My Voiceprint Is My Authenticator: A Two-layer Authentication Approach Using Voiceprint for Voice Assistants

Yun-Tai Chang
Computing and Software Systems
University of Washington
Bothell, Washington, USA
Yuntai.CYT@gmail.com

Marc Dupuis
Computing and Software Systems
University of Washington
Bothell, Washington, USA
marcjd@uw.edu

***Abstract*—Voice assistants are a ubiquitous service of contemporary daily life. Their intuitive use and 24-hour-a-day convenience make them popular with an increasing user base. However, the security of voice assistants not not increased commensurate with the rising number of users and their increasing technical abilities. The lack of an authentication mechanism gives attackers an opportunity to exploit voice assistants to control and get personal information from linked services. The goal of this research is to provide an authentication method that protects voice assistants from attacks without degrading their usability. We utilize Microsoft Azure Speaker Recognition API and Google Speech API to implement an Android application to examine the approach. The results indicate that the voice authentication method can resist replay attacks and be easily learned and used.**

***Index Terms*—voice assistants, two-layer authentication, privacy, security, usability, attacks, verification**

I. Introduction

In recent years, voice assistants have become a popular and ubiquitous technology. They are pre-installed in smartphones and have even become discrete devices, such as Amazon Echo and Google Home. In addition, voice assistants provide services through integrating with applications or smart devices. Although voice assistants make daily life more convenient, they are also vulnerable to voice attacks and may pose a unique threat to the personal information of end users [6], [7].

Demonstrations of bad and inaudible commands have indicated the ease with which voice assistants can be exploited and the importance of securing voice assistants by authentication [32], [33]. Various methods proposed for authenticating the user include asking the user to wear an additional token to prove her identity [9], or using voiceprint verification [27].

However, these methods typically cannot simultaneously fulfill the dual requirements of security and usability; they are either not robust enough or they are not easy to use. To reduce the gap between security and usability, this study proposes an authentication method that uses voiceprint to maintain usability and adds an additional challenge-response layer to enhance the security of voice assistants.

A. Voice assistants and security issues

Voice assistants fall into two main types: 1) an application in a smartphone, and 2) a discrete device placed in a house. Both types provide numerous services, from making a phone call, reading an email, or scheduling an event, to controlling various Internet of Things (IoT) devices [8], such as smart locks [2], [18].

For users, it is easy to understand how to use voice assistants. A user only needs to use a specific key phrase, e.g., “Ok Google” or “Alexa”, to trigger the voice assistant and then tell it what she wants the voice assistant to do, using natural language. Despite these advantages, they are insecure. Because voice assistants use voice as input, they are always listening, making them vulnerable since the voice is on a public channel. Attackers can exploit a voice assistant by making voices that can be received and interpreted by the voice assistant [28], [33]. And voice assistant connections to applications and other devices could harm the user’s financial accounts and privacy.

B. Limitations of current authentication methods for voice assistants

Some research methods, such as wearing a token or placing a motion sensor in the house to ensure the command comes from an authenticated user, have provided robust security but decreased usability [9], [19]. Users’ unwillingness to wear an additional gadget or to install sensors in the house hinders widespread adoption of these methods. Other methods, such as using a voiceprint to identify speakers, are more user-friendly but could be exploited by playing a recording of the victim’s voice (replay attacks) [1], [27]. It is easy to implement a method that provides robust protection but with imperfect usability, and vice versa. If the method is too hard to use, the user will abandon it; if a method is easy to use but cannot provide protection, the user will also not use it.

C. Motivation and research focus

Unlike the past, when users only used voice assistants to turn lights on or off, voice assistants now can manipulate things with more serious implications, such as bank accounts,

email, and online payment applications. This necessitates stronger authentication techniques to ensure that users do not face a financial loss or a violation of their privacy. The lack of usable and secure authentication motivates this study to create a mechanism that can provide robust security and maintain usability. Considering the intuitive interaction of voice, this study uses voiceprint as the authenticator to maintain usability while providing secure authentication.

D. Goals and criteria

The goals of this research are: 1) improving the security of voice assistants, and 2) maintaining the usability of the system. To achieve these goals, the method must satisfy the following:

- Criteria 1: Achieve false acceptance rate and false rejection rate of less than 1%.
- Criteria 2: Maintain ease of use with a two-layer authentication method.

E. Approach overview and contributions

Among voiceprint vulnerabilities, replay attacks are the most threatening and prolific [17], [30]. To mitigate replay attacks, this research proposes a two-layer authentication method using voiceprint wherein users' own voices authenticate their identities on voice assistants. With the proposed mechanism and implementation, the contributions of this research are:

1) Integrating automatic speaker verification systems into voice assistants;
2) Maintaining the usability of the proposed method, and
3) Improving the security of voice assistants.

II. Related works

A. Voice assistants

1) Abilities: Voice assistants are a type of human-computer interface with many capabilities, such as launching apps, setting schedules, making calls, sending messages or emails, and playing music [18].

2) Abuses and attacks: As voice assistants become more efficient and useful, the abuses of voice assistants can cause more serious loss to users. Audio input is an easy and common method since voice assistants are always listening.

3) Protections: Technology companies have been aware of the security issues of voice assistants and some companies have enhanced the ability of voice assistants to recognize individual speakers. In 2017, Google published Voice Match [1] to let users link their Google accounts and voices.

B. Authentication

1) Security: The purpose of authentication is to confirm the truth of the identity claimed by a person. There are three factors to authenticate a person: 1) what you know, 2) what you have, and 3) what you are. Passwords are an example of the first type; tokens, such as an access card, are a physical object that we have; and biometrics, such as fingerprint, are a means to present who we are.

Due to the limitations of the three authentication factors, researchers have invented multi-factor authentication to provide better protection. Two-factor authentication is currently the most common method, combining any two of the three authentication factors together. Among the combinations, many studies have combined biometrics and tokens [10], [12], [13].

Researchers also implemented two-step authentication methods that used two authenticators within the same factor [11], [29], such as using one's face and fingerprint as a pair to authenticate an individual [11]. These dual-biometric systems reduce the weakness of biometrics in security [4], [15].

2) Usability: Authentication methods require users to interact with them, thus, usability is a key factor. But the conflicts between usability and security are notorious. For instance, users tend to create passwords that violate secure password rules because they are easy to remember. Tokens could ease the pain of the memorizability of passwords [26], but such extra devices are inconvenient and users can lose them. Biometrics have the best usability among the three authentication factors because do not require memorizability and cannot be taken by others [3], [21]. In our study, we focus on voice authentication since voice is the only means which allows users to communicate with voice assistants.

C. Voice speaker verification systems

In the present study, voice is considered the best way to authenticate users since voice assistants use voice as the input. For all voice assistants, voice authentication needs no extra devices and can thus be smoothly integrated. Furthermore, speaker recognition systems are an active research field. Research has focused on increasing the recognition rate, which has improved to the point that the recognition quality is sufficient for banks and companies use it as an identification method. However, speaker recognition systems are still vulnerable to attacks.

1) Basic knowledge: Voice recognition is a part of speech processing. Speaker-specific recognition requires speaker identification and speaker verification. Speaker identification compares the speaker with all or a group of voiceprints in the database (one to many) to verify that the voice belongs to a particular user. In contrast, speaker verification compares the speaker with a specific user's voiceprint (one to one) to check if the voice is from her.

Speech types can be classified as text-independent or text-dependent in speaker verification.Text-dependent speaker verification requires the speaker to say the same sentence that she used in registration, requiring the user to remember the text. However, text-dependent systems have higher positive acceptance rates due to the dependent text. For text-independent systems, the user can say anything and the system will recognize her, but with a lower positive acceptance rate.

Before using a speaker-recognition system, users must enroll in it. Speaker enrollment includes two phases. One is an offline phase which generates a background model from the analog background voice. The other is an online phase which creates the target user model.

For the process of identification and verification, the unknown speaker's voice is transformed from analog to digital data by feature extraction and the digital data is compared with other models in pattern matching. For identification, the pattern matching will use models in the database for comparison. For verification, the pattern matching will use the hypothesized model to compare with the digital data. The comparison(s) outputs score(s), which are normalized and sent to the decision logic to decide if the score passes the criteria.

2) Technologies: The current technology to verify a speaker is the likelihood ratio test, which can compare two statistical models [23]. The two hypotheses for the likelihood ratio test of speaker verification are [HII-C2] Input speech Y is from the hypothesized speaker S and [HII-C2] input speech Y is from other speakers. Further, the likelihood ratio test to decide between these two hypotheses is:

$$\frac{p(Y|H0)}{p(Y|H1)} \begin{cases} \geq \theta, & \text{accept H0} \\ < \theta, & \text{accept H1} \end{cases} \tag{1}$$

where p(Y|Hi), i=0,1, is the probability density function for the hypothesis Hi. For the equation 1, Y can be replaced by the voice model of input speech, H0 can be the voice model of the hypothesized speaker, and H1 can be the voice model of the other speakers, known as the universal background model (UBM). Since all these voice models uses Gaussian mixture model (GMM) to present the features that are extracted from the input speech, and the verification system needs a universal background model (UBM) to compute the likelihood, the technology is called GMM-UBM.

3) Attacks: Ratha et al. [22] identified vulnerabilities of biometrics-based authentication systems. Based on Ratha's research, Wu et al. [31] discussed spoofing attacks of speaker verification systems. Spoofing attacks, also called direct attacks, utilized different techniques to make the input voice have a similar voice model as the target user's. Indirect attacks required hacking the speaker-verification system itself and thus were more difficult to launch.

The study presented a comprehensive examination of spoofing attacks and concluded that researchers should pay more attention to them. Spoofing attacks can be classified into four types: 1) impersonation; 2) replay attacks; 3) synthesis voice attacks, and 4) converted voice attacks. The first two types do not require strong technical skills. In an impersonation attack, the attacker imitates the target speaker's voice. In a replay attack, the attacker plays a pre-recorded voice of the target speaker. Synthesis attacks require more technical sophistication and involve techniques to collect samples of a target speaker's voice. With these samples, the attacker can extract the voice features of the victim and use the voice features to generate a speech that sounds like it is from the victim. For converted voice attacks, an attacker will use devices or software to convert his own voice to the target speaker's voice.

4) Countermeasures: Wu et al. [31] published countermeasures for replay attacks, synthesis voice attacks, and converted voice attacks. The study did not address impersonation attacks, since this method cannot easily break speaker verification systems.

Of the many ways to resist replay attacks, the easiest method is to compare the incoming recording to one or more stored ones [25]. If the incoming recording is similar to the stored ones and exceeds a defined threshold, the authentication system would consider the incoming recording as a replay attack. Synthesis and converted voice attacks shared some similarity since the vocoders use similar techniques to generate voices. Countermeasures for these attacks include examining the discriminative features or Mel-cepstral coefficients in order to discriminate between natural and synthetic voices [5], [24].

Although the countermeasures were reported to be effective, their efficiency was questionable without a standard dataset, protocols, and metrics for testing. The ASV Spoofing and Countermeasures (ASVspoof) initiative was created to solve this problem [30] but only achieved a detection equal error rate of 6.78% [17].

III. Two-layer authentication method with voiceprint

A. Assumptions and limitations

This authentication method is designed as a service to install in a voice-assistant-enabled device. It can identify and verify the user by voiceprint; further, no extra secure token or device installation is required. The goal of this authentication method is to protect voice assistants from replay attacks.

The use of voiceprint as the authentication method requires some assumptions of the system:

1) The voice-assistant-enabled device is only used by one person or fewer than six people.
2) The voice-assistant-enabled device is not compromised by malicious software that records the user's voice.
3) The voice-assistant-enabled device is not compromised by malicious software that records the user's voice through the connected IoT devices.

The first assumption prevents the system from losing its ability to authenticate a user. False match rate (FMR), also known as false acceptance rate (FAR), measures the authentication accuracy as the rate at which an invalid input biometric is matched with a record in the database. In a database with N records, the FMR can be shown as formula 3.1. When $N \rightarrow \infty$, the FMR(N) will close to one.

$$FMR(N) = 1.0 - [1.0 - FMR(1)]^N \tag{2}$$

In this situation, any input biometric can be authenticated. This means the biometric completely loses the ability to accurately authenticate. The second assumption ensures that attackers cannot record the user's voice through the voice assistant. The third assumption extends the second assumption from voice assistants to connected devices.

B. Authentication method

The proposed two-layer authentication method includes two processes: one for enrollment, and the other for user authentication.

1) Enrollment process: Before a user works with the system, she must first register her voice. The user is first asked to read a sentence composed of digits. The system records the user's voice and sends the recording to the backend. Next, the automatic speaker verification (ASV) system obtains the recording and extracts the user's voice model (i.e., voiceprint). Finally, the ASV system saves the user's voice model in the database.

2) Authentication process: The authentication process will be explored from two different perspectives: the user and the backend.

From the user perspective, the authentication process contains three actions: 1) trigger voice assistant; 2) speak service command, and 3) reply challenge text. The design of the user flow adds a step in the end, in order not to interrupt the original user flow of voice assistants. Steps 1 and 3 are the same as the steps of using voice assistants. Step 6 is the additional step for challenge-response protocol to mitigate replay attacks.

In Step 1, users have to speak the keyword, such as "OK Google" to trigger the voice assistant service. When the voice assistant is listening, speakers can go to Step 3 and say a service command, such as "buy me a cup of coffee." In Step 6, users then might receive a series of random numbers (challenge) which he must repeat within 5 seconds, and then wait for the system to execute or reject the demand.

From the backend perspective, the process contains six actions. Step 2 shows that the system triggers the voice assistant when it receives the trigger command and will emit a sound to let the speaker know it is working. While a user is speaking a service command, the backend is recording his voice.

In Step 4, the backend sends the recording to a speaker verification system after the user finishes the command. The speaker verification system will then determine if the voiceprint of the recording matches a registered user. If it finds a match, the system identifies the speaker as a specific user. If the speaker verification system cannot find a best match model among the enrolled users, the identification fails. The backend will terminate and tell the user that identification has failed.

In Step 5, if the identification is successful the backend will generate a series of random numbers as a challenge. The random numbers can be shown as text-to-speech (TTS) or as text on a screen. After prompting the speaker with the challenge, the backend records voice for five seconds and this will be considered the response to the challenge.

In Step 7, the backend sends the challenge response recording to the speaker verification system to determine whether the speaker is the same as the identified speaker in Step 4. If the result shows that the speaker in Step 7 is different from the one in Step 4, the backend will be terminated and not execute the service command. However, if the result indicates that the two identified speakers are the same, the backend will move to the next step.

Step 8 is compares the text of the challenge with the response. The recording is sent to the speech recognition system to transfer the voice to text. The backend uses this text to compare with the challenge. If the response text does not match the challenge, the authentication fails and the backend will terminate. If the response text is identical with the challenge, the backend will execute the service command.

The system is called two-layer authentication because it utilizes two voice inputs to identify and verify the user. The first layer is Step 4, where backend confirms the speaker is a registered user of the voice assistant. This mechanism prevents non-authorized speakers from accessing the voice assistant. Attacks discussed earlier will not pass this layer unless the attackers can get the enrolled users' voice recording. The second layer for this authentication is Step 7. If an attacker can get an enrolled user's voice recording and pass the first layer, the second layer utilizes random numbers to challenge the speaker. The 5-second time limitation makes it hard to generate the response with the enrolled user's voice.

C. System implementation

In order to evaluate the usability of the two-layer authentication with voiceprint, this study has to simulate the authentication process. In the simulated process, the user is using a voice assistant to transfer money and must pass the two-layer authentication to complete the task.

Since the voice assistant is pre-installed in Androids, the mobile platform with the largest market, this study implemented an Android application. The application applies Microsoft Azure Speaker Recognition API for speaker recognition and Google Cloud Speech API to accomplish speech recognition.

The application contains four managers to access different sources and handle different jobs. The voice recognition manager handles the enrollment, identification, and verification jobs for the speaker verification system, which is MS Azure Speaker Recognition.

The enrollment job records the user's voice and sends it to the verification speaker system to help enroll the user in the system. The identification job sends the recording to the speaker verification system. The speaker verification system, utilizing the verification job to verify the enrolled user, sends the user to our system. The speech recognition manager handles speech recognition and speech comparison jobs. Google Speech Recognition API converts speech to text, and the speech comparison manager compares the text against the challenge. The audio manager is responsible for recording voices via the smartphone's microphone and using the transferring-audio-format job to convert the audio. The log manager handles create, update, and delete log file jobs, which assist us with usability experiments and should not be used in commercial products. The log files record the date, participant id, identification and verification results, and challenge and response comparisons.

The application was deployed to a Huawei Nexus 6P phone running Android 8.0.0 on a Snapdragon 810 processor that provides 2.0GHz octa-core and 64-bit computing power. The decision to deploy on Android 8.0.0 was made because, as of this writing, it is the newest version of Android and will become the dominant version in Android phones in the near future. For the encoding format of audio files, the audio recorder was configured to a 16K sample rate, monophonic channel, and PCM 16-bit encoding format to fulfill the requirements of the MS Azure Speaker Recognition API.

IV. Usability experiment

A. Experimental design

The main goal of the experiment is to understand how users feel and think about the two-layer authentication. This experiment process contains 10 steps and takes 30 minutes.

Step 1 investigates the background of the participant. In addition to basic user demographics (e.g. gender, age, and major), the background questionnaire also asks the user about the frequency of voice assistant use, the awareness of the security of information and privacy in general, and the knowledge of voice assistant security. Then in Step 2, the participant is told that voice assistants can do many things (e.g., transfer money, buy products, unlock doors) when connecting to different applications and devices.

In Step 3, the participant is asked to give a service command to the Google voice assistant to complete each of four tasks. The participant will then fill out a questionnaire to evaluate the usability of the voice assistant. In Step 4, the participant is shown two voice assistant attacks via video: 1) a BadVoice [32] and 2) a DolphinAttack [33]. and then the participant is again asked to take the usability questionnaire.

In Step 5, two protections will be introduced and demonstrated to the participant. The two protections are designed to evaluate the difference between usability of one-layer and two-layer authentication. One-layer protection (A) only covers speaker identification. Two-layer protection (B) combines speaker identification and the challenge-response protocol for verification. Step 6 asks the participant to recite a series of numbers for system enrollment, which completes the setup of the two protections.

In Step 7, the participant starts to experience the two protections. There are two factors that can affect the participant's opinion on the usability of a protection: 1) first impression [16] and 2) habit [20]. For first impression, the participant might prefer the protection that he starts with. Therefore, this experiment controls for the influence of the first impression by separating the participants into two groups. Participants in group 1 will start with protection A and those in group 2 will start with protection B. For habit, if the participant repeats a protection, she might think that it is easier to use this protection than the other. To avoid the influence of habit, the participant will alternate the use of the two protections (i.e., AB or BA) and will repeat the order four times (i.e. ABABABAB or BABABABA).

Steps 8 and 9 asks the participant to experience both protections again. The order depends on the group of the participant. She will then fill out the usability questionnaires for the protection experienced. For example, if the participant begins with protection A, she will experience protection A again in Step 8 and fill out questionnaire for protection A; in Step 9, the participant will use protection B and fill out the questionnaire for it.

In Step 10, an interview is held. The participant was asked the following five questions:

1) What do you feel/think about the security of your information?
2) What do you feel/think about privacy?
3) What are you worried about when using this technique?
4) How do you feel when you interact with this technique?
5) What do you feel/think about using other biometrics (face recognition, fingerprint) on voice assistants?

V. Results

A. Participants' background

In this section, we analyze the background questionnaire, which used a 7-point Likert scale, representing either never to very often or strongly disagree to strongly agree. The average experiment time was 30 minutes and a total of 41 participants were invited. The experiment included 21 male and 19 female participants, with an additional participant preferring not to answer this question. Most (39 of 41) of the participants were 30 years of age or younger; 23 participants were STEM majors; and 22 participants had never or had rarely used a voice assistant before the experiment. Only 2 said they were never worried about the security of their information and privacy. For knowing the privacy risks and security threats of voice assistants, 10 of the participants chose neither agree nor disagree. For understanding the different approaches to protect users, 23 of the participants somewhat to strongly disagreed.

1) Group analysis: The participants were separated into 4 groups. The participants in group 1 (G1) and group 2 (G2) watched the two voice assistant attack videos, while the participants in group 3 (G3) and group 4 (G4) did not. Therefore, the participants in G1 and G2 filled out the usability questionnaire four times (QNR1, 2, 3, and 4); others in G3 and G4 filled it out three times (QNR1, 3, and 4). The participants in G1 and G3 started with protection A (1-layer protection: identification only) and then used protection B (two-layer protection: identification and verification); the participants in G2 and G4 started with protection B followed by protection A. The groups varied slightly in size: G1 had 11 participants, while G2, G3, and G4 each had 10 participants. This study also maintained a gender balance in each group.

B. Analysis method

To evaluate the questionnaire, this study used the independent-samples t-test and the paired-samples t-test to evaluate the mean differences [14]. In our questionnaire, we separated the questions into two types: Usability (Q1-Q10) and Sense of Security (Security) (Q11-Q14)

We calculated the usability and security means. The usability mean represents the mean of questions of the usability type (Q1-Q10), while the security mean is the mean of questions of the security type (Q11-Q14). The range of the mean values are from 1 to 7. This study used IBM SPSS Subscription (June, 2018) to perform statistical analyses. IBM SPSS provides the 2-tailed t-test calculation, and we used it to compare the usability and security means of groups and the questionnaires completed at various points during the experiment.

C. General analysis

In this section, we discuss the participants' opinions about the two protections. To evaluate the participants opinions about 1-layer protection, we compared questionnaire 1 and 3 (Pair 1: QNR1 vs. QNR3). To analyze the participants' opinions about 2-layer protection, we compared questionnaire 1 and 4 (Pair 2: QNR1 vs. QNR4). Additionally, to ascertain the difference between participants' views on 1-layer and 2-layer protection, we compared QNR3 and QNR4 (Pair 3).

From a usability standpoint, there was a statistically significant difference in perceptions about how usable the 2-layer protection was versus the 1-layer protection. One-layer protection was seen as more usable. In contrast to usability, perceptions related to security were increased with respect to the 2-layer protection (Pair 3).

D. The effect of revealing the security information

Since half of the participants watched the attack videos of voice assistants, this section discusses the effect of the security information on participants' responses. The means of QNR1 and QNR2 are used for the t-tests and comparison. The usability mean and the security mean of QNR1 and QRR2 are significantly different. This indicates that after the participants watched the attack videos, they changed their opinions about voice assistants in terms of usability and security.Thus, revealing security information decreases the participants' opinion of voice assistant's usability and sense of security.

One interesting result is that revealing the security information decreased the usability of voice assistants. Participants may have been dissuaded from voice assistants due to increased security concerns resulting from the video.

E. Participants without security information

Earlier, we found that revealing security information changes the participants' opinions. Thus, this study separates the participants into two groups based on whether the security information (videos) were shown and evaluates the two groups' respective opinions. In this section, we focus only on the group without security information. QNR1, QNR3, and QNR4 from G3 and G4 are used in the analysis.

From a usability standpoint, a statistically significant difference was found between 1-layer and 2-layer protection. The usability mean of the 2-layer protection is smaller than that of the 1-layer protection. Thus, the participants believe that the 2-layer protection is less usable, compared to the 1-layer protection. In contrast, the participants without security information feel the 2-layer protection is more secure than the 1-layer protection.

F. Participants with security information

This section discusses the opinions of the participants with security information. QNR1, 2, 3, and 4 of G1 and G2 are used for the evaluation. Pairs 2 (QNR1 vs. QNR 4) and 5 (QNR 3 vs. QNR 4) indicate that the participants with security information have significantly different usability means compared the participants' opinions of 2-layer protection to that of voice assistants before knowing the security information. In addition, after knowing the security information, the participants have different opinions about the usability when comparing 1-layer protection with 2-layer protection.

Before the participants learn about the security information, they feel that 2-layer protection decreases the usability to the voice assistant. When comparing the usability of the 1-layer protection and the 2-layer protection (Pair 5), participants feel 1-layer protection is more usable than 2-layer protection.

From a security perception standpoint, Pairs 3 (QNR2 vs. QNR3) and 4 (QNR2 vs. QNR4) indicate that the participants with security information have a different sense of security for the 1-layer protection and the 2-layer protection. The 1-layer protection and the 2-layer protection increase the sense of security after the participants have been exposed to the security information.

G. The questionnaire differences of participants with and without security information

This section evaluates the differences of voice assistants and protections between the participants with and without security information. QNR1, 3, and 4 of G1, G2, G3, and G4 are used in the analysis. Differences in QNR4 are statistically significant, which indicates that the participants with and without security information have significantly different opinions about 2-layer protections. Participants with security information (G1G2) perceive this voice assistant implementation as having less usability than the participants without security information (G3G4) in 2-layer protections.

The reason for the difference could be that the participants with security information have higher criteria on usability than the participants without security information. To examine this possibility, we used 1-tailed independent-samples t-test to assess the usability-mean differences of QNR1, 3, 4 between G1G2 and G3G4. QNR1, 3, and 4 have statistically significant values. Thus, participants with security information have higher criteria on usability than the participants without security information. From a security standpoint, the participants both with and without security information have a similar sense of security on voice assistants, 1-layer protection, and 2-layer protection.

H. Summary

For the entire participant pool and participants without security information, the usability and security of 1-layer protection does not change. However, after the participants

were shown the security information, their sense of the security of 1-layer protection increased.

For 2-layer protection, the participants perceive that usability decreases and the security does not change when comparing with voice assistants. But when we take a closer look, the participants with and without security information have different opinions. The participants without security information kept the same opinion about usability and security of 2-layer protection when comparing with voice assistants. For participants with security information, 2-layer protection has less usability and the same security before they knew the security information of voice assistants. However, after the participants with security information knew the security information, they perceived 2-layer protection as having the same usability with voice assistants and as being more secure.

Finally, all participants believe that 2-layer protection is less usable but more secure than 1-layer protection. However, the participants with security information felt that 2-layer protection has less usability and the same sense of security as 1-layer protection. Thus, when participants know the threats of voice assistants, they don't consider 2-layer protection to harm usability and also feel more secure.

I. Accuracy

In the experiment, we recorded the results identification and verification to logs. Since our experiment includes only benign users with no attackers, the logs can provide true acceptance rates (TAR) and false rejection rates (FRR), but cannot evaluate false acceptance and true rejection rates.

We have 568 records from the participants for the MS Azure Speaker Recognition System, of which 415 records shows that the system correctly recognized a participant. Thus, the TAR of the MS Azure Speaker Recognition System is 73.06% and the FRR is 26.94%. For 2-layer protection, we have 206 records, only 135 of which show that a participant was authenticated. Therefore, 2-layer protection gets only 65.53% TAR and 34.47% FRR.

The results do not meet our expectations. Since the accuracy of 2-layer protection depends on the speaker recognition system, the lower accuracy of the MS Azure Speaker Recognition System causes the low accuracy of 2-layer protection. Two-layer protection gets a lower TAR and higher FRR for the same reason. Since 2-layer protection terminates the authentication when the first layer fails, the high FRR of the speaker recognition system increases the FRR of 2-layer protection.

Finally, the factors that cause the high FRR of the MS Azure Speaker Recognition System may include: 1) the short training speech, and 2) the participant using a tone different from their normal speaking voice when creating the training speech. In the experiment, the participants were asked to say 10 digits to create the training speech, which took around 5 seconds. The MS Azure Speaker Recognition System suggests a 30-second training speech, without silence. However, we did not require the participants to generate a 30-second training speech on the grounds that the long speech without silence is tedious and could decrease usability.

The other possible reason for the high FRR of the MS Azure Speaker Recognition System is that some participants were nervous when they were recording the training speech. The nervousness changed the participants' tone and affected the quality of the voiceprint. After the enrollment phase, participants were relaxed and used his or her normal tone when using the 1-layer and 2-layer protections. This explains why the participants were rejected from the 1-layer and 2-layer protections, and increased the FRR of the MS Azure Speaker Recognition System.

J. Interview results

In the interviews, we found that numerous participants feel that their information and privacy is insecure. Only 2 participants said they do not care about privacy at all. Nearly half (20 of 41) of the participants are worried about the security of voiceprint, including the possibility that someone can mimic their voice and break the protection. Of the participants, about 39% of them said they worried about the accuracy of voiceprint.

Participants in G1 and G2 gave opinions about the user interfaces and the response speed of 1-layer and 2-layer protection, while participants in G3 and G4 did not mention this. Participants in G1 and G2 also have slightly higher criteria regarding security. Specifically, 4 participants in G1 and G2 said they would not use voice assistants to do financial tasks, while no one in G3 and G4 made this kind of statement. In addition, three participants in G1 and G2 said they do not trust biometrics as a way to authenticate themselves; only 1 participant in G3 and G4 has the opinion.

Lastly, we also received an interesting suggestion from some participants. They suggested we apply different protections to different situations, such as using 1-layer protection to simple tasks (i.e., texting) and 2-layer protection to financial tasks.

VI. Conclusion

This research has presented a two-layer authentication method to protect voice assistants and maintain their usability. By using a voiceprint and challenge-response protocol, the authentication method can recognize the speaker through the input voices and resist replay attacks by requiring the users to respond to the challenge within 5 seconds.

The results show that the false rejection rate (FRR) of MS Azure Speaker Recognition System is 26.94%. This FRR demonstrates that it cannot provide sufficient authentication to protect voice assistants. However, a major advantage of using voiceprint to authenticate speakers is that users are only required to enroll their voices into the system, rather than carry an additional token. Thus it is easy to integrate voiceprint because it only inserts three steps — identification, challenge and verification — before voice assistants execute a command.

A. Limitations

There are three primary limitations of this project. First, the security of our method is based on the security of speaker

recognition systems. The accuracy will affect the user's opinion regarding usability and sense of security. Therefore, in our experiment if the participant was wrongly rejected or accepted by the system, the usability of the two-layer authentication might be underestimated by the participant. Second, this study utilized the MS Azure Speaker Recognition System to verify speakers. It assumes that the communications between the authentication method and the remote speaker recognition system are secure. In the real world, network communications should be encrypted. Additionally, the method needs a quiet environment to perform well.

B. Future work

The speaker recognition library limitations should be reduced. This would involve finding a suitable training speech that can assist the speaker recognition system to create an ideal voiceprint and without irritating the user and reducing usability. Additionally, various countermeasures of speaker recognition systems could be applied to the 2-layer protection. In this way, we could get stronger evidence that the 2-layer authentication method can improve the security of voice assistants. Finally, other future work could involve cooperating with voice assistant companies and implementing the method with real services. This would enable us to gain more reliable data from users and thus have higher confidence about usability.

REFERENCES

[1] Voice Match and media on Google Home - Google Home Help. https://support.google.com/googlehome/answer/7342711?hl=en.

[2] Apple HomePod vs. Amazon Echo vs. Google Home, June 2017.

[3] Chandrasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywe, Lorrie Faith Cranor, and Marios Savvides. Biometric authentication on iphone and android: Usability, perceptions, and influences on adoption. *Proc. USEC*, pages 1–2, 2015.

[4] C. H. Chen and C. Y. Chen. Optimal fusion of multimodal biometric authentication using wavelet probabilistic neural network. In *2013 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 55–56, June 2013.

[5] L. W. Chen, W. Guo, and L. R. Dai. Speaker verification against synthetic speech. In *2010 7th International Symposium on Chinese Spoken Language Processing*, pages 309–312, November 2010.

[6] Marc Dupuis and Robert Crossler. The compromise of ones personal information: Trait affect as an antecedent in explaining the behavior of individuals. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. IEEE, 2019.

[7] Marc Dupuis, Robert Crossler, and Barbara Endicott-Popovsky. Measuring the human factor in information security and privacy. In *The 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2016.

[8] Marc Dupuis and Mercy Ebenezer. Help wanted: Consumer privacy behavior and smart home internet of things (iot) devices. In *Proceedings of ACM SIGITE conference (SIGITE '18)*, Oct 2018.

[9] Huan Feng, Kassem Fawaz, and Kang G. Shin. Continuous Authentication for Voice Assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, MobiCom '17, pages 343–355, Snowbird, Utah, USA, 2017. ACM.

[10] Purdy Ho and John Armington. A Dual-Factor Authentication System Featuring Speaker Verification and Token Technology. In *Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 128–136. Springer, Berlin, Heidelberg, June 2003.

[11] Lin Hong and Anil Jain. Integrating faces and fingerprints for personal identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1295–1307, December 1998.

[12] Y. Isobe, Y. Seto, and M. Kataoka. Development of personal authentication system using fingerprint with digital signature technologies. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pages 9 pp.–, January 2001.

[13] Andrew Teoh Beng Jin, David Ngo Chek Ling, and Alwyn Goh. Biohashing: Two factor authentication featuring fingerprint data and tokenised random number. *Pattern Recognition*, 37(11):2245–2255, November 2004.

[14] Harry N. Boone Jr and Deborah A. Boone. Analyzing Likert Data. *Journal of Extension*, 50(2), April 2012.

[15] D. S. Kim and K. S. Hong. Multimodal biometric authentication using teeth image and voice in mobile environment. *IEEE Transactions on Consumer Electronics*, 54(4):1790–1797, November 2008.

[16] Heejun Kim and Daniel R. Fesenmaier. Persuasive Design of Destination Web Sites: An Analysis of First Impression. *Journal of Travel Research*, 47(1):3–13, August 2008.

[17] Tomi Kinnunen, Md Sahidullah, Hector Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. 2017.

[18] Tyler Lacoma. Virtual assistant comparison: Cortana, Google Assistant, Siri, Alexa, Bixby. https://www.digitaltrends.com/computing/cortana-vs-siri-vs-google-now/, August 2017.

[19] Xinyu Lei, Guan-Hua Tu, Alex X. Liu, Chi-Yu Li, and Tian Xie. The Insecurity of Home Digital Voice Assistants - Amazon Alexa as a Case Study. December 2017.

[20] Chechen Liao, Prashant Palvia, and Hong-Nan Lin. The roles of habit and web site quality in e-commerce. *International Journal of Information Management*, 26(6):469–483, December 2006.

[21] Václav Matyáš and Zdeněk Říha. Biometric Authentication — Security and Usability. In *Advanced Communications and Multimedia Security*, IFIP — The International Federation for Information Processing, pages 227–239. Springer, Boston, MA, 2002.

[22] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001.

[23] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1):19–41, January 2000.

[24] Takayuki Satoh, Takashi Masuko, Takao Kobayashi, and Keiichi Tokuda. A robust speaker verification system against imposture using an HMM-based speech synthesis system. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[25] W. Shang and M. Stevenson. Score normalization in playback attack detection. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1678–1681, March 2010.

[26] Frank Stajano. Pico: No More Passwords! In *Proceedings of the 19th International Conference on Security Protocols*, SP'11, pages 49–81, Berlin, Heidelberg, 2011. Springer-Verlag.

[27] Fitz Tepper. Your voice is your password with Sesame's Alexa app, 2016.

[28] Aaron Tilley. How A Few Words To Apple's Siri Unlocked A Man's Front Door. https://www.forbes.com/sites/aarontilley/2016/09/21/apple-homekit-siri-security/, September 2016.

[29] P. C. van Oorschot and Tao Wan. TwoStep: An Authentication Method Combining Text and Graphical Passwords. In *E-Technologies: Innovation in an Open World*, Lecture Notes in Business Information Processing, pages 233–239. Springer, Berlin, Heidelberg, May 2009.

[30] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado. ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):588–604, June 2017.

[31] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66:130–153, February 2015.

[32] Park Joon Young, Jo Hyo Jin, Samuel Woo, and Dong Hoon Lee. BadVoice: Soundless voice-control replay attack on modern smartphones. In *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 882–887, July 2016.

[33] Guoming Zhang, Chen Yan, Xiaoyu Ji, Taimin Zhang, Tianchen Zhang, and Wenyuan Xu. Dolphinatack: Inaudible voice commands. *arXiv preprint arXiv:1708.09537*, 2017.