BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

RESILIENCE OF BIOMETRIC AUTHENTICATION OF VOICE ASSISTANTS AGAINST DEEPFAKES
ODOLNOST BIOMETRICKÉ AUTENTIZACE HLASOVÝCH ASISTENTŮ OPROTI DEEPFAKES

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR — Oskar Šandor
AUTOR PRÁCE

SUPERVISOR — Mgr. Kamil Malinka, Ph.D.
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

147932

Institut: Department of Intelligent Systems (UITS)
Student: **Šandor Oskar**
Programme: Information Technology
Specialization: Information Technology
Title: **Resilience of Biometric Authentication of Voice Assistants against Deepfakes**
Category: Security
Academic year: 2022/23

Assignment:

1. Study voice biometric authentication.
2. Get familiar with methods of creating voice deepfakes.
3. Verify the ability of selected voice assistants (e.g. Alexa) to perform biometric authentication.
4. Design an experiment to verify the robustness of at least 3 selected assistants to voice spoofing using deepfakes.
5. Implement and evaluate the experiment.
6. Discuss the results and propose possible defense methods.

Literature:

- FIRC Anton, MALINKA Kamil a HANÁČEK Petr. Creation and detection of malicious synthetic media - a preliminary survey on deepfakes. In: *Sborník příspevků z 54. konference EurOpen.CZ, 28.5.-1.6.2022*. Radešín, 2022, s. 125-145. ISBN 978-80-86583-34-1.
- FIRC Anton a MALINKA Kamil. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: *SAC '22: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. New York, NY: Association for Computing Machinery, 2022, s. 1646-1655.

Requirements for the semestral defence:
Items 1 to 4

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor: **Malinka Kamil, Mgr., Ph.D.**
Head of Department: Hanáček Petr, doc. Dr. Ing.
Beginning of work: 1.11.2022
Submission deadline: 10.5.2023
Approval date: 3.11.2023

Abstract

With the rise of deepfake technology, imitating the voice of strangers has become a lot easier. It is no longer necessary to have a professional impersonator to imitate the voice of a person to possibly deceive a human or machine. Attackers only need a few recordings of a person's voice, regardless of the content, to create a voice clone using online or open-source tools. In that case, he or she can create recordings with content that the person may have never said. These recordings can be misused, for example, for unauthorized use of voice-assistant devices. The aim of this work is to determine whether voice assistants can recognize synthetized recordings (deepfakes). Experiments conducted in this thesis show that deepfakes created in a matter of minutes can spoof speaker recognition in voice assistants and can be used to carry out several attacks.

Abstrakt

S rozvojom technológie deepfake sa napodobňovanie hlasu cudzích ľudí stalo oveľa jednoduchším. Na napodobnenie hlasu osoby a prípadné oklamanie človeka alebo stroja už nie je potrebné mať profesionálneho imitátora. Útočníkom stačí niekoľko nahrávok hlasu osoby bez ohľadu na obsah, aby vytvorili klon hlasu za pomoci online nástrojov. V takom prípade dokáže útočník vytvoriť syntetické nahrávky s obsahom, ktorý daná osoba možno nikdy nepovedala. Tieto nahrávky sa dajú zneužiť napríklad na neoprávnené používanie hlasových asistentov. Cieľom tejto práce je zistiť, či hlasoví asistenti dokážu rozpoznať tieto nahrávky. Vykonané experimenty ukazujú, že deepfakes vytvorené v priebehu niekoľkých minút dokážu obísť schopnosť hlasových asistentov rozpoznať hovoriaceho a môžu byť použité na uskutočnenie viacerých útokov.

Keywords

deepfake, voice biometrics, speaker recognition, voice assistant

Kľúčové slová

deepfake, hlasová biometria, rozpoznanie rečníka, hlasový asistent

Reference

ŠANDOR, Oskar. *Resilience of biometric authentication of voice assistants against deepfakes*. Brno, 2023. Bachelor’s thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.

Rozšírený abstrakt

Hlasoví asistenti, ako sú Alexa, Google Assistant a Siri, si v poslednej dobe získali popularitu. Títo asistenti sa nachádzajú v rôznych zariadeniach a vykonávajú úlohy, ako je telefonovanie, fotografovanie, vyhľadávanie informácií, prehrávanie hudby, správu predplatného, ovládanie zariadení internetu vecí a nastavovanie pripomienok. Niektoré z týchto schopností sa však dajú zneužit a preto je nutné ich zabezpečenie. Jedno z týchto zabezpečení je použitie technológie na rozpoznanie biometrických údajov.

Biometrické technológie využívaju na identifikáciu jedinečné fyziologické a behaviorálne znaky. Hlasová biometria identifikuje osoby na základe charakteristík hlasu, či už fyziologických (ako je sila hlasiviek) alebo behaviorálnych (ako je výška hlasu a intonácia). Tieto charakteristiky sa používajú na rozpoznávanie hovoriaceho, pričom hovoriaci musí byť "zaregistrovaný" alebo inak povedané, musí mať vytvorenú šablónu hlasu v databáze. Vo fáze registrácie sa zhromažďujú hlasové vzorky na vytvorenie šablón, pričom sa tieto šablóny neskôr využívajú na porovnávanie hlasu. Hlasová biometria, ktorá sa široko používa na autentifikáciu používateľov, má stále zraniteľnosti a nedostatky.

Útoky na hlasové systémy, ktoré spočívajú v napodobňovaní alebo krádeži cudzieho hlasu sa stali dostupnejšími vďaka pokroku v nahrávacích zariadeniach a umelej inteligencii. Predtým boli tieto útoky zriedkavé a vyžadovali si profesionálne technické znalosti alebo podobný hlas. Nedávne zlepšenia v oblasti hlbokých neurónových sietí a umelej inteligencie však umožnili vytvoriť umelé napodobeniny konkrétneho hlasu takzvané „hlasové kópie". Tieto kópie môžu byť použíté na vytvorenie syntetických nahrávok alebo inak nazyvané deepfakes.

Deepfakes je technológia založená na umelej inteligencii, ktorá dokáže vytvárať realistické videá a obrázky ľudí, ktorí robia alebo hovoria veci, ktoré nikdy nerobili. Deepfakes využívajú hlboké učenie a neurónové siete na napodobňovanie výrazov tváre, pohybov tela a spôsobov reči. Hlasové deepfakes, známe aj ako hlasová syntéza, vytvárajú realisticky znejúcu reč pomocou umelej inteligencie. Existuje niekoľko rôznych spôsobov ako vytvoriť syntetický hlas. Medzi nich patria syntéza textu na reč, ktorá premení obsah z textovej formy na formu hovorenú a konverzia hlasu, ktorá upravuje hlas osoby tak, aby znel ako hlas niekoho iného. Detekcia deepfakes je náročná, ale momentálne techniky na detekciu zahŕňajú analýzu akustických vlastností, lineárnu chybu predikcie a identifikáciu artefaktov na rozlíšenie pravej reči od deepfakes.

Na základe získaných informácií o deepfakes a o schopnostiach hlasových asistentov bol navrhnutý experiment pre overenie schopnosti hlasových asistentov rozpoznávať synteticky vytvorené nahrávky. Jedná sa o experiment, v ktorom je preskúmaná odolnosť hlasových asistentov voči útokom pri ktorých sa používaju deepfakes. Úspešnosť týchto útokov je porovnávaná s úspešnosťou útokov, pri ktorých figurujú klasické hlasové nahrávky a cudzie osoby. Na základe výsledkov z experimentu je možné vyvodiť, že hlasový asistenti nevedia s úplnou presnosťou rozpoznať, či sa jedná o deepfake. Taktiež bolo zistené, že úspešnosť útokov syntetických nahrávok, je vyššia, ako úspešnosť utokov cudzieho hlasu.

Resilience of biometric authentication of voice assistants against deepfakes

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Mgr. Kamil Malinka Ph.D. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

. . . . . .

Oskar Šandor

May 9, 2023

Acknowledgements

I would like to thank my supervisor Mgr. Kamil Malinka Ph.D. for his priceless ideas and advices.

I would also like to thank my parents for participating in the experiment as impostors.

Contents

1. Introduction

Importance of securing user information in voice assistants is rapidly increasing due to popularity of such assistants among people. More than a half of American adults use voice assistants on their devices. These voice assistants are capable of not only playing music and setting timers, but also controlling other household devices, ordering things from internet websites, making phone calls, set reservations and even capture photos with a camera. Additionally, voice assistant is capable of remembering anything it is told to remember. This can include user's schedule, front door pin or package delivery time. Information of this kind can be easily misused by fraudsters and criminals. Therefore, various security systems were created.

Speaker recognition or voice recognition is one of them. Individual's speech or voice as a biometric is enough to verify speaker's identity. After verification it is on the system to decide if the speaker should or should not be granted access to the device. However, this voice authentication is in constant race against other technologies that could result in breaching their protection.

Deepfake has only come to the attention of the general public in the last decade, which shows that it is a very new technology. During this time, it managed to raise many questions about the security of various systems and concerns about the reliability of the information found online.

With the widespread of deepfakes, various tools for creating this type of media started to appear on the Internet. Most of these online tools are based on a commercial basis. Creating relatively high-quality voice deepfakes is thus very easy even for someone who has no technical background. Thanks to these tools, deepfakes have started to spread on the Internet several times faster. These can be videos or images intended for entertainment, but it is also possible to create videos or voice recordings with malicious intent.

Synthetic media can thus be used not only for spreading disinformation, defaming individuals, but also for spoofing and phishing attacks. In this work, possible attacks on three different voice assistants on five different devices are presented and some of them are later executed. Findings of every attack are presented and discussed. Several possible methods of defense against such voice spoofing are also discussed.

The structure of this thesis goes as follows. Possible malicious purposes of voice deepfakes, attack vectors and attacker model are described in Chapter 2. Chapter 3 explains how voice biometric authentication works and its current uses in the world. Various techniques of creating a voice deepfake are described in Chapter 4. In Chapter 5 a brief overview of voice assistants is presented. Chapter 6 describes the design of the experiment and needed preparation for the execution. Lastly, the findings and the evaluation of the experiment are presented, and possible methods of defense are discussed in Chapter 7.

2. Malicious use of audio recordings

In the past, it took a mimicry professional or a person with a voice similar to the voice of a victim to mimic or steal their voice. This was very rare and therefore often times not possible to carry out such a form of attack. With the passage of time, however, recording devices have improved and have brought new forms of possible attacks. This unlocked a new way to "replicate" the voice for various illicit purposes. Until recently, replay attacks were the biggest threat to voice spoofing. In recent years, artificial intelligence has improved and deep neural networks have been used for various purposes. One of these purposes was the creation of an artificial voice that resembled human voice. Later on, imitations of real people's voices started to be created and now it is possible to clone a person's voice without any technical knowledge. This has resulted in huge increase in the number of potential attackers. There are various tools on the Internet for creating voice deepfakes. One of these tools is described in Chapter 6 and used in the experiment in Chapter 7.

2.1 Attack vectors on speaker verification systems

With the tremendous speed of development of authentication, various automatic speaker verification systems have started to develop. These systems have gradually made their way into various end-user devices. Some of these devices include mobile phones, smart speakers and now also TVs and cars. Devices such as smart speakers can do much more than just play music and find out what the weather is like. They can be used for banking, home automation or even logging into applications [19]. There are many other functions such as opening doors, setting your own schedule, making phone calls, sending text messages and even unlocking cellphones [20].

Even though voice biometrics for user authentication has become a part of everyday life for many people, that does not mean that it does not have its drawbacks. One of the disadvantages is the possibility of an attack on such authentication. Attacks can be of various nature, so we divide them into two categories logical-access and physical-access [21]. The most well-known and simplest voice spoofing of the physical-access category is using replay [21]. What makes this form of voice spoofing a high threat is the fact that the attacker does not need to have advanced technical experience [22]. All that is needed is a voice recording of user, which the attacker then plays back to a device with voice authentication enabled. It is also possible to create voice recordings using a cut and paste system. This system is based on cutting short pieces most of audio, commonly whole words and putting these words into a sentence needed for a text dependent system [36]. Another form of physical-access is impersonation [21]. This is a form of attack in which a professional impersonator tries to imitate the vocal characteristics of the target person. Examples of these include phonation, pitch, loudness, and speaking rate. If the impersonator is able to mimic the fundamental frequency and format frequencies of the target voice, there is a potential vulnerability to automatic speaker recognition systems that use spectral features to identify the speaker [23].

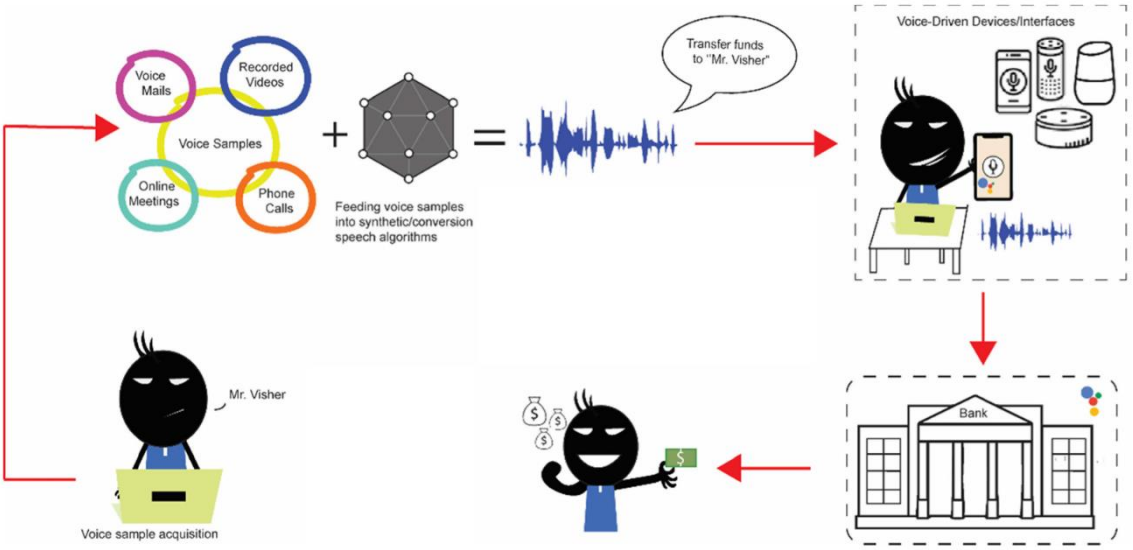

Figure 1: Voice phishing attack scenario proposed by Tuba Arif et al. [21]. Mr. Visher as an attacker obtains voice samples of a potential victim from various sources such as online meetings, phone calls, voice mails or videos. He stores these samples and later creates synthetic recordings from them, which he then plays to a device with access to the victim's bank account. If the recordings are of sufficient quality, the device can evaluate them as a genuine attempt to transfer payment. Retrieved from [21].

Another method of voice spoofing is text-to-speech (TTS) synthesis and voice conversion (VC). Both of these methods fall into the category of logical-access attacks [21] and their main difference is the input data. The text-to-speech system converts written text into spoken form using a speech synthesizer [24]. Possible scenario of text-to-speech method used for voice phishing is shown in Figure **1**. On the other hand, voice conversion is a process in which a speaker's identity is transformed to another speaker while preserving the content of their speech [25]. Both of these methods are described in more detail in Chapter 4. These methods can create a recording of the victim's voice, with content that the victim may never have said. In this case, it depends only on the skill of the attacker to decide, how he can misuse these recordings. In a typical case, the attacker could gain access to a bank account, a terminal device, or control of a home [21]. With the convenience of today's voice assistants, it is also possible to send text messages, make phone calls, and order items from the Internet. These actions are all chargeable and could be exploited for financial gain for the attacker or financial loss for the victim. Possible attacks on specific voice assistants are described in more detail in the Chapter 6 experiment design, where the main purpose of the experiment is to verify robustness of selected assistants against deepfake voice spoofing. Afterwards, the success rate of deepfake attacks is compared with other types of voice spoofing. Results of executed attacks are presented in Chapter 7 and possible defense methods are proposed.

2.2 Possible Threats

In his paper Jon Bateman [17] presents ten different scenarios that could harm an individual, a business, a global or national market and financial regulatory structures. Each of these events can affect more than one of these entities. He further divides synthetic media into narrowcast synthetic media and broadcast synthetic media. The primary difference between the two is that narrowcast synthetic media are created for the purpose of targeting individuals and are spread through private communication channels such as mail or phone calls. On the other hand, broadcast synthetic media are designed to influence a group of people such as investors, and are distributed through mass communication channels such as social media [17]. The main difference in the prevention of attacks is that social media and journalistic articles can be verified. Information is available to the public, so that in the case of fake news or other forms of misinformation, it is possible to moderate it. However, this is not the case for narrowcast communications as it is up to the recipient to check and verify private e-mails or text messages [17].

2.2.1 Narrowcast synthetic media

Using a synthetic voice to imitate someone else is one possible attack vector. An attacker can create a synthetic recording that mimics person's voice for harmful intentions. Deepfake recordings can thus fool individuals or gain access to systems that rely on voice authentication. In 2020 a huge fraud used deepfake voice to trick a bank manager in the Hong Kong. The bank manager received a call from a man whose voice he recognized. The voice that spoke to him was one of a director at a company with whom he had talked before. The director needed the bank to authorize some transfers to the tune of $35 million. The bank manager, believing everything appeared legitimate and genuine, started to transfer the money [32]. This is not the first time something like this has happened. In 2019, a similar attack in which thieves managed to trick the managing director of an unnamed British energy company to send hundreds of thousands of dollars to a secret account [34]. The managing director believed he was on the phone with his boss and even stated that he recognized his boss’s subtle German accent. This proves that the voice of a person can indeed be spoofed and used for malicious intent [34]. Another attempt to deceive a victim using a synthesized voice happened in early 2023. It was a fake kidnapping scam where the attacker cloned the voice of the victim's daughter and demanded one million dollars as ransom [35].

Synthetic voice can also be used in social engineering attacks. Techniques that are designed to manipulate and deceive people in order to gain private information or tricking them into opening doors is very powerful in combination with synthetic voice recordings. Created voice replicas that seem genuine enough to deceive individuals could bypass physical or digital security [33].

2.2.2 Broadcast synthetic media

The world wide web provides the ability to share information of all kinds. With the large number of people on social networks and uncensored content, this information is now becoming a weapon. Nowadays, deepfakes are appearing on social networks in both audio and video form. Some of these videos may be satirical in nature, but there are also those, that seek to spread misinformation. Studies have shown that the form of the information presented (video, voice recording or text) does not play a major role in a person's ability to recognize whether it is false information or truthful information [30]. Some suggest that even if deepfake is not good enough to deceive anyone, it still leads to uncertainty in consumer of such media and causes a distrust in other information found on the internet [31].

At the moment, it is not difficult to create an anonymous profile on the biggest social networks and upload content that can have huge consequences not only for individuals but also for groups of people. A well-crafted deepfake shared on the right platform at the right time can have several negative outcomes. Jon Bateman [17] in his paper states that it is possible to manipulate small-cap stocks or influence public opinion about bank weaknesses. Both of these situations could be exploited either for the profit of the perpetrator or to harm the company. Mika Westerlund [27] states that deepfakes also put some pressure on journalists who have to filter real news from fake news. Fake news of any kind can severely affect national security and interfere in elections.

2.3 Attacker model

A perpetrator can be considered a person, whose purpose is to harm someone, gain unwarranted access, or push his or her agenda. This can be accomplished in several ways. He or she may attempt to create or otherwise obtain a voice recording of the victim. He may play this recording directly to the voice system in an attempt to achieve his goal. He may also pay a professional impersonator to attempt to break into the system. Another way in which he might succeed is to create a synthetic voice or, in other words, a voice deepfake. In this case, it may be someone who has sufficient resources and skills to create a deepfake [16]. Such an attacker must know the procedures by which the victim authenticates himself or otherwise proves his identity. In the case of an automated system, the attacker must know all the steps necessary to access the system or bank account. This information can then be used to create the required deepfake recordings to successfully bypass the authentication.

3. Voice biometric authentication

Biometric technologies are used to recognize a person's identity based on their physiological and behavioral traits. These features are unique and distinct for each person. According to J. Wayman et al. [1], the beginning of the use of human physiological characteristics for their identification in scientific literature dates back to the seventies of the nineteenth century. Measurements such as skull diameter, and hand or foot length were used. The authors also mention that Henry Faulds, William Herschel and Sir Francis Galton proposed measurements such as fingerprint and facial measurements in the 1880s. These measurements are still used in today's world. Approximately 80 years later, thanks to the research on digital signals and their processing, it was possible to automate the process of human identification [1]. Voice recognition and fingerprint-based recognition were the first to be explored [1]. In later years, retina and signature recognition were added [1]. At the end of the twentieth century, automatic facial and iris recognition also started to be developed [1]. In today's world, biometrics are used to differentiate between individuals. Biometrics have the advantage of not being forgettable, as passwords and pins are. It is also often more complicated to steal someone's voice or fingerprint than it is to steal their password.

3.1 Principles of voice biometrics

Voice recognition works by identifying a person based on voice characteristics [5]. These voice characteristics can also be called voice biometrics. Voice biometrics consist of both physiological aspects (strength of the vocal cords, size and shape of the throat and mouth) and also behavioral patterns (pitch, speaking rate, intonation) [5]. However, the behavioral side can change based on the mood and health, but the physiological side changes very rarely [5].

Biometric authentication could be divided into two stages. The first stage being enrollment and the second stage being verification and identification [5]. During the enrollment person is required to follow a procedure that is designed to obtain all the necessary data (voice samples) from a given characteristic [5]. This data is the input to an algorithm that ensures the construction of a "template" or a “voiceprint” [5]. If the creation of the template fails due to a faulty input, the person is asked to repeat the procedure. A successful registration ends with the template being stored in the database. Data used to create the template are usually not stored in database to avoid unnecessary crowding of the data storage. These templates are then used in the second stage for voice matching on subsequent authentication attempts. The purpose of speaker verification is to determine whether the speaker is who he or she claims to be [3, 6]. System must have the ability to verify the speaker from a vast pool of possible deceivers [3]. On the other hand, the goal of speaker identification is recognizing speaker from a group of people with active templates in database [3, 6]. Templates do not need to be associated with any other identifier of the given person [1]. No name, date of birth or ID number is required [1]. This creates room for anonymous authentication which could be beneficial in some cases.

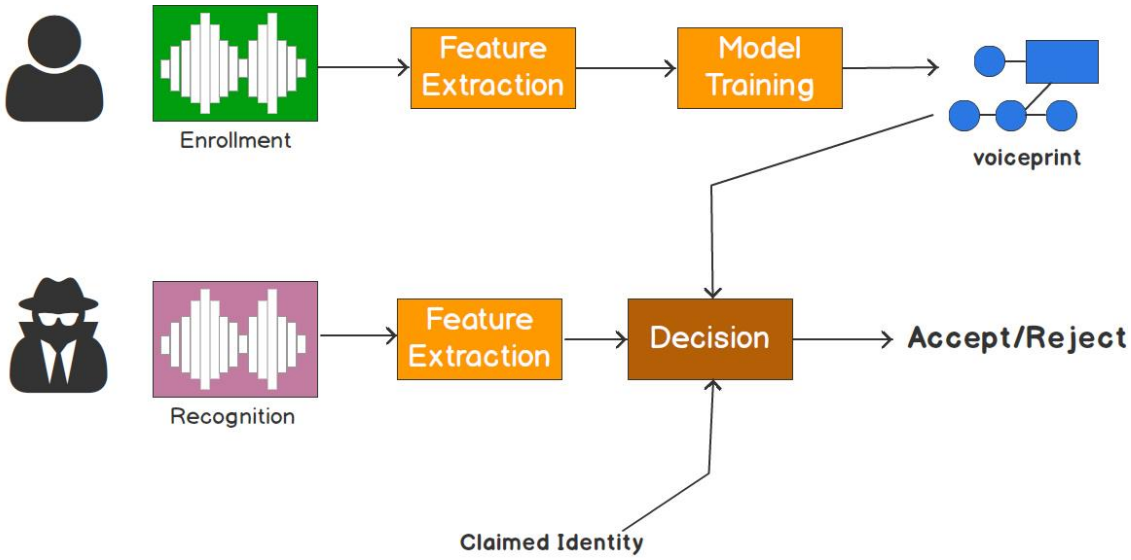

Figure 2: Process of enrollment and recognition. Retrieved from [9].

Further, voice authentication divides into text-independent and text-dependent [3]. As the name suggests, text-independent authentication should authenticate user based on his or hers voice only. Speaker does not need to use the same words or sentences used during the enrollment, as the identification algorithm learns the speaker's voice pattern, not just the combination of text and voice [3]. This forces the system to learn how the person speaks. On the other hand, text-dependent systems require the person to use the same phrases or passwords for identification as during enrollment [3]. A potential drawback of this approach is the increased risk of spoofing via replay attack or synthetized voice if no anti-spoofing measures are put in place [4].

3.2 Usage of voice biometrics

Voice based recognition systems are used in various fields including attendance systems, mobile texting, medicine and many more [2]. Voice authentication is also a significant assistance in data protection in the banking sector [4]. Similar to a fingerprint, the human voice is unique enough to differentiate people. Although no biometric system is 100% accurate [4], this gives companies the ability to facilitate customer access to their accounts. Currently, the majority of financial institutions all over the world provide telephone banking in some capacity [4]. These services simplify and help customers to perform various banking tasks. Many times, this includes creating a new bank account over the phone, which was once inaccessible [4]. Some of these services are partly automated, which means that a voice recording is played to the caller. Depending on the type of service, the caller is then prompted to respond accordingly. Usually, this is either by voice or by pressing a button on the phone (in the case of smartphones with a touch screen, by tapping a number on the screen). Thanks to this feature, services such as "check account balance" do not require a human employee in the call center. Usage of voice biometrics could be the first step to completely eliminate the need of the customer to remember any PIN, password, or other authentication means. Such authentication styles are called knowledge-based [4] (a person must

remember or otherwise know the correct order of letters or numbers). This is oftentimes burdensome and inconvenient. Thus, it becomes common practice to create simple passwords in the hope that the person can remember them [4]. Such passwords are easy to guess since they mostly consist of names and words. Another practice that threatens account theft is the repetition of the same passwords in different applications. The consequence of this practice is reduced protection across all accounts with the given password. The thief needs to figure out a single password to gain access to multiple applications or devices. Currently, two-factor authentication has become a popular method of authentication [4]. As the name suggests, it is a combination of two factors that a person must meet in order to gain access to an account. Usually, it is a password or pin from knowledge-based authentication and a PIN or a confirmation message sent to another device. This reduces the chance of misuse, but it also reduces the convenience of using the application [4]. When selecting a suitable biometric, or combination of biometrics, the main consideration is security followed up by user satisfaction. This is why voice biometrics is preferred among other biometrics to be used in telephone banking. It is a quick and simple way to authenticate users without unnecessarily burdening them. In practical terms, using voice biometrics does not change the process at all, but it does add another layer of security.

3.3 Strengths and shortcomings of voice authentication

Voice biometrics allow for hands-free authentication, making them a non-intrusive and convenient method for users to verify their identity [2]. Requiring physical contact with a device or a sensor makes biometric authentication somewhat intrusive and inconvenient. They are often very low-cost [15], which makes them a preferred alternative to other biometric technologies, such as fingerprint or facial recognition, which require special hardware to operate properly. Being able to authenticate large numbers of users simultaneously makes voice biometrics a scalable solution for use in call centers or other high-volume environments.

The big drawback of voice recognition is the imperfection of the surrounding conditions. The quality of the received audio can easily be degraded by surrounding noise, low-quality microphones, or malfunctions in the transmitting media. Furthermore, environment may cause change in tonality or ability of a person to speak clearly. In cases like these, the system must be robust enough to recognize different phonation types of a person. An example would be the handling of whispered phonation which is usually hard to collect [18] and not required in natural enrollment process. Another example would be recognition of persons voice while yelling from a different room in their house [3]. Things like illness can also affect the ability of a user to speak clearly. A quality system should recognize the person speaking despite of any mentioned circumstances. Another inconvenience that can happen during voice recognition is several people talking over each other [18]. This scenario could occur on a crowded bus or in other extreme conditions. The ability of the system to filter out other voices and ambient sounds is a reflection of its quality. Some users can have problems with using voice biometrics, especially older adults, people with certain disabilities, or non-native speakers. Finally, yet importantly, there is a possible threat with the use of voice authentication. Attacks known as voice spoofing attacks show vulnerability in voice authentication systems. Systems such as automatic speaker verification may

be compromised via voice recording [26]. Alternatively, an artificially created or otherwise called synthesized voice could be used [18]. Possible styles of creating synthetized audio recordings are further discussed in Chapter 4. Some techniques to prevent or at least reduce the success of such voice spoofing could be based on a new concept called voice texture where the texture of cloned voice varies from genuine ones [21]. Other techniques trying to stop voice recordings are based on so-called liveness detection [4]. Liveness detection works on the principle of analyzing the input and detecting the acoustic effect that could indicate, that the input may be a recording [4]. There are also other ways to reduce the risk of voice spoofing such as using phrases that have been used in the enrollment (text-dependent authentication) or two-factor authentication. This matter is looked upon in Chapter 7, which includes discussion on possible defenses against voice spoofing.

4. Deepfake

A deepfake is a type of artificial intelligence-based technology used to create realistic-looking videos and images that depict people doing and saying things they never did or said [27].
The name "deepfake" comes from the fact that the technology is based on deep learning, which is a type of machine learning that uses neural networks [16] with multiple layers to analyze and process large set of data samples [27]. Deepfakes are created by feeding a computer algorithm a large dataset of images and videos of a person, along with recordings of their voice [28]. The algorithm then learns how to mimic the person's facial expressions, body movements, and speech patterns [16]. Once the algorithm has learned enough about the person, it can be used to create new videos and images of the involved person without their consent [27].

At the moment, the word deepfake has no precise definition, but it is a combination of "deep learning" and "fakes" [16, 28]. It refers to synthetic media created through AI method called deep learning, which relies on a computing system called a deep neural network [17]. There are several forms of deepfakes such as shape-shifting (face-swap), synthetically created audio or text forms [17]. Some of these deepfakes are recognizable by humans, considering people can be trained and develop their ability to distinguish between what is fake and real [29]. Imperfections such as weird eye movement, unnatural glare from glasses and strange skin coloration [29] could indicate, that the video has been algorithmically manipulated. However, this does not apply to well-made deepfakes which are very difficult to recognize, whether by humans or machines. These carry the threat of either fraud or unauthorized access to devices.

Deepfakes have become a concern for a number of reasons. They can be used to spread misinformation, propaganda or change public opinion. They can also be used to deceive people into believing in false information or to cause damage to a person's reputation. Deepfakes can also be used to impersonate real people, to create audio hoaxes, or to impersonate people over the phone. Some of the other cases have already been covered in Chapter 2.

4.1 Voice deepfakes

Voice deepfakes, also known as "voice synthesis" or “speech deepfakes” is a type of deepfake technology that can create realistic-sounding speech using artificial intelligence. Voice created by this technology is called “synthetized voice”. The phrase speech synthesis (SS) refers to the creation of an artificial human-sounding voice using software and hardware system programs [36].

Important aspect of creating deepfake speech is the quality of target’s voice recordings. The better the quality of the recordings, the more realistic the deepfake speech will sound. Another important aspect is the size and diversity of the training dataset. The more data that is used to train the voice model, the better the model will perform. A different challenge is to improve the naturalness and speaker identity preservation of the deepfake speech [7]. Method that could address this problem is developing a model that is capable of capturing and preserving the speaker identity information of the original voice [25].

The detection of audio deepfakes is very important for the society, especially since in recent years criminal activities related to the use of audio deepfakes have been emerging. Examples of attempts of fraud related to deepfakes have already been described in Chapter 2.

4.1.1 Text-to-speech synthesis

All information contained in this section has been retrieved from *Review on Text-To-Speech Synthesizer* [24]. Text-to-Speech (TTS) is a technology that generates speech by converting written text into spoken words using speech synthesizer. Speech synthesizer takes in the text as input and outputs corresponding spoken waveform, trying to mimic the way a human would sound. The text processing component's goal is to analyze the provided input text and generate a suitable sequence of phonemic units. This system is dependent on the functioning of two components, namely text processing and voice generation.

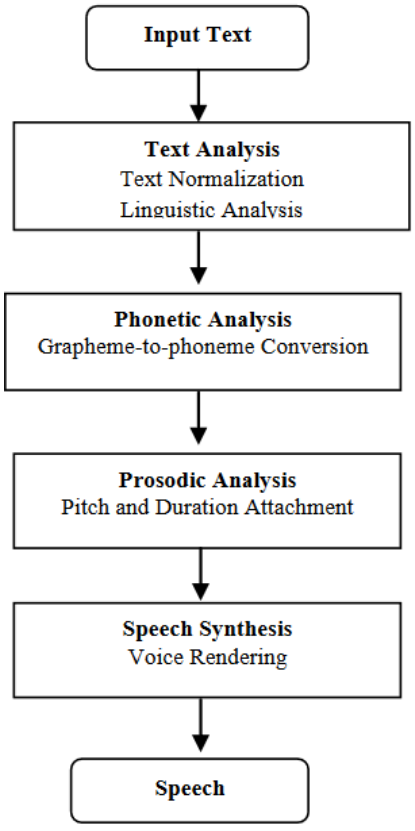

Figure 3: Typical text-to-speech pipeline. First the text is normalized, abbreviations are replaced with full words (for example Dr. to doctor) and morphological analysis ensures correct pronunciation of words. Afterwards, parameters such as speaking style and emphasis are detected. Then the intensity and duration of the individual frequencies are determined in the prosodic analysis. These parameters go into speech synthesis in which the final voice is created. Retrieved from [24].

In text processing, the input is first analyzed for correct document structure and later normalized. Normalization deals with cases where an abbreviation or acronym is found in the text. These are then matched with their correct representation for proper output. Lastly, a linguistic analysis is performed, which includes a morphological analysis for the correct pronunciation of words and the resolution of any ambiguities in text. The second part of the process which is speech generation consists of phonetic and prosodic analysis. These ensure not only the correct pronunciation of each word along with the speaker's emphasis and style, but also prosodic features such as accent, rhythm and intonation. The whole process is illustrated in Figure ***3***.

4.1.2 Voice conversion

Voice conversion is a technology that allows for the modification of a person's voice to sound like another person's voice [25]. This technology can be used to create deepfake speech, where a person's voice is replaced with the voice of another person in an audio recording. Creating a voice deepfake using voice conversion is usually divided into extracting information about the identity of the speaker and extracting linguistic content. Linguistic content describes the content of the sentence (the words that were spoken) and information about the speech, such as rhythm and intonation [25]. Both of these extractions are the work of a unit called an encoder. The primary work of the encoder is the integration and correct representation of identity and linguistic content extractions [25]. Tasks such as feeding information into the encoder and extracting information from linguistic content extractions are time-dependent, so they are often combined [25]. After processing this information, the encoder then sends it to the decoder and vocoder. These two process the obtained extractions and together create an appropriately manipulated soundtrack.

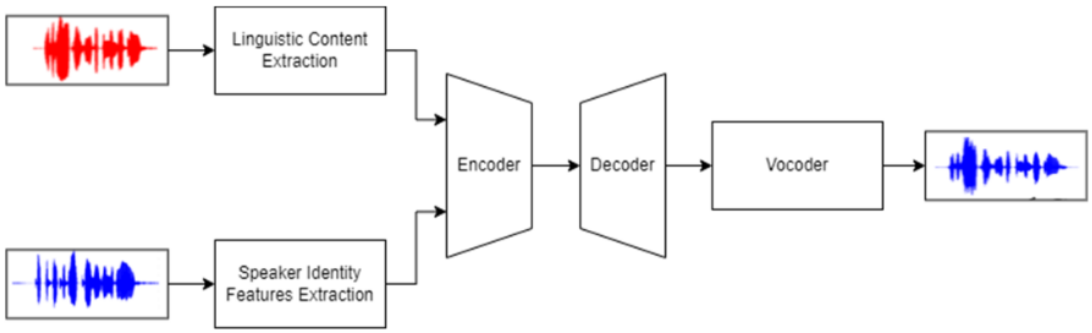

Figure 4: Simplified voice conversion pipeline for illustrative purposes. Input consisting of waveform or spectrogram is fed into the encoder which then extracts speaker's identity and linguistic content. Extractions are then forwarded into the decoder/vocoder where the final soundtrack is produced. Retrieved from [25].

According to Mohammadi and Kain [37] there are several ways to categorize voice conversion methods. These methods are divided based on:

- **Recordings used during training** to parallel and non-parallel.
- **Text dependency** to text-dependent and text-independent.
- **Language dependency** to language-dependent and language-independent.
- **Amount of training data used.**

Parallel training is a method in which the model is trained on a dataset of parallel data, meaning parallel sentences (including the same linguistic content) are used between the source and target speaker to train the system [37]. Another method is through the use of a technique called non-parallel training, where the model is trained on non-parallel recordings between the source and target speaker [37]. Text-dependent systems, unlike text-independent systems, need a phonetic transcript along with the recordings. The third division depends on whether it is a voice conversion system that should work between the same or different languages [37]. The last division is based on the size of the dataset used to train the given model. For larger training data methods that remember more are used, while for smaller data, methods that can generalize better are preferred [37].

4.1.3 Speech morphing

Speech morphing is a technology that allows for the modification of a person's voice to sound like another person's voice, similar to voice conversion, but the approach and the underlying technology is a bit different. The primary objective of the techniques for speech morphing that have been created is the seamless transition from one sound to another resulting in two sounds being combined to produce a new sound with an intermediate timbre [42]. Methods that can create synthetic audio using speech morphing are primarily based on the interpolation of sound parametrizations resulting from analysis techniques, such as the Short-time Fourier Trans-form (STFT), Linear Predictive Coding (LPC) or Sinusoidal Models synthesis (SMS) [43].

4.2 Voice deepfake detection

As with most technologies, the creation of deepfakes is improving much faster than their detection. Some deepfakes are almost indistinguishable from genuine voice recordings, which can lead to an increased threat of malicious uses. Therefore, synthetic speech detection has emerged as a crucial study area and several detection techniques have recently been developed.

These techniques are usually based on the extraction of acoustic features in the spectral domain [44]. Different technique of speaker verification spoofing countermeasure is based on analysis of linear prediction error [45]. Other methods focus on finding artefacts that could distinguish genuine speech from spoofed one [46]. There is currently no one right way to detect deepfakes, which is why different methods use different techniques.

5. Voice assistants

More than 62% of American adults already use voice assistants (VAs) on their devices [8] and this number is rapidly increasing. Voice assistants are not only found in smartphones, computers and smart speakers, but now also in TV, cars and household appliances. There are several voice assistants that have the ability to recognize speech input from users and respond with appropriate actions or information. Some of the most well-known examples include Amazon's Alexa, Google Assistant, Apple's Siri. Actions such as asking questions, controlling home automation devices, playing music, managing calendars and to-do lists are used every day by millions. Although each assistant has a few distinctive qualities, they all perform the same basic tasks such as [39, 40, 41]:

- Making calls and sending messages.
- Taking photos.
- Finding basic information online.
- Getting directions.
- Playing music from connected services.
- Managing subscriptions.
- Controlling Internet of Things devices.
- Setting reminders, schedules and building to-do lists.
- Basic math calculations.

The first voice assistant presented to the market was Apple’s Siri. Starting as standalone application in 2010 with later integration into the iOS in 2011. Another entry into the voice assistant market came in 2013 with Microsoft's voice assistant named Cortana. Right after that, Amazon joined the ranks of companies with its own voice assistant when it launched Alexa in 2014, along with an Echo-connected home speaker. Two years later, Google also joined in with an embedded app for Android devices and its own smart speaker. Later in 2017 a virtual assistant named Bixby which is developed by Samsung Electronics entered the market.

Compared to prior voice-activated technologies, today’s voice assistants can react to a considerably wider range of instructions and queries thanks to their internet connection [38]. Each interaction is transmitted back to a centralized computing system, which examines the user's vocal instructions and sends the appropriate answer to the assistant [38]. Voice assistants have also improved in their ability to recognize natural human speech in which different sentences can have the same meaning. An example of this may be the questions "where did I leave the car" or "do you remember where I parked" [38]. Both of these sentences trigger the expected response which prevents user’s frustration of earlier voice recognition systems, which were text-dependent and required specific phrases to get the correct response [38].

According to S. Subhash et al. [11] systems intended to converse with humans consist of six components which include Voice recognition, voice language apprehension, dialog manager, natural language generation, text to speech convertor, and knowledge base. Assistants used in current smart phones and smart speakers use "orders" to correctly interpret and understand the

request from the user [47]. The expressed order can be decomposed into several parts which the assistant identifies, isolates and then performs the correct function on their basis. A typical order consists of three parts which are wake word, invocation name and utterance [47]. Wake word literally wakes up the device and puts it into listening mode where it is ready to receive a request. Invocation name serves as a trigger to invoke a specific skill or action, which is usually followed by utterance. Utterance can be understood as an identifier of the user's intent on the basis of which the device decides what to say or what to do [47]. Assistants such as Alexa, Siri and Google send this information to a cloud-based back-end service, where the request is evaluated and information on how to respond to the request is sent back to the device [10, 12]. The devices thus require a constant connection to the Internet, otherwise they will not evaluate even simple requests.

Figure 5: Example of an "order" using Amazon Alexa. Voice assistant listens for a wake word, which is followed by information containing a specific skill with the addition of intent. Retrieved from [47].

Features	**Google Assistant**	**Siri**	**Alexa**	**Bixby**	**Cortana**
Device Compatibility	Android, iOS, Smart speakers and displays, smart home devices	iOS, macOS, HomePod, Apple Watch, smart home devices	Amazon Echo, Fire TV, smart home devices	Samsung Galaxy devices, smart home devices	Windows devices, smart home devices
Wake Word	"Hey Google"	"Hey Siri"	"Alexa"	"Hi Bixby"	"Cortana"
Speaker recognition	Yes	Yes	Yes	Yes	No[1]
Natural Language Processing	Yes	Yes	Yes	Yes	Yes
Software Used	Google Assistant app	Apple iOS, WatchOS	Amazon Alexa app	Samsung Bixby app	Microsoft Windows

Table 1: Comparison of individual voice assistants.

[1] https://answers.microsoft.com/en-us/windows/forum/all/no-try-to-respond-only-to-me-option-in-cortana/fad51ca8-0b4b-4350-bee7-c3d92b584d3d

6. Experiment design and preparation

As mentioned in Chapter 2, there are several ways to spoof someone else's voice and gain unauthorized access to a device or bank account. It is also possible to trick the person who is responsible for bank transactions or account management. This chapter describes the design and procedure used to create the experiment, which was subsequently carried out and the results are presented in Chapter 7. The proposed experiment consists of several parts. First, it was necessary to identify all possible vulnerabilities in the voice assistants. Since not all voice assistants work solely on mobile devices, it was necessary to analyze what each assistant is capable of. It was important to examine all the functions and find out which ones could be misused for malicious purposes. In order to be included in the experiment, the feature provided by the voice assistant had to meet one or more of the following attack goals

Attack Goal	**Description**
Provide the attacker with vulnerable information.	The attacker gains access to sensitive information that can be used later for malicious purposes.
Provide the attacker with a platform used for spreading messages.	The attacker gains access to the victim's platform to send messages in their name, which could be used to spread false information or cause reputational damage.
Provide the attacker with potential financial gain.	The attacker gains access to financial information or resources that can be used for fraudulent activities, theft, or extortion.
Provide the attacker with unauthorized access.	The attacker gains access to areas or applications that they would not otherwise have access to, which could be used for further attacks or data theft.
Provide the attacker with a tool for financial harm.	The attacker gains a tool that can be used to cause financial harm to the victim, such as stealing funds or performing unauthorized transactions.
Provide the attacker with a tool to tarnish reputation.	The attacker gains a tool to undermine the reputation of the victim, such as posting false information or spreading rumors.
Provide the attacker with a tool to harass the victim.	The attacker gains a tool to make the victim's life unpleasant, such as sending harassing messages.

Table 2: Table showing the attack goals that must be achievable by an attack to be included in the experiment.

These features have been tested by the owner of the device or in other words by someone who uses the device on a regular basis. The device "knows" his voice which has the ability to control all the features that the device provides. These features were then tested from the point of view of the attacker in several different forms, namely

- Replay attack.
- Voice of unauthorized person (female).
- Voice of unauthorized person (male).
- Deepfake.

The tests were performed on devices that have access to either Apple Siri, Amazon Alexa or Google Assistant. Namely iPhone 7 (Siri), Google Pixel 4a (Google), Xiaomi Mi Smart Speaker (Google), Google Nest Mini 2 (Google) and Amazon Echo Dot 3 (Alexa). All of these assistants provide some form of voice recognition. Individual companies refer to this feature differently. In the case of Siri, it is "Personal Requests", Google calls this feature "Voice Match" and Amazon refers to it as "Voice ID". Each voice assistant (VA) has its very own phone application, either downloadable or built-in, that allows enabling and disabling this feature. The process of setting up voice recognition is very simple. Reading a few sentences that appear in the application is enough to create a template according to which the following answers to the requests will be evaluated.

Since this thesis focuses on the resilience of voice assistants against deepfakes, it was essential to create synthetic voice recordings. In this work, a free version of the commercial tool Resemble AI is used to create deepfakes. It was chosen mainly due to the fact that the creation of an artificial recording with this tool is very easy. It does not cost any money and the tool can be used by anyone who has a basic knowledge of the English language.

As voice assistants are gradually being integrated into more and more devices and their number of users and their functionality is constantly expanding, it is necessary to prevent possible misuse of their functionality. The goal of this experiment is to create deepfake of sufficient quality to spoof the voice authentication of individual voice assistants. Play all pre-recorded phrases to voice assistants as “replay” attacks. Then play all deepfakes created to the voice assistants as “deepfake” attacks. Additionally, two impostors were used to test the speaker recognition feature available in each assistant. Finally, all results were recorded, compared and discussed with possible defense methods proposed in Section 7.6.

The attack design in this chapter assumes that the attacker has access to the device on which the voice assistant is enabled for the time required to execute each attack. It also assumes that the settings are correctly configured to enable all possible assistant features.

6.1 Identifying and classifying weaknesses

Even though these assistants largely share the same functionality, there are some differences between them. This section describes and then breaks down the functions that could be abused. The categorization is based on several features, namely

- **Difficulty** of execution
    - **Low** – short sentences
    - **Medium** – medium length sentences
    - **High** – long sentences and follow up questions

- **State** of the device
  - **Locked** – the device is locked
  - **Unlocked** – the device is unlocked
- **Damage** of the attack
  - **Low** – inconvenience for the victim or low information gain
  - **Medium** – exploitable information, low financial loss or defamation
  - **High** – unauthorized access to an object or significant financial loss
- **Voice Assistant (VA)** under threat
  - **Siri**
  - **Google** Assistant
  - **Alexa**

1) Phone calls.

Difficulty:	State:	Damage:	VA:
High	Locked	Medium – High	Alexa/Siri/Google

Table 3: Classification of phone call attacks.

The ability to make calls from a stranger's device can be abused in a number of ways, such as making scam calls from a stranger's number or calling premium rate numbers. In the case of smart speakers, it is not possible to exploit things like calls to emergency services since very few providers have this functionality enabled. Dialing premium rate numbers is only available on Siri and Google since Alexa does not support dialing these numbers[2]. The difficulty of the attack has been classified as medium to high because it is necessary to pronounce the whole number in quick succession. A slight pause in the pronunciation of the phone number will interrupt the action. Possible damages have been classified as medium to high since it is possible to make phone calls via a paid line and to dial premium numbers[3]. These conditions apply to the use of assistants via mobile phones. Smart speaker devices have this functionality limited to certain locations[4].

2) Sending messages

Difficulty:	State:	Damage:	VA:
High	Locked	Medium – High	Alexa/Siri/Google

Table 4: Classification of attacks based on sending messages.

Sending text or multimedia messages can be misused to transmit scam messages, to send advertisements or to send dangerous links that can be part of SMS phishing. These attacks are however very difficult to execute as the entire content of the message needs to be dictated to the device. In the case of dictation of clickable links, it would be necessary to dictate the link character by character. However, in the experiment it turned out that this form of dictation is very difficult to perform, as the voice assistant has trouble recognizing individual characters. This form of

[2] https://www.androidauthority.com/can-alexa-make-phone-calls-3242911/
[3] https://support.google.com/googlenest/answer/9465808
[4] https://support.google.com/googlenest/answer/7363847

attack can also be performed without unlocking the phone. Damage factor has been rated medium to high mainly due to the fact that messages can be sent to someone in contacts. This increases the chance that the recipient will be fooled by an SMS phishing message, considering that he or she will receive a message from someone he or she knows. Even if the SMS phishing is not successful, SMS is a paid service, so the victim can still be at a financial loss.

3) Reading notifications.

Difficulty:	State:	Damage:	VA:
Low	Locked	Low – Medium	Alexa/Siri/Google

Table 5: Classification of attacks based on reading notifications.

The possible misuse of reading notifications can vary widely, as it is possible to read all the content in notifications. Primarily it can be used to read personal conversations. However, it can be used to read messages that may also contain a verification code to log into a bank account or to read notifications from applications providing two-factor authentication. The difficulty of executing this attack is low simply because the sentence to trigger the action is very simple. Even though the state is categorized as locked, it depends on the settings of the phone, which must be set so that the content of the message would be displayed in the notification even on the locked screen. It is also possible to read the notification only once. The possible damage caused by this attack is categorized as low-medium because it depends on the settings of the phone and the stand-alone code that the attacker gets cannot be exploited. For a possible exploit, the attacker would have to trigger the notification himself via a bank login or use the code to receive a package in a stranger's name.

4) Reading text messages.

Difficulty:	State:	Damage:	VA:
Low	Unlocked	Low – Medium	Alexa/Siri/Google

Table 6: Classification of attacks based on reading text messages.

An attack working on the same principle as reading notifications, with the difference being that it is only possible to read text messages in the preconfigured application for sending and reading messages. Unlike reading notifications, this function is only available when the phone is in an unlocked state. It is possible to read older messages and read messages more than once. However, it is not possible to get information from other applications as in the case of reading notifications.

5) Taking pictures and recording videos.

Difficulty:	State:	Damage:	VA:
Low – Medium	Locked	Low – Medium	Alexa/Siri/Google

Table 7: Classification of attacks based on misusing camera features.

The use of camera functions can also be exploited by an attacker as it is possible to take photos and videos using only voice commands. With Google Assistant and Siri, this function can be invoked even in locked mode. Alexa also has this function, but it requires a specific kind of device called Echo Show, which has its own display and camera. Taking photos and videos in itself is not considered high-risk, but in conjunction with messaging, it could be a quite dangerous combination.

6) Misusing digital wallet.

Difficulty:	State:	Damage:	VA:
Medium – High	Unlocked	High	Siri

Table 8: Classification of attacks based on misusing Apple pay.

Misusing a digital wallet like apple pay can be very easy on devices using Siri, as all it takes is a short sentence to send a payment between known accounts. However, it is classified as medium to high in difficulty as this attack also requires confirmation of payment by tapping on the smart phone screen, making it so the attack cannot be carried out by voice alone.

7) Managing subscriptions.

Difficulty:	State:	Damage:	VA:
Low – Medium	Locked	Medium – High	Alexa

Table 9: Classification of attacks based on managing subscriptions.

If the user of a device with the Alexa voice assistant has filled in all the necessary details for payment, it is possible to subscribe to the Amazon Music app using voice only. As this is a pay-as-you-go plan with per-month billing, there is a possibility of a significant financial loss if this event goes unnoticed.

8) Using Smart Home or Internet of Things (IoT) devices.

Difficulty:	State:	Damage:	VA:
Low – Medium	Locked	Low – High	Alexa/Siri/Google

Table 10: Classification of attacks based on controlling Smart Home or IoT devices.

Smart home devices are also one of the possible directions of attack on the voice assistants. As more and more devices can be connected to voice-controlled systems, the following devices are under threat of being misused:

- Smart Televisions
- Thermostats
- Lights
- Locks
- Cameras

Because of the large variation of different devices with different functionalities, it is not possible to clearly determine the possible amount of damage that could be caused by such attack. A case in which an attacker lights a light bulb in a room might not have as many financial or other consequences as a case in which a perpetrator unlocks the front door of a house or sets the thermostat to the highest possible temperature.

9) Managing calendars, schedules, to-do lists, timers and routines.

Difficulty:	State:	Damage:	VA:
Low – Medium	Locked	Low – Medium	Alexa/Siri/Google

Table 11: Classification of attacks based on managing daily events and timers.

All assistants can store and manage large amounts of information that may be of little to no value to an attacker. Most of these functionalities would probably only be an inconvenience to the victim, but there are some that could be considered vulnerable. For example, information about a person's schedule could provide his whereabouts, which could then be exploited by the criminal.

10) Making online purchases.

Difficulty:	State:	Damage:	VA:
High	Unlocked	High	Alexa/Google

Table 12: Classification of attacks based on making purchases.

Making online purchases is a broad term and for each voice assistant it can mean something different. In some countries, Google Assistant allows users to authorize payments and make in-app purchases through Google Play. Alexa allows users to manage their shopping cart and make purchases through amazon shop. In the case of Siri, Apple has decided not to provide purchases through the voice assistant due to privacy concerns and unreliability of authentication[5].

11) Retrieving information.

Difficulty:	State:	Damage:	VA:
Low – Medium	Unlocked/Locked	Low – Medium	Alexa/Siri/Google

Table 13: Classification of attacks based retrieving information.

The information provided to the assistant is stored differently by different systems. Google Assistant can remember specific information such as the front door code or package shipments [14]. The process of storing and retrieving information is as follows

"Hey Google, remember that my front door code is 1110."

"Hey Google, what's my front door code."

[5] https://iphoneislam.com/en/2022/04/apple-prevent-siri-purshare-because-of-privacy-concerns/104126

With this request it is possible to get a response containing the code from the front door. This function is also available when the phone is locked. Alexa stores the same information in its notes and therefore this information cannot be obtained by asking

"Alexa, what's my front door code."

but it is necessary to use the question

"Alexa, read my front door code note." or *"Alexa, read my notes."*

Siri stores this information in the same way as Alexa, so reading the notes is required to retrieve the information. However, this is not possible on the iPhone 7 from a locked state and therefore the device must be unlocked.

6.2 Deepfake Tools

Synthetic recordings can be created by several different techniques. Some of these are mentioned in Section 4.1. For this experiment, I decided to use an online tool called Resemble AI[6], which allows the creation of text-to-speech, speech-to-speech recordings. This web application provides a simple interface for creating deepfakes but mainly allows the user to clone his/her voice. The reasons why this tool was used are as follows:

- **Simplicity** – User friendly interface and no technical knowledge needed.
- **Quality** – The deepfakes created were of sufficient quality for this experiment.
- **Effects** – Tool offers the use of effects that increase the quality of the created deepfakes.
- **Time** – Computer learned to speak like me very quickly.
- **Cost** – Tool is free but also provides paid features.

The tool is very simple to use and voice cloning can be handled by anyone with a basic knowledge of the English language. User Interface is very simple to use and intuitively guides the user through the whole process. Even though this is a free version of the application, the quality of the deepfakes is sufficient to create requests or orders for voice assistants. The fact that the tool includes the addition of various effects to the speech such as emphasis, phoneme or emotion also helps. The most important effects are pause and the effect that allows to read the text character by character. Another advantage was the speed at which the computer was able to learn to speak like me in a very short time.

[6] https://www.resemble.ai/

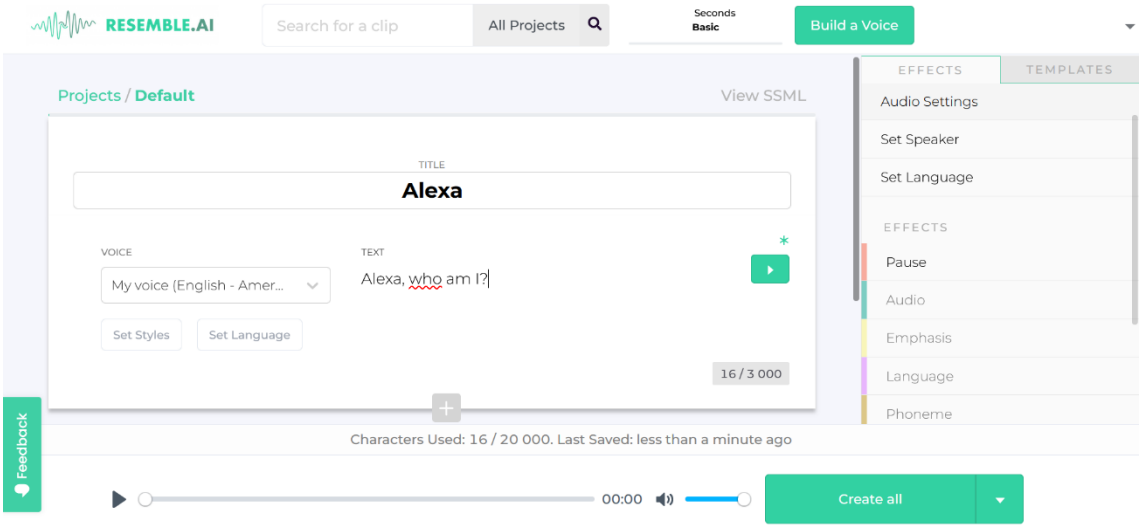

Figure 6: Resemble.ai user interface for creation of a text-to-speech deepfake. The controller in the middle part of the figure is used for text inputs and voice selection. On the right of the figure there are effects that can be used to enrich the output voice.

6.3 Voice Assistant setup

All voice assistants used in the experiment have some form of voice-based verification of identity. This feature needs to be turned on and the user undergoes a short enrollment process in which the user says a few sentences displayed in the mobile application. These sentences will then be used to create a voice template of the user, according to which the voice will be recognized.

6.3.1 Google Assistant

Google Assistant uses so-called "Voice Match" to verify people, where Google claims that

"When you turn on Voice Match, you can teach Google Assistant to recognize your voice so it can verify who you are before it gives you personal results.[7]*"*

Google uses the following phrases to create a Voice Match

"Ok Google, what's the weather tomorrow?"
"Ok Google, where is the nearest post office."
"Hey Google, remind me to buy flowers."
"Hey Google what time is the sunrise."

[7] https://support.google.com/assistant/answer/9071681

6.3.2 Siri

Similar to Google, Siri uses voice recognition for "Personal Requests" for which Apple claims that

"When you set up Personal Requests, you can do even more with voice recognition—like send and read messages, check your calendar, make phone calls, and more.[8]*"*

The voice recognition also required a short training with the following sentences

"Hey Siri."
"Hey Siri, send a message."
"Hey Siri, what's the weather like today."
"Hey Siri, set a timer for three minutes."
"Hey Siri, play some music."

6.3.3 Alexa

Amazon, as the developer of the Alexa voice assistant, has chosen the name "Alexa voice ID" for voice recognition. This feature should also recognize speaker's voice and Amazon states that

"You can create an Alexa voice ID for a personalized experience when Alexa recognizes your voice. With Alexa voice ID, Alexa can call you by name and provide enhanced personalization.[9]*"*

Alexa voice ID also required several phrases to set up

"Alexa"
"Alexa, what's the temperature outside."
"Alexa, add milk to my shopping list."
"Alexa, ignore the incoming call."

6.4 Test environment

This section is intended to clearly define the versions of the software on which the experiment has been performed because of possible version-based improvements. It should also be said that prior to the start of the experiment, all devices had been in use for approximately 10 days with an estimated number of requests at 5 per day. If the voice assistant improves with each request [10] the results could be affected by the length of usage.

[8] https://support.apple.com/sk-sk/guide/homepod/apd1841a8f81/homepod
[9] https://www.amazon.com/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X

Device:	Software:	Version:
Xiaomi Mi Smart Speaker	Google Home	2.67.1.8
Google Nest Mini 2	Google Home	2.67.1.8
iPhone 7	iOS	15.7.1
Amazon Echo Dot 3	Amazon Alexa	2023.9
Google Pixel 4a	Android	13

Table 14: Table showing versions of software, that is used to connect to device.

7. Experiment execution

This chapter describes the process of cloning of my voice and later the creation of deepfakes. These deepfakes are then used to verify that different devices with the same voice assistant are responding and recognizing the voice in the same way. Afterwards, the process of testing the voice verification that the voice assistant should have is described. Next, an attempt is made to perform some of the attacks mentioned in Section 6.1. Finally, possible defenses that could reduce the success of voice spoofing are suggested.

Attacks were divided into 2 parts, wake word and request. Each test was executed ten times and the results were recorded and used to create the graphs in this chapter. Tests include different sources of human voice whether genuine, replay or deepfake. One male and one female voice, that the device has never heard before are also used. The results may not be completely accurate as all recordings and voices belong to people who are not native English speakers. This has occasionally led to mispronunciation of phrases. Also, all attempts that have failed to trigger the wake word were considered unsuccessful.

7.1 Creating deepfakes

As mentioned in Section 6.2, the online tool Resemble AI was used to create deepfakes. The first thing to do was to create a copy of my voice. The procedure in Resemble AI was very simple and intuitive. In the web application, all one had to do was click on the "Build a Voice" button and follow the instructions. The main focus of the procedure was reading sentences displayed on the screen. After reading each sentence, the tool evaluated the quality of the given recording, which could be replaced if necessary. All the sentences that produced my voice were rated by the application as being of high quality. Only 25 recordings were needed to create the voice, which took about 7 minutes to record. My voice was ready within 20 minutes of uploading the recordings.

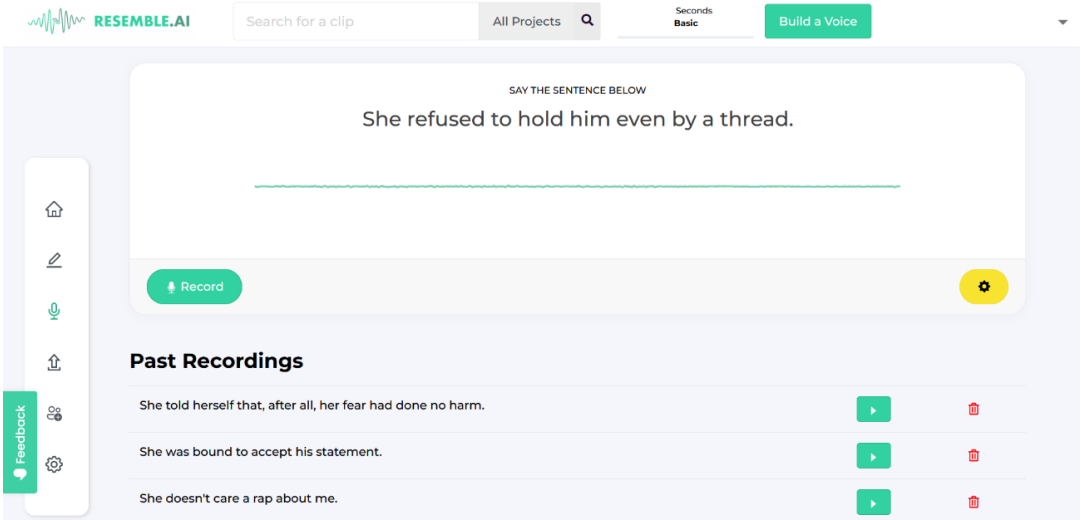

Figure 7: Creation of a voice using the Resemble.ai tool. In the middle of the figure, a sentence to be read. At the bottom a transcript of sentences from previous recordings can be seen.

After the voice was created, it became possible to create the first deepfake. Thanks to the clear user interface, this step was also very trivial. All that was needed was to select the voice that would be used to create the synthesized recording, the content of which was located in a text field to the right. To make the recordings more successful, pause effects have been added to the sentences. For some deepfakes, the "Say each character" effect was also used.

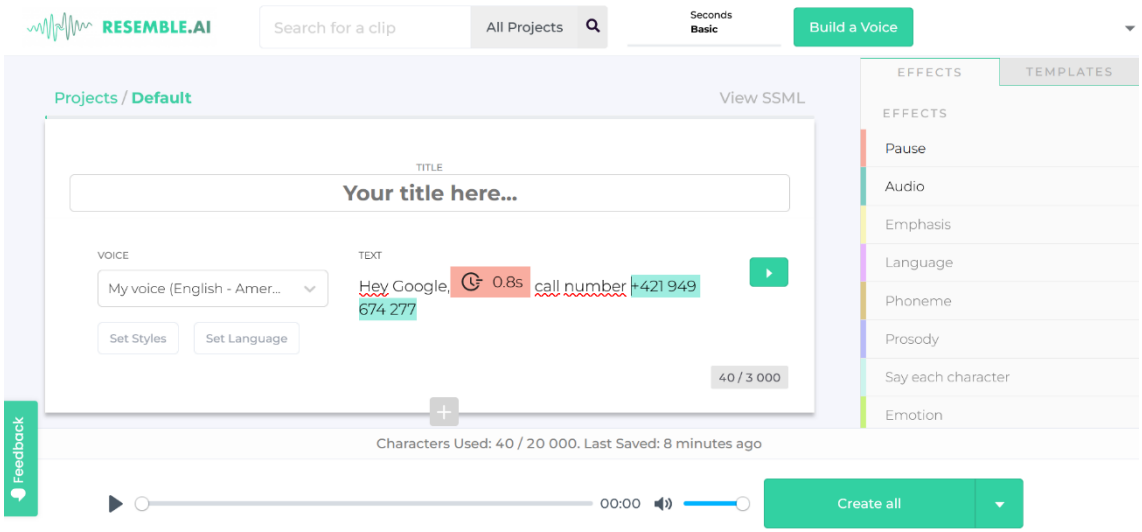

Figure 8: Creation of a sentence containing "Hey Google, call number +421 949 674 277". In the middle of the figure on the left, select the voice with the name "My voice". To the right of the voice, the sentence to be synthesized with a pause effect added after the word "Google" and the "Say each character" effect applied to the phone number for a better chance of being picked up by the voice assistant.

7.2 Comparing different devices with the same Voice Assistant

Before testing individual attacks, it was necessary to see if different devices using the same voice assistant react the same way. It was also to be determined whether they were equally capable of countering attacks from different sources. Google Pixel 4a, Google Nest Mini 2, Xiaomi Mi Smart Speaker, which use Google Assistant, were compared.

The following phrase was used to test these devices

"Hey Google, what time is it"

The devices recognized genuine requests with a high success rate of 90%, with Google Nest Mini 2 device recognizing all 10 attempts. Replay attacks were equally or in one case more successful than genuine request mainly due to the fact that the device did not recognize the wake word. Not recognizing the wake word was the main problem with deepfake recordings and impostors. Other attempts might have failed because of a very short pause between the wake word and the request.

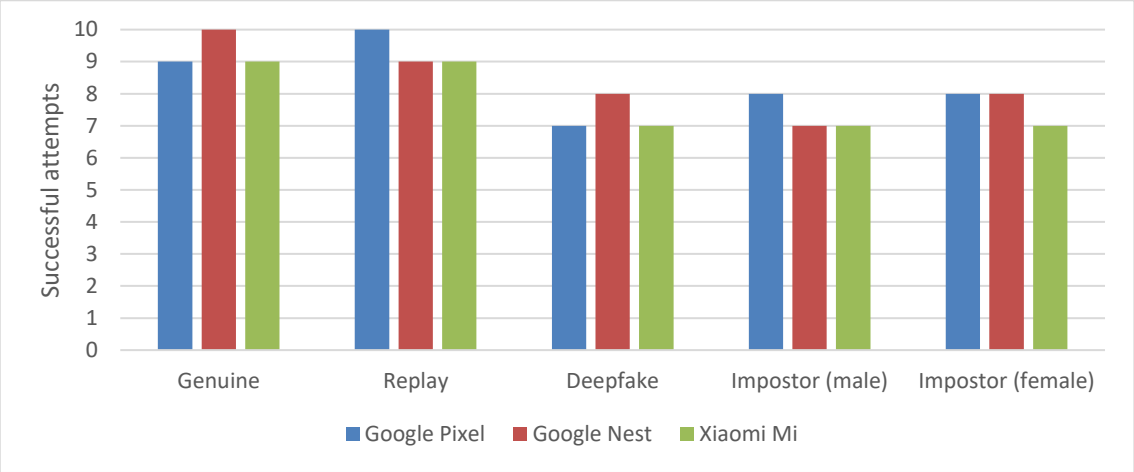

Figure 9: Graph showing the success rate of individual attacks on different devices that use the same voice assistant.

Based on the data obtained from this test, it can be argued that there is no significant difference in the success rate of voice recognition as the differences in the data are minimal.

7.3 Testing voice verification

With each device offering some form of voice-based authentication, it was necessary to verify how robust this technology is. During testing, a following sentence was played to each voice assistant

"Who am I." with the correct wake word.

The assistants responded as follows

Siri: *"You are Oskar."*
Google Assistant: *"Your name is Oskar."*
Alexa: *"I'm talking to Oskar."*

The only assistant who showed any doubt about the speaker's identity was Alexa responding to a voice from Impostor (male) with

"You are probably Oskar." or *"This is Oskar's account."*

To the voice of Impostor (female) Alexa replied

"I think I am talking to Oskar."

Since the sentences use words like "probably" and "I think", one might assume that the device is not quite sure if it is my voice. Even though Alexa showed some uncertainty the attacks were considered successful. Neither Siri nor Google showed any concern about the person's identity.

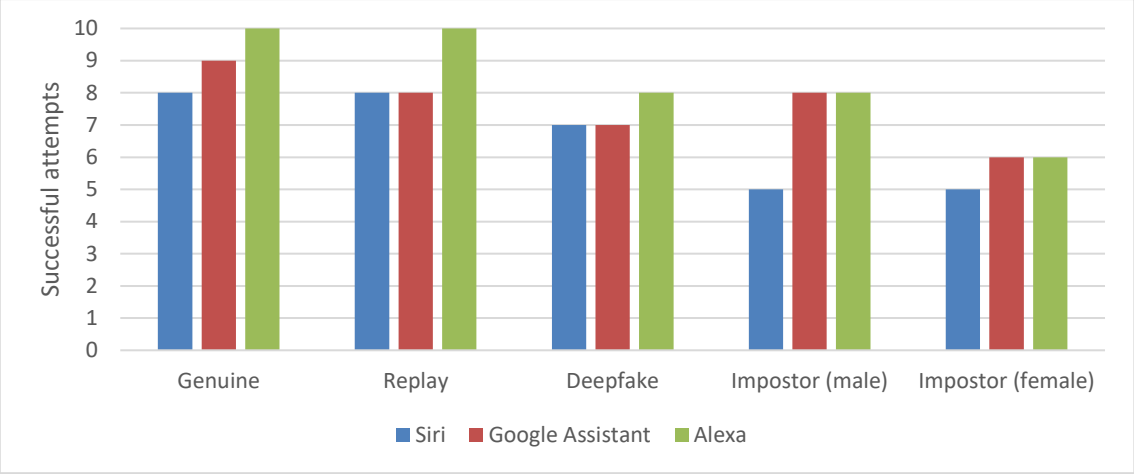

Figure 10: Graph showing the number of responses where the device responded that it was talking to Oskar or claimed the speaker was Oskar.

Based on these findings, it can be concluded that even though the devices use some form of verification, this verification is not very robust. Alexa suspected that it was not my voice but did not reject the potential attacker. Devices with Google Assistant and Siri responded to the unfamiliar voice in the same way as to a genuine request.

7.4 Executing attacks

In this section, execution of some of the attacks mentioned in Section 6.2 is discussed.

7.4.1 Phone calls

Phone calls have only been performed on Google Pixel 4a and iPhone 7 as Alexa has certain restrictions in some countries. The following sentence was played to mobile devices after wake word

"Call number 0949 674 277."

This attack had a relatively low success rate due to a number recognition problem. When dictating numbers character by character, assistants had problems with adding the number 0 between 2 and 7 or interpreting the number 2 as "to". Therefore, a method of saying the number in hundreds was used. This increased the success rate to at least 50% for genuine request.

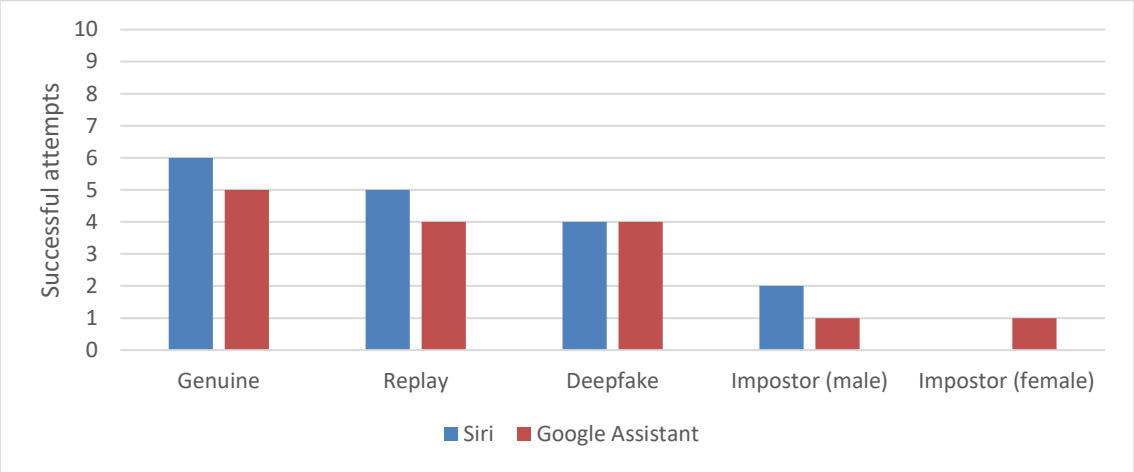

Figure 11: Graph showing the success rate of attacks focused on dialing telephone numbers.

7.4.2 Text messages

The attack via text messages was very difficult to execute as it was necessary to split the message into 3 parts. In addition to specifying the recipient of the message in the second part, it was also necessary to create a third part that contained the content of the message. The attack in which the link to the web page was misspelled was also considered successful, since all attacks failed to spell the link correctly. The transcript for this attack was a wake word followed by

"Send message to number 0949 674 277"

After that assistant asks for the content of the message which was

"Hi, it's me, click on this link www.vut.cz"

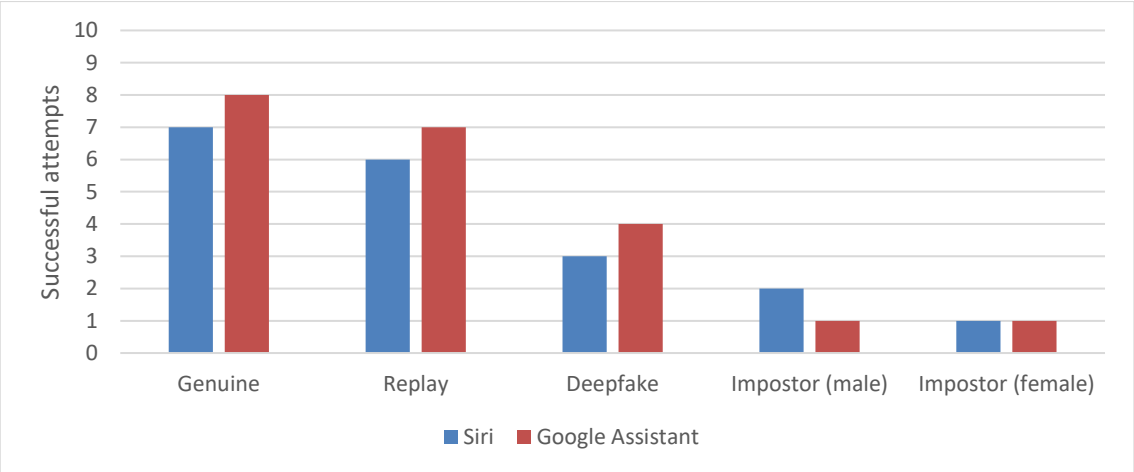

Figure 12: Graph of the success rate of attacks based on text messaging.

During this attack the same issue occurred as with the phone calls attack as devices had trouble recognizing the mobile number correctly. After entering the phone number correctly, it had no further problem recognizing the spoken text except for the link at the end of the message.

7.4.3 Smart home

As there are several Smart Home devices, in this test, smart lights were used to represent Smart devices. Since only one smart device was connected to the assistants, it was straightforward to control. In the case of multiple devices, it would be necessary to specify more precisely which of them should be controlled. To perform this test, the following phrase was used after the wake word

"Turn the light on."

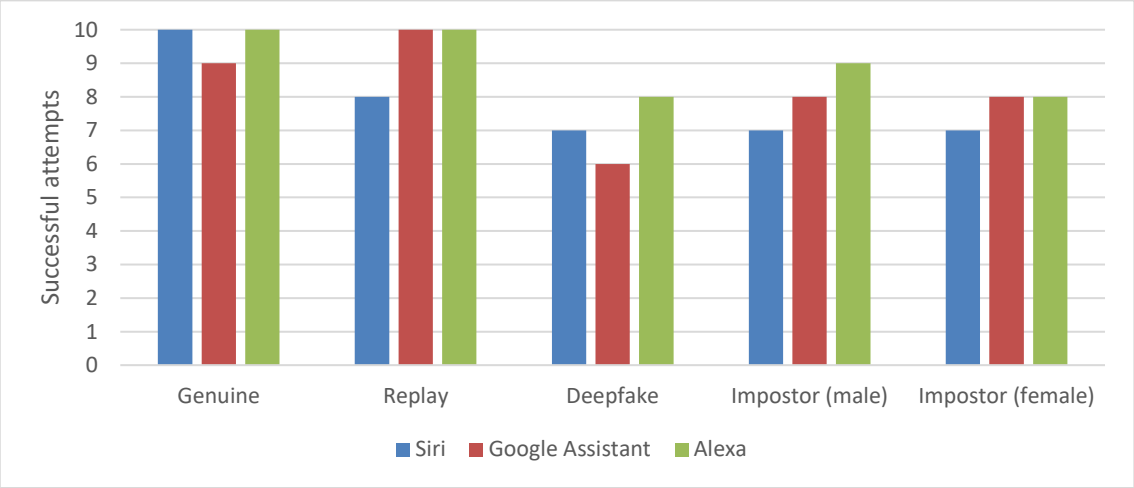

Figure 13: Graph showing the high success rate of attacks on Smart Home devices.

The connected Smart Home device was controlled with a very high success rate across all voice sources. Deepfake has the lowest success rate of all attacks, but all of the failed attempts were due to the failure to recognize the wake word. This means that deepfake recordings that got through the wake word had a 100% success rate.

7.4.4 Retrieving information

Retrieving information from devices is also one of the possible ways to exploit access to a device with voice assistant. Voice assistants like Siri and Alexa store all information in notes. Google Assistant can likewise store information in notes, but it can also remember specific information. In this case, we assume that the attacker knows exactly how to question the device to get the information. For example, if a user were to store information about the code needed to open the front door to the device in the following way

"Hey Google, remember that my front door code is 1110."

A perpetrator would be able to obtain this information by saying

"Hey Google, what is my front door code"

The results of the tests of this feature showed that 7 out of 10 deepfake attacks were successful. When testing a phrase

"Read my notes."

The results were similar with a 63% success rate across all attacks.

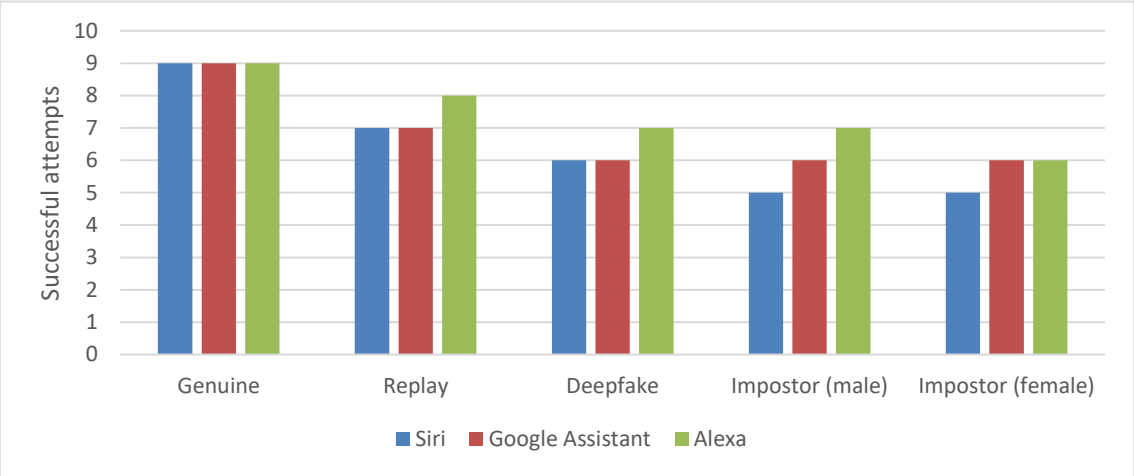

Figure 14: Graph showing the success rate of attacks focusing on extracting information from notes.

7.5 Results

Experiment showed that voice assistants can be deceived by a synthetic voice. It only took 25 sentences to create a copy of my voice, which is considered a relatively small training sample size. On the other hand, the voice was created from high quality sentences without background noise. Also, these were specific sentences selected to properly produce a copy of the voice. This quality of recordings is not likely to be accessible to the attacker, and therefore it can be assumed that a voice made of sentences of poorer quality would not perform as well. To perform an attack on voice assistants, it was not strictly necessary to create a clone of my voice, since the tested voice assistants appeared to have a problem with speaker recognition. Factor that may have influenced the test results is for example pause between requests. Also, the actors used in the experiment are not native English speakers, but they still managed to fool some device at least once in each attack.

	Genuine	Replay	Deepfake	Impostor(M)	Impostor (F)	Sum (%) (ex. gen.)
Voice verification	27/30 (90%)	26/30 (87%)	22/30 (73%)	21/30 (70%)	17/30 (57%)	86/120 (72%)
Phone calls	11/20 (55%)	9/20 (45%)	8/20 (40%)	3/20 (15%)	1/20 (5%)	21/80 (26%)
Text messages	15/20 (75%)	13/20 (65%)	7/20 (35%)	3/20 (15%)	2/20 (10%)	25/80 (31%)
Smart Home	29/30 (97%)	28/30 (93%)	21/30 (70%)	24/30 (80%)	23/30 (77%)	96/120 (80%)
Retrieving information	27/30 (90%)	22/30 (73%)	19/30 (63%)	18/30 (60%)	17/30 (57%)	76/120 (63%)
Sum (%)	109/130 (84%)	98/130 (75%)	77/130 (59%)	69/130 (53%)	60/130 (46%)	

Table 15: Table with complete information about the success rate of all requests. Last column is the sum of all attacks (excluding genuine requests). All percentages are rounded.

	Siri	Google Assistant	Alexa	Sum (%)
Voice verification	25/40 (63%)	29/40 (73%)	32/40 (80%)	86/120 (72%)
Phone calls	11/40 (28%)	10/40 (25%)	x	21/80 (26%)
Text messages	12/40 (30%)	13/40 (33%)	x	25/80 (31%)
Smart Home	29/40 (73%)	32/40 (80%)	35/40 (88%)	96/120 (80%)
Retrieving information	23/40 (58%)	25/40 (63%)	28/40 (70%)	76/120 (63%)
Sum (%)	100/200 (50%)	109/200 (55%)	95/120 (79%)	

Table 16: Table with complete information about the success rate of all attacks (excluding genuine requests) on individual assistants. All percentages are rounded.

	Siri	Google Assistant	Alexa	Sum (%)
Genuine	40/50 (80%)	68/80 (85%)	29/30 (97%)	137/160 (86%)

Table 17: Table showing success rate of genuine requests on voice assistants (including comparison of different devices with the same voice assistants). All percentages are rounded.

The results of the tests have shown that Alexa, which was the only one that showed suspicion about the identity of the attacker, is the most vulnerable to all types of attacks. This could be due to the fact that it is also the best at recognizing all requests since it had the highest success rate with genuine requests. This trend can also be observed with Siri and Google Assistant. Based on the experience gained during the tests, Alexa seemed to be the best at recognizing wake words. Google Assistant sometimes asked for additional information, which would have improved performance in regular usage. Similarly, Siri would ask for additional information or a repetition

of the request on some occasions. However, these additional questions were not answered and the attempt was evaluated as unsuccessful.

7.6 Possible defense methods

From the results of the experiment, it is reasonable to conclude that the protection currently used by voice assistants against voice spoofing is insufficient. As described in Section 4.2 there are several methods to detect voice recordings or to detect synthesized voice.

In paper from 2018, S. Mochizuki et al. [12] propose a phoneme-based pop-noise detection algorithm for voice liveness detection and automatic speaker verification systems, that could be used as a countermeasure against replay and speech synthesis attacks. This algorithm is based on specific characteristics of phonemes related to pop-noise phenomenon, which occurs in natural speech. When pronouncing words, there are specific phoneme groups that differ from each other in their breathing patterns. One requires almost no breathing while the other requires much breathing. The algorithm detects whether the speech contains any pop-noise periods, which are compared with the individual phoneme groups, and based on these it determines whether it is genuine speech or a recording.

H. Tak et al. [46] describe novel graph neural network approach to detection of speech deepfakes based on model-level spectro-temporal attention. This system is called RawGAT-ST and is designed to use a self- attention mechanism to learn the relationship between different spectro-temporal estimates and the most discriminative nodes within the resulting graph. The authors also claim that this method works directly with the raw waveform and was among the least complex of all solutions at the time of publication. In the case of voice assistants, where all requests are sent to a remote server where they are evaluated, a less complex solution could mean faster evaluation times. The speed of request evaluation is very important when using voice assistants, as it is one of the parameters that users look for when purchasing them.

Another method that could be used for audio spoofing detection designed and implemented by M. Alzantot et al. [48] is based on deep residual convolutional. Various feature extraction algorithms can be used to transform the input signal into a 2D representation. The results of these transformations are then sent as input to the convolutional model. This fusion of models has proven successful in comparison with evaluation dataset scores and could be used as one of the possible defense mechanisms against attacks on voice assistants.

Since voice assistants have shown very low protection against voice spoofing, all of the above methods could be implemented in the evaluation server to reduce the success rate of attacks using replay or synthetic recordings.

8. Conclusion

Voice authentication uses a person's voice biometrics to distinguish between people. Such form of authentication can be a significant assistance in data protection. Currently there are several different ways to spoof this type of authentication. Some are based on imitating the voice of the victim by another person or by directly recording the victim's voice. With technical advances, new ways to spoof such authentication have emerged. One of them is the creation of a synthetic voice of a person or in other words deepfake. There are several different ways to create deepfakes, the most famous being text-to-speech and voice conversion.

As the use of voice assistants, which are vulnerable to possible voice spoofing, is rapidly expanding, the question arises about the security level of devices with these assistants. The majority of the most widely used assistants use a speaker recognition system for voice recognition, which serves to personalize the response. In this work, possible attacks on individual voice assistants are described, and some of them are executed. The experiment is based on testing individual attacks on individual assistants using replay attacks, attacks using deepfake recordings, and attempts to spoof voice verification using voices unfamiliar to the device.

The aim of this work was to create deepfakes and test the robustness of voice assistants to these synthetically created media. The fact that the robustness of voice assistants to deepfake attacks is low has been demonstrated in the experiment, since every single one of the selected attacks was successfully performed. Deepfake attacks performed better than the Impostor attacks but still lagged behind the replay attacks. Since all types of attacks could mean a possible financial loss for the victim, the functionality of the voice assistants should be secured. Lastly, some of the possible methods that could be used to protect against deepfakes or other attacks are suggested.

Bibliography

[1] J. Wayman, A. Jain, D. Maltoni, and D. Maio, “An introduction to biometric authentication systems,” in Biometric Systems, pp. 1–20, Springer, 2005.

[2] Asaolu, O. S. & Folorunso, Comfort & Popoola, Oluwatoyin. (2019). A Review of Voice-Base Person Identification: State-of-the-Art. Journal of Engineering Technology. 3. 36-57. doi:10.20370/2cdk-7y54

[3] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," 2017 12th System of Systems Engineering Conference (SoSE), Waikoloa, HI, USA, 2017, pp. 1-6, doi: 10.1109/SYSOSE.2017.7994971.

[4] Gunson, Nancie & Marshall, Diarmid & Mcinnes, Fergus & Jack, Mervyn. (2011). Usability evaluation of voiceprint authentication in automated telephone banking: Sentences versus digits. Interacting with Computers. 23. 57-69. doi: 10.1016/j.intcom.2010.10.001.

[5] A. N. Kataria, D. M. Adhyaru, A. K. Sharma and T. H. Zaveri, "A survey of automated biometric authentication techniques," *2013 Nirma University International Conference on Engineering (NUiCONE)*, Ahmedabad, India, 2013, pp. 1-6, doi: 10.1109/NUiCONE.2013.6780190

[6] Krawczyk, S., Jain, A.K. (2005). Securing Electronic Medical Records Using Biometric Authentication. In: Kanade, T., Jain, A., Ratha, N.K. (eds) Audio- and Video-Based Biometric Person Authentication. AVBPA 2005. Lecture Notes in Computer Science, vol 3546. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/11527923_115

[7] Y. Tabet and M. Boughazi, "Speech synthesis techniques. A survey," *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, Tipaza, Algeria, 2011, pp. 67-70, doi: 10.1109/WOSSPA.2011.5931414.

[8] Edison Research and NPR *The Smart Audio Report 2022*. [cit. 2023-1-9]. Available at: https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/

[9] jack souma (2023). speaker recognition biometric system matlab code (https://github.com/SamiHagrai/speaker-recognition-biometric-system-matlab-code), GitHub. Retrieved January 10, 2023.

[10] P. Cheng and U. Roedig, "*Personal Voice Assistant Security and Privacy—A Survey*," in *Proceedings of the IEEE*, vol. 110, no. 4, pp. 476-507, April 2022, doi: 10.1109/JPROC.2022.3153167.

[11] S. Subhash, P. N. Srivatsa, S. Siddesh, A. Ullas and B. Santhosh, "*Artificial Intelligence-based Voice Assistant*," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, UK, 2020, pp. 593-596, doi: 10.1109/WorldS450073.2020.9210344.

[12] Mochizuki, S., Shiota, S., & Kiya, H. (2018). Voice liveness detection using phoneme-based pop-noise detector for speaker verification. In *Proc. Odyssey Speaker Lang. Recognit. Workshop*.

[13] A. Qais, A. Rastogi, A. Saxena, A. Rana and D. Sinha, "Deepfake Audio Detection with Neural Networks Using Audio Features," *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, Hyderabad, India, 2022, pp. 1-6, doi: 10.1109/ICICCSP53532.2022.9862519.

[14] Sattelberg W. *Assistant Memory looks like it's about to pick up some great new ways to keep track of stuff* [online]. 2021-3-24 [cit. 2023-4-3]. Available at: *https://www.androidpolice.com/2021/03/24/assistant-memory-looks-like-its-about-to-pick-up-some-great-new-ways-to-keep-track-of-stuff/*

[15] Kounoudes, Anastasis & Kekatos, Vassilis & Mavromoustakos, Stephanos. (2006). *Voice Biometric Authentication for Enhancing Internet Service Security*. 1020 - 1025. doi: 10.1109/ICTTA.2006.1684514.

[16] FIRC, Anton. *Applicability of Deepfakes in the Field of Cyber Security*. Brno, 2021. Master's Thesis. Brno University of Technology, Faculty of Information Technology. 2021-06-22. Supervised by Malinka Kamil. Available at: https://www.fit.vut.cz/study/thesis/23761/

[17] Bateman, J. (2020*). Deepfakes and synthetic media in the financial system: Assessing threat scenarios.* Carnegie Endowment for International Peace.

[18] Beigi, Homayoon. (2012). *Speaker Recognition: Advancements and Challenges.* doi: 10.5772/52023.

[19] Voiceprint: The New WeChat Password [online]. 2015-5-21 [cit. 2023-3-27] Available at: https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/

[20] S. Millward. Open Sesame: Baidu Helps Lenovo Use Voice Recognition to Unlock Android Phones [online]. 2012-11-30 [cit. 2023-3-27] Available at: https://www.techinasia.com/baidu-lenovo-voice-recognition-android-unlock

[21] T. Arif, A. Javed, M. Alhameed, F. Jeribi and A. Tahir, "Voice Spoofing Countermeasure for Logical Access Attacks Detection," in IEEE Access, vol. 9, pp. 162857-162868, 2021, doi: 10.1109/ACCESS.2021.3133134.

[22] Gunendradasan, Tharshini & Irtza, Saad & Ambikairajah, Eliathamby & Epps, Julien. (2019). Transmission Line Cochlear Model Based AM-FM Features for Replay Attack Detection. 6136-6140. doi: 10.1109/ICASSP.2019.8682771.

[23] Hautamäki, Rosa & Kinnunen, Tomi & Hautamäki, Ville & Leino, Timo & Laukkanen, Anne-Maria. (2013). I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. doi: 10.21437/Interspeech.2013-289.

[24] Mache, Suhas & Baheti, Manasi & Mahender, C. & Professor, Asst. (2015). Review on Text-To-Speech Synthesizer. International Journal of Advanced Research in Computer and Communication Engineering. 4. 54-59. doi: 10.17148/IJARCCE.2015.4812.

[25] Walczyna T, Piotrowski Z. Overview of Voice Conversion Methods Based on Deep Learning. Applied Sciences. 2023; 13(5):3100.
Available at: https://doi.org/10.3390/app13053100

[26] Zhou, J., Hai, T., Jawawi, D.N.A. et al. Voice spoofing countermeasure for voice replay attacks using deep learning. J Cloud Comp 11, 51 (2022).
Available at: https://doi.org/10.1186/s13677-022-00306-5

[27] Westerlund, Mika. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review. 9. 39-52. doi: 10.22215/timreview/1282.

[28] Alattas, Khalid & Bayoumi, Magdy. (2020). Artificial Intelligence in Deepfake Technologies Based on Supply Chain Strategy.

[29] Groh, M. Detect DeepFakes: How to counteract misinformation created by AI [online]. 2023 [cit. 2023-4-6]. Available at: https://www.media.mit.edu/projects/detect-fakes/overview/

[30] Hancock JT., Woodworth MT, Goorha S. See no evil: the effect of communication medium and motivation on deception detection. Group Decision Negotiation 2010; 19:
327–343

[31] Vaccari C., Chadwick A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news.
Social Media and Society 2020; 6:1–13.

[32] Brewster T. *Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find* [online]. 2021-10-14 [cit. 2023-4-7]. Available at: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

[33] Elgan M. *Synthetic Media Creates New Social Engineering Threats* [online]. 2023-1-19 [cit. 2023-4-12]. Available at: https://securityintelligence.com/articles/synthetic-media-new-social-engineering-threats/

[34] Damiani J. *A Voice Deepfake Was Used To Scam A CEO Out Of $243,000* [online]. 2019-9-3 [cit. 2023-4-12]. Available at: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000

[35] Cuthbertson A. *AI clones child's voice in fake kidnapping scam* [online]. 2023-4-13 [cit. 2023-4-16]. Available at: https://www.independent.co.uk/tech/ai-voice-clone-scam-kidnapping-b2319083.html

[36] Khanjani, Z., Watson, G., & Janeja, V.P. (2021). How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey.

[37] Seyed Hamidreza Mohammadi and Alexander Kain. 2017. An overview of voice conversion systems. Speech Communication 88 (2017), 65–82. Available at: https://doi.org/10.1016/j.specom.2017.01.008

[38] Matthew B. Hoy (2018) Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, Medical Reference Services Quarterly, 37:1, 81-88, DOI: 10.1080/02763869.2018.1404391 Available at: https://doi.org/10.1080/02763869.2018.1404391

[39] Google. Explore what you can do with Google Nest or Home devices [online]. 2023 [cit. 2023-4-17]. Available at: https://support.google.com/googlenest/answer/7130274

[40] Apple Inc. Siri [online] 2023 [cit. 2023-4-17]. Available at: https://www.apple.com/siri/

[41] Amazon.com. Alexa features [online] 2023 [cit. 2023-4-17]. Available at: https://www.amazon.com/b?ie=UTF8&node=21576558011

[42] Cano, Pedro & Loscos, Alex & Bonada, Jordi & Boer, Maarten & Serra, Xavier. (2002). Voice Morphing System for Impersonating in Karaoke Applications.

[43] Huang, Dong-Yan & Rahardja, Susanto & Ong, Ee. (2010). High Level Emotional Speech Morphing Using STRAIGHT.

[44] Pianese, A., Cozzolino, D., Poggi, G., & Verdoliva, L. (2022, December). Deepfake audio detection by speaker verification. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-6). IEEE.

[45] Janicki, A. (2015) Spoofing countermeasure based on analysis of linear prediction error. Proc. Interspeech 2015, 2077-2081, doi: 10.21437/Interspeech.2015-470

[46] Tak, Hemlata & Jung, Jee-Weon & Patino, Jose & Kamble, Madhu & Todisco, Massimiliano & Evans, Nicholas. (2021). End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. 1-8.
doi: 10.21437/ASVSPOOF.2021-1.

[47] Gonfalonieri A. *How Amazon Alexa works? Your guide to Natural Language Processing (AI)* [online]. 2018-11-21 [cit. 2023-4-17]. Available at: https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3

[48] Alzantot, M., Wang, Z., & Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. arXiv preprint arXiv:1907.00501.

Appendix A

Contents of storage media

Included storage media contains following contents:

- replay_recordings: replay recordings used in the experiment
- synthetized_recordings: deepfakes used in the experiment
- src: docx used to create this pdf

Appendix B

Recordings used in the experiment

Following phrases were recorded and synthetized for the experiment:

- “Alexa”
- “Call number 0949 674 277”
- “What’s my front door code”
- “Hey Google”
- “Hey Siri”
- “Hi, it’s me, please click on this link www.vut.cz”
- “Read my notes”
- “Read my notifications”
- “Send message to number 0949 674 277”
- “Turn the light on”
- “What time is it”
- “Who am I?”