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 PROTI DEEPFAKES

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **PETR KAŠKA**
AUTOR PRÁCE

SUPERVISOR **Mgr. KAMIL MALINKA, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2024

Bachelor's Thesis Assignment

154457

Institut:	Department of Intelligent Systems (DITS)
Student:	**Kaška Petr**
Programme:	Information Technology
Title:	**Resilience of Biometric Authentication of Voice Assistants Against Deepfakes**
Category:	Security
Academic year:	2023/24

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 three selected assistants to voice spoofing using deepfakes.
5. Implement and evaluate the experiment.
6. Discuss the results and propose possible defense methods.

Literature:

- Fake It: Attacking Privacy Through Exploiting Digital Assistants Using Voice Deepfakes Ubert, Justin. Marymount University ProQuest Dissertations Publishing, 2023. 30486781.
- FIRC Anton, MALINKA Kamil a HANÁČEK Petr. Deepfakes as a threat to a speaker and facial recognition: an overview of tools and attack vectors. *Heliyon*, roč. 9, č. 4, 2023, s. 1-33. ISSN 2405-8440.
- 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:	1.11.2023
Submission deadline:	9.5.2024
Approval date:	6.11.2023

Abstract

Voice assistants (Apple Siri, Amazon Alexa, Google-assistant, Samsung Bixby) supporting voice control offer more and more possibilities to make all our daily activities easier. People give them access to data and information to take full advantage of all these features. Along with the rapidly developing voice deepfake technology, there is a big threat in the area of misusing deepfakes to trick smart voice assistants. An attacker can record the victim's voice, synthesize the voice and create a recording of some command to trick the assistant in order to harm the victim. The aim of this work is to design an experiment that will simulate attacks, performed by synthetic voice, on voice assistants and then evaluate their defensiveness. The conducted experiment confirms the initial hypothesis of the vulnerability of voice assistants to deepfake attacks and the results are very alarming with an overall success rate of 90% indicating insufficient defense of voice assistants and require the implementation of additional countermeasures to prevent the risk of misuse as the number of voice assistants in active use is rapidly increasing.

Abstrakt

Hlasoví asistenti (Apple Siri, Amazon Alexa, Google-assistant, Samsung Bixby) podporující hlasové ovládání nabízejí stále více možností, jak nám usnadnit všechny každodenní činnosti. Lidé jim umožňují přístup k datům a informacím, aby mohli všechny tyto funkce plně využívat. Spolu s rychle se rozvíjející technologií hlasového deepfake se objevuje velká hrozba v oblasti zneužití deepfakes k oklamání chytrých hlasových asistentů. Útočník může nahrávat hlas oběti, syntetizovat hlas a vytvořit nahrávku nějakého příkazu, aby oklamal asistenta za cílem poškodit oběť. Cílem této práce je navrhnout experiment, který bude simulovat útoky, provedené syntitickým hlasem, na hlasové asistenty a následně vyhodnotit jejich obranyschopnost. Provedený experiment potvrzuje výchozí hypotézu o zranitelnosti hlasových asistentů vůči deepfake útokům a výsledky jsou velice alarmující s celkovou úspěšností 90% naznačující nedostatečnou obranu hlasových asistentů a vyžadují zavedení dalších protiopatření, která by zabránila riziku zneužití, protože počet hlasových asistentů v aktivním používání rychle roste.

Keywords

Deepfakes, Speaker Recognition, Voice Assistants, Cyber Security, Security Analysis, Spoofing Attacks

Klíčová slova

Deepfakes, Rozpoznávání Mluvčího, Hlasoví Asistenti, Kybernetická Bezpečnost, Bezpečnostní Analýza, Spoofing Útoky

Reference

KAŠKA, Petr. *Resilience of Biometric Authentication of Voice Assistants against Deepfakes.* Brno, 2024. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.

Rozšířený abstrakt

Digitální asistenti, známí také jako virtuální asistenti, inteligentní osobní asistenti nebo asistenti s umělou inteligencí, jsou díky své rostoucí sofistikovanosti a schopnostem stále populárnější. Tito asistenti jsou integrováni do zařízení, jako jsou reproduktory, mobilní telefony a webové služby, a využívají pokročilou umělou inteligenci a algoritmické přístupy k provádění úkolů za uživatele, zodpovídání otázek, vedení konverzace s uživateli a uchovávání informací pro vydávání připomínek a varování na základě omezení prostředí, jako je čas a poloha.

Virtuální asistenti jsou nastaveni tak, že čím více osobních informací jim uživatel o sobě poskytne, tím více funkcí a vlastností odemkne. Další významnou oblastí, v níž se virtuální asistenti posouvají vpřed, je jejich integrace s internetem věcí (IoT), která umožňuje vytvořit inteligentní domácnost. Taková domácnost automatizuje a tím zjednodušuje pro uživatele mnoho každodenních činností.

Někteří hlasoví asistenti jsou vybaveni technologií rozpoznávání hlasu, kterou může uživatel volitelně zapnout a tím zjednoduší použití systému. Taková technologie umožňuje systému rozpoznat identitu uživatele a následně rozhodnout, zda tomuto uživateli povolí nebo odepře přístup. Tato technologie však neustále soupeří s různými metodami útočníků, kteří se pokoušejí o neoprávněný přístup do systému. Jednou z těchto metod jsou útoky využívající deepfakes.

Deepfake je syntetické médium, které je vytvořeno metodou umělé inteligence zvanou deep learning. Spočívá v tom, že umělá inteligence zpracovává skutečné vstupy a upravuje je do podoby, která se nikdy ve skutečnosti nestala. Postupem času je technologie deepfake stále dostupnější veřejnosti a dnes je v takové podobě, že i člověk s minimálními znalostmi IT je schopen vytvořit vlastní deepfake během několika desítek minut a následně tento deepfake šířit po sociálních sítích s cílem zmást společnost nebo zaútočit na biometrický systém, jako je hlasový asistent.

Z těchto důvodů vyvstává otázka odolnosti hlasových asistentů vůči současným technologiím deepfake. Cílem této práce je vytvořit empirickou studii, která vyhodnotí odolnost čtyř nejpoužívanějších hlasových asistentů vůči deepfakes. Samostatně jsme zkoumali odolnost každého asistenta proti útokům replay a deepfake spoofing.

Útoky replay jsme použili pro srovnání s útoky deepfake a proto, že jsou nejsnáze realizovatelné.

Pro deepfake útoky jsme vybrali 4 nástroje pro jejich vytvoření. Dva z nich byly komerční a dva byly free software. V prvním kroku jsme shromáždili hlasové vzorky od 72 respondentů, z nichž jsme pak vytvořili syntetické příkazy z každého ze 4 nástrojů. Poté jsme každého účastníka zaregistrovali do každého hlasového asistenta. Dále jsme hlasovým asistentům přehráli syntetickou nahrávku z každého nástroje. Nakonec jsme pro lepší srovnání výsledků provedli útok přehrání nahrávky .

Přínosy této bakalářské práce jsou uvedeny zde:

- Byla prokázána zranitelnost čtyř hlasových asistentů vůči útokům založeným na deepfake a přehráváním hlasu.
- Vyhodnocení vhodnosti vybraných nástrojů pro syntézu řeči pro tento typ útoku.
- Byl vytvořen soubor dat deepfake obsahující anglickou řeč.
- Byl navržen a proveden experiment k ověření vlastností hlasových asistentů.

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. Kamila Malinky Ph.D. The supplementary information was provided by Firc Anton, Ing. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

. . . .
Petr Kaška
April 25, 2024

Acknowledgements

I would like to thank my supervisor Mgr. Kamil Malinka Ph.D. for his valuable guidance and expert advice throughout the process of writing this thesis.

I would also like to express my gratitude to my family and girlfriend for their support and patience during not only my studies but also throughout my life.

Finally, I would like to thank everyone involved in the experiment for their willingness and SCHOTT for providing some of the respondents.

Contents

1 Introduction **3**

2 Deepfake **5**
2.1 What are deepfakes 5
2.2 Voice deepfakes creation 6
2.2.1 Text to speech 6
2.2.2 Voice conversion 8
2.2.3 Zero-shot voice cloning 8
2.2.4 End-to-end voice cloning 9
2.3 Use of deepfakes 9
2.3.1 Positive application 9
2.3.2 Negative application 10

3 Voice Assistants **12**
3.1 Voice assistant setup architecture 12
3.2 Analyzing the voice assistant command 14
3.3 Generic workflow 15
3.4 Authentication 15
3.4.1 Authentication process 16
3.4.2 Voice assistant authentication 16
3.5 Representatives of voice assistants 17
3.5.1 Amazon Alexa 17
3.5.2 Apple Siri 18
3.5.3 Google Assistant 19
3.5.4 Bixby 19

4 Related Work **20**
4.0.1 Spoofing Voice Assistants 20
4.0.2 Spofing attacks on biometrics systems 21

5 Experiment Design **23**
5.1 Attacker Model 23
5.2 Used voice Assistants 25
5.3 Used tools for creating voice deepfakes 25
5.3.1 ResembleAI 26
5.3.2 CoquiAI 27
5.3.3 TorToiSe 27
5.3.4 XTTS 28

5.4 Experiment participants . . . 29
5.5 Details of the setup . . . 29
5.5.1 Experiment to verify assistant characteristics . . . 29
5.6 Sample collection . . . 30

6 Course of the Experiment **32**
6.1 Participants in the experiment . . . 32
6.2 Demographic distribution of participants . . . 32
6.3 Course of the experiment . . . 33
6.4 Metrics used . . . 34

7 Experiment Evaluation **36**
7.1 Main results . . . 36
7.2 Secondary results and their interpretation . . . 37
7.3 Defence Strategies . . . 38

8 Discussion **40**
8.1 Observations . . . 40

9 Conclusion **41**

Bibliography **42**

A Contents of the enclosed memory Media **49**

B List of recordings that were used in the experiment **50**

Chapter 1

Introduction

Digital assistants, also known as Virtual Assistants, Intelligent Personal Assistants, and Artificial Intelligence Assistants (VAs), are becoming increasingly popular due to their growing sophistication and capabilities. These assistants are integrated into devices such as speakers, mobile phones and web services and use advanced AI and algorithmic approaches to perform tasks for individuals, answer questions, maintain conversations with users, and retain information for issuing reminders and warnings based on environmental constraints like time and location.

VAs are set up so that the more personal information a user gives them about themselves, the more functionality and features they unlock. Another significant area in which VAs are moving forward is their integration with the Internet of Things (IoT) to create a smart home. Such a household will automate a lot of daily activities for the user.

Some VAs have speaker recognition technology that the user can optionally turn on to increase the system's performance. Such technology allows the system to recognize the identity of the user and then decide whether to grant or deny access to that user. However, this technology is in a constant race with various methods of attackers for unauthorized access to the system. One of these methods are attacks utilizing deepfakes.

Deepfakes are a synthetic medium that is created by an artificial intelligence method called deep learning. It involves artificial intelligence processing real input and modifying it into things that never really happened. As time goes on, deepfake technology is becoming more and more accessible to the public, and today it's in such a form that even a person with minimal IT knowledge is able to create their own deepfake within tens of minutes and then spread the deepfake across social media to confuse society or to target a biometric system like the VA.

These reasons raise the question of the resilience of voice assistants to current deepfake technologies, this thesis is designed to create an empirical study to evaluate the deepfake resilience of four of the most widely used voice assistants. Separately, we investigated the resilience of each assistant against replay and deepfake spoofing attacks.

We used replay attacks to compare with deepfake attacks and because they are the easiest to implement.

For deepfake attacks, we selected 4 tools to craft them. Two of them were commercial and two were free software. We then collected voice samples from 72 participants in the experiment from which we then created synthetic commands from each tool. Next, we registered each participant to each voice assistant. We then played the voice assistants a synthetic recording from each instrument. Then, for better comparison, we also played the voice assistants the recording to perform a replay attack .

The contributions of this bachelor thesis are listed here:

- The vulnerability of four voice assistants to voice-based deepfake and replay attacks was demonstrated.
- Evaluation of the suitability of the selected speech synthesis tools for this type of attack.
- A deepfake dataset containing English speech was created.
- An experiment to verify the property of voice assistants was designed and carried out.
- The main source of data from the experiment for the Esorics 2024 [1] conference paper.

An explanation of what deepfakes are, the creation of voice deepfakes, and their positive and negative effects on society are described in Chapter 2. In Chapter 3 the general theory on voice assistants, the process of voice authentication, and the individual assistants used in the experiment are described in detail. In Chapter 4 we introduce papers dealing with similar topics and describe how this thesis extends them. Chapter 5 describes the design of the experiment. Chapter 6 describes the implementation of the proposed experiment. Chapter 7 discusses the flow of the experiment, the results and suggests some defensive methods for the voice assistant. The Chapter 8 discusses the impact of our experiment on society. Finally, Chapter 9 summarizes the results and findings of this work.

[1] https://esorics2024.org/

Chapter 2

Deepfake

This chapter focuses on a deeper exploration of deepfakes. First, we will look at the very nature of deepfake technology. We then discuss a particular type of deepfake technology, namely voice deepfakes. Finally, we will look at various applications of deepfakes and show their positive and negative applications in the world.

2.1 What are deepfakes

Deepfakes are a combination of the words „deep-learning“ and „fake“ and represent artificially created content, namely fake videos, audio recordings or images that are created by artificial intelligence and depict events that never happened [43]. These artificial media have the ability to appear authentic in front of ordinary humans or even biometric systems. Previously, creating quality deepfakes was very time consuming and required the efforts of an entire team. However, with the advancement of technology, creating them has become easier and faster, even for individuals with minimal knowledge [3, 40]. A breakthrough in the popularity of deepfakes came in 2017, when a user on the online platform Reddit posted pornographic content in which actors' faces were mistaken for famous celebrities [63]. This method is called face-swapping and is probably the most well-known method of creating deepfakes. The proof of the boom is the Figure 2.1, which summarises the years and the number of articles created on this topic [39]. Further evidence of rapid growth is the statistic showing that the number of deepfakes on the Internet roughly doubles every six months, which raises increased concerns about their misuse. This sharp increase in the proliferation of deepfakes highlights the urgent need for technological countermeasures and increased public awareness [49]. Therefore, it can be said that the main intention of these synthetic media is to create misinformation that can lead to mass panic or to ridicule someone. However, deepfakes can also be created with positive intentions, they are used extensively and many times we don't even notice that we have come into contact with a deepfake. In Section 2.3, the potentially beneficial and harmful aspects of deepfakes are discussed in detail.

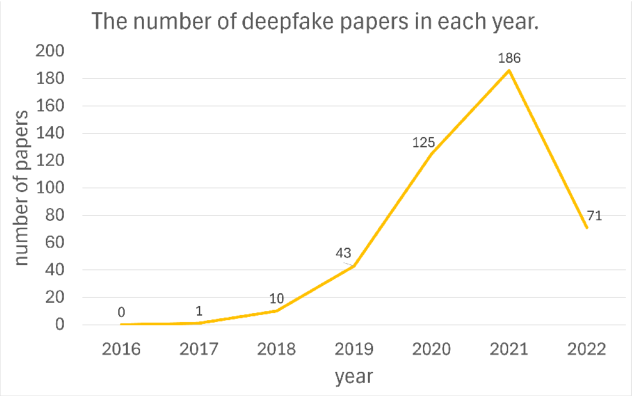


Figure 2.1: The number of scientific papers in deepfake research, broken down by year of publication [39].

2.2 Voice deepfakes creation

A voice deepfake is an artificially created voice recording, mainly using deep learning techniques to create it. However, the process of creating them involves multiple disciplines such as acoustics, signal engineering and further processing, linguistics or statistics. This technology makes it possible to mimic the voice patterns of specific voice characteristics of individuals and is becoming increasingly common in everyday human life [46]. Voice deepfakes can be divided into two categories according to Text-to-speech and Voice conversion. However, there are extensions of the mentioned methods that are worth mentioning zero-shot or end-to-end [55].

2.2.1 Text to speech

Text-to-speech (TTS) is one of the methods of creating voice deepfakes, which consists of converting text input into audio output that aims to sound as similar as possible to the human spoken word. Nowadays, there are many works that deal with this problem and each one tries to come up with something innovative [1].

Sasirekha [53] described the basic architecture of TTS well and its individual modules architecture is shown on Figure 2.2.

Text Analysis & Text Detection

the aim of this section is to identify and pre-process the input text into a clear list of words. This section can accept input text either in plain text format or this section can extract text from PDF, photos or even video. It is necessary to be able to handle different syntaxes of different languages and thus be able to recognize the ends of sentences.

Text Normalization & Text Linearization text normalization transcribes text into a pronounceable form, i.e. the main goal is to identify punctuation marks or spaces. Very often this process is used to convert capital letters into lower case and to convert them into a uniform format, that is, to convert, for example, ungrammatical words or number conversion. This is taken care of by the modules - number converter, abbreviation converter, acronym converter and word segmentation. Subsequently, the Linearization process helps the user to better navigate the page by creating hyperlinks.

Phonetic Analysis sound, defined by its characteristic shape in the form of a sound wave, is called a „phone“. This smallest unit of sound plays a key role in the linguistic world.

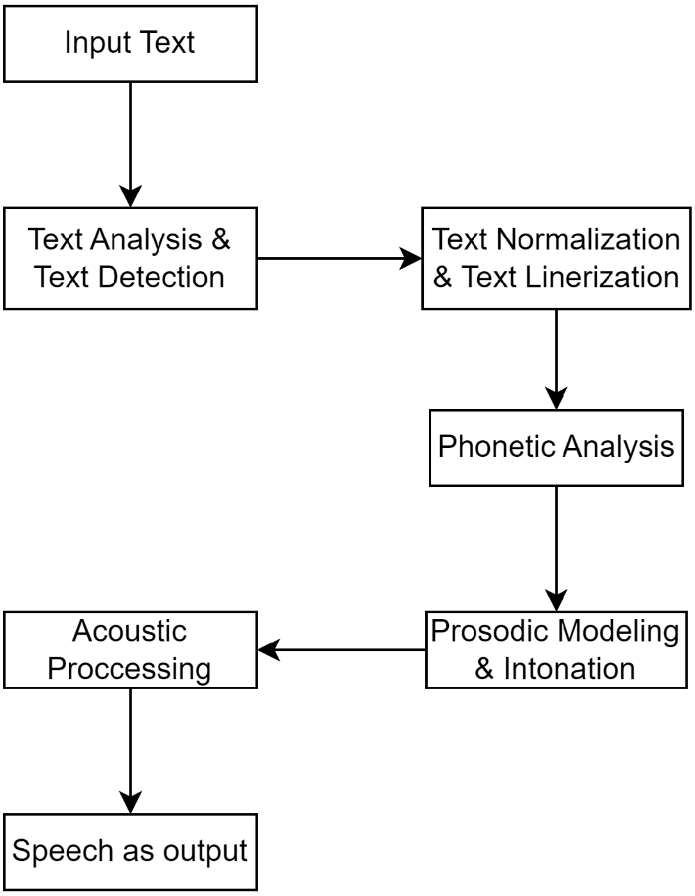

Input Text
Text Analysis & Text Detection
Text Normalization & Text Linerization
Phonetic Analysis
Prosodic Modeling & Intonation
Acoustic Proccessing
Speech as output

Figure 2.2: TTS Architecture Overview [53].

The combination of these phones produces phonemes, which are the smallest unit of speech that allows us to distinguish individual words from each other and which together form the language system. The number of phonemes is smaller than the number of graphemes, more precisely there are only 44 of them. It must also be taken into account that each word may be pronounced differently from the way it is written. Therefore, there are two approaches. The first is the dictionary approach, which consists of a huge database in which the pronunciation of each word is recorded. So the advantage is that we can very quickly find a pronunciation whose quality will be significantly better than in the second approach, which is rule-based and therefore needs no database and works with any input.

Prosodic Modeling & Intonation prosodic modelling and intonation focuses on making the human voice sound as authentic as possible and brings the speaker's emotions into the speech production process through stress patterns, rhythm or intonation. For example, using rising or falling intonation, the resulting pattern can be made to reflect feelings such as joy, sadness, and others.

Acoustic Processing acoustic processing aims to create a resulting voice in accordance with the characteristics of a particular person. There are 3 types. The first is Concatenative Synthesis, which has a pre-dialed database with pre-formed words that it combines together to form speech. The second method is the Format Method, which the output sounds artificial and robotic, but there is no need for a database and the last method is Articulatory Method, in which techniques and algorithms are created based on the human vocal tract to synthesize speech. This method is usually based on mathematical models and its output is a complete synthetic speech.

There are other works that build on these individual modules. In this work Mirjam Ernestus is working on creating models for acoustic reduction, the main goal is to adjust the speed of synthetic speech to match human speech as closely as possible, and she is starting to use comprehension models [18].

2.2.2 Voice conversion

Voice conversion (VC) is the second type of deepfakes, which consists in modifying the speech of the source speaker to sound as if it were uttered by the target speaker [14]. In these works [41, 19], the general architecture of the VC system is described and the system can thus be divided into the following modules described in the Figure 2.3.

The architecture is divided into two parts. The first one is the *Training phase*, in which recordings of utterances from both source and target speakers are used. These recordings are then analyzed, and based on this analysis, the system decides which kind of transformation is most optimal for converting the source utterance to the target utterance. The general flow of the training phase can be briefly described as follows. First, the individual frames of the input signals are analyzed. Each frame is then transformed into a suitable vector of numbers using the appropriate algorithm for (VC). After creating both the source and target vectors, a step called *alignment* follows, where correlations between the two vectors are sought, such as similar phonetic and acoustic information. Finally, in the *Training phase*, the system learns, based on the previous calculations, which conversion function to apply in the *Conversion phase*. The second part of the system is the *Conversion Phase*, during which the recording of the source speaker's utterance is parameterized and analyzed. Subsequently, the conversion function that was previously selected is applied to the parameterized recording of the source speaker. The product of this operation is the resulting synthesized recording.

2.2.3 Zero-shot voice cloning

This method only requires very small recordings to produce synthetic speech and is also able to produce synthetic voice for many source speakers who may not have been present in previous training sessions of the model and are therefore unknown to the synthesizer. In most cases, the voice of the target speaker is created based on information obtained from only a single voice segment. This is done through a „speaker encoder“, which is a network that extracts a vector embedding containing information about a particular speaker in a given voice segment. There are two ways in which this Speaker Encoder can work. The first is that it is pre-trained on the tasks associated with identification or can be trained simultaneously with other parts of the voice conversion network [33].

Zero-shot voice conversion has only recently begun to be addressed, and there are even fewer models that achieve satisfactory results. It can be said that the first model that demonstrated reasonable performance in this area was an autoencoder-based model called AutoVC and is described by Kaizhi Qian [52]. AutoVC uses a pre-trained speaker encoder together with the proposed bottleneck layers.

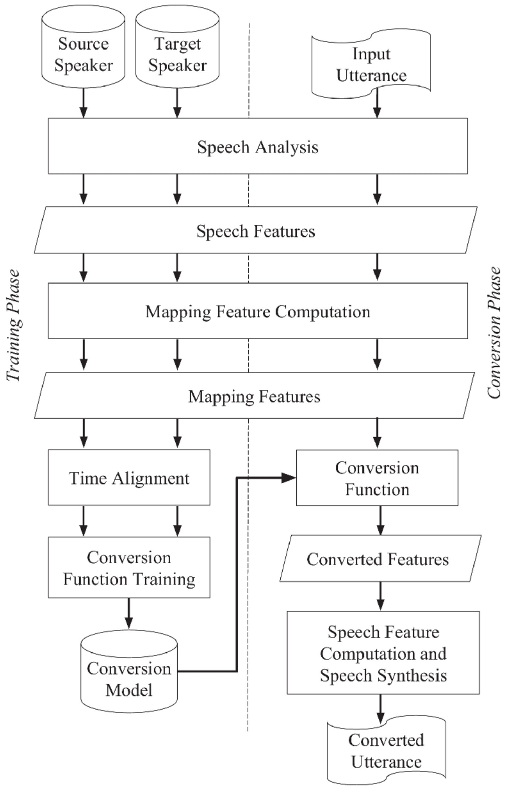


Figure 2.3: General VC system architecture diagram [14].

2.2.4 End-to-end voice cloning

End-to-end voice cloning is based on the idea that all modules of the learning system should be differentiable and that the whole system is trained as a whole using gradient-descent and backpropagation. Also, these models do not need a separate Vocoder module. This technique is often called straightforward brute-force [33].

2.3 Use of deepfakes

This section looks at the positive and negative uses of deepfake technology in the world and also presents some real attacks that have been carried out.

2.3.1 Positive application

There are plenty of opportunities for deepfake technology in the film industry. It provides filmmakers with greater flexibility in filming while saving money. Instead of reshooting, scenes can be reshaped according to the director's vision [35]. Furthermore, this technology also allows the voices of deceased actors to be revived, opening up further possibilities. Deepfake technology also allows filmmakers to work with a created model of the actor, which can eliminate the involvement of expensive actors [42]. This not only presents financial advantages, but also provides greater freedom in the selection of actors and their performance within films. However, for actors, it is a huge threat to their craft [51].

Another use of deepfakes in the film industry is in dubbing. Artificial intelligence is used to train an actor's voice model, which allows the original film to be dubbed into another language. Compared to traditional dubbing via another actor, this method brings more natural lip movements while speaking or the ability to avoid the need for live actors with similar voices and the subsequent reduction in the time of creating the dub itself, so that the dubbing time is now limited only by computing power [63]. In addition, technology could be extended to support deaf people, where actors' voices could be dubbed into sign language. An extended view would be to integrate a special module into televisions containing a pretrained model that would allow real-time translation of actors' words into sign language. In this way, the deaf could watch any TV station not just those that provide such a service as a default.

In the fashion industry, the aim is to make the purchase as simple and quick as possible for the customer. Some fashion brands are trying to create a way in which they allow customers to create their own personalized avatars online. They also create an assortment of online models of their products and customers can virtually try on clothes on their avatars. By doing this, manufacturers want to eliminate the physical need for a visitor in the store.[17]

2.3.2 Negative application

The misuse of deepfake technology poses a danger to society, individuals, political systems and companies. Deepfakes pose a greater threat than fake information because they are more difficult to discern in the mass of information that the average person consumes [63].

Deepfakes have the ability to influence political events in the country by taking the level of gossip to a new level. Some of the fears of influencing American elections have arisen in the context of deepfakes. Specifically, that words could be put into candidates' mouths that they would not utter in the real world, which could influence voters' decisions [58].

Deepfake video of a live podcast of Elon Musk smoking marijuana led to a significant drop in investor confidence. This event also damaged public confidence in Elon Musk and Tesla, and it caused Tesla's stock to drop by 6%, causing a large financial loss [32]. However, deepfake technologies can work in the opposite direction as well. That is, fake videos can be created featuring celebrities promoting specific products. In this way, consumers can be deceived and the reputation of celebrities [7] .

However, the porn industry clearly has the largest representation of synthetic media, as this article says that up to 98% of all deepfake videos circulating on the internet are pornography [61]. The most commonly used technique is the face-swap, where the face of a porn actress is swapped with a target person. The Figure 2.4 above is an example of the high-quality output used by this method. It is known that 99% of these attacks target women and only 1% target men. Despite legal action taken by some states in the US, England and Wales against deepfakes in the pornography industry, it is difficult to eradicate this form of attack completely.

An interesting example of a recent attack of this kind is the Taylor Swift [45] case that took place earlier this year. This incident garnered massive attention, with nearly 45 million users enough to view the post before it was deleted. While the experience was certainly unpleasant for the victim herself, the event had an extreme impact, bringing awareness of the existence and potential dangers of deepfake technology to many users who were encountering the issue for the first time.

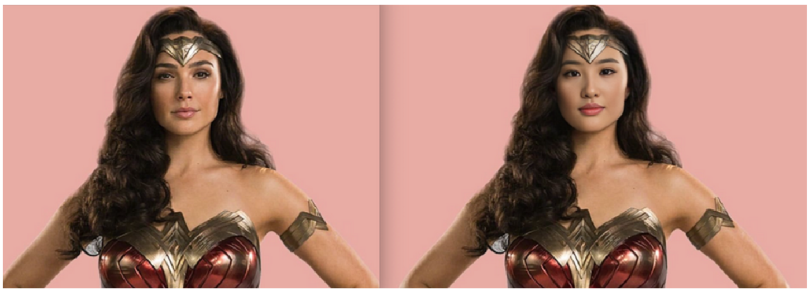


Figure 2.4: Example of a deepfake created using the face-swapping method (https://medium.com/@ryf123/unleashing-deep-fake-face-swapping-exploring-google-colab-and-roop-for-seamless-transformations-fed0f1bab788)

There are many tools that are free, online, which is why this type of attack is spreading even among high school teenagers. In the hands of immature individuals, this tool pushes the boundaries of bullying to a new level. There are already several known cases of the dissemination of nude photos created by artificial intelligence (deepfake) in schools. In most cases, children are bullied in this way, but there are also a few cases where a teacher has been attacked in this way [64].

Real-time deep-faking is a threat that has emerged in recent years. Voices or faces are generated in real-time, which allows an attacker to achieve much higher success rates when performing an attack, thus allowing attackers to create convincing fake voices or streams to manipulate their victims. Several such attacks have already occurred. And the victim has suffered no small harm. For example, in 2019, an attacker phoned the CEO's voice to his company to forward $243,000 to a given account. Or another example where in 2021 a Hong Kong company's banker was tricked in this way and sent $35 million to the attacker's account. It may not only be large companies that are being fooled in this way, but as deepfake technology becomes more and more familiar to people, attacks where an attacker calls a pensioner in the voice of a grandson and demands that he forward money could become commonplace [23].

Chapter 3

Voice Assistants

Voice is becoming a very popular interface for human communication with modern technology. Mainly because of the growing trend of wirelessness. Around 3.25 billion VA devices were purchased globally in 2019. Projections indicate that by the end of 2024, the number of VA devices will increase to approximately 8.4 billion units, a figure equivalent to the world's population [16]. That is why better and better voice assistants are being developed to form an intermediate layer between the user and technology. Many activities are greatly simplified thanks to them, such as hands-free in cars, sending messages, reading emails, reminding appointments in the calendar or checking bank balance [48]. In some cases, the voice interface is safer, such as when driving in a car. Voice assistants belong to the voice-user interface (VUI) category and are software applications that run in the background of voice command devices and are activated on the signal of special phrases. Voice assistants are most often used with a smart speaker combination. When these speakers are connected to other home appliances, a Smart Home is created. The Figure 3.1 shows examples of IoT devices that can be controlled via a smart assistant. This type of home allows the user to use a phone or other input device to remotely control home appliances through an internet connection. Thus, the user can control, for example, the temperature in the house, the lights on or the security access to the house [36].

In the first part of this chapter, we will focus on the architecture and setup of voice assistants. The second part of this chapter will focus on the analysis of voice commands and how voice commands are recognized and interpreted. The third part of this chapter will be devoted to the general workflow of voice assistants. The next part of this chapter will deal with authentication. The last part of this chapter will discuss some selected voice assistants and their key capabilities. We will discuss their advantages and disadvantages and how they differ from each other.

3.1 Voice assistant setup architecture

The architecture of the most commonly used voice assistants [10] could be described by a generic diagram, which you can see in Figure 3.2. The legend of the following diagram is described below [66, 29, 15].

Human user is an individual who actively uses the Voice Assistant. The individual utters voice commands to interact with the system and control the behavior of connected devices.

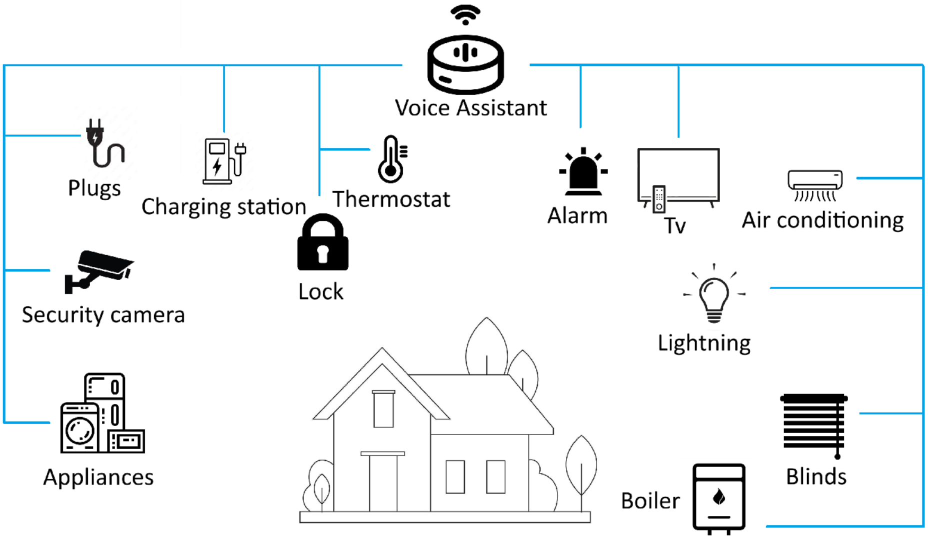


Figure 3.1: Devices that can be controlled using the assistant [48].

Microphone is a physical device that serves as an input device for voice commands from users. The microphone picks up audio signals that are then processed by the voice assistant application.

Speaker is a component that is used to output audio information generated by the voice assistant. The speaker allows users to hear responses to their voice commands or other audio outputs such as alarms, phone calls or alerts.

Voice assistant application is a software application that receives and processes voice commands from users. This application is used to perform desired actions such as searching for information, launching applications or controlling IoT devices.

Voice assistant server is part of the system that processes voice commands from users. This part converts voice commands into service requests and forwards some of them to appropriate cloud servers for further processing and execution.

Cloud services / Data sources ensure the processing of voice commands and the storage of relevant information. Cloud services provide the means to analyze and process voice data, while data includes information transmitted and processed within the system, such as user voice commands, voice assistant responses and status information about IoT devices.

IoT controller is a device that provides management and control of IoT devices. This controller receives control commands from the voice assistant and sends responses back to influence the behavior of connected IoT devices.

IoT Devices are a physical devices connected to the IoT that can be controlled by a voice assistant and an IoT controller. Typical examples are smart home appliances, sensors or security systems.

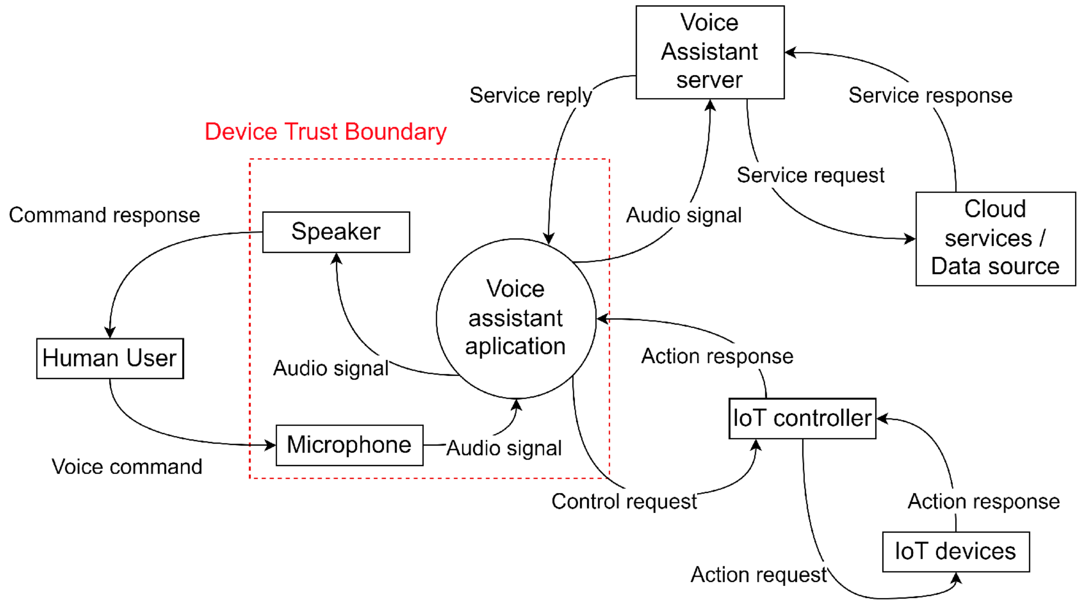


Figure 3.2: An example of a high-level architecture for a typical AI powered Virtual Assistant [15].

3.2 Analyzing the voice assistant command

As mentioned before, an important aspect of communicating with the voice assistant are the commands. In the following Figure 3.3 is the example command for the Amazon Alexa voice assistant, which will explain the different parts of the command [25]. This is how almost every command for any voice assistant is built except for the Invocation Phrase, which is different for each assistant.

Wake word is a spoken phrase that puts the assistant into a so-called listening mode, in which it waits for the user's next request. Typically, the VA signals the transition to listening mode audiovisually. Before switching to listening mode, that is, when the invocation phrase is uttered, the assistant performs voice authentication; we verified this property in our experiment described in Subsection 5.5.1. This phase is also text-dependent, where the assistant waits for a specific invocation phrase such as „Alexa“ or „Hey Siri“. Based on the Wake word phrase spoken by the user, the assistant will segment the Wake word phrase and create a Wake word fingerprint. Then it compares the fingerprint against its database of user voiceprints. When it matches the fingerprint already in the database, it verifies the user and the user can now make his/her request to the assistant.

Phrase is the part of the command to the voice assistant that the assistant does not need to perform the task correctly. The phrase is only used by humans to correctly understand the sentence. Thus, the phrase is irrelevant to the assistant. We get the same result when we give it a command without or with phrases.

Invocation Phrase is the main part of the voice assistant command. Keywords or words that ask the assistant to perform an action. This is typically some command, question, or user-created command. This invocation phrase is necessary for the assistant to understand what to do. Thus, each request must consist of at least one invocation phrase.

Utterance is the part of the command that allows the assistant to recognize the user's specific intentions. Therefore, this part of the command is important for correctly routing commands and providing relevant responses. It should be clear, concise and contain key words or phrases that characterize the user's intent. Its absence does not prevent the assistant from providing answers, but it may limit the assistant's ability to interpret commands with greater accuracy.

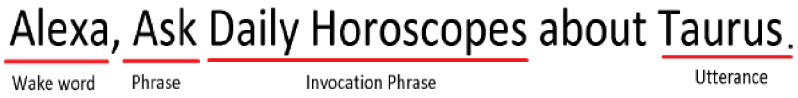


Figure 3.3: Example command for a voice assistant, specifically Amazon Alexa, asking for the daily horoscope of the sign of the bull.

3.3 Generic workflow

Regardless of the manufacturer, voice assistants have a general workflow shown in Figure 3.4. The user interacts with the voice assistant using a voice command. The voice assistant has „keyword spotting“ technology that allows it to recognize its wake-up command from ordinary speech. Once it detects the command it sends a request to the server. Next, the server interprets the command and sends a corresponding request to the cloud server, which evaluates the request and returns a response to the request.

3.4 Authentication

Authentication is the process by which the identity of a subject is verified. That is, verifying that the subject is really who they say they are.

The rise of authentication occurred with the emergence of the Internet and the rise of computer technology, when multiple computers began to communicate with each other, and in order to increase the security of communication, they had to authenticate with each other. Thus, the oldest and simplest method is the password, but as time went on and more sophisticated systems were created and more and more important information was entrusted

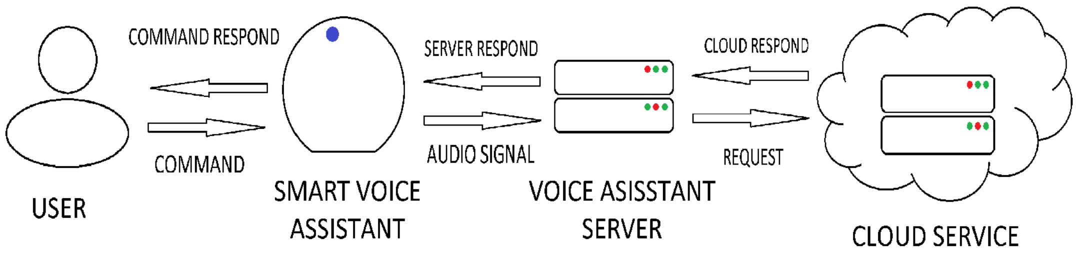


Figure 3.4: Generic Workflow of voice assistant

to them, more rigorous approaches to achieve authentication had to be devised. According to [11, 37], one of the following 4 properties is needed to achieve strong authentication. The first property is something only the subject can know (password). The second property is something that is part of the subject (biometrics - fingerprint, voice, face). The third property requires the subject to possess something that only the subject can possess (ID card). The last property requires the subject to identify with something that only the subject can produce (a signature). By combining all these properties, we can achieve strong authentication [11].

We could further divide authentication into two essential technologies. The first is Speaker Verification, in which it decides whether the subject is really who he/she claims to be. And the second is Speaker Identification and that means whether the subject is in some database of authenticated subjects. Speaker verification can then be further divided into text-dependent and text-independent. The difference between the two is that text-dependent requires the subject to be identified each time by some pre-known text (password). This combines multiple properties (password + voice) to create stronger security [11].

3.4.1 Authentication process

Biometric systems work on the basis of extracting characteristics that distinguish individuals from each other, in the case of voice it is a voiceprint or voice template.

The voice recognition system initially records the speech of the desired subject and creates an initial template for him. Very often it is practiced to combine multiple templates to improve the quality of the resulting print and the print is entered into a database. Subsequently, when the identity of the person is already verified, the biometric system records the voice samples of the person and creates a template from them. And it compares this newly obtained template with the already enrolled template Figure 3.5. If the result of the template comparison is very positive it means that there is a high probability that the person from whose voice the second template was created is the one whose original voice template is already in the database [38, 37].

3.4.2 Voice assistant authentication

Voice biometrics enables contactless identity verification, a noninvasive and convenient method for users. The ability to authenticate a large number of users simultaneously makes voice authentication ideal for use in voice assistants. In addition, there is no need for multiple hardware to use voice authentication and everything is implemented on one single device [6, 50].

One of the disadvantages of using this method arises from the variable environmental conditions. For example, wind, ambient noise, or other external factors can distort the sound of the human voice, negatively affecting the ability of the voice biometrics system to correctly identify the speaker. Using better hardware can reduce this problem, but can never be eliminated 100% and, in addition, the price will increase with better hardware, which may discourage customers from purchasing [31].

Other situations in which voice recognition may fail are due to subtle changes in speech patterns, accents or vocal characteristics of the speaker. These changes can be caused by a variety of factors such as emotional states or fatigue. Also, an individual's accent can affect the ability of the voice assistant to recognize the user, especially if the accent differs from standard language patterns. Furthermore, common health problems such as sore throats,

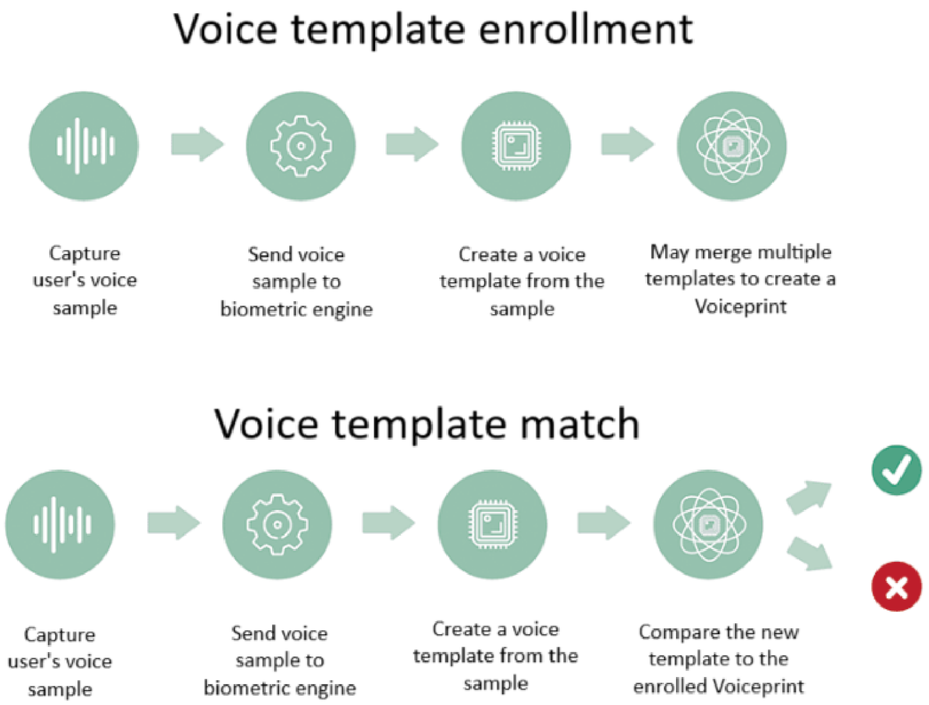

Voice template enrollment
Capture user's voice sample
Send voice sample to biometric engine
Create a voice template from the sample
May merge multiple templates to create a Voiceprint
Voice template match
Capture user's voice sample
Send voice sample to biometric engine
Create a voice template from the sample
Compare the new template to the enrolled Voiceprint

Figure 3.5: The process of enrolling templates and the process of comparing templates with subsequent rejection or validation [11].

colds or allergies can affect the quality of speech and thus significantly reduce the success rate of voice recognition [2].

A considerable risk arises with the advent of synthetic voice creates an additional challenge for voice authentication. Voice assistants that rely on ASV face new opportunities for attack and misuse. As the persuasiveness of deepfake voices increases, the risk that attackers will be able to override the authentication system and gain unauthorized access to sensitive information or perform malicious operations increases [34]. Some techniques to defend against synthetic voice are discussed in Section 7.3.

3.5 Representatives of voice assistants

This section discusses in detail the voice assistants that were used in our experiment, that is, Siri, Alexa, Google and Bixby. Furthermore, all voice assistants that currently support user authentication and their availability on the Czech market are mentioned here in Table 3.1.

3.5.1 Amazon Alexa

Alexa[1] is most commonly used through a smart speaker known as Echo. However, Alexa's services can also be used through a mobile app and is starting to appear on various smart devices such as headphones. Originally, Alexa was created to facilitate online shopping,

[1] https://developer.amazon.com/en-US/alexa

VA	Voice	ASV	Available in the Czech Republic
Siri	✓	✓	✓
Alexa	✓	✓	✓
Google Assistant	✓	✓	✓
Bixby	✓	✓	✓
Cortana	✓	✓	✓
Alibaba AliGenie	✓	✓	×
Tencent Xiaowe	✓	✓	×

Table 3.1: Table of voice assistants supporting user authentication and if they are available in the Czech Republic.

but this functionality has expanded and is now able to work with a wide range of home appliances, not just those from Amazon.

Essentially, Alexa is just a speaker, but her real use is to connect with other smart home devices. To maximize the personalization of communication with the user, Alexa retains relevant information from previous queries on the server. This information is then used to assess further queries with maximum accuracy. When the user is unlikely to ask further questions after a certain period of time, the information is deleted from the server [27].

The data is temporarily stored in DynamoDB and encrypted before storage. The system works with a double table. The first table is used to store system events such as requests to transcribe user utterances or instructions to synthesize responses. Information is stored as strings that contain references to the second table. The second table contains encrypted user utterances, responses from Alexa, and other contextual information. Each entry has its own record, so the information does not need to be decrypted on each access, speeding up access to the data.

Alexa creates a context vector that it uses to evaluate whether the next user query is related to some encrypted information in the database. This calculation is performed when responding to the previous task [5]. This approach allows for a more personal interaction with the user and a more pleasant interaction with Alexa, while ensuring maximum information security.

3.5.2 Apple Siri

Siri[2] is included in most of Apple's operating systems. Originally, the virtual assistant was only on the phone, but over time its use has grown. Now we can also use it to control appliances that are part of a smart home. Only appliances in the HomeKit family can be controlled via Siri. This makes the circle communicating directly with Siri narrower and therefore promotes greater security. Siri can also work and respond to user requests offline, but its responses in this case are very clipped and many functionalities are not supported. Siri does not temporarily store the user's answers anywhere, so her answers can often be repeated. Siri uses Natural Language Processing (NLP), which is a combination of statistical, machine learning, and deep learning models. Together, these technologies allow computers to understand human speech [36].

[2] https://www.apple.com/siri/

3.5.3 Google Assistant

Google Assistant[3] has the advantage of being able to be installed as a standalone third-party application that can be downloaded from the online store. This means that Google Assistant is not limited to a specific operating system and can be used on a wide range of devices that support the app. This flexibility allows users to use Google Assistant on any device they want to install it, regardless of the operating system. Unlike previous assistants, Google Assistant does not offer any other significant differences [30].

3.5.4 Bixby

Bixby[4], the voice assistant developed by Samsung, is pre-installed on every new phone and other IoT devices from the company. Its uniqueness lies in its deep integration with the operating system and third-party applications, allowing users to fully control almost all applications using only simple voice commands. Unlike previous assistants, the user has the ability to set the phone to unlock by voice and make other complex adjustments to settings that previous voice assistants did not allow. Bixby also includes a feature called Bixby Vision, which allows users to easily get information about objects using the camera. With this feature, the user can point the camera at an object and Bixby will then tell them relevant information about the object [47].

[3] https://assistant.google.com/

[4] https://www.samsung.com/cz/apps/bixby/

Chapter 4

Related Work

The main differences that distinguish this work from Related work are the number of voice assistants participating in the experiment and the number of respondents included in the experiment.

This bachelor's thesis was the main source for a research paper accepted at the Esorics 2024 conference [1] together with the work of Šandor [69], whose thesis lacks a large-scale experiment. In his thesis he analyses the functions of voice assistants and the harm that an attacker can cause if a successful attack on a voice assistant is carried out. In his work, he classifies attacks according to the following parameters. The first one is the *Difficulty* which means how long the command is and hence the longer the command is the bigger is the difficulty of executing the attack. The next parameter is *State*, which indicates whether the device must be unlocked or locked to allow the user to use the corresponding function. And the last parameter is *Damage*, which indicates how much damage the perpetrator can inflict on the victim when accessing this functionality. Due to the existence of this classification of voice assistant vulnerabilities, it was no longer necessary to examine this classification in this thesis.

Furthermore, there are several works that test the resilience of voice assistants or the resilience of biometric systems, and for the sake of clarity, we divide them into two logical units of spoofing voice assistants and spoofing attacks on biometric systems.

4.0.1 Spoofing Voice Assistants

In a paper by Bilika [9], the resilience mechanism of two voice assistants (VAs), namely Google Assistant and Siri, which are currently the most widely used voice assistants, was investigated. The work focused on the security shortcomings of VAs that are used in smart devices as they are integrated into the daily tasks and the control of smart home devices. The purpose of the work was to determine the extent to which the resilience of protection mechanisms that are designed to restrict sensitive operations only to the device owner works. The study involved participants training these VAs to recognize their voices. That is, each participant created a voiceprint, which they then retrained and then attempted to penetrate the systems using deepfake voice commands generated from voice recordings provided by the participants. The results of this work revealed some disturbing results, namely that more than 30% of the deepfake attacks successfully manipulated VAs to perform potentially risky tasks. The effectiveness of these attacks revealed significant differences between the two tested assistants, suggesting different levels of vulnerability. In

[1] https://esorics2024.org/

addition to this research, other results highlight the gender imbalance observed in one case, suggesting possible differences in how these VAs process and respond to voice commands.

Ubert's thesis [62] focused on exploring the misuse of virtual assistants using voice deepfakes. The aim of this thesis was to demonstrate how easily a malicious attacker could gain access to sensitive data through virtual assistants such as Google Assistant, Alexa and Siri. As part of this work, an experiment was conducted to train a model for voice deepfake using voice samples from 12 participants. The success of this model was then tested against the three digital assistants mentioned above. The results of the experiment showed that using voice deepfakes can successfully extract sensitive information from VAs such as birth dates, addresses, personal contacts and notes. And it further presents that such vulnerability requires security and privacy enhancements for these voice assistants.

4.0.2 Spofing attacks on biometrics systems

Generic biometric system can become very vulnerable to voice synthesis, voice conversion, impersonation or replay attack vectors as Alegre et al. in [4, 28]. During an impersonation attack, an individual mimics someone else's voice to bypass biometric authentication. Research has shown that an attacker does not necessarily need advanced IT skills, specifically voice mimicking, to overcome Automatic Speaker Verification (ASV) technology [57].

Evans et al.[20] tested the resistance of different ASV systems to synthetic voice and replay attacks. And they showed how easy it is to carry out a successful attack on a biometric system. They also come up with proposed methods that could reduce the attacker's success rate, but they do not test these methods in their work.

Wu et al.[65] show that an attacker could achieve an attack success rate of 78.36% false acceptance rate (FAR) by playing a recording of a male voice and a female voice with an FAR of 65.28%, which is extremely high. Playback attacks were previously considered a major threat to ASV because they are simple, the attacker only needs to record the victim's voice and replay it. However, this is no longer true today as the complexity of creating a synthetic voice has been simplified by various online tools that are simple to use. Also, as the population's awareness of deepfakes grows, people are learning about all the possibilities they can do with them, and the methods of creating them are becoming more public [22].

These studies [21, 54] have shown that attacks using synthetic media fraud in biometric systems are possible and thus it is only a matter of time before they start happening more frequently. This is also helped by the fact that, thanks to paid services that can quickly, easily and reliably create synthetic media using voice cloning [22], and advances in hardware, creating high-quality deepfakes is currently a matter of minutes. As the aforementioned studies suggest, biometric systems do not have a standard to prevent such attacks.

Furthermore, in this work [24] by Gernot, a method for generating one-time biometric templates for user authentication applications is presented. The proposed system effectively mitigates replay attacks in which an attacker intentionally resends captured proof of user identity repeatedly. Biometric characteristics are extracted from the media using advanced deep learning and then protected by biohashing. Furthermore, the application of cryptographic hashing and symmetric encryption ensures the generation of one-time, non-replicable templates. The experimental results demonstrated the high effectiveness of this defense mechanism, as the biometric system successfully resisted repeated replay attacks once the method was incorporated into the system.

The work [44] by Nacimiento-García investigated spoofing attacks on Amazon Alexa devices and the ability of the voice assistant to detect the attack. His work used YourTTS, a text-to-speech (TTS) synthesis system, which was then integrated into the Telegram bot. In this way, they were able to generate synthesized voice samples using a voice cloning technique. These artificially created voice samples were then used to attack VAs in order to mimic a real Alexa user, thereby bypassing the voice profile-based identification mechanisms. Finally, the ultimate goal of the experiments was to test whether it is possible to perform unauthorized activities in this way. This work highlighted the vulnerability and imperfect security of voice assistants such as Amazon Alexa.

This article [66] by Yana presents an extensive look at the security of voice assistants, also focusing on potential attacks and various mitigation measures. The conclusion of this paper is that the main safeguard of a voice assistant should be the intelligence of users. Thus, they should prefer to activate speaker authentication for their voice assistants. And further, that they should refrain from enabling high-risk voice assistant features such as banking or home unlocking. As additional measures to increase protection, they suggest preventing unattended device abandonment, disabling voice assistant when the device is locked, or even completely disabling wake-word password detection. Finally, it divides the measures against attacks into detection and prevention categories, which are organized on the basis of shared defense strategies.

This thesis significantly advances the field by expanding the number of respondents to 72 individuals, which exceeds existing work in this area and meets the guidelines for qualitative studies of this type [7]. Furthermore, we surveyed a wider range of voice assistants, i.e., the experiment was conducted on the four most commonly used models of [10].

Chapter 5

Experiment Design

The previous chapters introduced deepfake technology and voice assistants. This chapter presents the main objective of this thesis which is an experiment. The main goal of the experiment is to investigate whether voice assistants using automatic speaker recognition can be tricked by deepfakes to reveal personal information.

There are already several papers dealing with this topic, the individual papers are discussed in Related work in Chapter 4. It is important to mention that the main difference that this paper carries over from others of a similar nature is the scale of the experiment. In all previous studies, the experiments were carried out on small samples of test subjects. In contrast, this bachelor thesis is characterized by a broader and more robust methodology and presents a large-scale experiment consisting of **72** respondents.

5.1 Attacker Model

An attacker is someone who is able to create a voice deepfake by synthesizing the victim's voice and penetrate the automatic speaker authentication (ASV) of selected voice assistants (VAs) to command the voice assistant to perform malicious actions or obtain the victim's personal information, such as a voice assistant that is part of a smart home. Thus, the attacker has samples of the victim's voice, knows the necessary information and procedures to create voice deepfakes, may know some details of the ASV system [26], its functionality or parameters, but does not have access to specific information. In this case we talk about *Grey-Box* attack. The second type of attack is called *Black-Box* in which the attacker has no information about the system, does not know the structure, any parameters or functionality at all. Given the characteristics of this experiment, a potential attacker could attack the biometric system in both ways.

However, both methods of attack require that the potential attacker is able to collect the necessary number of samples of the victim's voice to create his own synthesized model of the victim's voice and thus control the VA.[66]

The targets that an attacker can target with his attack can be as follows

- **Security** - The attacker wants to manipulate devices that support Smart-Home and cause damage. For example manipulate door locks, window locks.
- **Privacy** - The attacker wants to obtain information about the victim in order to exploit it, blackmailing the victim. For example, obtain information about the victim's daily routine through a calendar shared among VAs, have VAs call a toll-free

line, insert some malicious event into the victim's calendar, or send messages with the victim's name (emails,... linking to the card).

- **Other** - Attacker's ways to harm the victim are to order the assistant to visit some malicious site or buy things on the internet from the victim's account.

Further classification of attacker's targets can be based on the difficulty of the attack and the damage that the attacker can cause with a particular attack. In this way, we can categorize attacks as follows [69]. Some selected targets are summarized below.

- **Phone calls** - These attacks allow an attacker to use a foreign device to make fraudulent calls or calls to expensive numbers. The difficulty of this attack is described as moderate as it requires a certain level of technical expertise and the ability to make a call quickly and accurately. Damage that can result from misuse of this feature is classified as medium to high, as it may include financial losses from unauthorized calls to toll-free numbers or commissions from fraudulent transactions.

- **Messaging** - These attacks allow you to send fraudulent messages, advertisements or malicious links. The difficulty of this attack is high as it requires dictating the entire content of the message character by character, which can be challenging and time consuming. Damage from this attack is rated moderate to high, as it can lead to fraud or financial loss from paid SMS services.

- **Smart Home Device Control** - The target of this attack is smart home elements, these elements may include smart TVs, thermostats, lighting, locks or cameras. The attack capabilities range from low to high depending on the specific security measures of the household in question. For example, if devices are well password-protected and have up-to-date software updates, they may be less susceptible to misuse. However, if there is inadequate security, smart home devices can be misused to carry out a variety of attacks, such as unlocking doors, turning lights on or off, or monitoring via cameras. The consequences of this misuse can be manifold and can include loss of privacy, financial loss or even physical harm if the devices being controlled are linked to home security.

- **Manage calendars, schedules, to-do lists, timers and routines** - The possibilities of an attack are generally considered low to moderate because access to this information is not usually password protected by default and is not difficult to gain access to. However, the level of protection may depend on the specific device settings and security measures used. The damage from this form of attack is generally rated as low to medium because tampering with calendars, to-do lists or timers may not have direct financial consequences but may cause unplanned changes to the victim's time management.

The only problem for the potential attacker is to obtain samples of the victim's voice, but these can be obtained by manipulating the victim [66].

Features	**Google Assistant**	**Siri**	**Alexa**	**Bixby**
Wake Word	„Hey Google“	„Hey Siri“	„Alexa“	„Hi Bixby“
Speaker Recognition	Yes	Yes	Yes	Yes
NLP	Yes	Yes	Yes	Yes
Software	Google Assistant	iOS, WatchOS	Alexa app	Bixby app

Table 5.1: Voice Assistants and Experiment Parameters

Voice Assistant	**Device**	**Software Version**	**Device Type**
Google Assistant	Google Nest Mini 2nd Gen.	2.57.375114	Speaker
Siri	iPhone SE	iOS 16.6.1	Mobile Phone
Bixby	Samsung Galaxy A53 5G	Android 13	Mobile Phone
Alexa	Echo Dot 4th Gen. (2020)	9295801732	Speaker

Table 5.2: Hardware and Software Versions used in Experiments for Individual Assistants.

5.2 Used voice Assistants

In selecting the VAs who were included in our experiment, the main criteria were which voice assitants are most commonly used in the population and their popularity [59, 10]. Assistants are discussed in more detail in Section 3.5. For this reason, Apple Siri, Google, Alexa and Bixby were selected and their parameters are shown in Table 5.1. Some of these assistants run on multiple platforms, but we have chosen one platform per assistant. The reason for choosing one platform per assistant was that it would make the experiment exponentially more time consuming, since we use four tools to create voice deepfakes. The reasons why these four tools were chosen are described in the following section Section 5.3.

Throughout the experiment, a consistent software version was an important issue, as we found in the experiment design phase that different software versions had different effects on the behavior of the voice assistants and can have a significant impact on the results of the experiment. This precaution was necessary in order to ensure a constant environment throughout the research. In this way, we could achieve reliable and comparable results. Maintaining the stability of the software environment contributed to the validity of our findings and allowed us to more accurately analyze the reactions and behavior of the tested elements. The individual assistants and their parameters are listed in Table 5.2.

5.3 Used tools for creating voice deepfakes

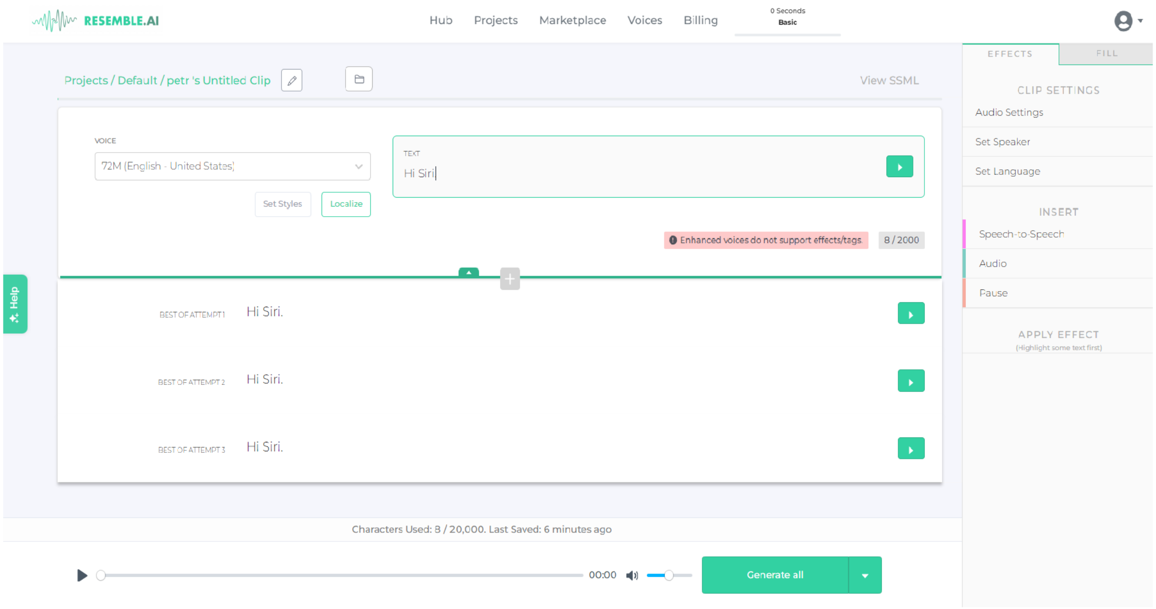
As mentioned in the previous section in the context of the experiment, four tools were used, ResembleAI[1], CoquiAI[2], TorToiSe[3] [8] and XTTS[4]. This selection of synthesis tools was made with respect to criteria that even a potential attacker would take into account, such as ease of use, availability and speed of deepfake creation. The number of synthesis

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

[2] Discontinued in 12/2023.

[3] https://github.com/neonbjb/tortoise-tts

[4] https://github.com/coqui-ai/TTS



Projects / Default / petr 's Untitled Clip

VOICE: 72M (English - United States)

TEXT: Hi Siri

Enhanced voices do not support effects/tags. 8 / 2000

BEST OF ATTEMPT 1 Hi Siri.

BEST OF ATTEMPT 2 Hi Siri.

BEST OF ATTEMPT 3 Hi Siri.

Characters Used: 8 / 20,000. Last Saved: 6 minutes ago

Figure 5.1: Screenshot of the main interface for voice synthesis in ResembleAI with an open pop-up window that allows you to choose between alternative recordings.

tools was also deliberately chosen, because we wanted to test different synthesis models while keeping the experiment size within an adequate range, as the experiment time would increase exponentially as the number of synthesizers increased. So, in the end, we selected four instruments, two of which were commercial and two of which were open source. Each of these tools will be discussed in detail in the following subsections. The lengths of the used recordings that were inserted into the instruments are written in Table 5.3.

5.3.1 ResembleAI

Is a web-based commercial text-to-speech and voice conversion tool that uses artificial intelligence and generative speech models to create realistic deepfakes.

We decided to choose this particular tool because of its easy-to-use interface and the good quality of the deepfakes produced. This tool also has a free version, but we decided not to use it because of the reduced resources for creating deepfakes and the extended time of their creation. The process of creating a custom model simply requires reading at least 25 predefined sentences in the user-friendly interface. Each sentence is read sequentially and a quality measurement is performed after each reading. These measurements determine whether the recordings need to be repeated, for example if the speaker speaks too softly or if there is significant background noise. After the last recording is uploaded, all recordings are sent and the process of training the model begins, which usually takes between ten minutes and one hour. The duration of the process depends on the current load on the ResembleAI servers. Once the model training is complete, we could start creating the first deepfakes. After creating a deepfake, the application offers the possibility of creating alternative recordings, as shown in Figure 5.1. This allows us to select the best quality version of the deepfake.

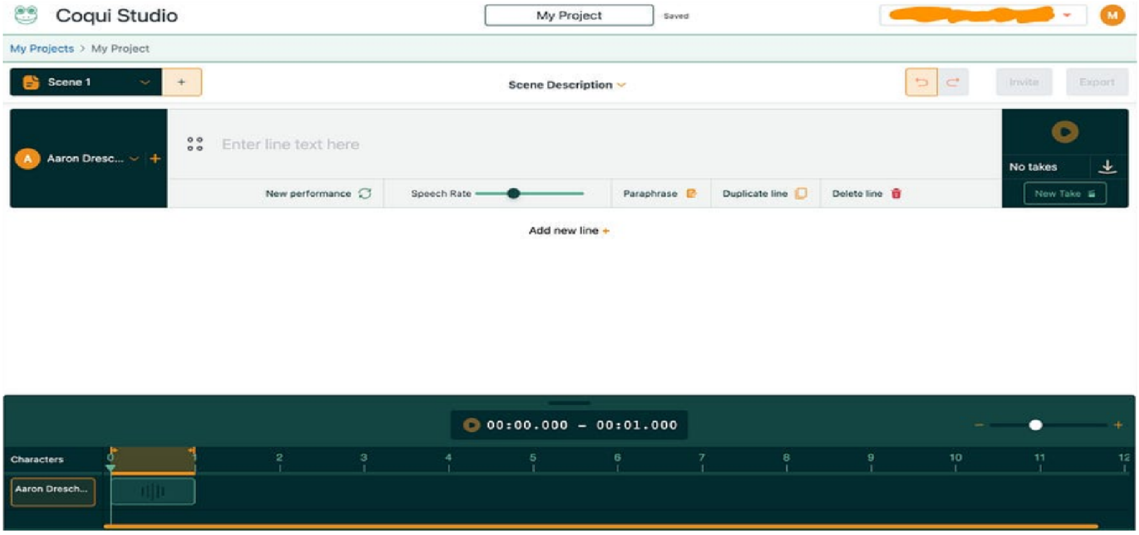


Figure 5.2: Screenshot of the interface Coqui in a state where the program waits for text input and then the user can start synthesis. The target speaker is selected on the left and the target text is inserted into the box in the center of the screen.

5.3.2 CoquiAI

It is also a web-based commercial tool that uses artificial intelligence to generate authentic voice deepfakes.

We chose this tool because of its ease of use and because of the sufficient quality of the resulting deepfakes for our experiment, which are produced in a very short time. Creating deepfake recordings was straightforward. The input recording had to be recorded separately and we used Audacity[5] for this. Next, all we had to do was insert a recording into the tool, which could be three seconds or more long, and then start the training, which usually took around ten seconds. The length of the training depended on the length of the text entered. Once training was complete, the resulting deepfake could be edited using the built-in tools in the app. These included speeding up the recording, slowing down the recording, increasing the depth of the voice, etc. Specifically, in most cases we had to use the slow-down option to adjust the recording to make it sound realistic, because for unknown reasons, the tool produced output records that sounded sped up by default. The Figure 5.2 shows the main page on which the voice was synthesized.

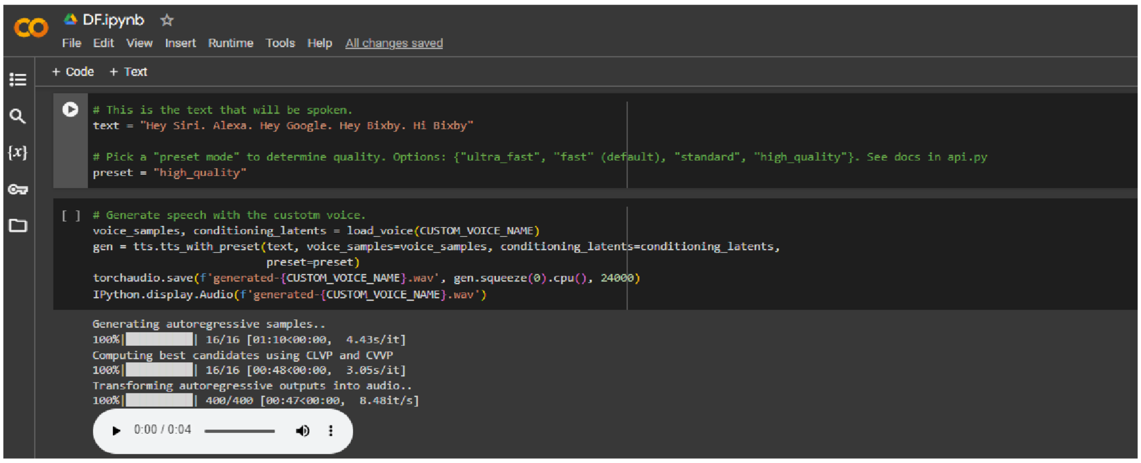
5.3.3 TorToiSe

It is an open source text-to-speech tool that we have been using through Google Colab[6] in Figure 5.3.

We chose this tool because we wanted to include an open source tool in the experiment. Because the tool works primarily as a script and the author of the tool recommends inserting six voice recordings lasting 10 seconds, we decided to pre-record the recordings again using Audacity. But because we wanted to get the best possible deepfake from the instrument

[5] Audacity is free and open-source audio editor and recording software that allows you to record, edit and mix audio files in formats such as WAV, AIFF, MP3 and more. https://www.audacityteam.org/

[6] Google Colab is a free service from Google that allows you to write and run Python code in a web browser. https://colab.google/



```
# This is the text that will be spoken.
text = "Hey Siri. Alexa. Hey Google. Hey Bixby. Hi Bixby"

# Pick a "preset mode" to determine quality. Options: {"ultra_fast", "fast" (default), "standard", "high_quality"}. See docs in api.py
preset = "high_quality"
```

```
# Generate speech with the custotm voice.
voice_samples, conditioning_latents = load_voice(CUSTOM_VOICE_NAME)
gen = tts.tts_with_preset(text, voice_samples=voice_samples, conditioning_latents=conditioning_latents,
                          preset=preset)
torchaudio.save(f'generated-{CUSTOM_VOICE_NAME}.wav', gen.squeeze(0).cpu(), 24000)
IPython.display.Audio(f'generated-{CUSTOM_VOICE_NAME}.wav')
```

```
Generating autoregressive samples..
100%|██████████| 16/16 [01:10<00:00,  4.43s/it]
Computing best candidates using CLVP and CVVP
100%|██████████| 16/16 [00:48<00:00,  3.05s/it]
Transforming autoregressive outputs into audio..
100%|██████████| 400/400 [00:47<00:00,  8.48it/s]
```

Figure 5.3: Screenshot of Google Colab running TorToiSe. Specifically, the situation after synthesis is captured and it is already possible to listen to the generated deepfake. In the upper code cell, we can issue a mode selection and insert the text we want the tool to pronounce.

and at the same time we wanted to keep the length of the experiment reasonable so that the responders would not find the recording annoying, we decided to do eight recordings of twenty seconds each. The tool was intuitive to use and was triggered by turning on the script. The tool loaded the necessary libraries and pre-trained models and then prompted us to insert our recordings, which we selected through a file dialog. Next, the tool allows us to type the text we want it to read in a synthetic voice in our case it was the wake-word phrases of each assistant. After that, the overhead was in the hands of the tool, which produced the final deepfake for about a minute. Sometimes it also happened that the synthesis result was insufficient and no words could be discerned. To solve this problem, we only had to re-run the script and in most cases we got a satisfactory recording on the second attempt, in a few cases even re-running it was not enough and therefore we kept running the script until we reached the desired quality. The maximum number of script restarts for a single recording was four attempts.

5.3.4 XTTS

It is a multilingual text-to-speech tool that uses zero-shot voice cloning. Again, the tool was used through Google Colab and requires only one recording, with a minimum of three seconds.

However, we did opt for a twenty-second recording length, which was also used for the previous tool. Specifically, it was always the first of eight. The tool was run as a classic Python script, but before it could be run, the input recording had to be inserted into the required folder, which was located in `/content/`, and the text field we required the tool to generate had to be filled in.

Since it is a zero-shot, it performs synthesis in a very short time, making its output mostly unintelligible to the human ear. To solve this problem, we again used multiple generation, so we created several recordings and selected the best one of them.

Name	Type	Min. Enrollment Sample	Used Enrollment Sample
CoquiAI	Paid	3 seconds	20 seconds
ResembleAI	Paid	25 sentences	25 sentences
TorToiSe	Open-source	6 * 10 seconds	8 * 20 seconds
XTTS	Open-source	3 seconds	20 seconds

Table 5.3: Overview of Used Speech Synthesizers

5.4 Experiment participants

The experimental subjects are randomly selected from the population and are individuals who are proficient in English and have reached the age of 18 years or older.

5.5 Details of the setup

In this section, the experimental design and its specific parameters are analyzed in detail. Given the use of four different tools, each requiring a specific input data format and four individual voice assistants, it is crucial to develop the most efficient procedure. At the same time, it is a priority to minimize the time spent with each respondent in order to conduct the experiment on a mass scale.

Another important aspect was the registration of the respondents in the voice assistant systems, with each assistant only being able to keep a single voiceprint. This limited number of registered voice identities put limitations on us in terms of running the experiment in parallel. After recording the voiceprint of one respondent, the entire experiment process had to be completed until the testing phase before we could proceed to register the next respondent. Therefore, it was not possible to record all the respondents and perform the tests retrospectively.

Next, a special experiment that was used to test a specific feature of voice assistants, which is automatic speaker recognition (ASR) is described in Subsection 5.5.1. This feature was tested to verify whether the voice assistant only uses ASR on the wake-word phrase and thus whether subsequent voice commands can be entered by anyone.

5.5.1 Experiment to verify assistant characteristics

The main purpose of conducting this experiment was to confirm the property mentioned previously, that is, the voice assistant performs ASR only on the invoking phrase. Once the assistant responded to this phrase, it would be clear that the subsequent voice command could be given by anyone.

The experiment was carried out with six people, from which we made three pairs. We registered the voiceprint of person A in the voice assistant. Person A uttered an invocation phrase to activate the voice assistant, while person B uttered an arbitrary command. This procedure was repeated five times for each pair. In all fifteen cases, the voice assistant's response to the command of Person B (who was not registered) was 100% successful. This confirmed our hypothesis of exclusive use of ASR only in the invoking phrase.

This verified property subsequently allowed us to test the voice assistants' response to the invocation phrase only in the main experiment. In practice, this meant that in the main experiment it was not necessary to create a deepfake for the invocation phrase associated with the command and only the invocation phrase was sufficient, which greatly simplified the process of generating artificial voice data.

5.6 Sample collection

The main experiment was initiated with the arrival of a participant who first provided informed consent (blank informed consent is stored on the storage medium attached to this thesis, the structure of which is in Appendix A). This consent included a provision regarding the anonymization of the voice data collected. Subsequently, the participant's voice was registered into the voice assistant. Once the registration was complete, each participant was subjected to 30 authentication trials with each of the assistants.

The next step was to upload the participant's voice into the ResembleAI tool. Since, as mentioned in Subsection 5.3.1, this tool has its own 25 predefined sentences that must be sequentially read directly into the tool, we decided to take advantage of this and in parallel uploaded the participant's voice into the Audacity tool. So, the capture was done with two tools at the same time. After this, the first recording block was completed. This was followed by the recording of 75 sentences we made up and displayed in Appendix B. This seemingly high number of recorded phrases was chosen as a compromise of the minimal input needs of the individual synthesizers. After the recording was finished, the presence of the respondent was no longer necessary and we further trimmed the recordings to the necessary lengths to match the input lengths of the individual instruments.

We have combined these sentences into nine recordings. The first recording consisted of five sentences that lasted approximately seven seconds. The first sentence was used to test-activate the microphone during the recording and was then removed from the final recording, followed by the four invocation phrases of each assistant. The remaining eight recordings of twenty seconds each were used for the TorToiSe instrument. Finally, only the second recording of the nine recordings was used for the CoquiAI and XTTS instruments. We synthesized one recording from each instrument that contained the wake-word phrases of each assistant. Thus, in the end, 16 recordings (four assistants and four synthesizers) were created. The Figure 5.4 visually shows the temporal distribution of the individual recordings and what they were used for.

This completed the preparation of the test data and the testing phase began. First, we tested the replay attack by playing the original sentences with the invocation phrases for each assistant. After synthesis was completed on all instruments, we started playing individual deepfake recordings from different instruments to all assistants. Each recording was played a total of thirty times. Thus, one test block took 150 playbacks of the recordings (120 playbacks of the output of each instrument and 30 playbacks of the original recording). The Figure 5.5 shows the progress of testing a respondent and the individual time requirements for different test blocks.

The entire testing, including the part when the respondent was involved in the experiment, took approximately **an hour and thirty minutes**. So, if we break the final time down into individual parts it was: registering a person into the assistants took 10 minutes, recording the respondent's voice took 15 minutes, creating deepfakes took from ten minutes to sixty minutes. Furthermore, performing replay attacks took ten minutes, mainly the length depended on ResembleAI and the load on its servers. Finally, deepfake spoofing attacks took forty five minutes. We decided to omit bonafide tests from our research for two reasons. The first reason was their time-consuming nature. These tests require the

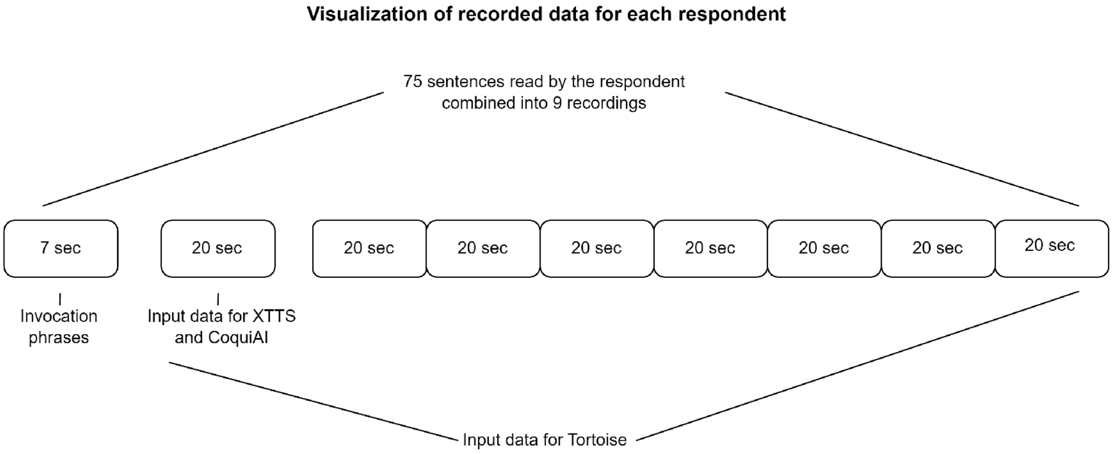

Visualization of recorded data for each respondent
75 sentences read by the respondent combined into 9 recordings
7 sec
20 sec
20 sec
20 sec
20 sec
20 sec
20 sec
20 sec
20 sec
Invocation phrases
Input data for XTTS and CoquiAI
Input data for Tortoise

Figure 5.4: Visualisation of the content of each recording, the number of sentences it contains and the purpose of each recording.

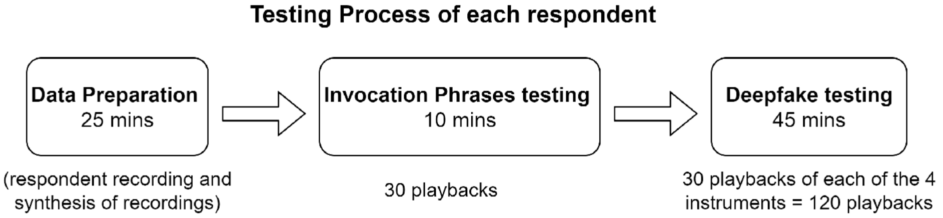

Testing Process of each respondent
Data Preparation
25 mins
Invocation Phrases testing
10 mins
Deepfake testing
45 mins
(respondent recording and synthesis of recordings)
30 playbacks
30 playbacks of each of the 4 instruments = 120 playbacks

Figure 5.5: Averaged time complexity of the experiment for testing one respondent.

respondent to be present at all times, which would have increased the respondent's time with the experiment from 25 minutes to 40 minutes. The second reason was the lack of variability in the results of authentic tests. These tests often provide consistent results without significant variation. During the initial tests with 36 respondents, we repeatedly observed high success rates, with accuracies exceeding 95%. This suggested to us that bofide tests in the overall experiment would not be very informative for our research objectives. Given that our main goal was to test deepfake attacks and to make the study as convenient as possible for participants, we decided to omit these tests, saving respondents fifteen minutes. This decision will not affect the validity of our research in any way. Rather, it should reflect the fact that voice assistants are primarily commercial products whose design and features are optimized for user-friendliness and often favor comfort over security.

Chapter 6

Course of the Experiment

We conducted the experiment according to the developed design. The experiment was carried out on **72** test subjects. The total duration of the experiment was **2 months**. Evaluation and description of the result of the experiment is described in this chapter.

6.1 Participants in the experiment

The experiment was initiated by searching for people who were interested and willing to participate in this research. Participants were contacted one by one mainly by personal encounter, where the principle of the experiment was explained to them, the time their personal presence is required and the reward in the form of VUT merch they would receive for their participation. Most of the people agreed and knowingly of their own volition helped us in the experiment.

6.2 Demographic distribution of participants

As mentioned earlier in the experiment, **seventy-two** participants took part, from whom we collected information about their age and gender. And hence in this section we will take a closer look at the proportions of the respondents. All subjects met the predetermined requirements set out in the section 5.4.

If we look at the age distribution of the respondents we can see that the main age distribution of the respondents is specifically between the ages of twenty-two to twenty-five years and it is 41.6%. The second largest representation is between the ages of eighteen to twenty-one and it is exactly 25%. We further calculated the average age of the respondent and it was 29.8 years. More clearly, the age distribution of the respondents is recorded in Figure 6.1.

The demographic chart of the group of respondents examined in this paper reflects interesting and significant aspects of the gender distribution. The Figure 6.2 shows that of the 72 respondents, 72.22% are male and 27.78% are female. These distributions can provide deeper insights into the gendered aspects of the issue under study, which will be addressed in Chapter 7.

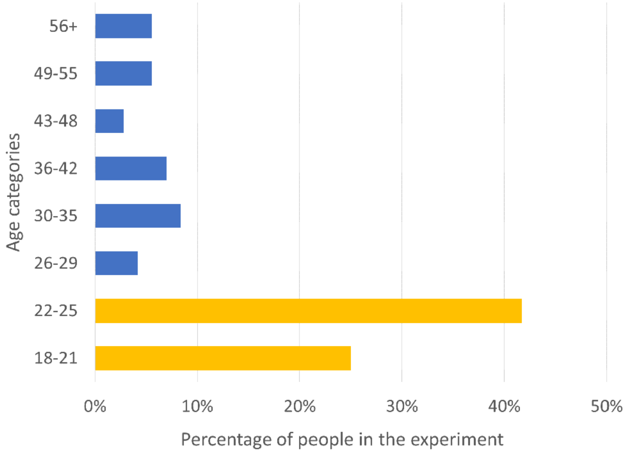


Figure 6.1: Percentage age distribution of participants in the experiment.

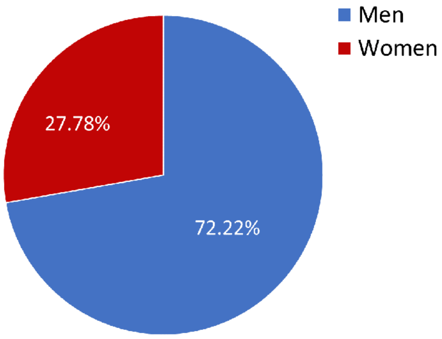


Figure 6.2: Gender representation of participants in our experiment

6.3 Course of the experiment

The experiment went very smoothly, without any major problems. It was even completed in a very short period of **2 months**. In the beginning, we tried to schedule respondents to come in at the time they were supposed to come in. But this did not work very well, as participants often did not respect the agreed dates or did not come at all at the appointed time, which, given the nature of the experiment, caused us considerable organizational and time difficulties.

In response to these problems, we abandoned scheduling respondents in advance. Instead, after completing one respondent, we immediately sought another respondent, whom we then tested. On average, we tested three respondents each day, which amounted to approximately three to four hours per day. Only the experimental time is included in this time, so it does not include the time of searching for responedents, so the resulting time may still be slightly different.

To maintain consistency in the recording environment for all participants, only the integrated microphone in the Lenovo ThinkPad E590[1] laptop was used. During the course of the experiment, a technical malfunction occurred on this recording laptop, which took

[1] https://www.lenovo.com/cz/cs/laptops/thinkpad/edge-series/E590/p/22TP2TEE590

us 5 days to repair. However, it is important to note that the repair of this glitch had no effect on the microphone's function. So the recording environment remained consistent and did not change.

Most of the experiment took place in a quiet room in the student residence halls, which explains the characteristics of the age group of the participants Figure 6.1. Another setting for the experiment was the spacious and quiet meeting room of firm SCHOTT[2], which is illustrated by the photograph of the meeting room Figure 6.3. We asked the company to participate in the experiment so that we could have representatives from other age groups than eighteen to twenty-five, and the company agreed.

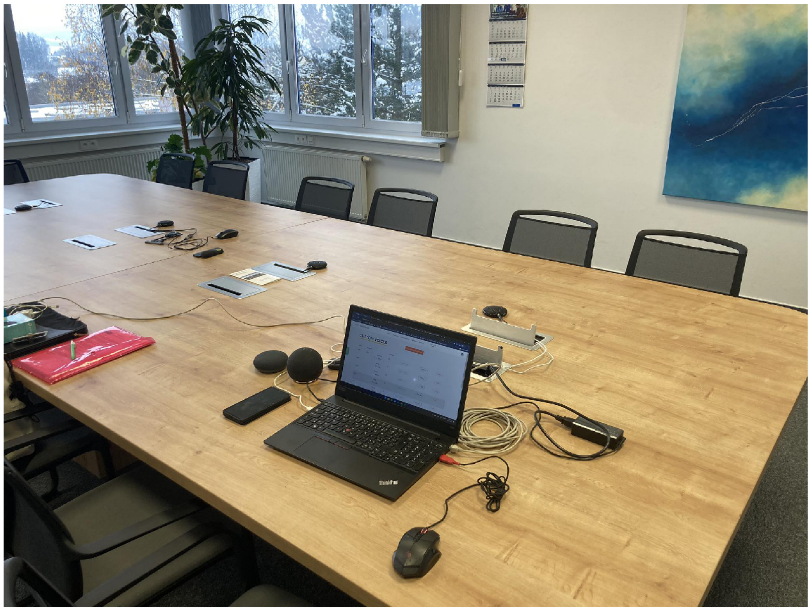


Figure 6.3: Demonstration Setup - For the standard setup, normal IT equipment was used in a normal room and at a standard distance of the user from the technical resources.

6.4 Metrics used

In order to quantify the effectiveness of each verification, a success rate was introduced based on verification attempts. For each unique combination of participant, voice assistant, and speech synthesizer, the ratio of successful attempts to the total number of attempts made (30 in total) was calculated. This ratio was then converted to a percentage representing the proportion of successful trials to the total number of experiments.

The success rate is a key metric, with an optimal value approaching 100% for bonafide matching trials reflecting high reliability. On the contrary, a value approaching zero for attempts to resist deepfake fraud attacks indicates a high level of security and robustness to manipulation.

Mathematically, the success rate for each participant, voice assistant, and speech synthesizer can be expressed by the following formula.

$$\textit{success rate (\%)} = \left(\frac{\text{number of successful trials}}{30} \right) * 100$$

[2] https://www.schott.com/en-ro/about-us/company/regions-and-locations/lanskroun

This methodology provides an objective and quantifiable way to assess the success of verification and robustness of the system.

However, it is important to note that this metric was created primarily to facilitate the evaluation of our results in this thesis. In the real world, an attacker need only achieve one successful attempt to cause potential damage. If the attacker fails to successfully execute the attack the first time, he can still repeat his attempts indefinitely. In fact, voice assistants place no limit on the number of attempts to enter a command, creating ideal conditions for an attacker to potentially launch a brute force attack.

Chapter 7

Experiment Evaluation

Baseline data from our research show that bonafide attempts were collected from **72 respondents**, and the success rate of these attempts was unremarkably above the respectable 95%. In the absence of any significant fluctuations in observed success rates, we decided to discontinue bonafide testing, with an emphasis on saving respondents time and effort.

The results suggest that the assistants we included in our experiment are designed primarily to maximize usability for users and are not overly concerned with defensive properties against potential deepfake or replay attacks, as confirmed by the high success rates we observed during testing.

7.1 Main results

This section provides a detailed and comprehensive overview of the main results of our experiment. In addition, this section presents the findings that we have achieved through data analysis and evaluation of the experimental results.

The detailed distribution of the success rates for different types of attacks is illustrated in Figure 7.1. An interesting observation is that the replay attacks achieved success approximately every second attempt, while some deepfake models were able to reproduce the bonafide success rate with over 90% accuracy. We believe that this result was due to the fact that some respondents did not understandably pronounce the commands to the assistants. Hence, the assistant did not respond. On the other hand, the deepfake recordings were made from more recordings and hence the resulting voice was more intelligible and more acceptable to the assistants as the result suggests.

Furthermore, the results of our analysis reveal that Bixby systematically resisted most attempted attacks, potentially making it one of the most secure assistants compared to the others tested. On the contrary, the other assistants were mostly fooled, considering most of the attacks legitimate attempts. However, given the proprietary nature of Bixby's internal mechanisms and the lack of published details about the parameters of its deep learning model, we find ourselves in a difficult situation when trying to identify the specific factors that contribute to its significantly improved security performance.

Paid synthesizers excel at achieving significantly high success rates, which eloquently demonstrates the real likelihood of such an attack. Open-source XTTS is an exception in this regard, deviating from the excellent trend of paid tools, yet it still retains the potential for manipulation and its success rate extends across the spectrum. Unfortunately, the data collected do not reflect a clear pattern that explains this distribution. It may be influenced

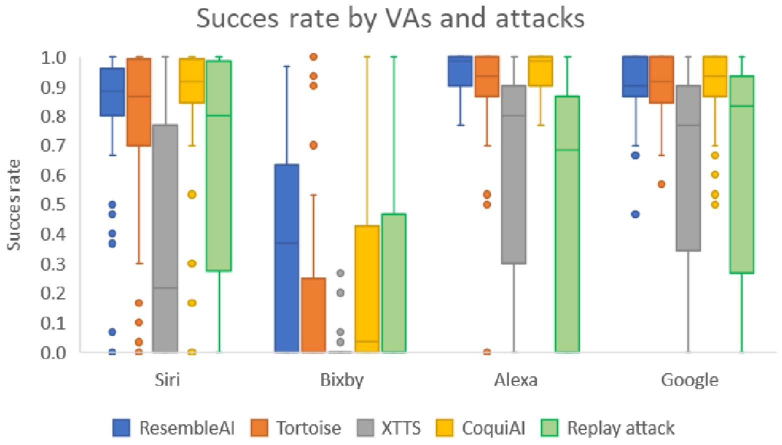


Figure 7.1: Attack success rates targeting voice assistants.

by the individual voice characteristics of each respondent; some may possess a voice similar to that used in XTTS training, contributing to their increased deepfake success rate and vice versa.

Given the generally higher speech quality of paid synthesizers, it is clear that these synthesizers achieve optimal results. Nevertheless, even open-source TorToiSe has managed to approach the level of paid synthesizers. However, even the least powerful tool (XTTS) achieved success at least once for most of the respondents. Due to the lack of a set limit for verification attempts, an attacker may spend more time on repeated attempts until one of them succeeds. The lower success rate also increases the complexity of the whole attack in time.

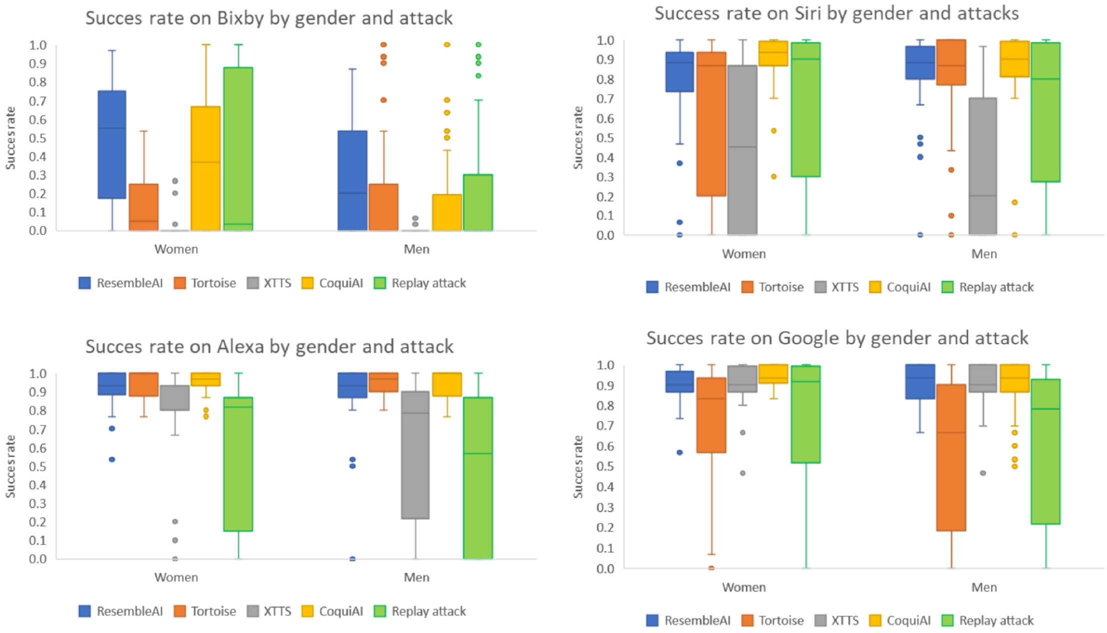
In general, the success rate of deepfake spoofing attacks is significantly high. Thus, the manipulation of voice assistants appears to be a relatively accessible process, causing legitimate security concerns.

7.2 Secondary results and their interpretation

In this section, we focus on the analysis and observations that arise outside the main framework of our experiment. These results bring with them interesting and unpredictable aspects.

Specifically, they are concerned with the gender distribution of attacks. That is, on the success rate of female versus male deepfakes. However, it should be taken into account that the number of representatives of both genders was not equal, as seen in the figure Figure 6.2, and thus the results may be biased due to this numerical difference. Since we have no way to control this parameter retrospectively, this section serves rather to point out and reflect on the possibilities to further investigate these aspects in future experiments. At the same time, we discuss the possible impact of these differences on the interpretation of the results.

A detailed evaluation of the success rate of attacks on individual voice assistants, broken down by gender, is shown in Figure 7.2. An interesting observation is the performance of the Bixby assistant, which, ignoring the different ratio of male and female participants, shows a higher success rate for female deepfakes in both the ResembleAI and CoquiAI commercial



Succes rate on Bixby by gender and attack

Success rate on Siri by gender and attacks

Succes rate on Alexa by gender and attack

Succes rate on Google by gender and attack

Figure 7.2: Attack success rates for individual assistants split by gender.

tools and replay attacks. This remarkable behavior contrasts with other assistants, where the success rate of male deepfakes is higher than that of female deepfakes.

Possible reasons for this behavior may be due to the data on which the assistant model is trained (that it was trained primarily on male voices and therefore detects more male deepfake attacks), the version of the assistant software, or some security feature of the assistant. Unfortunately, we are unable to determine the causes of this behavior because we do not have access to information about the specific algorithms and features of the assistant. More in-depth analysis and collaboration with assistant developers would be necessary to clarify these issues.

7.3 Defence Strategies

Our experiment shows that existing voice assistant defenses against deepfake as well as recording attacks are unsatisfactory. There are already some suggestions on how to secure voice asistents. In this section we will discuss some of them.

The Challenge-Response scheme represents one possible defense against unauthorized access or misuse of voice assistants [12, 13, 60, 67]. This defense consists of the voice assistant posing a question, i.e., a challenge, to the user and requiring the user to provide the correct answer within a limited time frame. If the user does not respond correctly to the prompt, the voice assistant will not execute the received command.

Similar to the well-known CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart), this method can use audio prompts and responses to verify that the speaker is indeed human. This layer of security aims to prevent automated attacks and ensure that the interaction with the voice assistant is with a human person.

However, it is important to note that while these methods may provide a level of security, it may also impose additional usability costs. Adding a step that requires a user response can slow down the interaction process and can be perceived as limiting. This could be addressed by adding security only to the most critical services. Additionally, any security measures can be bypassed if an attacker finds a way to adequately respond to the challenge, which highlights the need for continuous innovation in voice assistant security.

Another way to protect voice assistants is through liveness detection. This article presents the LiveEar [68] system, which focuses on detecting voice authenticity when using voice assistants. It exploits differences in phoneme phases between live human voices and voices reproduced by a speaker. Specifically, it calculates the time difference of arrival (TDoA) in the sequence of phoneme sounds to the voice assistant microphone. Subsequently, an SVM-based classification model is trained with the extracted TDoA features.

The last proposed defense [56] of voice assistants mentioned in this paper is based on an approach that specifically proposes protection of automatic speaker verification (ASV) against deepfake recording attacks. This approach focuses on the analysis of sound artifacts, in particular the so-called pop noise. This sound artifact is produced by the interaction between the speaker's breath and the microphone and is a key element for the presented detection method to identify synthesized recordings. This paper [56] presents two detection methods that differ in the number of channels used in pop noise analysis and in their combination.

The single-channel approach focuses on the analysis of low-frequency energy in voice signals, which helps to identify areas associated with pop noise. In contrast, the two-channel approach uses two microphones, one with a pop noise filter and one without, and then works with the subtracted signal, reducing the influence of general noise and allowing only pop noise to be captured.

The most significant feature of the paper is the innovative tandem algorithm that combines both methods into one. The tandem algorithm first processes the input signal using dual-channel pop noise detection and then analyzes the subtracted signal using single-channel pop noise detection. This combination provides significantly better results compared to individual methods and increases the overall pop noise detection capability. Thus, the tandem approach more effectively identifies irregularities in the readout signal, ultimately increasing the ability to detect fraudulent materials.

Chapter 8

Discussion

Our research demonstrates how vulnerable voice assistants are to deepfake attacks. Although a tool like XTTS was not particularly successful and the Bixby voice assistant resisted most attacks, it is crucial to highlight a few key facts that we found.

The first point is that the lower success rate of an attack actually only increases the time it takes to execute it, but not its effective defense. For example, achieving the approach in only one of 30 attempts is sufficient to consider the attack successful. Thus, even if the voice assistant rejects most attempts, it is only a matter of time before the attacker achieves success. Bixby, which is considered the most secure of the voice assistants, resisted access attempts in only seven of 72 subjects, where it rejected access in all 30 attempts. On the contrary, for the other voice assistants, every subject gained access at least once. Thus, it is clear that while a higher success rate of an attack indicates its effectiveness, any success rate greater than zero ultimately leads to the same result of the attacker gaining access.

The second fact is that voice assistants are not designed to limit the number of access attempts. This means that they constantly listen for activating phrases like „Hey Siri" without making a distinction between the device owner and others. This characteristic means that even a single successful attempt can be considered effective for an attacker, as unlimited attempts are available.

8.1 Observations

The observations we made during our evaluation could potentially affect the ultimate outcomes of our experiment. Yet, to comprehend these outcomes fully and their potential effect on our experiment, these factors would need to be examined in detail in a subsequent experiment. However, such an investigation would exceed the boundaries of our current research and could be a recommendation for future research.

The aforementioned observations are:

O1: Sometimes, when attempting bonafide or deepfake activations, Google Assistant responds to the wake word „Hey Siri".

O2: Bixby's success rate improves when there's a longer pause between saying „Hey" and „Bixby.

O3: When recording respondents, some read sentences unnaturally fast, which led to a reduction in the quality of the resulting deepfake.

O4: "Some participants incorrectly pronounced the wake words, like saying "Hey Siiiiiiri-iiiiii".

Chapter 9

Conclusion

We conducted an extensive experiment as part of this work, which clearly confirms the lack of robustness of current voice assistants to replay and deepfake attacks. The findings show that there is a real danger that a potential attacker can relatively easily record the victim's voice, synthesize his speech, and then use this fake speech to manipulate the victim's voice assistant. In this way, the attacker could reveal sensitive personal information or cause financial harm.

In this context, the choice of appropriate use cases for voice assistants is a key element of decision-making. It is necessary to consider in detail when it is appropriate to use voice authentication and when it is necessary to implement more robust security measures. A through analysis of potential threats reveals the potential risks of privacy breaches and the potential financial damage that can result from inadequate protection against these types of attacks. It is imperative that organizations carefully assess risks and take measures that address the specific security challenges associated with the use of voice assistants to minimize potential privacy and financial damage.

At the same time, it is important to stress that most voice assistant developers are fully aware of the security risks associated with the speaker recognition functionality implemented in their systems. For this reason, they actively limit the ability of assistants to operate critical functions through voice commands. However, in the case of third-party developers or end-users, these speech recognition features may be used to control devices or authorize online payments. This practice poses significant security challenges and potential risks, as uncontrolled use of these features can result in loss of privacy and exposure to security threats. Organizations and developers should therefore be vigilant and implement measures to control access and secure these key speech recognition features.

Furthermore, this thesis proposes security measures that could support the protection of assistants against deepfake and recording attacks.

Bibliography

[1] Adam, D. E. E. B. Deep learning based NLP techniques in text to speech synthesis for communication recognition. *Journal of Soft Computing Paradigm*. december 2020, vol. 2, no. 4, p. 209–215. DOI: 10.36548/jscp.2020.4.002. ISSN 2582-2640. Available at: https://irojournals.com/jscp/article/view/2/4/2.

[2] Alam, R. and Akilarasu, G. Evaluating use of biometric authentication for face and voice recognition. *Advances and Applications in Mathematical Sciences*. Mili Publications, India. 2022, vol. 21, no. 8, p. 4747–4760. 2020 Mathematics Subject Classification: 68T01.

[3] Alattas, K. and Bayoumi, M. Artificial Intelligence in Deepfake Technologies Based on Supply Chain Strategy. *Int. J. Sup. Chain. Mgt*. Inderscience Enterprises Ltd. October 2020, vol. 9, no. 5, p. 411. ISSN 2050-7399. Available at: https://www.ojs.excelingtech.co.uk/index.php/IJSCM/article/viewFile/5671/2936.

[4] Alegre, F., Janicki, A. and Evans, N. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In: EURECOM and Warsaw University of Technology. *Proceedings of the Conference Name*. Sophia Antipolis, France and Warsaw, Poland: [b.n.], 2023. ISBN 978-3-88579-624-4.

[5] Amazon. *How Alexa Responds How to speak so that people can easily understand and respond* [Amazon Developer]. 2023. Available at: https://developer.amazon.com/fr/designing-for-voice/what-alexa-says/.

[6] Asaolu, O. S., Folorunso, C. and Popoola, O. A Review of Voice-Base Person Identification: State-of-the-Art. *Journal of Engineering Technology*. june 2019, vol. 3, p. 36–57. DOI: 10.20370/2cdk-7y54. Available at: https://journals.covenantuniversity.edu.ng/index.php/cjet/article/view/1635.

[7] Bateman, J. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Carnegie Endowment for International Peace, 2020. i–ii p. Available at: http://www.jstor.org/stable/resrep25783.1.

[8] Betker, J. *Better speech synthesis through scaling*. 2023. DOI: 10.48550/arXiv.2305.07243. Available at: https://arxiv.org/abs/2305.07243.

[9] Bilika, D., Michopoulou, N., Alepis, E. and Patsakis, C. Hello me, meet the real me: Voice synthesis attacks on voice assistants. *Computers and Security*. 2024, vol. 137, p. 103617. DOI: 10.1016/j.cose.2023.103617. ISSN 0167-4048. Available at: https://www.sciencedirect.com/science/article/pii/S0167404823005278.

[10] BotPenguin. *Which are the 7 best voice assistants of 2023?* Nov 2023. Available at: https://botpenguin.com/blogs/which-are-the-7-best-voice-assistants-of-2023.

[11] Cankaya, E. Authentication. In: Tilborg, H. C. A. van and Jajodia, S., ed. *Encyclopedia of Cryptography and Security*. Boston, MA: Springer US, 2011, p. 61–62. DOI: 10.1007/978-1-4419-5906-5_772. ISBN 978-1-4419-5906-5. Available at: https://doi.org/10.1007/978-1-4419-5906-5_772.

[12] Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M. et al. Hidden Voice Commands. In: *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, August 2016, p. 513–530. ISBN 978-1-931971-32-4. Available at: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini.

[13] Chang, Y.-T. and Dupuis, M. J. My Voiceprint Is My Authenticator: A Two-Layer Authentication Approach Using Voiceprint for Voice Assistants. In: *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*. 2019, p. 1318–1325. DOI: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00243. ISBN 978-1-7281-4035-3.

[14] Childers, D., Wu, K., Hicks, D. and Yegnanarayana, B. Voice conversion. *Speech Communication*. 1989, vol. 8, no. 2, p. 147–158. DOI: https://doi.org/10.1016/0167-6393(89)90041-1. ISSN 0167-6393. Available at: https://www.sciencedirect.com/science/article/pii/0167639389900411.

[15] Cho, G., Choi, J., Kim, H., Hyun, S. and Ryoo, J. Threat modeling and analysis of voice assistant applications. In: Kang, B. B. and Jang, J., ed. *Information Security Applications - 19th International Conference, WISA 2018, Revised Selected Papers*. Springer Verlag, 2019, 11402 LNCS, p. 197–209. Lecture Notes in Computer Science. ISBN 9783030179816.

[16] Daniel Ruby. *65 Voice Search Statistics For 2023 (Updated Data)*. 2023. Available at: https://www.demandsage.com/voice-search-statistics/.

[17] Dietmar, J. GANs and deepfakes could revolutionize the fashion industry. *Forbes*. 2019. Available at: https://www.forbes.com/sites/forbestechcouncil/2019/05/21/gans-and-deepfakes-could-revolutionize-the-fashion-industry/.

[18] Ernestus, M. Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*. 2014, vol. 142, p. 27–41. DOI: https://doi.org/10.1016/j.lingua.2012.12.006. ISSN 0024-3841. SI: Usage-Based and Rule-Based Approaches to Phonological Variation. Available at: https://www.sciencedirect.com/science/article/pii/S0024384113001307.

[19] Erro, D., Moreno, A. and Bonafonte, A. INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora. *IEEE Transactions on Audio, Speech, and Language Processing*. 2010, vol. 18, no. 5, p. 944–953. DOI: 10.1109/TASL.2009.2038669.

[20] Evans, N., Kinnunen, T. and Yamagishi, J. Spoofing and countermeasures for automatic speaker verification. In: *Proceedings of INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*. Lyon, France: [b.n.], August 2013. DOI: 10.21437/Interspeech.2013-288.

[21] Firc, A. and Malinka, K. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. New York, NY, USA: Association for Computing Machinery, 2022, p. 1646–1655. SAC '22. DOI: 10.1145/3477314.3507013. ISBN 9781450387132. Available at: https://doi.org/10.1145/3477314.3507013.

[22] Firc, A., Malinka, K. and Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*. 2023, vol. 9, no. 4, p. e15090. DOI: https://doi.org/10.1016/j.heliyon.2023.e15090. ISSN 2405-8440. Available at: https://www.sciencedirect.com/science/article/pii/S2405844023022971.

[23] Frankovits, G. and Mirsky, Y. Discussion Paper: The Threat of Real Time Deepfakes. In: *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*. Ithaca: Association for Computing Machinery, 2023, p. 20–23. WDC '23. DOI: 10.1145/3595353.3595881. ISBN 9798400702037. Available at: https://doi.org/10.1145/3595353.3595881.

[24] Gernot, T. and Rosenberger, C. Robust biometric scheme against replay attacks using one-time biometric templates. *Computers and Security*. 2024, vol. 137, p. 103586. DOI: 10.1016/j.cose.2023.103586. ISSN 0167-4048. Available at: https://www.sciencedirect.com/science/article/pii/S0167404823004960.

[25] Gonfalonieri, A. *How Amazon Alexa works? Your guide to Natural Language Processing (AI)*. 21. november 2018. [online]. Available at: https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-languageprocessing-ai-7506004709d3.

[26] Gupta, P., Patil, H. and Guido, R. Vulnerability issues in Automatic Speaker Verification (ASV) systems. *J AUDIO SPEECH MUSIC PROC*. 2024, vol. 10, p. 10. DOI: 10.1186/s13636-024-00328-8. Available at: https://doi.org/10.1186/s13636-024-00328-8.

[27] Hoy, M. B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*. 2018, vol. 37, no. 1, p. 81–88. DOI: 10.1080/02763869.2018.1404391. Available at: https://doi.org/10.1080/02763869.2018.1404391.

[28] Huang, Y., Li, Z., Zhou, F. and Zhu, R. Robust AN-Aided Beamforming Design for Secure MISO Cognitive Radio Based on a Practical Nonlinear EH Model. *IEEE Access*. 2017, vol. 5, p. 14011–14019. DOI: 10.1109/ACCESS.2017.2730220. ISSN 2169-3536.

[29] James, J., Fatima, S., Avasthi, S., Nalawade, C., James, J. et al. A Mobile Application for Voice Enabled Virtual Bot. *International Journal of Applied Engineering Research*. august 2018, vol. 13, p. 13118–13122. ISSN 0973-4562. Available at: https://www.ripublication.com/ijaer18/ijaerv13n17_20.pdf.

[30] Jansen, M. *What is Google Assistant? Here's the guide you need to get started.* March 19 2023. Available at: https://www.digitaltrends.com/mobile/what-is-google-assistant/.

[31] Jiřík, P. *7 Benefits of Voice Biometric Authentication in Call Centers.* 25. březen 2021. Available at: https://www.phonexia.com/blog/7-benefits-of-voice-biometric-authentication-in-call-centers/.

[32] Kalaiarasu, S., Rahman, N. A. A. and Harun, K. S. Deepfake impact, security threats and potential preventions. *AIP Conference Proceedings.* 2024, vol. 2802, no. 1. DOI: 10.1063/5.0183097. ISSN 0094-243X. Available at: https://doi.org/10.1063/5.0183097.

[33] Kang, W. *Speaker Anonymization using End-to-End Zero-Shot Voice Conversion.* 2022. Master's thesis. Massachusetts Institute of Technology. Available at: https://hdl.handle.net/1721.1/144662.

[34] Khan, A. and Malik, K. M. Securing Voice Biometrics: One-Shot Learning Approach for Audio Deepfake Detection. In: *2023 IEEE International Workshop on Information Forensics and Security (WIFS).* 2023, p. 1–6. DOI: 10.1109/WIFS58808.2023.10374968. ISBN 9798350324914.

[35] Kietzmann, J., Lee, L. W., McCarthy, I. P. and Kietzmann, T. C. Deepfakes: Trick or treat? *Business Horizons.* 2020, vol. 63, no. 2, p. 135–146. DOI: 10.1016/j.bushor.2019.11.006. ISSN 0007-6813. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING. Available at: https://www.sciencedirect.com/science/article/pii/S0007681319301600.

[36] Këpuska, V. and Bohouta, G. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC).* 2018, p. 99–103. DOI: 10.1109/CCWC.2018.8301638. ISBN 978-1-5386-4649-6.

[37] Malathi R. and Jeberson Retna Raj R.. An Integrated Approach of Physical Biometric Authentication System. *Procedia Computer Science.* 2016, vol. 85, p. 820–826. DOI: 10.1016/j.procs.2016.05.271. ISSN 1877-0509. International Conference on Computational Modelling and Security (CMS 2016). Available at: https://www.sciencedirect.com/science/article/pii/S1877050916306214.

[38] Markowitz, J. Voice biometrics. *Communications of the ACM.* New York: ACM. 2000, vol. 43, no. 9, p. 66–73. ISSN 0001-0782. Available at: https://www.researchgate.net/publication/27293606_Voice_Biometrics.

[39] Masood, M., Nawaz, M., Malik, K., Javed, A., Irtaza, A. et al. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence.* june 2022, vol. 53, no. 4, p. 1–53. DOI: 10.1007/s10489-022-03766-z. ISSN 0924-669X.

[40] Mirsky, Y. and Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* New York, NY, USA: Association for Computing Machinery. jan 2021, vol. 54, no. 1. DOI: 10.1145/3425780. ISSN 0360-0300. Available at: https://doi.org/10.1145/3425780.

[41] Mohammadi, S. H. and Kain, A. An overview of voice conversion systems. *Speech Communication*. 2017, vol. 88, p. 65–82. DOI: 10.1016/j.specom.2017.01.008. ISSN 0167-6393. Available at: https://www.sciencedirect.com/science/article/pii/S0167639315300698.

[42] Murphy, G., Ching, D., Twomey, J. and Linehan, C. Face/Off: Changing the face of movies with deepfakes. *PLOS ONE*. Public Library of Science. july 2023, vol. 18, no. 7, p. 1–19. DOI: 10.1371/journal.pone.0287503. ISSN 1932-6203. Available at: https://doi.org/10.1371/journal.pone.0287503.

[43] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y. K. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*. 2023, vol. 154, p. 113368. DOI: https://doi.org/10.1016/j.jbusres.2022.113368. ISSN 0148-2963. Available at: https://www.sciencedirect.com/science/article/pii/S0148296322008335.

[44] Nacimiento García, E., Caballero Gil, C., Nacimiento García, A. and González González, C. Alexa, Do What I Want To. Implementing a Voice Spoofing Attack Tool for Virtual Voice Assistants. In: Bravo, J., Ochoa, S. and Favela, J., ed. *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022)*. Cham: Springer International Publishing, 2023, p. 413–418. ISBN 978-3-031-21333-5.

[45] Navlakha, M. Deepfakes of Taylor Swift have gone viral. How does this keep happening? *Name of the Publication (if available)*. January 2024. Article on the rise of AI-generated porn and the need for legal and societal change. Available at: https://mashable.com/article/taylor-swift-viral-deepfake-porn-explainer.

[46] Ning, Y., He, S., Wu, Z., Xing, C. and Zhang, L.-J. A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*. 2019, vol. 9, no. 19. DOI: 10.3390/app9194050. ISSN 2076-3417. Available at: https://www.mdpi.com/2076-3417/9/19/4050.

[47] Pandey, R. and Shah, P. *Samsung Bixby: What it is and how to use it*. September 3 2023. Available at: https://www.androidpolice.com/what-is-samsung-bixby/.

[48] Patrizi, M., Vernuccio, M. and Pastore, A. "Hey, voice assistant!" How do users perceive you? An exploratory study. *Sinergie Italian Journal of Management*. 2021, vol. 39, no. 1, p. 173–192. DOI: 10.7433/s114.2021.10. ISSN 0393-5108.

[49] Petkauskas, V. Report: Number of Expert-Crafted Video Deepfakes Double Every Six Months. November 15 2023. Deputy Editor. Available at: https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months/.

[50] Pinto, R. *Voice Authentication: How It Works & Is It Secure?* July 28 2021. Available at: https://www.1kosmos.com/biometric-authentication/voice-authentication/.

[51] Preminger, A. and Kugler, M. B. The Right of Publicity Can Save Actors from Deepfake Armageddon. *Berkeley Technology Law Journal*. 2023, Northwestern Public

Law Research Paper No. 23-52, p. 44. Forthcoming. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4563774.

[52] Qian, K., Zhang, Y., Chang, S., Yang, X. and Hasegawa Johnson, M. Autovc: Zero-shot voice style transfer with only autoencoder loss. In: PMLR. *International Conference on Machine Learning*. Cornell University Library, arXiv.org, 2019, p. 5210–5219. DOI: 10.48550/arXiv.1905.05879. ISSN 2331-8422.

[53] Sasirekha, D. and Chandra, E. Text to speech: a simple tutorial. *International Journal of Soft Computing and Engineering (IJSCE)*. Citeseer. 2012, vol. 2, no. 1, p. 275–278. ISSN 2231-2307.

[54] Seymour, J. and Aqil, A. *Your Voice is My Passport*. 2018. Available at: https://www.blackhat.com/us-18/briefings/schedule/#your-voice-is-my-passport-11395.

[55] Shaaban, O. A., Yildirim, R. and Alguttar, A. A. Audio Deepfake Approaches. *IEEE Access*. 2023, vol. 11, p. 132652–132682. DOI: 10.1109/ACCESS.2023.3333866. ISSN 2169-3536.

[56] Shiota, S., Villavicencio, F., Yamagishi, J., Ono, N., Echizen, I. et al. Voice Liveness Detection for Speaker Verification based on a Tandem. In: *Odyssey 2016: The Speaker and Language Recognition Workshop*. International Speech Communication Association, June 2016, p. 259–263. DOI: 10.21437/Odyssey.2016-37. Available at: https://www.isca-archive.org/odyssey_2016/shiota16_odyssey.pdf.

[57] Simmons, D. *BBC News*. BBC, May 2017. Available at: https://www.bbc.com/news/technology-39965545.

[58] Singh, D. Google, Facebook, Twitter put on notice about deepfakes in 2020 election. *CNET*. July 2019. Available at: https://www.cnet.com/tech/mobile/google-facebook-and-twitter-sent-letters-about-deepfakes-by-rep-schiff/.

[59] Staff, R. *The best voice assistant*. Sep 2021. Available at: https://www.zdnet.com/home-and-office/smart-home/the-best-voice-assistant/.

[60] Sugawara, T., Cyr, B., Rampazzi, S., Genkin, D. and Fu, K. Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems. *ArXiv.org*. Ithaca: Cornell University Library, arXiv.org. 2020. ISSN 2331-8422.

[61] Team, S. O. D. *2023 STATE OF DEEPFAKES*. 2023. Available at: https://www.homesecurityheroes.com/state-of-deepfakes/#key-findings.

[62] Ubert, J. *Fake It: Attacking Privacy Through Exploiting Digital Assistants Using Voice Deepfakes*. 2023. 166 p. Dissertation. Marymount University College of Business, Inovation, Leadership and Technology (BILT). ISBN 9798379504199. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-05-18. Available at: https://www.proquest.com/dissertations-theses/fake-attacking-privacy-through-exploiting-digital/docview/2811176534/se-2.

[63] Westerlund, M. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*. Ottawa: Talent First Network. 11/2019 2019,

vol. 9, p. 40–53. DOI: http://doi.org/10.22215/timreview/1282. ISSN 1927-0321. Available at: http://timreview.ca/article/1282.

[64] Winnard, N. *The Rise of Deepfakes in Schools*. 17. January 2024. Available at: https://www.tieonline.com/article/3632/the-rise-of-deepfakes-in-schools.

[65] Wu, Z., Gao, S., Cling, E. and Li, H. A study on replay attack and anti-spoofing for text-dependent speaker verification. *2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014*. february 2015. DOI: 10.1109/APSIPA.2014.7041636.

[66] Yan, C., Ji, X., Wang, K., Jiang, Q., Jin, Z. et al. A Survey on Voice Assistant Security: Attacks and Countermeasures. *ACM Comput. Surv.* New York, NY, USA: Association for Computing Machinery. nov 2022, vol. 55, no. 4. DOI: 10.1145/3527153. ISSN 0360-0300. Available at: https://doi.org/10.1145/3527153.

[67] Yuan, X., Chen, Y., Wang, A., Chen, K., Zhang, S. et al. All Your Alexa Are Belong to Us: A Remote Voice Control Attack against Echo. In: *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, p. 1–6. DOI: 10.1109/GLOCOM.2018.8647762. ISBN 1538647273.

[68] Yue, L., Cao, C., Li, Y., Li, J. and Liu, Q. LiveEar: An Efficient and Easy-to-use Liveness Detection System for Voice Assistants. *Journal of Physics: Conference Series*. 2021, vol. 1871, no. 1, p. 012046. DOI: 10.1088/1742-6596/1871/1/012046. Published under licence by IOP Publishing Ltd.

[69] Šandor, O. *Resilience of Biometric Authentication of Voice Assistants against Deepfakes*. Brno, Czech Republic, 2023. Bachelor's Thesis. Brno University of Technology. Available at: https://www.vut.cz/www_base/zav_prace_soubor_verejne.php?file_id=251953.

Appendix A

Contents of the enclosed memory Media

The content of the storage media is structured as follows:

- **thesis.zip**: contains source files for the text of the bachelor thesis.
- **informed_consent.pdf**: a blank informed consent form signed by each participant in the experiment.

Appendix B

List of recordings that were used in the experiment

List of predefined text that was read by the respondents and from which we created input recordings of deepfakes tools.

1. I like books.
2. Hey, Siri.
3. Hey, Google.
4. Hey, Bixby.
5. Hey, Alexa.
6. The sun is shining brightly.
7. She danced gracefully across the stage.
8. I love eating pizza with extra cheese.
9. The dog barked loudly at the mailman.
10. He ran as fast as he could to catch the bus.
11. The book I'm reading is really interesting.
12. We went for a walk in the park.
13. The movie was thrilling from start to finish.
14. The flowers in the garden are blooming beautifully.
15. She smiled warmly at her friend.
16. The car broke down in the middle of the road.
17. They celebrated their anniversary with a romantic dinner.
18. He plays the guitar skillfully.

19. I enjoy listening to classical music.

20. The children laughed happily at the clown's antics.

21. She painted a breathtaking landscape.

22. The raindrops fell gently on the roof.

23. The chef prepared a delicious three-course meal.

24. The students eagerly raised their hands to answer the question.

25. The mountain peak was covered in snow.

26. The baby slept peacefully in her crib.

27. He greeted his guests with a warm hug.

28. I'm craving a refreshing glass of lemonade.

29. The athlete sprinted across the finish line.

30. She carefully wrapped the gift with colorful paper.

31. The concert was sold out within minutes.

32. They explored the ancient ruins of a civilization.

33. The clock ticked loudly in the silent room.

34. The teacher explained the lesson clearly.

35. The ocean waves crashed against the shore.

36. The museum displayed stunning works of art.

37. He wrote a heartfelt letter to his loved one.

38. The thunder rumbled in the distance.

39. She wore a stunning gown to the gala.

40. The computer screen flickered for a moment.

41. They built a sandcastle on the beach.

42. The airplane soared through the clouds.

43. He cooked a delicious meal for his family.

44. She solved the puzzle effortlessly.

45. The fireworks lit up the night sky.

46. The hiker reached the summit of the mountain.

47. They laughed uncontrollably at the joke.

48. The baby took her first steps.

49. The artist created a masterpiece with his brush strokes.

50. He sang a beautiful melody.

51. The teacher praised the student for his hard work.

52. The wind rustled through the leaves of the trees.

53. She typed quickly on the keyboard.

54. The basketball player scored a three-point shot.

55. They embarked on a thrilling adventure.

56. The sun rises in the east and sets in the west.

57. She loves to read books and explore new places.

58. The cat meowed loudly outside the window.

59. The marathon runner trained hard for months to achieve victory.

60. The majestic mountains stood tall against the blue sky.

61. The smell of freshly baked bread filled the air.

62. He smiled warmly, showing his appreciation for the kind gesture.

63. The children played happily in the park, laughing and running around.

64. The scientist conducted experiments to test their hypothesis.

65. After a long day at work, he enjoyed a relaxing bath to unwind.