voice may be due to fading off in the voice sample. Poor performance on the Voice+Face condition may be due to the cognitive demand of a task involving two disparate modalities. Practice may reduce these dual condition error rates, but this remains to be empirically tested.

In future work, we will examine the quality and consistency of biometric samples provided by the participants, and the performance of verification algorithms on this data set.

5.3 Impact on the memory recall task

In contrast to prior work that examined password typing time on a mobile device [7], this study presented authentication within a task that demanded short term memory recall. Authentication 'failure' due to a poor quality sample, led to a steep drop in task success, from 74% to 47%, confirming the challenge of the task and the disruptive nature of authentication. Perhaps because of this cost of failure, participants actively employed memory recall strategies to boost their task performance.

Face authentication, the only condition that involved no password or PIN, supported the highest memory task perfor-

mance. Using the same authentication prompt in all other conditions, no significant difference was found between voice and gesture modalities. Combination modalities produced significantly poorer performance.

Participants spent significantly longer on the trial screen that presented the memory task in the Voice condition, compared to Gesture or Face. This may be indicative of additional effort invested in memorization of the values when in conditions that involve speech. These results underscore the importance of carefully choosing authentication points that least interfere with user task flow.

Further work should examine the impact of using different kinds of spoken/gestural material such as spoken phrases, or abstract gestures, and user-selected vs. system-selected items. This would separate users' reactions to the method of authentication from the content of the authentication prompt. Although system generated prompts may increase the cognitive load on the user.

One possibility would be to allow users to combine prompted speech with any other speech of their choosing. Participants could, for example, have chosen to say something like "526mg 35793579 526mg", ensuring liveness while allowing them to verbalize any information in working memory. This may actually help with their task, rather than hinder it. In contexts where the task is known, prompts should be designed so as not to interfere with the task content.

5.4 User reactions

User responses to the SUS were low, with grades ranging from C to F. As one participant put it "Authentication is never fun". Interestingly, the Voice condition was faster, less error prone, did not suffer very long server delays, and supported relatively high task success, yet received only a 'D' grade for usability from participants. Although participants perceived it as interfering with their ability to perform the memory task, this was not reflected in their results. Authentication prompts that are very different in nature to the task context may reduce such interference to some extent, and should be explored in future studies.

From observations of users during the study, many were not comfortable with the speech volume required for sample acceptance. Sample quality and naturalness of speech need to be carefully balanced.

User reactions to Face authentication were mixed, with some commenting that the process of taking a photograph was cumbersome, while others found it easy. Further work into appropriate user feedback to make it easier to take a good quality photograph with a tablet device in varied locations is needed.

Dual biometric conditions were considered fatiguing and less usable by participants. However, these conditions also involved variable, and sometimes long, server delays. Server response time should be more tightly controlled in future work, to allow for separation of the impact of user action times, modalities and prompts.

6. CONCLUSIONS

We report a laboratory study of the usability of three biometric authentication modalities on a tablet device within the context of a memory task, independent of the performance of biometric verification algorithms. Speaker, face and gesture verification, as well as password entry, were compared using 8-digit written and spoken PIN codes, under six single and dual-biometric conditions. The study identifies usability issues and biometric performance requirements that can serve as a focus for research.

Each biometric modality has unique strengths and weaknesses, and has the potential to improve on the Password approach. Face and Voice are fast but not universally usable. Gesture is reliably performed and worked for everyone, but a much shorter gesture would be needed to achieve a competitive time, posing a challenge to gesture recognition algorithms. The memory task context provides further insight into the broader impact of authentication, and demonstrates a significant advantage for Face, and a lesser advantage for Voice in supporting memory task performance.

However, the Voice condition was considered less usable than Password, Face and Gesture. Speaking at a comfortable level did not always meet the voice sample quality threshold, indicating a requirement to operate with a lower threshold. Participants also reported interference with the memory task that was not reflected in their performance. They maintained high performance by using sophisticated memorization strategies, as indicated by their comments and differences in authentication preparation time.

Using face recognition also posed challenges for participants, even in good conditions. Careful user interface design is needed to overcome issues with screen reflection and provide feedback for proper alignment.

The conditions that combined two biometric authentication modalities were disliked by the participants, had higher Failure To Acquire and lower performance on the memory recall task. This suggests that combined sample collection for biometric fusion is not necessarily preferable to collecting individual samples.

Providing a face or voice biometric to a mobile device seems to be a natural extension of normal device usage requiring no special setup or extra hardware. Software developments such as built-in face recognition are opening further opportunities to streamline the user experience of mobile authentication. This study demonstrates a complex set of trade-offs in selecting and using biometric authentication methods on mobile devices, even in quiet, well-lit conditions. Studies like this one can help to identify critical research challenges for biometric verification algorithms, in addition to design challenges for mobile authentication user interfaces. The goal is to improve on the notoriously cumbersome password method, leading to mobile biometric authentication that is both secure and usable.

7. ACKNOWLEDGEMENTS

We thank the study participants, and Bonnie E. John, Rachel L. K. Bellamy, John C. Thomas, Nalini Ratha, David Nahamoo, Ron Hoory, Hagai Aronowitz, and Amir Geva for valuable feedback, and technical contributions.

8. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy: Why users compromise computer security mechanisms and how to take remedial measures. *Communications of the ACM*, 42(12):40–46, Dec. 1999.

[2] Adobe Systems Inc. PhoneGap. `http://phonegap.com`.

[3] G. Aggarwal, N. K. Ratha, R. M. Bolle, and R. Chellappa. Multi-biometric cohort analysis for

biometric fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, 2008.

[4] E. Altmann and G. Trafton. Task interruption: Disruptive effects and the role of cues. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, IL, 2004.

[5] A. Baddeley and G. Hitch. Working memory. In G. Bower, editor, *Recent Advances in Learning and Motivation*. Academic Press, 1974.

[6] A. Bangor, P. T. Kortum, and J. T. Miller. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 2008.

[7] P. Bao, J. Pierce, S. Whittaker, and S. Zhai. Smart phone use by non-mobile business users. In *MobileHCI*, Stockholm, Sweden, 2011.

[8] J. Basak, K. Kate, V. Tyagi, and N. Ratha. QPLC : A novel multimodal biometric score fusion method. *CVPR Workshop on Biometrics*, 2010.

[9] J. Brooke. *SUS: A quick and dirty usability scale*, pages 189–194. Taylor and Francis, 1996.

[10] L. Coventry. Usable biometrics. In L. F. Cranor and S. Garfinkel, editors, *Security and Usability: Designing Secure Systems that People can Use*. O'Reilly Books, 2005.

[11] P. Dunphy, A. P. Heiner, and N. Asokan. A closer look at recognition-based graphical passwords on mobile devices. In *SOUPS*, Redmond, WA, 2010.

[12] D. Florencio and C. Herley. A large-scale study of web password habits. In *WWW*, Banff, Canada, 2007.

[13] D. Florêncio and C. Herley. Where do security policies come from? In *SOUPS*, Redmond, WA, 2010.

[14] N. Gunson, D. Marshall, F. McInnes, and M. Jack. Usability evaluation of voiceprint authentication in automated telephone banking: Sentences versus digits. *Interacting with Computers*, 23(1):57–69, Jan. 2011.

[15] T. J. Hazen, E. Weinstein, B. Heisele, A. Park, and J. Ming. Multimodal face and speaker identification for mobile devices. In R. I. Hammoud, B. R. Abidi, and M. A. Abidi, editors, *Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems*. Springer, 2007.

[16] Y. Ijiri, M. Sakuragi, and S. Lao. Security management for mobile devices by face recognition. In *Proceedings of the 7th International Conference on Mobile Data Management (MDM)*, Nara, Japan, 2006.

[17] N. Jackson. Infographic: How Mobile Phones Are Replacing Our Credit Cards, 2011. `http://www.theatlantic.com/technology/archive/2011/07/infographic-how-mobile-phones-are-replacing-our-credit-cards/241703/`.

[18] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *HotSec*, Montreal, Canada, 2009.

[19] L. A. Jones, A. I. Antón, and J. B. Earp. Towards understanding user perceptions of authentication technologies. In *Proceedings of the ACM Workshop on Privacy in Electronic Society*, Alexandria, VA, 2007.

[20] S. Krawczyk and A. K. Jain. Securing electronic medical records using biometric authentication. In *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Hilton Rye Town, NY, 2005.

[21] S. Kurkovsky, T. Carpenter, and C. MacDonald. Experiments with simple iris recognition for mobile phones. In *Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations (ITNG)*, Las Vegas, NV, 2010.

[22] M. Lee. Google Turns to Face Detection With Samsung to Take On Apple Speech Parser, 2011. `http://www.bloomberg.com/news/2011-10-19/google-turns-to-face-detection-to-take-on-apple-iphone-s-speech-technology.html`.

[23] M. Lennon. One in Three Experience Mobile Device Loss or Theft. Do People in 'Party Cities' Lose More Phones?, 2011. `http://www.securityweek.com/one-three-experience-mobile-device-loss-or-theft-do-people-party-cities-lose-more-phones`.

[24] S. F. Nagata. Multitasking and interruptions during mobile web tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Denver, CO, 2003.

[25] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci. In *CHI*, Portland, OR, 2005.

[26] N. K. Ratha, S. Chikkerur, J. H. Connell, and R. M. Bolle. Generating cancelable fingerprint templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):561–572, Apr. 2007.

[27] J. Sauro. Measuring usability with the System Usability Scale (SUS), 2011. `http://www.measuringusability.com/sus.php`.

[28] S. A. Schuckers, R. Derakhshani, S. Parthasardhi, and L. A. Hornak. Liveness detection in biometric devices. In *Electrical Engineering Handbook, 3rd edition*. CRC Press, 2006.

[29] Q. Tao and R. N. J. Veldhuis. Biometric authentication for a mobile personal device. In *Proceedings of the 3rd Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services*, San Jose, CA, July 2006.

[30] R. Tassabehji and M. A. Kamala. Improving e-banking security with biometrics: modelling user attitudes and acceptance. In *Proceedings of the 3rd International Conference on New Technologies, Mobility and Security (NTMS)*, Cairo, Egypt, 2009.

[31] B. Tognazzini. Design for usability. In L. F. Cranor and S. Garfinkel, editors, *Security and Usability: Designing Secure Systems that People can Use*. O'Reilly Books, 2005.

[32] D. T. Toledano, R. Fernández Pozo, A. Hernández Trapote, and L. Hernández Gómez. Usability evaluation of multi-modal biometric verification systems. *Interacting with Computers*, 18(5):1101–1122, Sept. 2006.

[33] J. G. Trafton and C. M. Monk. Task interruptions. In D. A. Boehm-Davis, editor, *Reviews of Human Factors and Ergonomics*. 2008.

[34] V. Tyagi and N. Ratha. Biometrics score fusion through discriminative training. *CVPR Workshop on Biometrics*, 2011.