A Two-layer Authentication Using Voiceprint for Voice Assistants

Yun-Tai Chang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Cyber Security Engineering

University of Washington

2018

Committee:

Marc Dupuis, Chair

Brent Lagesse

Arnie Lund

Program Authorized to Offer Degree:
Computing & Software Systems

University of Washington

Abstract

A Two-layer Authentication Using Voiceprint
for Voice Assistants

Yun-Tai Chang

Chair of the Supervisory Committee:
Assistant Professor Marc Dupuis
Computing & Software Systems

Voice assistants are a ubiquitous service of contemporary daily life. Their intuitive use and 24-hour-a-day convenience make them popular and have more users. However, the security of voice assistants does not increase as much as the rising amount of users and increasing 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 thesis is to provide an authentication method that protects voice assistants from attacks without degrading their usability. We utilize Microsoft cognitive speaker recognition API and Google speech API to implement an Android application to examine the approach. The result indicates that the voice authentication method can resist replay attacks and it is easy to use and learn for users.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

"Thanks be to God for his indescribable gift!" 2 Corinthians 9:15

My sincere thanks to University of Washington Bothell, Professor Marc Dupuis, Professor Brent Lagesse, and Professor Arnie Lund. Professor Marc gave me invaluable help with the research and the experiment, Professor Brent assisted and encouraged me to find out and study into the security of voice assistants, and Professor Arnie provided precious advice to the questionnaires.

I would also like to thank Professor Nancy Kool, she spent a lot of time to read my thesis draft and help me to write better. Finally, heartfelt thanks go to my family, friends, and others who have helped me. The help I recevied is too much to detail.

This research would have been impossible to complete without any of them. I appreciate all of them and I am profoundly grateful to have them in my life.

Chapter 1

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. Their intuitive operation makes voice assistants the most wanted gift in recent years [1]. In addition, voice assistants provide services through integrating with applications or smart devices. For example, Paypal users can send money through Apple Siri [2]. Although voice assistants make daily life more convenient, they are also vulnerable to voice attacks. The demonstrations of bad and inaudible commands [52, 53] have indicated the ease with which voice assistants can be exploited and the importance of securing voice assistants by authentication. Thus, researchers and developers have proposed various methods for authenticating the user, such as asking the user to wear an additional token to prove her identity [16], or using a voiceprint to verify herself [42]. 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 the usability and adds an additional challenge-response layer to enhance the security of voice assistants.

In this chapter, section 1.1 details the abilities and the security issues of voice assistants. Section 1.2 introduces the limitations of current authentication methods for voice assistants. Section 1.3 includes the motivation and research focus. Section 1.4 discusses goals and criteria of this study. Section 1.5 briefly introduces the approach and the contribution of this study.

1.1 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 of voice assistants can provide numerous services. Depending on the specific applications and devices voice assistants connect to, they can assist the user, from making a phone call, reading an email, or scheduling an event, to controlling various Internet of Things (IoT), devices such as smart locks [27, 4].

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. For instance, a user can ask the Google voice assistant to schedule a meeting by saying, "Ok, Google, set a meeting with Samuel on Friday." If the voice assistant works correctly, it will create an event on the user's Google calendar.

With their multifold advantages of voice assistants, voice assistants are insecure. Because voice assistants use voice as input, they are always listening. This feature makes voice assistants vulnerable since the voice is on a public channel. Attackers can exploit a voice assistant by making voices that can be received and be interpreted by the voice assistant [43, 53]. For instance, after installing a smart lock linked to Siri to his front door, a man found that Siri would unlock the door to anyone who could give Siri the voice command "unlock the door" [43]. This example suggests that the voice assistant should confirm the identity of the speaker before executing the command.

Because voice assistants can connect to applications and devices that are related to the user's financial accounts and privacy (e.g., Paypal and door locks), it is necessary to integrate authentication into voice assistants to improve security.

1.2 Limitations of current authentication methods for voice assistants

Researchers have utilized various methods to increase the security of voice assistants. 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 [16, 28]. Users' unwillingness to wear an additional gadget or to install sensors in the house hinders the widespread adoption of these methods. Other methods, such as using a voiceprint to identify speakers, are more user-friendly but could be exploited by sending a pre-recorded victim's voice (replay attacks) [42, 3].

This conflict between security and usability is well-known [14, 51]. It is easy to implement a method that provides robust protection but with imperfect usability, and vice versa. Losing the balance of security and usability will lead users to abandon the method. 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. Therefore, researchers need to maintain the balance between security and usability to make the user will to use the authentication method.

1.3 Motivation and research focus

Voice assistants need to be secured. Unlike the past, when users only used voice assistants to turn on/off lights, voice assistants now can manipulate things with more serious implications, such as bank accounts, email, and online payment applications. This indicates that voice assistants need to introduce authentication to ensure that users do not face a financial loss or a violation of their privacy.

However, the current authentication methods to voice assistants decrease usability and are not deployed in the real world. This motivates this study to create an authentication mechanism that can provide robust security and maintain usability. Considering the intuitive interaction of voice, this study uses voiceprint as authenticator to maintain usability. Thus, the research focuses on utilizing voiceprint to provide an easy-to-use and secure authentication method for voice assistants.

1.4 Goals and criteria

Since this study uses voiceprint as authenticator to maintain usability, the authentication mechanism has to take the vulnerabilities of voiceprint into consideration. Among voice

attacks, replay attacks are the most threatening and prolific type [48, 26]. To mitigate replay attacks, this research proposes a two-layer authentication method using voiceprint. The two-layer authentication method allows users to use their own voices to authenticate their identities on voice assistants.

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:

- Criteria 1: Achieve false acceptance rate and false rejection rate of less than 1%.
- Criteria 2: Maintain the ease to use with which this can be accomplished: the two-layer authentication method should make the users feel willing to use.

1.5 Approach overview and contributions

The two layers in the authentication method have distinct functions. The purpose of the first layer is to identify the speaker so that only registered users can be accepted. The purpose of the second layer is to verify the speaker by a challenge-response protocol; in this way, only the real user can respond to the challenge in time and authorize the request. This study integrates an automatic speaker verification system and a speech recognition system to implement the two-layer authentication method. Further, the study also simulates the process by which a user transfers money to another person through a voice assistant with the two-layer authentication method.

With the proposed mechanism and implementation, the contributions of this research are:

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

Chapter 2
RELATED WORKS

This chapter reviews studies of voice assistants, authentication, and voice speaker recognition systems. Section 2.1 discusses abilities, attacks, and protections of voice assistants. Section 2.2 details the usability and security. Section 2.3 explains the technology, attacks, and countermeasures of voice speaker recognition systems.

2.1 Voice assistants

Voice assistants have become pervasive in our daily lives. Thus, it is important for us to understand voice assistants. This section discusses the abilities, abuses, attacks, and protections of voice assistants.

2.1.1 Abilities

Voice assistants are a type of human-computer interface with many capabilities. Lacoma compared popular voice assistants by different functions, such as launching apps, setting schedule, making calls, sending messages or emails, and playing music [27]. Beyond the applications functions of voice assistants, researchers have also integrated voice assistants to the Internet of Things (IoTs) to provide more efficient systems in different fields such as smart homes [7], assistant robot [33], or drug delivery [12].

2.1.2 Abuses and attacks

While voice assistants become more and more efficient and useful, the abuses of voice assistant can cause serious loss to users. For instance, a little girl asked Alexa in her home, "Can you play dollhouse with me and get me a dollhouse?" Alexa responded to the request and ordered

an expensive dollhouse through Amazon [30]. In another example, a man spent thousands of dollars to build a smart home and found that anyone can unlock his front door by stating outside his house, simply by saying, "Hey Siri, unlock the front door." The voice assistant in the house would receive the command and execute it without confirmation [43]. These cases show that voice assistants are vulnerable and need protections.

To attack voice assistants, audio input is an easy method since voice assistants are always listening. Researchers have demonstrated that voice assistants are easy to attack and exploit through malicious voice input. Diao et al. [15] launched an attack that utilized a text-to-speech (TTS) system and the voice assistant (Google search app) on a smartphone. Once the user installed the malicious app, the app would generate text commands, such as "OK, Google, call 1234 5678". The text command would be converted to a voice command through TTS. While the app plays the voice command through the speaker of the smartphone, the voice assistant of the smartphone would receive the voice command and execute it. The phone number could be a malicious number; the attacker could gain sensitive data through this attack if the victim was in a private meeting. It is difficult for anti-virus scannings to discover this malicious app since the app only required the permission of the audio speaker and did not ask for permission to access sensitive data. The limitations of the attack module proposed by Diao et al. was that the attack could only launch in the middle of the night and use the minimum volume to play the malicious voice command since the user could easily hear the attack if the malicious voice command was too loud.

While the above attack may seem unrealistic, due to the limitations, Young et al. [52] introduced a hardware device to attack the Siri voice assistant on a smartphone. The device physically accessed the victim's smartphone and spent less than three minutes to compromise the voice assistant. The attack could hijack the victim's accounts, such as Facebook and Google. Additionally, the attack device cost less than one hundred dollars. The short attack time and low-cost device demonstrated the vulnerability of voice assistants to attacks.

The above attack can be relatively easy for the user to discover, because the attackers need to physically access the smartphone in order to play voice commands. However, there is

a gap in the speech recognition system between human and machine [45]. This showed that distorted voice commands that sound like noise to a human, could be interpreted by speech recognition system. Attackers could utilize the gap to attack voice assistants. Carlini et al. [9] have demonstrated this attack and showed that, while a person could not understand the obfuscated voice commands, machines could interpret 82% of these voice commands. This attack was particularly effective in public areas since it sounded like background noises for users.

In addition to the mangled voice commands, researchers have also demonstrated ultrasound voice attacks [53, 40]. By shifting the frequency of voice commands to a level above the human range of hearing, the attack thus can be launched in any environment and has had a 100% success rate in quiet places (e.g., oce).

2.1.3 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 [3] to let users link their Google accounts and voices. Thus, Googles voice assistant could recognize the speaker through her voice and would only reveal the speaker's information. Google utilized a speaker's voiceprint to identify the user but claimed that the identification function cannot provide security protection. Some developers and researchers, aware of the lack of authentication for voice assistants, have done work to protect voice assistants.

Sesame [42] is an application of Amazon Alexa. This app utilizes the speaker's voice to authenticate Alexa and eable the transfer of money from the user to her friends. The application asks the user to record a phrase as her voice password; when the user wants to send money to a friend, the system will ask the user to say her voice password. After the system receives the replied password, it compares the voice features of the replied audio and the pre-recorded audio. If the features of the replied audio are similar to the pre-recorded one, the system will allow Alexa do its job and send money out. However, the protection could

be broken if the attacker recorded the voice passphrase and replayed it while the system is asking.

To prevent a replay attack, it is important for voice assistants to confirm that the user is near the voice assistant. Thus, researchers have considered checking the user's physical presence [16, 28]. Feng et al. [16] proposed a secure token that could continuously authenticate the user while she was using the voice assistant. The token requires that it touches the user's skin. When the user is speaking, the skin vibrates and the token will get the vibration data. The system uses the vibration data to calculate acceleration values. With the acceleration values, voice assistants can guarantee the voice command was generated by the user. The research indicates that the secure token can successfully protect voice assistants from replay attacks, mangled voice attacks, and impersonation attacks. Lei et al. [28] created a virtual security button to check if the user is in the house by detecting his motions. The study shows that the protection could prevent replay attacks.

2.2 Authentication

2.2.1 Security

Authentication is an important mechanism to protect security and privacy since its purpose is to confirm the truth of the identity claimed by a person. With authentication, a machine or the security department can let the user access data or execute actions that are authorized for her. 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.

O'Gorman [34] depicted the protocols of the three authentication factors and compared the three authentication factors with respect to security. This study concluded that although passwords are excellent authenticators, they could not detect compromise and repudiation. Tokens provide strong compromise detection but they are inconvenient and involve extra

cost. Biometrics can resist repudiation, since the user cannot lend their biometrics to others, but biometrics were unrecoverable once the feature data, representing the biometric, are breached. The study also indicated that challenge-response protocols could be a means to resist replay attacks and non-repudiation. Based on the challenge-response concept, Yan et al. [50] utilized a challenge-response protocol to authenticate the access of online or cloud services. The work implemented a system and showed that the challenge-response protocol was an efficient method of voice authentication.

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 among multi-factor authentication methods. The concept of two-factor authentication involves combining any two of the three authentication factors together. Among the combinations, many studies have combined biometrics and tokens [20, 18, 21].

Beyond combining two different factors, researchers also implemented two-step authentication methods that used two authenticators within the same factor [19, 46]. Hong and Jain [19] proposed to use one's face and fingerprint as a pair to authenticate an individual. The research used the user's face feature to identify and fingerprint to verify the user. U.S. Customs and Border Protection (CBP) is using this mechanism since it is a quick and accurate way to authenticate a person. These dual-biometric systems reduce the weakness of biometrics in security [24, 10].

2.2.2 Usability

Authentication methods require users to interact with them, thus, usability is a key factor for authentication. 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. Zviran et al. [54] did a study of password security and found that almost 50% of the 860 participants used five or fewer characters, 80% used only alphabetic characters to compose passwords, and 80% never changed their password. The goal of memorizability of passwords negatively influenced the security of passwords.

Tokens could ease the pain of the memorizability of passwords [41]. The study proposed a hardware token that was works-for-all, which could replace many types of passwords such as web login passwords, screen saver passwords, or PINs of standalone devices. Although tokens can provide stronger security, the extra devices are inconvenient and users can lose them. The risk of losing the devices reduces the usability of tokens.

Biometrics have the best usability among the three authentication factors. Unlike passwords, biometrics do not require memorizability. And unlike tokens, biometrics cannot be taken by others [5, 32]. Bhagavatula et al. [5] compared fingerprint and face unlock with PIN on smartphones. The results showed that users perceived fingerprint as more secure and convenient than a PIN. Matyáš et al. [32] claimed that biometrics are a effective way to authenticate users but should be an additional method to an existing one.

To compare the usability of different biometrics, Trewin et al. [44] did an examination focused on smartphones. The study compared six authentication methods: password, voice, face, gesture, face and voice, and gesture and voice. The results indicated, compared with typing the PIN and performing gesture, speaking the PIN was the fastest means to provide a sample to authentication method. The study also showed that it is easy to get a voice with sufficient quality when getting a sample of voice. For dual-biometric systems, users disliked them because it is more difficult to provide sufficient-quality sample entries. Additionally, users thought that dual-biometric systems were a barrier for their tasks since they had to recall them after the authentication.

Trewin et al. [44] administered a test scenario that involved authentication on smartphones. Before the user did her task, the system would first use voice to authenticate the user. Although this scenario interrup the user's tasks, it is useful to combine voice and face authentication when users cannot use hands. Thus, each biometric has its own advantages and disadvantages. The usability of biometrics depends on the use scenario.

In our study, we focus on voice authentication since voice is the only means by which allows users to communicate with voice assistants. Gunson et al. [17] evaluated different contents of voice authentication. The study found that users preferred digits to sentences.

Further, a series of random numbers made the user feel as secure as using sentences.

2.3 Voice speaker verification systems

From the discussion in subsections 2.1 and 2.2, it is clear that voice assistants need an authentication method, but the conflicts between security and usability have made it hard to find a proper authentication method for voice assistants. In the present study, voice is considered the best way to authenticate users since voice assistants use voice as the input. For all types of voice assistants, voice authentication needs no extra devices. This suggests that voice authentication method can be smoothly integrated into existing voice assistants. 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. It is worth to discuss the attacks on the speaker recognition system and the countermeasures.

The subsections discuss speaker recognition system, including basic background knowledge, technologies, attacks, and countermeasures.

2.3.1 Basic knowledge

Voice recognition is a part of speech processing. To identify the state of speaker recognition, Campbell [8] depicted an overview of speech processing, including speaker recognition, to show the components in speech processing.

Figure 2.1 shows that speaker recognition includes speaker identification, speaker detection, and speaker verification. For the input speech, it requires the speaker's cooperation to gain a high-quality input speech.

Speaker identification and speaker verification are different functions [36]. Speaker identification is a one-to-many comparison, while speaker verification is a one-to-one comparison. Speaker identification compares the speaker with all or a group of voiceprints in the database (1:N) to verify that the voice belongs to a particular user. In contrast, speaker verification compares the speaker with a specific user's voiceprint (1:1) to check if the voice is from her.

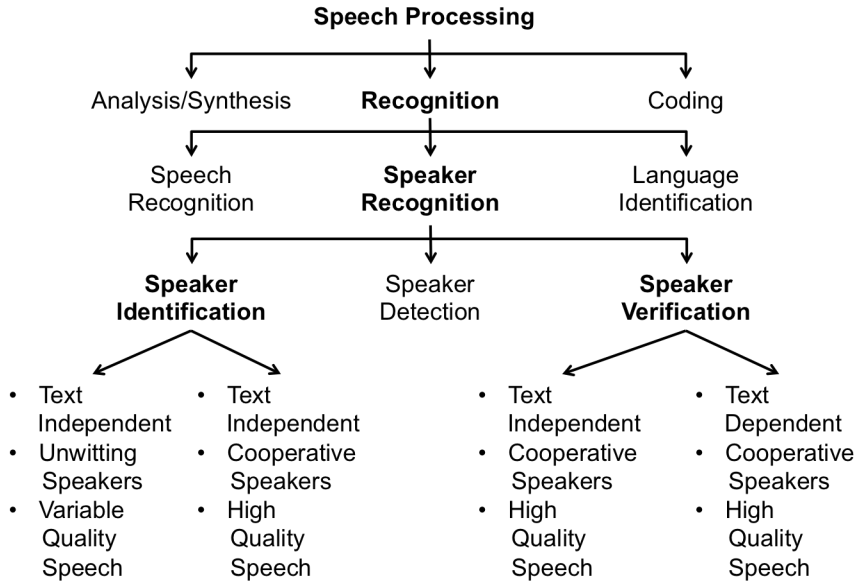


Figure 2.1: Speech processing [8]

Besides this difference between identification and verification, 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. If the speaker uses different sentences, the text-dependent system will not recognize her. This requires 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 have to enroll in it. Reynolds [36] depicts the components of speaker enrollment and speaker verification and identification shown in Figure 2.2 and Figure 2.3.

Figure 2.2 shows that speaker enrollment includes two phases. One is an offline phase and the other is online. The purpose of the offline phase is to generate a background model from the analog background voice. On the other hand, the online phase is to create the target user model. Feature extraction is used to transform the voice from analog to digital. After the

system receives the digital data, the data will be modified by a training algorithm to form a background model in the offline phase. For the online phase, the digital data will be adapted based on the background model to form a target model.

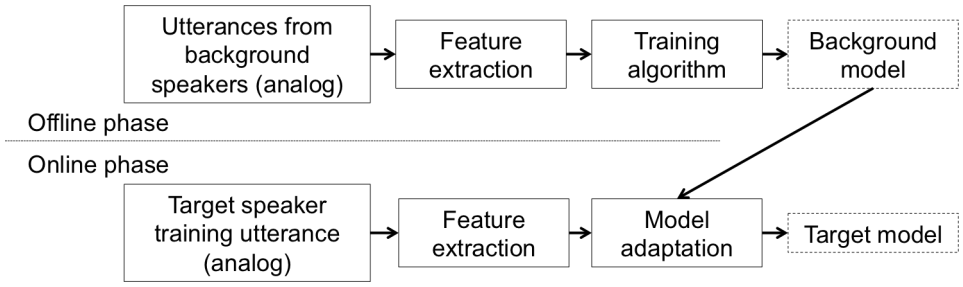


Figure 2.2: Speaker enrollment [36]

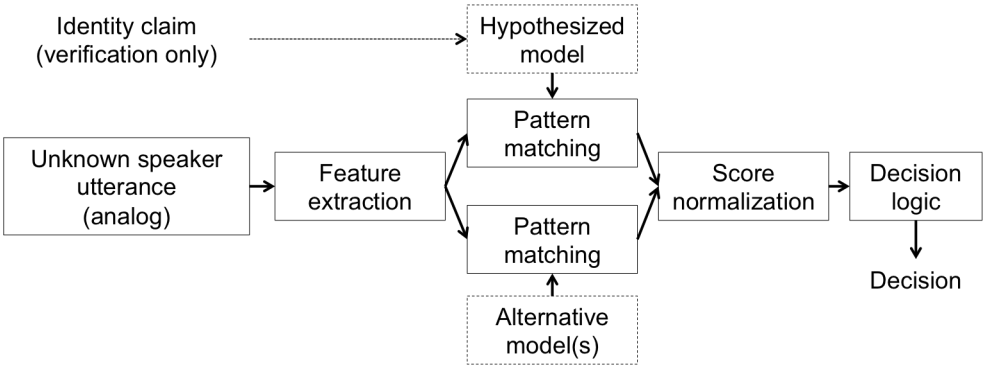


Figure 2.3: Speaker verification/identification [36]

Figure 2.3 depicts the process of identification and verification. The unknown speaker's voice in analog form will be transformed to digital data by feature extraction and this digital data will then be compared with other models in pattern matching. The difference between identification and verification can be seen here. For verification, the pattern matching will use the hypothesized model to compare with the digital data; for identification, the pattern matching will use models in the database for comparison. The comparison(s) gives score(s)

as output; the score(s) will be normalized in score normalization and sent to the decision logic. The decision logic uses a threshold to decide if the score achieves the pass criteria.

2.3.2 Technologies

The current technology to verify a speaker is the likelihood ratio test, which can compare two statistical models [37]. The hypotheses for the likelihood ratio test of speaker verification are

Hypothesis 0 (H0): *Input speech Y is from the hypothesized speaker S.*

Hypothesis 1 (H1): *Input speech Y is from other speakers.*

Further, the likelihood ratio test to decide between these two hypotheses is

$$\frac{p(Y|H0)}{p(Y|H1)} \begin{cases} \geq \theta, & \text{accept Ho} \\ < \theta, & \text{accept H1} \end{cases} \tag{2.1}$$

where p(Y|Hi), i=0,1, is the probability density function for the hypothesis Hi. For the equation 2.1, 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.

However, a speaker can produce multiple recordings, each distinctly different, because there are many factors that can affect the quality of recorded voices. The factors, for example, can be channel effects or the change of speaker's voice due to age. To solve the problem, researchers developed JFA [23] and i-vector [13] and consequently improved the accuracy of speaker verification systems. Both JFA and i-vector were based on GMM-UBM technology but calculated the channel effects into GMM by different means. Dehak et al. [13] achieved

a equal error rate (EER) of 1.12% by using the male English trials of the core condition of the NIST 2008 Speaker Recognition Evaluation dataset.

2.3.3 Attacks

Ratha et al. [35] identified vulnerabilities of biometrics-based authentication systems. Based on Ratha's research, Wu et al. [49] discussed spoofing attacks of speaker verification systems. Figure 2.4. shows the attack points of speaker verification systems.

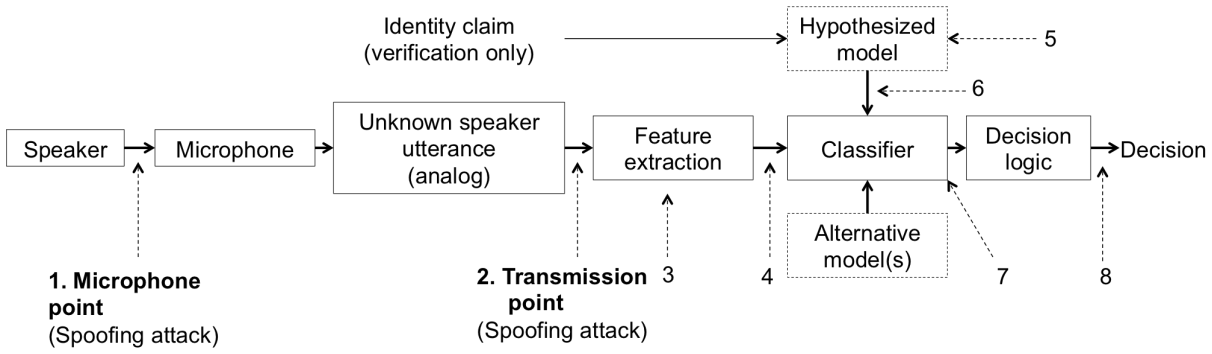


Figure 2.4: Speaker verification system attack points. [49]

Spoofing attacks, also called direct attacks, are labeled as attack points 1 and 2 in Figure 2.4. These attacks utilized different techniques to mimic or modify the input voice as the target user's voice and thus deceived the system. In other words, the purpose of direct attacks was to make the input voice have a similar voice model as the target user's. The attack points 3 to 8 labeled in Figure 2.4 were categorized as indirect attacks. These 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 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. Both synthesis and converted voice attacks need a target speaker's speech to train the systems. After training, the synthesis-voice system can generate speech with a target speaker's voice from text, and the converted-voice system will transform the voiceprint of input speech to target speaker's voiceprint.

2.3.4 Countermeasures

Researchers have shown that direct attacks can significantly impact the accuracy of speaker verification systems. A number of these studies were dedicated to helping the system resist direct attacks. Wu et al. [49] published countermeasures for replay attacks, synthesis voice attacks, and converted voice attacks. The study, however, 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 [39]. 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. Another method to counter a replay attack is to examine the channel noises of an incoming recording [47]. The replay recording incurs additional noises from the recording device and loudspeaker, while a benign recording does not. This difference gives the system a means to distinguish replay attacks from real speakers.

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-ceptral cepstral coefficients. Since synthetic voices had more variance in Mel-ceptral, the countermeasures thus could discriminate between natural and synthetic voices [38, 11].

Although the countermeasures were reported to be effective, the efficiency of countermea-

sures was questionable without a standard dataset, protocols, and metrics for testing. The ASV Spoofing and Countermeasures (ASVspoof) initiative was created to solve this problem [48]. The study post-evaluated several spoofing countermeasures reported in the literature and found the average detection equal error rate (EER) for synthesis and converted voice attacks was less than 0.3%. This result shows that speaker verification systems now have robust countermeasures to resist against synthesis and converted voice attacks. In the meantime, countermeasures for replay attacks do not provide strong protection to speaker-verification systems. The ASVSpoof 2017 challenge evaluated countermeasures for replay attacks and achieved the best average detection EER of 6.78% [26].

Chapter 3

TWO-LAYER AUTHENTICATION METHOD WITH VOICEPRINT

This study proposes to use voiceprint to identify and verify users in order to both secure and maintain the usability of voice assistants. Text-dependent voiceprint recognition, however, is vulnerable to replay attacks. In order to protect against replay attacks in our authentication method, this study integrates a challenge-response protocol to confirm the presence of the authorized user.

This chapter describes the details of the two-layer authentication method. Section 3.1 illustrates the assumptions and limitations of the method. Section 3.2 demonstrates the authentication method and the process. Section 3.3 shows the system implementation.

3.1 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 replayed 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), can be utilized to measure the authentication accuracy since it means the rate of 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 $N \to \infty$, the FMR(N) will close to one.

$$FMR(N) = 1.0 - [1.0 - FMR(1)]^N \tag{3.1}$$

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. It ensures that attackers cannot record the user's voice through any connected device. The second and third assumptions ensure that attackers cannot create the user's voiceprint by recording the user's voice.

3.2 Authentication method

The proposed two-layer authentication method includes two processes: one is for enrollment, and the other is for user authentication. The subsections illustrate the attack model of the two-layer authentication and explain these two processes.

3.2.1 Attack model

The attack model for the two-layer authentication is depicted in Figure 3.1. Steps 1 and 2 represent that the attacker is triggering the voice assistant. In Steps 3 and 4 the attacker speaks a service command (e.g., "transfer money") to the voice assistant and the voice assistant requires the user to send a voice-passphrase response. Step 5 indicates that the

attacker is launching a replay attack. Step 6 shows that the authentication system is verifying the speaker. Once the pre-recorded passphrase successfully passes the verification, the attack has broken the protection and the voice assistant will execute the service command in Step 9.

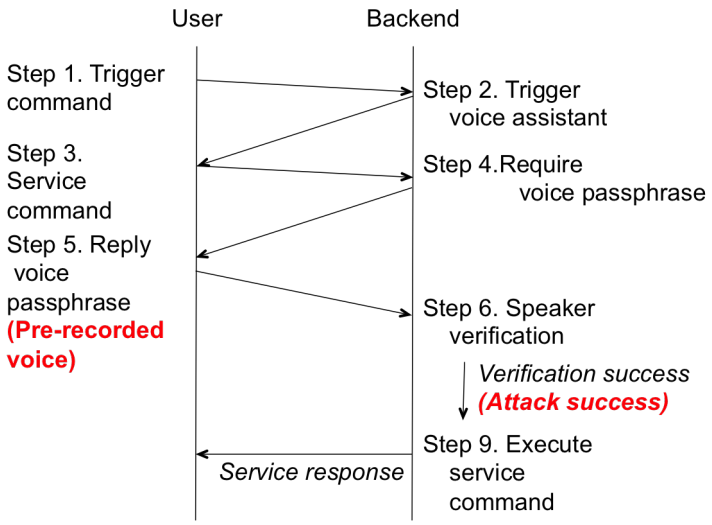


Figure 3.1: The attack model for the two-layer authentication.

3.2.2 Enrollment process

Before a user works with the system, she has to first register her voice. The enrollment process is depicted in Figure 3.2. During enrollment, the user is asked to read a sentence composed of digits during Step 1, and the system records the user's voice. After the user finishes her speaking, the recording is sent to the backend. In Step 2, the automatic speaker verification (ASV) system obtains the recording and extracts the user's voice model (i.e., voiceprint). In Step 3, the ASV system saves the user's voice model in the database.

After these three steps, the user successfully enrolls into the system and can use her voice to authenticate voice commands in the voice assistant.

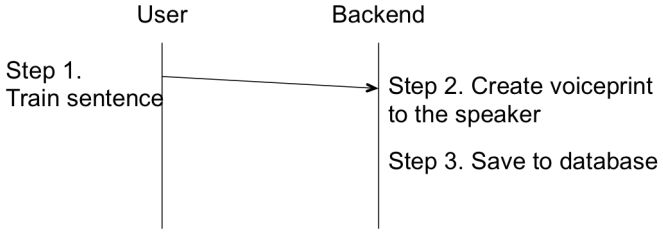


Figure 3.2: The process of the two-layer authentication enrollment.

3.2.3 Authentication process

The authentication process will be explored from two different perspectives: the user and the backend. Figure 3.3 depicts the process of the two-layer authentication method.

From the user perspective, the authentication process contains three actions: 1) trigger voice assistant; 2) speak service command, and 3) reply challenge text. The three steps are respectively shown as Steps 1, 3, and 6 in Figure 3.3. The design of the user flow adds a step in the end, in order not to interrupt the original user flow of voice assistants. Step 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 out the keywords, such as "OK Google" or "Alexa", to trigger the voice assistant service. When the voice assistant is listening, speakers can go to Step 3 in the user flow. In Step 3, users can say a service command, such as "buy me a cup of coffee," to make the voice assistant know his demand. Users then might receive a series of random numbers (challenge). Step 6 shows that if a user receives the challenge, he needs to repeat the numbers within 5 seconds and then wait for the system to execute or reject the demand.

From the backend perspective, the process contains six actions shown in Figure 3.3. Step 2 shows that the system triggers the voice assistant when it receives the trigger command and will give a response (i.e., a sound) to let the speaker know it is working. While a user is saying a service command, the backend is recording his voice.

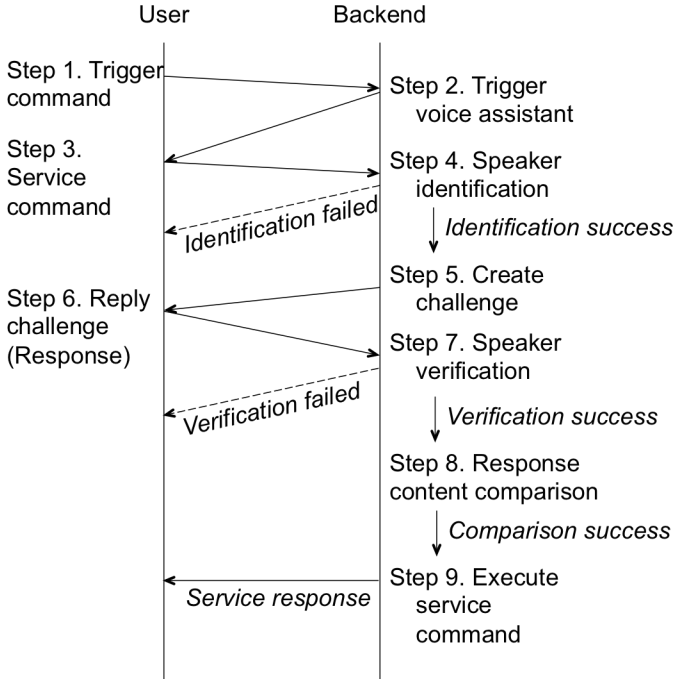


Figure 3.3: Two layer authentication flow.

In Step 4, the backend sends the recording to a speaker verification system after the user finishes the service command. The speaker verification system will then determine if the voiceprint of the recording matches a registered user. If the spekaer verification system finds a matched model, 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.

Step 5 shows that if the identification is successful the backend will generate a series of random numbers to the user. The random numbers are a challenge for the user in the authentication method. The random numbers can be shown in two ways: text-to-speech (TTS) and text on the screen. After prompting the speaker with the challenge, the backend starts to record voices for five seconds. This recording will be considered as the response to the challenge.

In Step 7, the backend sends the recording to the speaker verification system to identify

the speaker. The system checks 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 responsible for comparing the text of the challenge and the response. In this step, 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. On the other hand, if the response text is identical with the challenge, the backend ensures that the speaker is a benign user and will execute his 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 in Chapter 2.1.2 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. Attackers can simulate a valid user's voice and generate the response by voice synthesis and voice conversion libraries. However, speaker verification systems have efficient countermeasures to deal with these attacks. The equal error rate (EER) of speaker verification systems for synthesis and converted voice attacks is less than 0.3%; thus, it is difficult to break the challenge-response protocol through synthesis and converted voices.

3.3 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 smart phones and the Android platform has the most users in the market, this study implemented an Android application. The application applies Microsoft Azure speaker recognition API and Google cloud speech API to accomplish speaker recognition and speech recognition. The application uses MS Azure speaker recognition API due to its accuracy. As it is a commercial service provided by a company with a good reputation, we are reasonably confident about its accuracy. For the speech recognition library, this study uses Google cloud speech API, which has the lowest words error rate (WER) among libraries [6]. The Google cloud speech API has a reported 9% WER, while Microsoft API has 18% WER and CMU Sphinx has 37% WER. This shows that Google cloud speech API is the most accurate speech recognition system.

Using remote systems causes security issues, since attackers can get the data of the user's voice by eavesdropping. We considered using ALIZE, a speaker verification system that can be installed on the Android platform, but did not get satisfactory results and therefore chose the MS library. For speech recognition, we also considered the local speech recognition system in the Android phone, but the APIs do not support a pre-recorded voice input; accordingly, this study turns to the remote server solution.

Figure 3.4 shows the system design of the application. 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. In this project, the speaker verification system source is the MS Azure speaker recognition system. The enrollment job is called when the user wants to register to the system. The job records the user's voice and sends it to the verification speaker system to assist the user to enroll in the system. The identification job is responsible for sending the recording to the speaker verification system. The speaker verification system will identify which enrolled user the voice belongs to and return the result to our system. Similarly, the verification job sends the recording to the speaker verification system in order to verify that the voice belongs a specific user.

The speech recognition manager handles speech recognition and speech comparison jobs. The speech recognition job sends the recording of the response to a speech recognition library, which in our application is the Google speech recognition API. The result from the speech recognition library is the text of the recording. The speech comparison is responsible for comparing the text of the recording (response) and the challenge.

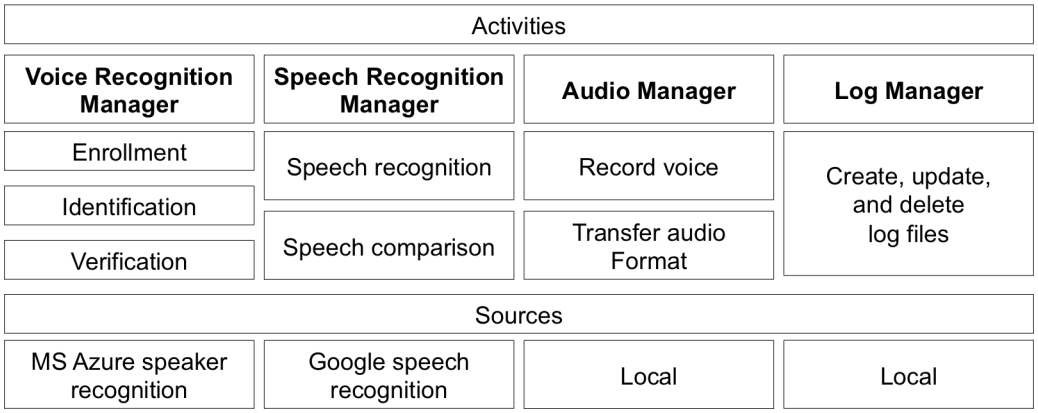


Figure 3.4: System design of the two-layer authentication application.

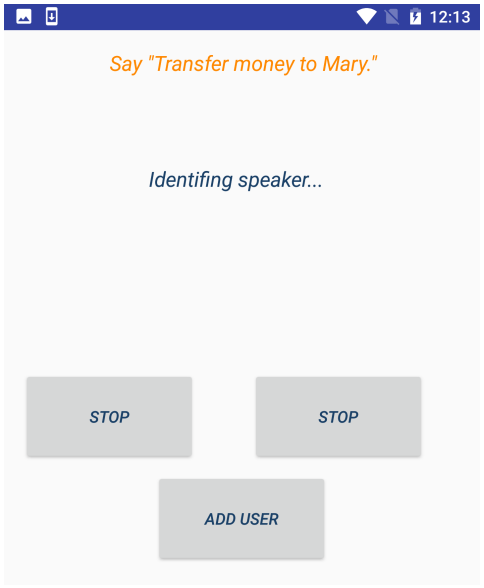
The audio manager is responsible for recording voices and transferring audio file format jobs. The recording voice job accesses the smart phone's microphone to start and stop voice recording. Since the recording-voice job saves raw audio data, the transferring-audio-format job will help to transfer the raw audio data to a different audio format (e.g., wav).

The log manager is created for usability experiments. It handles create, update, and delete log file jobs. The log files record the date, participant id, the ids from identification and verification results, and the result of the challenge and response comparison. Again, these log files are only for usability experiments and should not be used in commercial products.

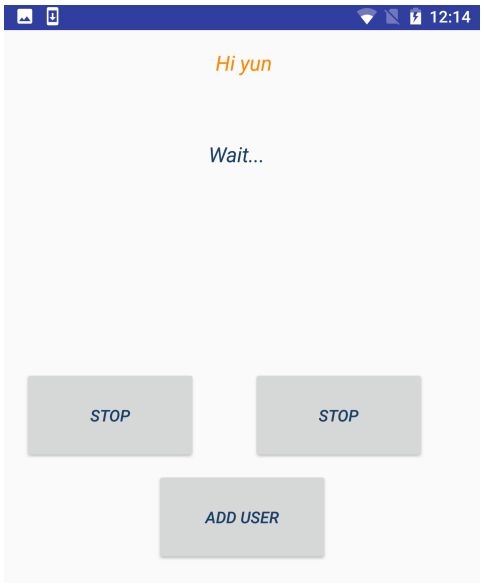
Finally, the application was deployed to a Huawei Nexus 6P phone running Android 8.0.0 on a Snapdragon 810 processor that provides 2.0GHz octa-core and 64-bit computing power. The decision to deploy on Android 8.0.0 was made because, as of this writing, it is the newest version of Android and will become 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.

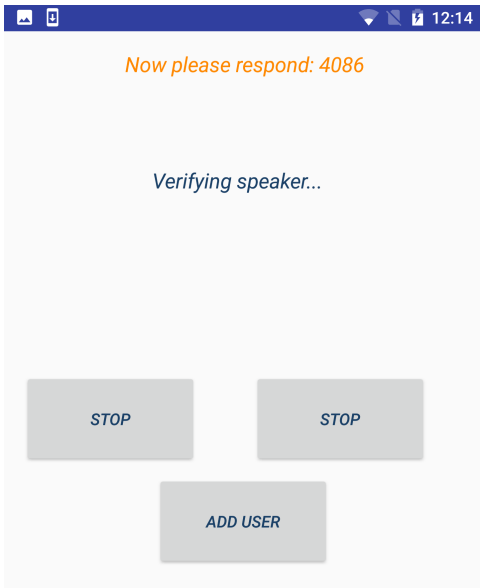
Figure 3.5 shows the screenshots of the application. Figure 3.5a, which corresponds to Step 4 in Figure 3.3, is the screenshot taken while the system was identifying the speaker after he said the "Transfer money to Mary" service command. Figure 3.5b can be mapped to Step 5 in Figure 3.3. At this point in the process, the system has identified the speaker and is creating a challenge. Figure 3.5c represents Steps 7 and 8 in Figure 3.3. The system has received the speaker's response and is verifying the speaker by the voiceprint and the content of the response. Figure 3.5d is the view taken when the user successfully passes the two-layer authentication.



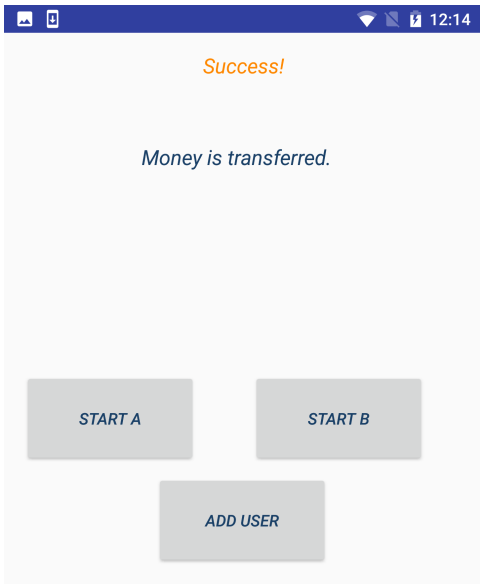
(a) Identifying the speaker after receiving service command



(b) Identified the speaker and generating random numbers



(c) Verifying the speaker after receiving response



(d) Authentication succeed

Figure 3.5: Application screenshots

Chapter 4
USABILITY EXPERIMENT

To evaluate the proposed authentication method in Chapter 3, the project implemented an Android application to simulate the process. We also conducted an experiment to prove the method cannot only provide protection to voice assistant but also maintain usability of voice assistants. Section 4.1 illustrates the details of the experiment and goals. Section 4.2 discusses the design of the usability questionnaire.

4.1 Experiment design

The main goal of the experiment is to understand how users feel and think about the two-layer authentication. This experiment process is depicted in Table 4.1. It contains 10 steps and requires 30 minutes to complete.

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. The background questionnaire is referenced in Appendix A.

Steps 2 and 3 are designed to confirm that the participant has basic knowledge and experience with voice assistants. During 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 use the Google voice assistant to complete four tasks. In each task, the participant needs to give a service command to interact with the Google voice assistant. In Step 3.1, the participant will fill out a questionnaire to evaluate

the usability of the voice assistant. The usability questionnaire shown in Appendix B will be explained in the next section. After the participant has acquired the basic knowledge and experience of the voice assistant, this experiment moves to the next step.

In Step 4, the participant is shown two voice assistant attacks via video: 1) a badvoice [52] and 2) a dolphin attack [53]. The two attacks are detailed in chapter 2.1.2. Step 4 assures that the participant understands some of the security issues related to voice assistants. In Step 4.1, the participant will fill out the usability questionnaire.

In Step 5, two protections will be introduced 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.

After the introduction and demonstration of the two protections, Step 6 asks the participant to recite a series of numbers for system enrollment. After the participant has successfully completed this step, the participant is ready to use the two protections.

In Step 7, the participant starts to experience the two protections. There are two factors that can effect the participant's opinion on the usability of a protection: 1) first impression [25] and 2) habit [29]. For first impression, the participant might prefer the protection that he starts with. Therefore, this experiment separates the participants into two groups. Participants in group 1 will start with protection A and those in group 2 will start with protection B. Thus, this experiment controls for the influence of the first impression. 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).

Step 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

Table 4.1: Experiment process

Step	Assignment
1.	Fill out the background part.
2.	Introduction of abilities of voice assistants.
3.	Let the participant use voice assistant. Task1: Transfer money. Task2: Buy me a pizza. Task3: Read my email. Task4: Create a meeting at 9 oclock on Friday morning.
3.1.	Fill out the questionnaire (QNR1).
4.	Introduction of security of voice assistants.
4.1.	Fill out the questionnaire (QNR2).
5.	Introduction of using the 2 protections. A. voice assistant can only identify the speaker B. voice assistant with our authentication method
6.	Register voice authentication.
7.	Alternatively using the 2 protections. Group 1: ABABABAB Group 2: BABABABA
8.	Using protection A or B.
8.1.	Fill out the questionnaire (QNR3 or QNR4).
9.	Using protection B or A.
9.1.	Fill out the questionnaire (QNR4 or QNR3).
10.	Interview.

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?

The questions are designed to learn more about the participants perspective about information security and privacy, their feelings about the protections, and their opinions when using different biometrics to protect voice assistants.

4.2 Usability questionnaire design

The design of the usability questionnaire is based on Lund's USE questionnaire [31]. It is one of the more commonly used usability questionnaires in usability research. This study deletes the questions in the questionnaire that are not suitable for our scenario. For instance, "It is flexible" is deleted since authentication does not provide flexibility in this experiment.

On the other hand, this study adds questions to the questionnaire in order to understand if the authentication method can increase the users sense of security, such as I feel good using it. To summarize the design of the questionnaire, questions 1 to 10 are related to usability and questions 11 to 14 are designed to evaluate the confidence of security.

During the experiment, the participant is asked to fill out the usability questionnaire four times (Step 3, 4, 8, and 9). The four questionnaires use the same scale to evaluate different experiences. As shown in Table 4.1, QNR1 is the questionnaire that the participant fills out

after the introduction of the voice assistant's abilities and having had some experience using the Google voice assistant. QNR2 is collected immediately after the participant knows about some of the vulnerabilities related to voice assistants. QNR3 and QNR4 are filled out after the participant alternatively tested the two protections; QNR3 is for one-layer protection and QNR4 is for the proposed mechanism. The result of QNR1 can be the base data to compare QNR2, QNR3, and QNR4 and can evaluate how participants change their opinion on usability and security and privacy after knowing about some of the vulnerabilities related to voice assistants and applying different protections.

Chapter 5
RESULT EVALUATION

In this chapter, we detail and discuss the experiment results. In section 5.1, we discuss the participants' background. In section 5.2, we introduce the methods used to analyze usability. In sections 5.3 through 5.8, we detail the different results of analysis. Section 5.9 presents our discussion of system accuracy. In section 5.10, we present and interpret the participant interviews.

5.1 Participants background

In this section, we analyze the background questionnaire of the participants to understand the population in our experiment. Section 5.1.1 demonstrates the population details. Section 5.1.2 discusses the differences between groups.

5.1.1 General analysis

The average experiment time was 30 minutes; 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. Of the participants, 56.1% (23 of 41) are STEM majors. In addition, 53.7% (22 of 41) of the participants had never used voice assistants or only used them a few times before the experiment.

In the background investigation questionnaire, we used a 7-point Likert scale. The scale ranges from one to seven and represents never to very often or strongly disagree to strongly agree. Of the participants, 56.1% (23 of 41) had never used speaker verification systems or rarely used them. Only 4.9% (2 of 41) of the participants said they were never worried

about the security of their information and privacy; the others are distributed almost equally among the options. For knowing the privacy risks and security threats of voice assistants, 24.4% (10) of the participants chose 4 (Neither agree or disagree) and others are distributed in different values. For understanding the different approaches to protect user, 56.1% (23) of the participants chose values from 1 (Strongly disagree) to 3 (Somewhat disagree).

In summary, most participants were younger than 30 and approximately half of them were in a STEM program. Further, slightly over half of them had little or no experience with voice assistants and speaker verification systems before the experiment.

Overall, the participants concern over security was not extreme: they were neither extremely skeptical nor overly confident about security. In addition, their knowledge about the privacy risks and security threats were also moderate: they were neither extremely illiterate nor expert about risks and threats. Finally, half of the participants did not understand the different technical approaches that might be used to increase the level of security of a user of a system.

5.1.2 Group analysis

The participants were separated into 4 groups. Table 5.1 shows the details of the 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 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.

We examined the differences of the background questionnaire between groups by independent-samples t-test. The results indicate that there is no significant difference of the background

questionnaire between any two groups, i.e., the four groups have participants with similar backgrounds based on our investigation.

Table 5.1: Groups of participants.

Group	Reveal security info.	Questionnaire	Order	# ofparticipants
G1	Yes	QNR 1,2,3,4	AB	11
G2	Yes	QNR 1,2,3,4	BA	10
G3	No	QNR 1,2,3,4	AB	10
G4	No	QNR 1,2,3,4	BA	10
QNR 1,2,3, and 4 use the same scale to evaluate different experiences as follows: QNR1: voice assistant, QNR2: voice assistant after knowing security information, QNR3: 1-layer protection, QNR4: 2-layer protection.				

5.2 Analysis method

To evaluate the questionnaire, this study uses the independent-samples t-test and the paired-samples t-test to evaluate the mean differences [22]. In our questionnaire, we separated the questions into two types:

1. Usability: Questions related to usability (Q1-Q10) in the questionnaire are included in this type.

2. Sense of security (Security): Questions related to the sense of security (Q11-Q14) in the questionnaire are included in this type.

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 is from 1 to 7 since our Likert scale is from 1 to 7.

This study uses IBM SPSS Subscription (June, 2018) to perform statistical analyses. IBM SPSS provides the 2-tailed t-test calculation, and we use it to compare the usability and security means of groups and the questionnaires completed at various points during the experiment.

5.3 General analysis

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

5.3.1 Usability

Table 5.2 shows the results of the paired-samples t-test for usability. Under 95% confidence, the significance-value is 0.05. Since pair 2 and pair 3 have significance-values that are less than 0.05, we can be confident that the participants have different opinions when comparing 2-layer protection with voice assistants versus 1-layer protection. Figure 5.1 shows the differences between QNR1 to QNR4 and QNR3 to QNR4. For QNR1 to QNR4, the mean drops from 5.47 to 4.89. For QNR3 to QNR4, the mean decreases from 5.43 to 4.89. The t-test results and the means indicate that the 2-layer protection has less usability than voice assistant and 1-layer protection.

Table 5.2: Paired-samples t-test results for general analysis: usability

Usability (general)		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	**QNR1 - QNR3**	0.04146	1.06536	0.16638	-0.29481	0.37773	0.249	40	0.804
Pair 2	**QNR1 QNR4**	0.58293	1.28858	0.20124	0.17620	0.98965	2.897	40	0.006
Pair 3	**QNR3 QNR4**	0.54146	1.01439	0.15842	0.22128	0.86164	3.418	40	0.001

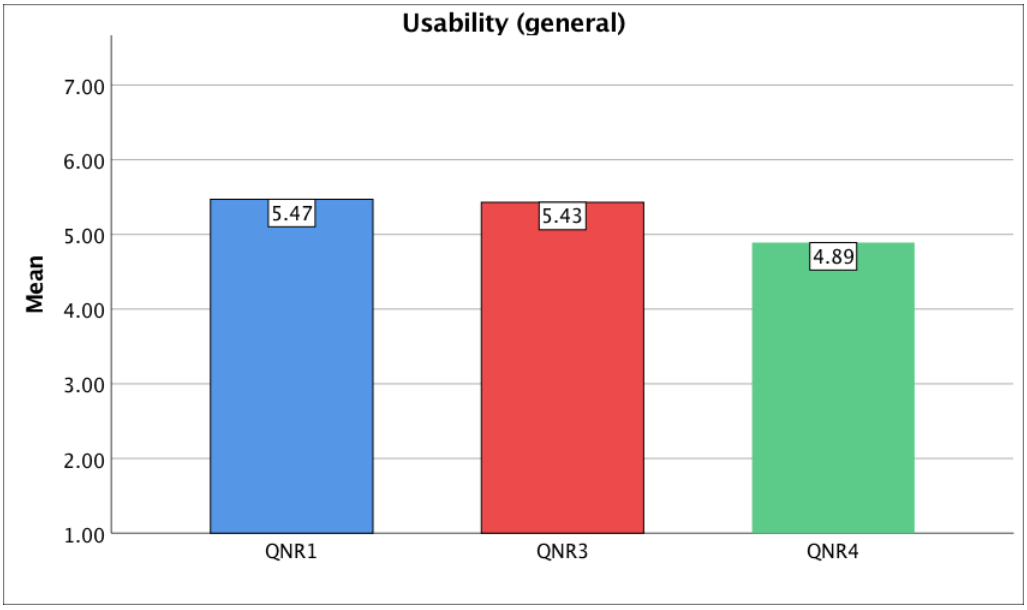


Figure 5.1: Usability

5.3.2 Security

Table 5.3 shows the paired-samples t-test results of security means comparison. For Pair 3, the means-difference of QNR3 and QNR4 is statistically significant since its significance-

value is less than 0.05. Figure 5.2 depicts the security means of the three questionnaires. We can see that the mean of QNR3 is smaller than QNR4 in Figure 5.3. Based on the t-test results, we are confident that the 2-layer protection provides a higher sense of security than the 1-layer protection.

Table 5.3: Paired-samples t-test results for general analysis: security

Security (general)		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	**QNR1 - QNR3**	0.12195	1.26999	0.19834	-0.27891	0.52281	0.615	40	0.542
Pair 2	**QNR1 - QNR4**	-0.20122	1.31840	0.20590	-0.61736	0.21492	-0.977	40	0.334
Pair 3	**QNR3 - QNR4**	-0.32317	0.86285	0.13475	-0.59552	-0.05082	-2.398	40	0.021

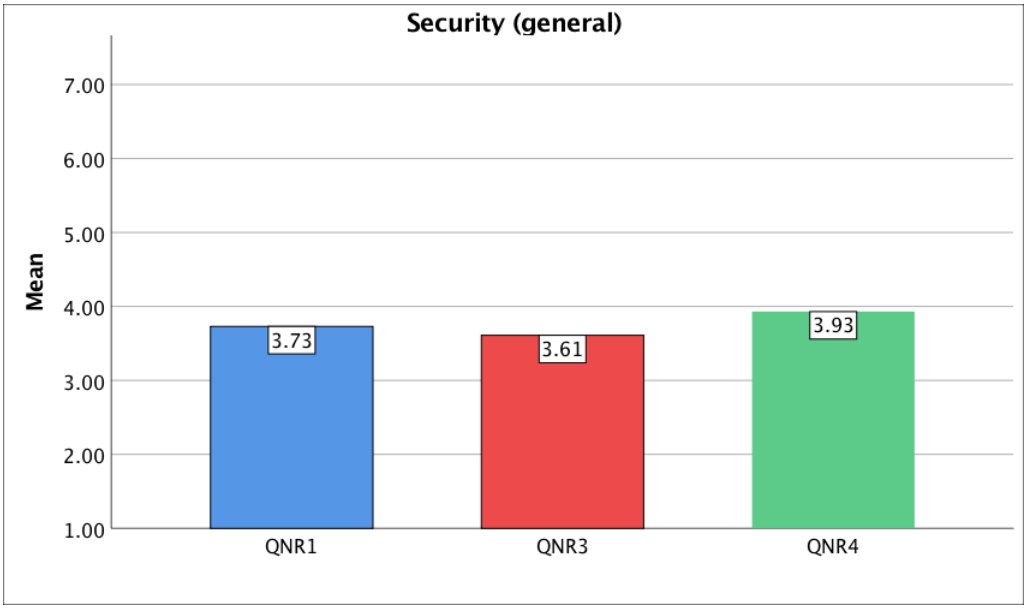


Figure 5.2: security

5.4 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.

Table 5.4 shows the results of the paired-samples t-test for usability and security. We can see that the usability mean and the security mean of QNR1 and QRR2 are significantly different since the significance-values are less than 0.05. This indicates that after the participants watched the attack videos, they changed their opinions about voice assistants in terms of usability and security. We can observe this tendency in Figure 5.3. With the figure and the t-test results, we are confident that 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. This may be because the attack videos, by increasing security concerns, made the participants feel they do not need to use voice assistants.

Table 5.4: Paired-Samples t-test results for revealing the security information

Revealing security information (G1G2)		**Paired Differences**					**t**	**df**	**Sig. (2-tailed)**
		Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval of the Difference**				
					Lower	**Upper**			
Usability	**QNR1 - QNR2**	0.58571	0.58076	0.12673	0.32135	0.85007	4.622	20	0.000
Security	**QNR1 - QNR2**	1.25000	1.16994	0.25530	0.71745	1.78255	4.896	20	0.000

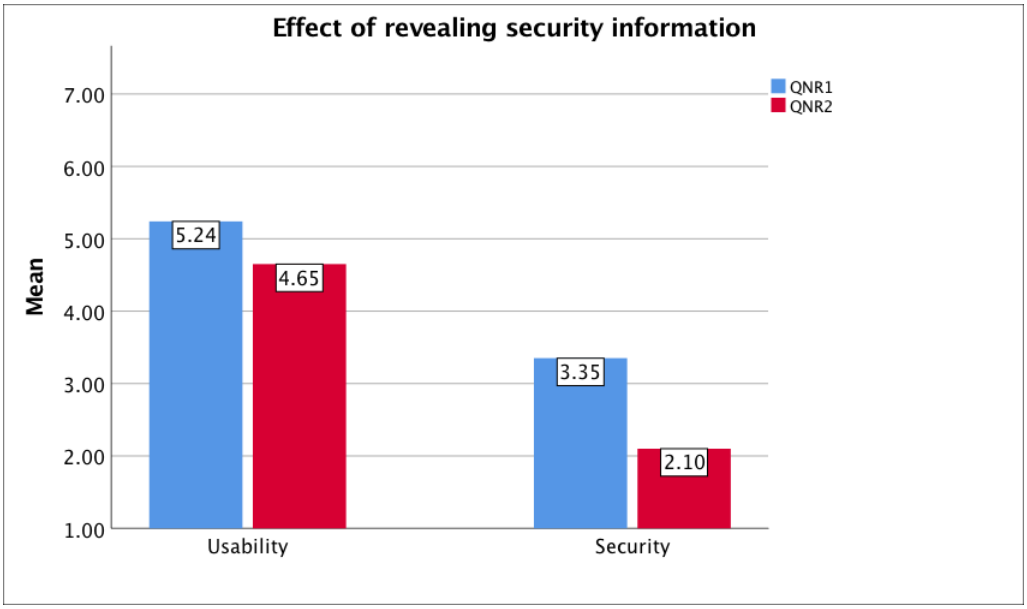


Figure 5.3: Effect of revealing security information

5.5 Participants without security information

In section 5.4, 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) was shown and evaluates the two groups' respective opinions.

In this section, we focus on the group without security information. QNR1, QNR3, and QNR4 from G3 and G4 are used in the analysis.

5.5.1 Usability

Table 5.5 shows the paired-samples t-test results of the opinions about usability of the participants without security information. Since the significance-value of Pair 3 is less than 0.05, the difference in the usability means between the 1-layer protection and 2-layer protection is significant. Figure 5.4 indicates that the usability mean of the 2-layer protection is smaller than that of the 1-layer protection. With the t-test result, we are therefore confident that the par-

ticipants believe that the 2-layer protection is less useable, compared to the 1-layer protection.

Table 5.5: Paired-samples t-test results for participants without security information: usability

Usability (G3G4)		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	**QNR1 - QNR3**	-0.03000	0.99372	0.22220	-0.49507	0.43507	-0.135	19	0.894
Pair 2	**QNR1 QNR4**	0.25500	0.93891	0.20995	-0.18442	0.69442	1.215	19	0.239
Pair 3	**QNR3 QNR4**	0.28500	0.50604	0.11315	0.04816	0.52184	2.519	19	0.021

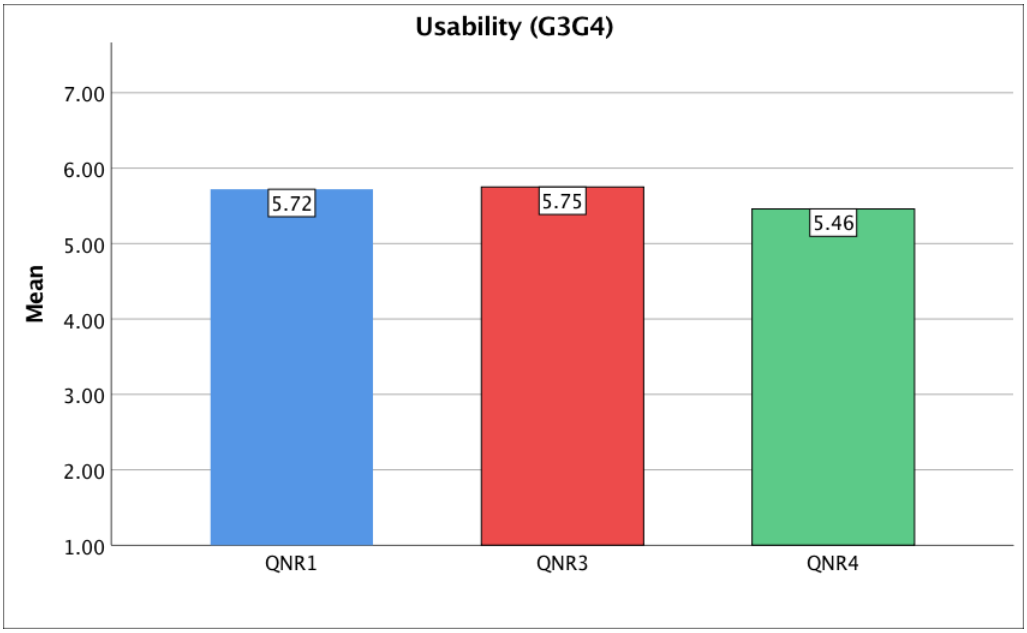


Figure 5.4: Usability mean of participants without security information.

5.5.2 Security

Table 5.6 shows the paired-samples t-test results of security means. The significance-value of Pair 3 is less than 0.05. This indicates that the participants without security information have different opinions when comparing the usability of the 1-layer protection and the 2-layer protection. Figure 5.5 displays the security means of QNR1, 3, and 4. With the t-test results and the mean tendency, we can be confident that the participants without security information feel the 2-layer protection is more secure than the 1-layer protection.

Table 5.6: Paired-samples t-test results for participants without security information: security

Security (G3G4)		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	**QNR1 - QNR3**	0.33750	1.34574	0.30092	-0.29233	0.96733	1.122	19	0.276
Pair 2	**QNR1 - QNR4**	-0.28750	1.30856	0.29260	-0.89993	0.32493	-0.983	19	0.338
Pair 3	**QNR3 - QNR4**	-0.62500	0.83705	0.18717	-1.01675	-0.23325	-3.339	19	0.003

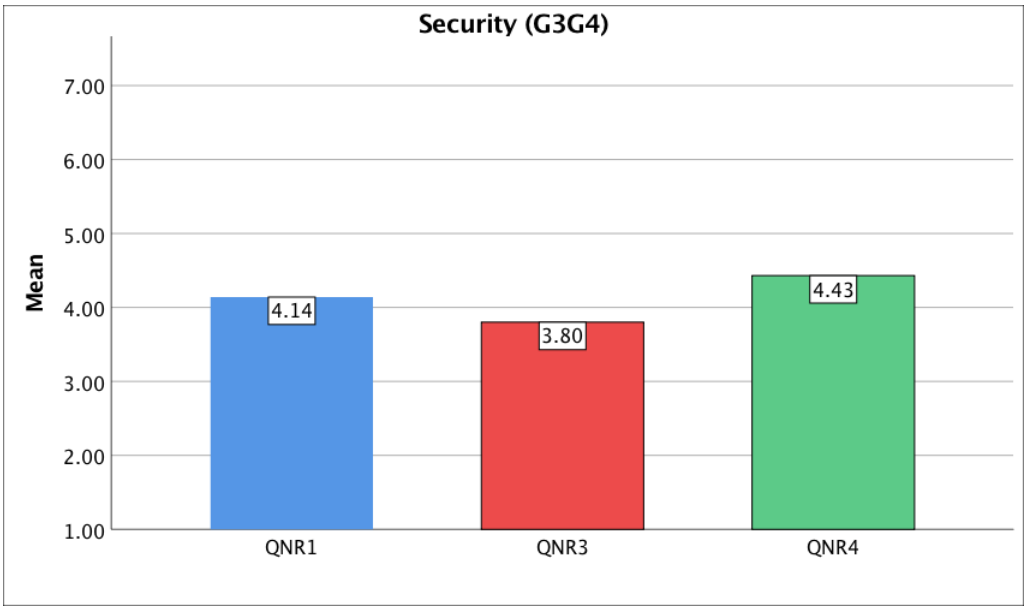


Figure 5.5: Security mean of participants without security information.

5.6 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.

5.6.1 Usability

Table 5.7 shows the paired-samples t-test results of the usability mean comparisons for the participants with security information. As we see, pairs 2 and 5 have a significance-value less than 0.05. This indicates 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.

Figure 5.6 shows that before the participants learn about the security information, they feel that 2-layer protection decreases the usability to the voice assistant (Pair 2 in Table 5.7). When comparing the usability of the 1-layer protection and the 2-layer protection (Pair 5 in Table 5.7), we can see that the participants feel 1-layer protection is more usable than 2-layer protection.

5.6.2 Security

Table 5.8 shows the paired-samples t-test results of the security mean comparisons. Pair 3 and Pair 4 have significance-values less than 0.05. This indicates that the participants with security information have a different sense of security for the 1-layer protection and the 2-layer protection. Figure 5.7 displays the security means. With the t-test results and the means, we are confident that both the 1-layer protection and the 2-layer protection increase the sense of security after the participants know the security information.

Table 5.7: Paired-samples t-test results for participants with revealing the security information: usability

Usability (G1G2)		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	**QNR1 - QNR3**	0.10952	1.14974	0.25089	-0.41383	0.63288	0.437	20	0.667
Pair 2	**QNR1 - QNR4**	0.89524	1.50781	0.32903	0.20889	1.58158	2.721	20	0.013
Pair 3	**QNR2 - QNR3**	-0.47619	1.19118	0.25994	-1.01841	0.06603	-1.832	20	0.082
Pair 4	**QNR2 - QNR4**	0.30952	1.68579	0.36787	-0.45784	1.07689	0.841	20	0.410
Pair 5	**QNR3 - QNR4**	0.78571	1.29857	0.28337	0.19461	1.37682	2.773	20	0.012

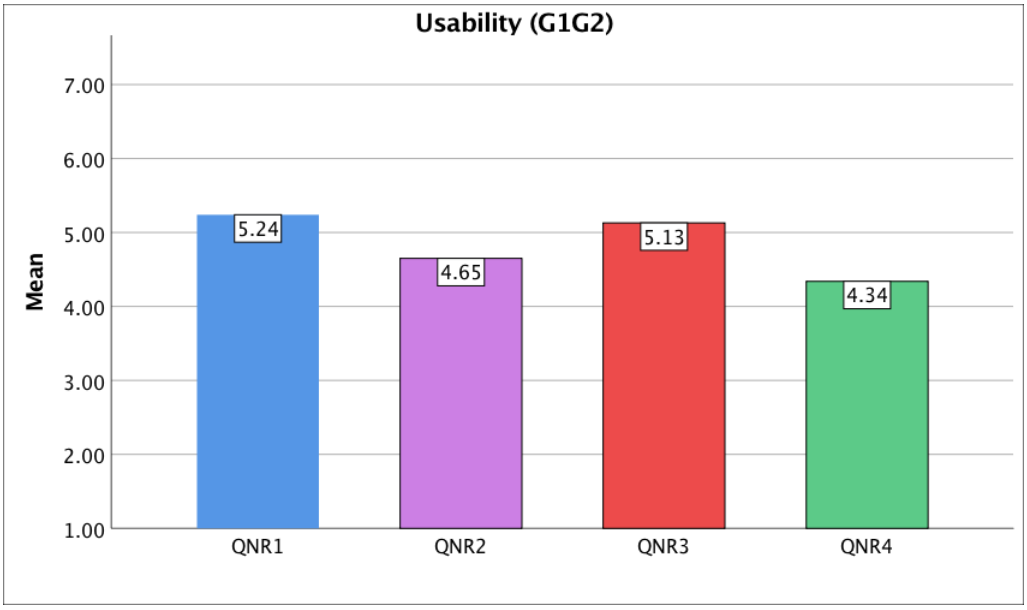


Figure 5.6: Usability mean of participants with security information.

Table 5.8: Paired-samples t-test results for participants with revealing the security information: security

Security (G1G2)		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	**QNR1 - QNR3**	-0.08333	1.18936	0.25954	-0.62472	0.45806	-0.321	20	0.751
Pair 2	**QNR1 - QNR4**	-0.11905	1.35467	0.29561	-0.73568	0.49759	-0.403	20	0.691
Pair 3	**QNR2 - QNR3**	-1.33333	0.87440	0.19081	-1.73136	-0.93531	-6.988	20	0.000
Pair 4	**QNR2 - QNR4**	-1.36905	1.22632	0.26761	-1.92726	-0.81083	-5.116	20	0.000
Pair 5	**QNR3 - QNR4**	-0.03571	0.80345	0.17533	-0.40144	0.33001	-0.204	20	0.841

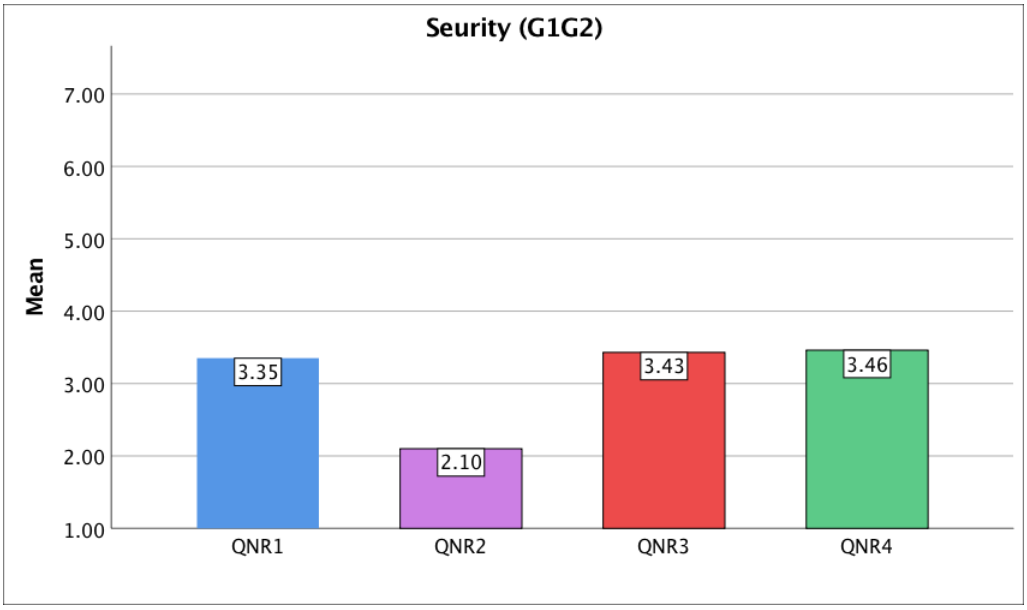


Figure 5.7: Security mean of participants with security information.

5.7 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.

5.7.1 Usability

Table 5.9 shows the independent-samples t-test results of usability mean comparisons between participants with and without security information. QNR4 has lower significance-values than 0.05, which indicates that the participants with and without security information have significantly different opinions about 2-layer protections.

Figure 5.8 displays the means of QNR1, 3, 4 of G1G2 and G3G4. The figure shows that the 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, i.e., G1G2 has a lower usability mean than G3G4. To examine this possibility, we use 1-tailed independent-samples t-test to assess the usability-mean differences of QNR1, 3, 4 between G1G2 and G3G4. We can use the t-values of Table 5.9 to deduce the results of 1-tailed t-test. Under 95% confidence with 39 degree of freedom, the critical value (CV) of 1-tailed t-test is 1.685. Comparing the absolute t-values in Table 5.9, QNR1, 3, and 4 have greater values than the CV, which indicates the differences are significant. Thus we are confident that the participants with security information have higher criteria on usability than the participants without security information.

Table 5.9: Independent-samples t-test results: usability comparisons of participants with and without security information.

Usability (G1G2 v.s. G3G4)		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% Confidence Interval of the Difference	
									Lower	Upper
QNR1	Equal variances assumed	0.001	0.975	-1.701	39	0.097	-0.47690	0.28033	-1.04392	0.09011
	Equal variances not assumed			-1.697	38.146	0.098	-0.47690	0.28103	-1.04575	0.09194
QNR3	Equal variances assumed	0.861	0.359	-1.710	39	0.095	-0.61643	0.36047	-1.34555	0.11269
	Equal variances not assumed			-1.706	38.162	0.096	-0.61643	0.36136	-1.34786	0.11500
QNR4	Equal variances assumed	0.017	0.898	-2.862	39	0.007	-1.11714	0.39030	-1.90660	-0.32769
	Equal variances not assumed			-2.864	38.960	0.007	-1.11714	0.39013	-1.90627	-0.32801

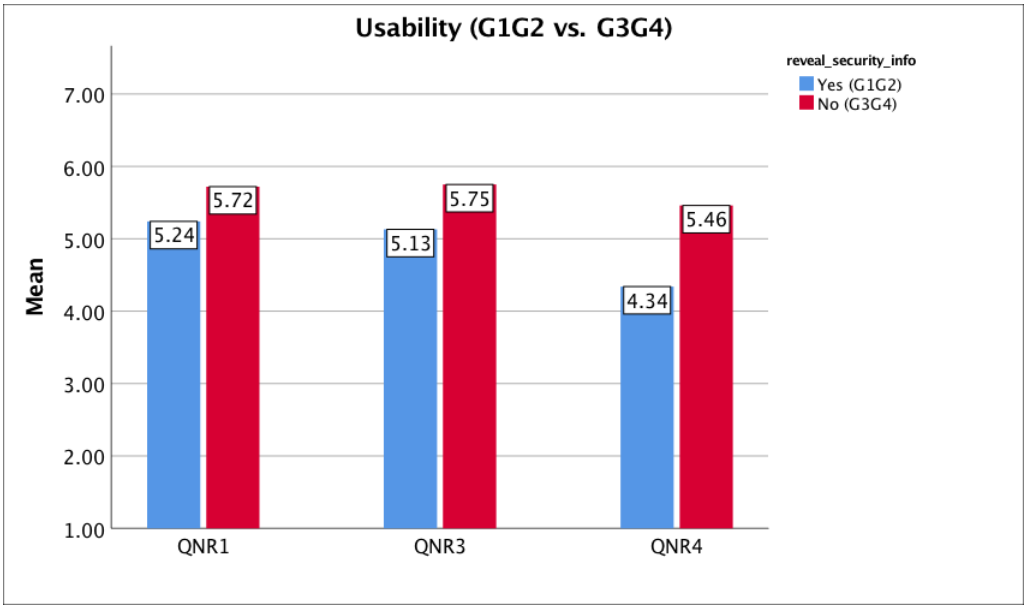


Figure 5.8: Usability mean of participants with and without security information.

5.7.2 Security

Table 5.10 indicates that the participants with security information do not have a significantly different opinion from those of the participants without security information. Although Figure 5.9 shows that the participants with security information perceive less security, the insignificant differences suggest that the participants with and without security information have a similar sense of security on voice assistant, 1-layer protection, and 2-layer protection.

Table 5.10: Independent-samples t-test results: security comparisons of participants with and without security information.

Security (G1G2 v.s. G3G4)		**Levene's Test for Equality of Variances**		**t-test for Equality of Means**						
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Diff.**	**Std. Error Diff.**	**95% Confidence Interval of the Difference**	
									Lower	**Upper**
QNR1	**Equal variances assumed**	0.447	0.508	-1.693	39	0.098	-0.79226	0.46798	-1.73883	0.15431
	Equal variances not assumed			-1.690	38.547	0.099	-0.79226	0.46866	-1.74057	0.15605
QNR3	**Equal variances assumed**	5.471	0.025	-0.787	39	0.436	-0.37143	0.47223	-1.32661	0.58375
	Equal variances not assumed			-0.780	33.965	0.441	-0.37143	0.47624	-1.33931	0.59645
QNR4	**Equal variances assumed**	3.194	0.082	-1.842	39	0.073	-0.96071	0.52155	-2.01565	0.09422
	Equal variances not assumed			-1.828	34.489	0.076	-0.96071	0.52567	-2.02844	0.10701

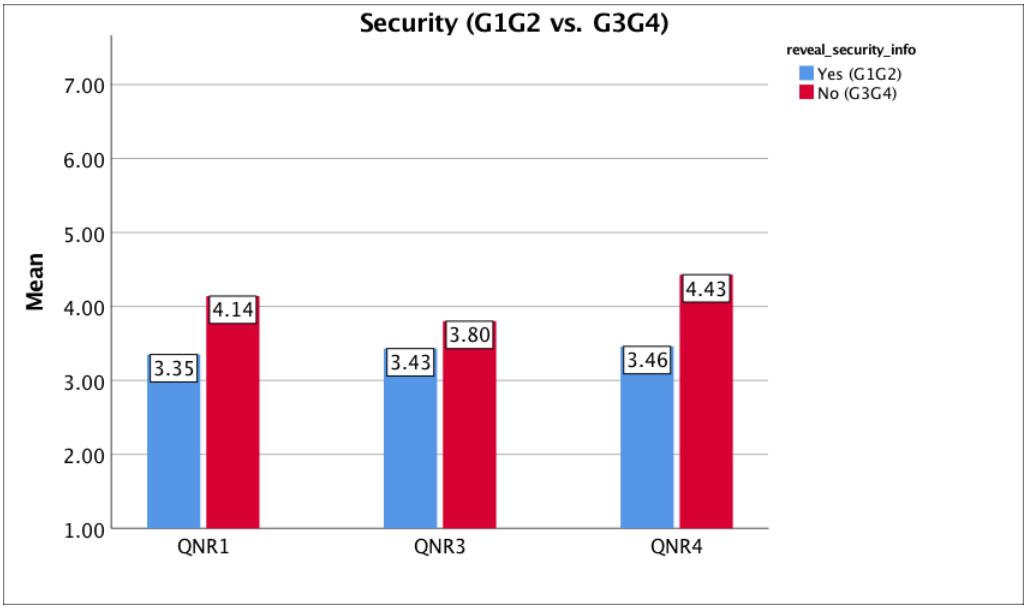


Figure 5.9: Security mean of participants with and without security information.

5.8 Summary

Table 5.11 depicts the summary of participants' opinions about 1-layer protection, 2-layer protection, and the comparison of 1-layer and 2-layer protection.

We can see that 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 increases.

For 2-layer protection, the participants perceive that the 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 perceive 2-layer protection as having the same usability with voice assistants and as being more secure.

Table 5.11: Summary of usability and security.

Summary		voice assistants	1-layer		2-layer		1-layer	2-layer	
			Usability	Security	Usability	Security		Usability	Security
whole			-	-	↓	-		↓	↑
without security info.			-	-	-	-		↓	↑
with security info.	Q1		-	-	↓	-		↓	-
	Q2		-	↑	-	↑			

Finally, all participants believe that 2-layer protection is less usable but is more secure than 1-layer protection. Taking a closer look, the participants without security information have the same opinions. However, the participants with security information perceive that 2-layer protection has less usability and the same sense of security as 1-layer protection.

In conclusion, when participants know the threats of voice assistants, they don't consider 2-layer protection to harm the usability of the voice assistant and also feel more secure.

5.9 Accuracy

In this section, we discuss the accuracy of the MS Azure speaker recognition system and our 2-layer protection.

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 rate (TAR) and false rejection rate (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 relatively low 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. Thus, the training speech is only around 5 seconds long. The MS Azure speaker recognition system suggests a 30-second training speech, without silence, for a high-quality voiceprint. 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.

5.10 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.

In section 5.7, we discussed that G1 and G2 have higher usability criteria than G3 and G4. The interviews also reflected this. 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 about 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, 3 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.

Chapter 6

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.

In this research, the two-layer authentication method has been implemented as a mobile application. A 30-minute experiment was used to evaluate the usability and security of this approach. The experiment compared the difference in usability between voice assistants with/without voice authentication. A total of 41 individuals participated in the experiment and evaluated the authentication with reference to the usability questionnaire. With the evaluation of the questionnaire, the results indicate that security information affects the participants' opinions about usability and the sense of security. With security information, the participants do not perceive that 2-layer protection decreases the usability and increases the sense of security. In addition, this study also collected data to evaluate the speaker-recognition system. The results show that the false rejection rate (FRR) of MS Azure speaker recognition system is 26.94% This FRR demonstrates that the MS azure speaker recognition system cannot provide sufficient authentication to protect voice assistants.

However, the major advantage of using voiceprint to authenticate speakers is that it does not require users to wear or install an additional token. Users are only required to enroll their voices into the system. They can then use the secured voice assistant as usual without worrying about missing the token. Thus, for all voice assistants, it is easy to integrate this method because it only inserts three steps before voice assistants execute a command. The first step utilizes the voice command to identify the speaker, the second step is creating a

challenge and prompting to the speaker, and the third step is using the voice response to verify the speaker. The three steps only need minor software modification and do not require any device collaboration (i.e., tokens).

6.1 Limitations

The limitations of this project include the following:

- The security of our method is based on the security of speaker recognition systems. In other words, the speaker recognition system's accuracy is related to our method. 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.

- This study utilizes 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.

- Similar to the devices that use voice input, the method needs a quiet environment to perform well (i.e., to have the best recognition result).

6.2 Future work

Work that can be done in the future includes reducing the speaker recognition library limitations. 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. Other future areas of research could be applying various countermeasures of speaker recognition system to the 2-layer protection. In this way, we could get stronger evidence that the two-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.

BIBLIOGRAPHY

[1] 49 Best Tech Gifts in 2018 For Men & Women - Top Tech Gift Ideas for Christmas. https://www.brostrick.com/tech/best-tech-gifts-this-year/.

[2] Siri can now send money via PayPal — TechCrunch. https://techcrunch.com/2016/11/10/siri-can-now-send-money-via-paypal/.

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

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

[5] 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.

[6] Gamal Bohouta and Veton Këpuska. Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). *Int. Journal of Engineering Research and Application*, 2248-9622:20–24, March 2017.

[7] J. Cabrera, M. Mena, A. Parra, and E. Pinos. Intelligent assistant to control home power network. In *2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, pages 1–6, November 2016.

[8] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, September 1997.

[9] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden Voice Commands. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 513–530, Austin, TX, 2016. USENIX Association.

[10] 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.

[11] 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.

[12] Marty Coyne, L Matchstick, and Chris Franzese. The Promise of Voice: Connecting Drug Delivery Through Voice-Activated Technology. 2017.

[13] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[14] R. Dhamija and L. Dusseault. The Seven Flaws of Identity Management: Usability and Security Challenges. *IEEE Security Privacy*, 6(2):24–29, March 2008.

[15] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, SPSM '14, pages 63–74, New York, NY, USA, 2014. ACM.

[16] 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.

[17] Nancie Gunson, Diarmid Marshall, Fergus McInnes, and Mervyn Jack. Usability evaluation of voiceprint authentication in automated telephone banking: Sentences versus digits. *Interacting with Computers*, 23(1):57–69, January 2011.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, May 2007.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] Andrew Liptak. Amazon's Alexa started ordering people dollhouses after hearing its name on TV. https://www.theverge.com/2017/1/7/14200210/amazon-alexa-tech-news-anchor-order-dollhouse, January 2017.

[31] Arnold M Lund. Measuring usability with the use questionnaire. *Usability interface*, 8(2):3–6, 2001.

[32] 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.

[33] A. Mishra, P. Makula, A. Kumar, K. Karan, and V. K. Mittal. A voice-controlled personal assistant robot. In *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, pages 523–528, May 2015.

[34] L. O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, December 2003.

[35] 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.

[36] D. A. Reynolds. An overview of automatic speaker recognition technology. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4072–IV–4075, May 2002.

[37] 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.

[38] 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.

[39] 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.

[40] Liwei Song and Prateek Mittal. Inaudible Voice Commands. August 2017.

[41] 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.

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

[43] 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.

[44] Shari Trewin, Cal Swart, Larry Koved, Jacquelyn Martino, Kapil Singh, and Shay Ben-David. Biometric Authentication on a Mobile Device: A Study of User Effort, Error and Task Disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC '12, pages 159–168, New York, NY, USA, 2012. ACM.

[45] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, Washington, D.C., 2015. USENIX Association.

[46] 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.

[47] Z. F. Wang, G. Wei, and Q. H. He. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *2011 International Conference on Machine Learning and Cybernetics*, volume 4, pages 1708–1713, July 2011.

[48] 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.

[49] 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.

[50] Z. Yan and S. Zhao. A Usable Authentication System Based on Personal Voice Challenge. In *2016 International Conference on Advanced Cloud and Big Data (CBD)*, pages 194–199, August 2016.

[51] Ka-Ping Yee. Aligning security and usability. *IEEE Security Privacy*, 2(5):48–55, September 2004.

[52] 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.

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

[54] Moshe Zviran and William J. Haga. Password Security: An Empirical Study. *Journal of Management Information Systems*, 15(4):161–185, 1999.

Appendix A

BACKGROUND QUESTIONNAIRE

Participant Background

Gender: □ male □ female □ prefer not to answer

Age: □ 18-20 □ 21-30 □ 31-40 □ 41-50

Major:
□ computer science
□ others in Science, Technology, Engineering, Mathematics (STEM)
□ others not in STEM

Frequency of voice assistant use:
□ never □ just a few times □ every day
□ at least once a week □ at least once a month

		Never			----------		**Very often**	
		1	2	3	4	5	6	7
1.	**How often do you use voice assistants such as Google Home, Alexa, Siri, etc.?**	□	□	□	□	□	□	□
2.	**How often do you use speaker verification systems?**	□	□	□	□	□	□	□
3.	**How often do you worry about your *security of information* when you use technical devices or applications?**	□	□	□	□	□	□	□
4.	**How often do you worry about your *privacy* when you use technical devices or applications?**	□	□	□	□	□	□	□

		Strongly disagree			----------		**Strongly Agree**	
		1	2	3	4	5	6	7
1.	**I know the privacy risks when I use voice assistant.**	□	□	□	□	□	□	□
2.	**I know the security threats when I use voice assistant.**	□	□	□	□	□	□	□
3.	**I understand the different approaches that may be used to increase the level of security of a user of a system.**	□	□	□	□	□	□	□

Appendix B

USABILITY QUESTIONNAIRE

		Strongly Disagree			----------			Strongly Agree
		1	2	3	4	5	6	7
1.	It is easy and simple to use.	□	□	□	□	□	□	□
2.	Using it is effortless.	□	□	□	□	□	□	□
3.	Occasional users would like it.	□	□	□	□	□	□	□
4.	Regular users would like it.	□	□	□	□	□	□	□
5.	I can use it successfully every time.	□	□	□	□	□	□	□
		1	2	3	4	5	6	7
6.	I learned to use it quickly.	□	□	□	□	□	□	□
7.	I am satisfied with it.	□	□	□	□	□	□	□
8.	I would recommend it to a friend.	□	□	□	□	□	□	□
9.	It works the way I want it to work.	□	□	□	□	□	□	□
10.	I feel I need to have it.	□	□	□	□	□	□	□
		1	2	3	4	5	6	7
11.	I feel secure using it.	□	□	□	□	□	□	□
12.	I feel it can protect the security of my information.	□	□	□	□	□	□	□
13.	I feel it can protect my privacy.	□	□	□	□	□	□	□
14.	I feel it is trustworthy.	□	□	□	□	□	□	□