BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

ASSESSING THE HUMAN ABILITY TO RECOGNIZE SYNTHETIC SPEECH

HODNOCENÍ LIDSKÉ SCHOPNOSTI ROZPOZNÁVAT SYNTETICKOU ŘEČ

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **DANIEL PRUDKÝ**
AUTOR PRÁCE

SUPERVISOR **Ing. ANTON FIRC**
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

140541

Institut:	Department of Intelligent Systems (UITS)
Student:	**Prudký Daniel**
Programme:	Information Technology
Specialization:	Information Technology
Title:	**Assessing the Human Ability to Recognize Synthetic Speech**
Category:	Security
Academic year:	2022/23

Assignment:

1. Get familiar with synthetic speech (deepfakes) and their threats to humans.
2. Learn about the experiments conducted to test the human ability to recognize deepfakes in audio and video formats.
3. Based on the knowledge gained, design your experiment to assess the human ability to recognize deepfakes.
4. Execute the proposed experiment.
5. Based on the results, evaluate the human ability to recognize deepfakes, suggest ways to train and improve human recognition of deepfakes, and ways to raise awareness of deepfakes.

Literature:

- FIRC, Anton. *Applicability of Deepfakes in the Field of Cyber Security*. Brno, 2021. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.
- Matthew Groh, Ziv Epstein, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. 2021. Human detection of machine-manipulated media. Commun. ACM 64, 10 (October 2021), 40–47. DOI:https://doi.org/10.1145/3445972

Requirements for the semestral defence:
1 - 3

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

Supervisor:	**Firc Anton, Ing.**
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2022
Submission deadline:	10.5.2023
Approval date:	3.11.2022

Abstract

This work responds to the development of artificial intelligence and its potential misuse in the field of cybersecurity. It aims to test and evaluate the human ability to recognize a subset of synthetic speech, called voice deepfake. This paper describes an experiment in which we communicated with respondents using voice messages. We presented the respondents with a cover story about testing the user-friendliness of voice messages while secretly sending them a pre-prepared deepfake recording during the conversation and looked at things like their reactions, their knowledge of deepfakes, or how many respondents correctly identified which message was manipulated. The results of the work showed that none of the respondents reacted in any way to the fraudulent deepfake message and only one retrospectively admitted to noticing something specific. On the other hand, a voicemail message that contained a deepfake was correctly identified by 96.8% of respondents after the experiment. Thus, the results show that although the deepfake recording was clearly identifiable among others, no one reacted to it. And so the whole thesis says that the human ability to recognize voice deepfakes is not at a level we can trust. It is very difficult for people to distinguish between real and fake voices, especially if they are not expecting them.

Abstrakt

Tato práce reaguje na vývoj umělé inteligence a jejího potencionálního zneužití v oblasti kybernetické bezpečnosti. Jejím cílem je otestovat a ohodnotit lidskou schopnost rozpoznávat podmnožinu syntetické řeči, zvanou hlasový deepfake. Práce popisuje experiment, ve kterém jsme s respondenty komunikovali pomocí hlasových zpráv. Respondentům jsme odprezentovali krycí příběh o tom, že testujeme uživatelskou přívětivost hlasových zpráv a přitom jim tajně během konverzace poslali předpřipravenou deepfake nahrávku a sledovali například jejich reakce, znalosti o deepfakes nebo kolik z respondentů správně určí, která zpráva byla upravená. Výsledky práce ukázali, že žádný z respondentů nezareagoval na podvodnou deepfake zprávu a pouze jeden zpětně přiznal, že si všiml něčeho konkrétního. Na druhou stranu, hlasovou zprávu, která obsahovala deepfake, po experimentu správně označilo 96,8% respondentů. Z výsledků tedy vyplývá, že ačkoli byla deepfake nahrávka snadno identifikovatelná mezi ostatními, nikdo na ni nezareagoval. Práce ukazuje, že lidská schopnost rozpoznávat hlasové deepfakes není na takové úrovni, abychom jí mohli důvěřovat. Pro lidi je velmi obtížné rozlišit mezi skutečnými a falešnými nahrávkami, zvláště pokud je nečekají.

Keywords

deepfake, voice deepfake, synthetic speech, artificial intelligence, cybersecurity, deepfake detection

Klíčová slova

deepfake, hlasový deepfake, syntetická řeč, umělá inteligence, kybernetická bezpečnost, detekce deepfake

Reference

PRUDKÝ, Daniel. *Assessing the Human Ability to Recognize Synthetic Speech.* Brno, 2023. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Anton Firc

Rozšířený abstrakt

Umělá inteligence se vyvíjí obrovskou rychlostí a přináší nám obrovské množství možností a věcí, které nám mohou usnadnit život. Využívá se v mnoha oblastech včetně zdravotnictví, letectví a bezpečnosti. Je však také spojena s hrozbami všeho druhu a deepfakes jsou jednou z nich.

Deepfakes jsou média vytvořená umělou inteligencí, konkrétně pomocí hlubokých neuronových sítí prostřednictvím metod hlubokého učení. Při jejich tvorbě umělá inteligence spojuje, kombinuje, nahrazuje nebo překrývá prvky médií a vytváří tak nové falešné reprezentace věcí, které se nikdy nestaly.

V dnešní době tato falešná média dosahují takového standardu, že je nerozpoznají ani stroje, nemluvě o lidech, kteří si existenci takových hrozeb v dnešním digitálním světě nemusí vůbec uvědomovat. V rámci audia se navíc již nejedná pouze o anglické modely. Vzniká mnoho vícejazyčných nástrojů pro vytváření hlasových deepfakes, které se mohou objevit téměř v jakémkoli jazyce.

Existuje mnoho scénářů útoků, při nichž byly deepfakes použity. Mohou to být například útoky zaměřené na konkrétní osoby nebo instituce v podobě vishingu, tedy útoku jehož cílem je vylákat od oběti osobní nebo platební údaje prostřednictvím telefonního hovoru, nebo plošné dezinformace za účelem šíření propagandy a podobně. Lidé by se měli umět proti tomuto šíření podvodných informací a médií bránit, měli by vědět, jak takové věci ověřovat a jak se s nimi vypořádat. Nevíme však, zda toho někdy budeme schopni.

Proto je cílem této práce otestovat a následně vyhodnotit schopnost člověka rozpoznat syntetickou řeč. Pokusů o posouzení toho, zda lidé dokáží rozeznat deepfake od skutečné řeči, bylo již několik. Tyto pokusy však účastníky nejprve seznámily s problematikou deepfake a teprve poté jim předložily médium a požádaly je o jeho identifikaci. Jejich výsledky jsou poměrně variabilní a liší se především v závislosti na metodice. Hlasové deepfakes testovali například Müller a spol., kteří uvádějí, že přesnost identifikace deepfake a pravé nahrávky je 80%.

My se inspirovali experimentem *Authorizing Card Payments with PINs*, ve kterém autoři replikovali skutečný útok skrytý za krycím příběhem, a tuto myšlenku jsme přenesli do oblasti hlasových deepfake.

Vytvořili experiment, ve kterém byli respondenti konfrontováni s hlasovými deepfakes, aniž by o nich byli informováni, a sledovali, zda poznají že jde o deepfake, nebo si alespoň všimnou něčeho zvláštního.

Celý experiment jsme ukryli za krycí příběh a pod touto oponou jsme účastníkům pokládali jednoduché otázky, v podobě faktů do hry *Dvě pravdy jedna lež*, pomocí hlasových zpráv v klasickém chatu. Stejně tak nám hlasovými zprávami odpovídali respondenti, toto mělo podpořit odprezentovaný krycí příběh. Mezi skutečnými hlasovými zprávami jsme také poslali jednu předem připravenou deepfake nahrávku a pak jsme sledovali, zda účastníci deepfake poznají nebo si alespoň všimnou něčeho zvláštního. Na konci experimentu jsme jim poslali dotazník, v němž jsme se ptali na jejich znalosti a postoj k deepfake, odhalili jsme jim krycí příběh a zeptali se, zda dokážou identifikovat deepfake zprávu v chatu.

Výsledky experimentu jsou zcela jednoznačné. Během konverzace žádný z 31 respondentů nijak nereagoval na podvodnou deepfake zprávu. Při vyhodnocování dotazníku na otázku, zda si něčeho všimli, jsme se shodli, že pouze jedna osoba si všimla a popsala něco specifického pro deepfakes. Na druhou stranu při identifikaci deepfake zprávy ji správně označilo 96,8% respondentů. A 83.9% respondentů správně označilo pouze deepfake zprávu, nikoliv ostatní pravé zprávy. Z výsledků tedy vyplývá, že ačkoli byla deepfake nahrávka jasně identifikovatelná mezi ostatními, nikdo na ni nijak nezareagoval. Samozřejmě je na

místě diskuse o tom, jak by to vypadalo, kdyby obsahem rozhovoru nebyly běžné otázky, ale například téma financí a podobně citlivá témata, kde jsou lidé mnohem opatrnější. V tomto jsme však byli vázáni etickým přístupem k práci.

Experiment dále zjistil, že 83,9% respondentů o deepfakes alespoň slyšelo, a to především ze sociálních sítí, vzdělávacích videí nebo na ně jednoduše narazilo na internetu. V dotazníku jsme se respondentů také ptali, jak moc si věří, že by deepfake odhalili. Tuto otázku jsme jim položili dvakrát, na začátku a na konci dotazníku, a nechali jsme je ohodnotit jejich pohled na škále od 1 do 5 bodů. Průměr byl na začátku 2,29 a po odhalení experimentu, na konci dotazníku se průměr zvýšil na 2,94. Přičemž hodnotu zvyšovali především mladší respondenti.

V závěru práce popisujeme vlastní návrhy na trénink rozpoznávání deepfakes a šíření povědomí o nich. Tyto návrhy vycházejí z výsledků naší a souvisejících prací nebo například z doporučení FBI. Navrhujeme zde edukační platformu, která by obsahovala interaktivní ukázky různých přístupů k výcviku, návrhy nástrojů pro rozpoznávání deepfakes a také odkazy na podporu obětí této technologie.

Hlavní přínosy této práce lze shrnout následovně:

- Práce ukazuje, že lidská schopnost rozpoznávat hlasové deepfakes není na takové úrovni, abychom jí mohli důvěřovat, a pro lidi je velmi obtížné rozlišit mezi skutečnými a falešnými nahrávkami, zvláště pokud je nečekají.
- Navrhuje platformu pro práci s deepfakes a tréninkem jeho rozpoznávání, společně s doporučením na užitečné nástroje nebo odbornou pomoc.
- Odhaluje, že povědomí lidí o deepfakes je poměrně vysoké, zejména z informativních videí, článků a podobně.
- Ukazuje, že produkce a kvalita hlasových deepfakes v neanglických jazycích nepředstavuje pro umělou inteligenci problém. A zároveň k jejich vytvoření nikdo nepotřebuje profesionální znalosti.

Assessing the Human Ability to Recognize Synthetic Speech

Declaration

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

.....
Daniel Prudký
May 7, 2023

Acknowledgements

I would like to thank my supervisor Ing. Anton Firc for his insightful suggestions for the work, consultation and other help in the development of the experiment and thesis.

I also would like to thank RNDr. Agata Kružíková for her advice in designing the experiment and all the respondents who volunteered to participate in it.

Last thanks to my closest ones, my girlfriend and my parents for the support they gave me for this work.

Contents

Chapter 1

Introduction

Artificial intelligence is evolving at a tremendous speed, bringing us a huge number of possibilities and things that can make our lives easier. It is used in many fields including healthcare, aviation and security. But it is also associated with threats of all types, and deepfakes are one of them.

Deepfakes are media created by artificial intelligence, specifically by using deep neural networks through deep learning methods. In their production, artificial intelligence merges, combines, replaces, or overlays elements of media to create new false representations of things that never happened.

Nowadays, these fake media are reaching a stage where they are not even recognizable by machines, let alone humans who may not even be aware of the existence of such threats in today's digital world. Moreover, within audio, it is no longer just about English models. A lot of multi-language tools for creating voice deepfakes are being developed, and they can appear in almost any language.

There are many attack scenarios in which deepfakes have been used. For example, they could be attacks targeting specific individuals or institutions in the form of vishing or widespread disinformation to spread propaganda, and so on. People should be able to defend themselves against this spread of fraudulent information and media, they should know how to verify such things and how to deal with them. But we don't know if we will ever be able to do that.

Therefore, the goal of this work is to test and then evaluate the human ability to recognize synthetic speech. There have been several attempts to assess whether people can distinguish a deepfake from a real one. However, these experiments first introduced participants to the deepfake problem before presenting them with a medium and asking them to identify it. Their results are quite variable and vary mainly depending on the methodology. Voice deepfakes have been tested, for example, in a survey by Müller et al. [36] who report that the accuracy of identifying a deepfake and a genuine recording is 80%.

We were inspired by the *Authorizing Card Payments with PINs* [33] experiment, in which the authors replicated an actual attack hidden behind the cover story and we transferred this idea to the voice deepfake field.

Created an experiment in which respondents were confronted with voice deepfakes without being told about them and observe whether they recognize it as a deepfake or at least notice something strange.

We hid the whole experiment behind a cover story and under this curtain, we asked the participants simple questions, in the form of facts for the game *Two Truths One Lie*, using

voice messages in a classic chat. Similarly, respondents answered us with voice messages, this was to support the cover story presented. In between the real voice messages, we also sent one pre-prepared deepfake recording of my voice and then watched to see if the participants recognized the deepfake or at least noticed anything odd. At the end of the experiment, we sent them a questionnaire asking about their knowledge of and attitude towards deepfake, revealed the cover story, and asked if they could identify the deepfake message in the chat.

At the end of the thesis, we describe our own suggestions for training people in recognizing deepfakes and spreading awareness about them. These suggestions are based on the results of our and related work, or for example, the recommendations of the FBI. We propose here an educative platform that would include interactive training demos of different training approaches, suggestions for tools to detect deepfakes, as well as links to support victims of this technology.

The main contributions of this work can be described as follows:

- The work shows that the human ability to recognize voice deepfakes is not at such a level that we can trust it, and it is very difficult for people to distinguish between real and fake recordings.
- It suggests a platform for working with deepfakes and training its recognition, along with recommendations for useful tools or expert help.
- Displays that people's awareness of deepfakes is quite high, especially from informative videos, articles and similar.
- Shows that the production and quality of voice deepfakes in non-English languages is not a problem for artificial intelligence. And at the same time, no one needs professional expertise to create one.

In Chapter 2, the thesis describes deepfake as it is, its definition, methods of use and creation. Works that discuss a similar problem and evaluate the human ability to detect deepfake in different types of media are described in Chapter 3. The design of the experiment is described in Chapter 4. The course of the experiment, its execution and its results are discussed in Chapter 5. Chapter 6 describes suggestions for training people in detection and ways to spread awareness of deepfakes, based on the results of our experiment and the findings of other papers. Finally, Chapter 7 evaluates the results of the thesis and suggests further work.

Chapter 2

Deepfake

This chapter describes deepfakes technology, including its definition, history, beneficial and dangerous applications, people's approach, and the technical background of voice deepfake creation.

2.1 What's deepfake

Mirsky and Lee [35] define a deepfake simply as a **"Believable media generated by a deep neural network."** A more expansive definition is that it is media created by artificial intelligence (AI), specifically using deep neural networks through deep learning (DL) methods. In their production, artificial intelligence merges, combines, replaces, or overlays features of the media to create new fake representations of things that never happened. This media can be practically unnoticeable from authentic ones, as shown in Figure 2.1. Deepfake technology brings many benefits, it can be used for entertainment, but it can also be used for revenge porn, bullying, spread fake news, political sabotage and more [12, 47].

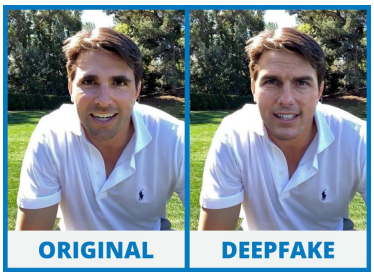

Figure 2.1: Example of video deepfake quality (https://www.trymaverick.com/blog-posts/are-deep-fakes-all-evil-when-can-they-be-used-for-good).

Deepfakes are just a subset of "Synthetic Media". Which is media of any type (image, video, audio and text) that has been created or modified by artificial intelligence (AI)/machine learning (ML), especially if the process is automated [3].

They are based on these advanced technologies, unlike other subsets of synthetic media such as "Cheap fakes", which are sometimes confused with deepfakes. Cheap fakes are much easier to produce and are based on techniques such as photoshopping, lookalikes, speeding and slowing video, or face swapping, for example, using the rotoscope method, a non-automated method performed by a human [39].

As mentioned deepfakes are created using deep neural networks through deep learning (DL) methods. This is also where the name deepfakes comes from, as a combination of the words **"deep"** from "deep learning" and **"fake"** [12]. Deep learning methods are representation learning methods with multiple levels of representation, i.e., sets of methods that allow to pass raw data to the machine and automatically discover representations for further processing. Between levels of representation, the data is transformed to higher, less abstract levels, and by composing enough of these transformations, the network can learn very complex functions. The main aspect of deep learning, therefore, is the feature layers are designed using a general learning procedure and are not designed by human engineers. This is a fundamental difference from conventional machine learning, which is limited in its ability to process natural data in its raw form and needs human design of the feature extractor that transforms the raw data and hence considerable author expertise [27].

The first mentions of deepfakes appeared on Reddit at the end of 2017 [35]. The user „deepfakes" was supposed to post pornographic videos where, using deep learning, he swapped the faces of porn actresses for those of celebrities. Since then, the number of identified deepfakes on the internet has been growing rapidly. There is information from the source referenced in the article [40], which we have not been able to trace back and confirm the authenticity of this information. This data claims that in June 2020, 49081 deepfakes were identified on the Internet and that this number is doubling every 6 months. This can be seen in Figure 2.2. Unfortunately, there is no more recent information available, if this trend continues today, there would be more than 3 million deepfake media on the internet by the middle of 2023.

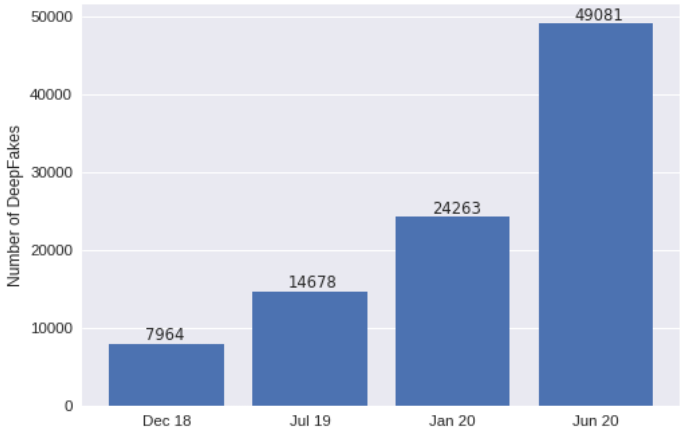

Figure 2.2: Number of deepfakes online relative to date [40].

2.2 Use of deepfakes

Deepfake media is created for different purposes. This technology can be useful in many ways, but as is often the case with artificial intelligence, there is also plenty of potential ways for abuse.

2.2.1 Beneficial ways of use

One positive case can be made that a revolutionary approach to protecting privacy and user data by creating your own fake identity to go online can be a good way to preserve the real one and completely separate those two realities [18]. An interesting idea is the use of deepfakes in dubbing[1][2], nowadays dubbing is a lengthy process and not always perfectly lip-synced to the actor. Deepfake technology could bring instant translation to all languages, along with the actor's lip-synced voice to their original voice. Deepfake brings us a great opportunity in the field of art, for example by animating dead characters or creating satire [47]. Like the nowadays popular pictures of the Pope in a puffer jacket, shown in Figure 2.3, which were created by the AI app Midjourney[3] and published on Reddit. This image was of very high quality and fooled millions of people [11].

Figure 2.3: Deepfake of the Pope in a puffer jacket [11].

2.2.2 Dangerous ways of use

On the other hand, this leads to the manipulation of evidence and audio and video recordings may not be valid evidence in court [47]. They can also be used, for example, for disinformation, fraud with destabilising political impact, terrorist propaganda or as a basis for all kinds of scams.

For now, pornography is clearly still the biggest field of action. The connection between pornography and deepfakes is further discussed in the article [5], which states that 96% of all deepfakes online are pornographic content. Of those deepfakes on pornographic sites, 100% depict women and 99% of them are actresses, musicians or working in the entertainment sector. The article also says that 94% of deepfake pornography videos are found on sites dedicated to deepfake porn and 6% of these videos are found on 8 out of 10 mainstream pornography websites. The article *Increasing Threats of Deepfake Identities* [3] cites

[1] https://www.respeecher.com/blog/synthetic-film-dubbing-ai-deepfake-technology-explained
[2] https://variety.com/2019/biz/news/ai-dubbing-david-beckham-multilingual-1203309213/
[3] www.midjourney.com

Karen Hao [2], senior editor of MIT Technology Review, who reportedly said that *"The biggest threat is to women and vulnerable populations. By far 95% of deepfakes are of non-consensual porn of women"*. It is the non-consensual pornography that is a huge problem on top of which more and more possible ways of cyberbullying, humiliation, degradation and so on are being added. Nowadays, it's no longer a problem to create a pornographic deepfake video, there are even articles that compare apps for creating them, under titles like „7 Best Deepnude Apps 2023 (50 Nude Generators Ranked)". These sites will rank the best online apps for creating non-consensual pornography, tell you how much they cost, what are their advantages or disadvantages, even for example a recommendation for a more affordable Yearly plan. Most of the sites don't mention pornography at all, posing as ordinary face swap apps, some claim to create pornographic images, but none write about what these images or videos can do to the victims. There are many such victims, for example, the famous case of Noella Martin [43]. Noella Martin claims that when she was 18 years old she used reverse image searching on a photo of herself posted on Facebook. She discovered dozens of fake pornographic images with comments such as "She sent me this", the name of her childhood best friend or her home address. This happened in 2012, at that time these were photoshopped images, but her face appeared in various pornographic images until the advent of deepfakes and even then it did not stop, her face was also used to create pornographic deepfake videos. Noella claims that she approached the site administrators asking them to delete it, sometimes the material was deleted and soon reappeared, once even blackmailing her with the idea that they would only delete the content if she sent them nude photos within 24 hours. Noella is now working on a campaign for laws against "revenge porn" and has spoken about her experience for example at the TED conference in Perth[4]. Revenge porn is just another term in this growing theme. It is the online posting of explicit photographs or videos of an individual without their permission for the purpose of degradation, according to Kamal and Newman [23]. Most of the time this is media provided by the victim during an intimate relationship, but with today's advances in deepfake technology and how easy it is to create, as described above, there is an opportunity to degrade someone and get revenge for all sorts of things with just one simple photo of their face. Victims then have to deal with the long-term personal and psychological consequences that can haunt them for life. The article by Kamal and Newman [23] refers to a study that claims 49% of victims have experienced cyber-harassment or cyberstalking. And 80 to 93 percent have suffered emotional distress, which can include anger, guilt, paranoia, feelings of isolation, deterioration in personal relationships, depression and even suicide. Deepfakes may be responsible for all of this. In the Czech Republic, counsellors and experts from www.napisnam.cz can help these victims.

Another scenario is an attack on a financial institution [3]. In this scenario, the attacker found the private information of his victim, who was active on social media, on the dark web. Through their posted videos, he was able to train a model and create audio deepfakes. He then called the financial institution where using deepfakes, he passed through the voice authentication system and, after telling them the information he found on the dark web, he requested a password change. He thus gained access to the victim's accounts.

Very widespread method is vishing, which is derived from the two words that define it, namely "voice" and "phishing". It is a version of phishing in which identity theft is carried out using voice devices such as the telephone, voice assistant, etc. Its use is described by Firc, Malinka and Hanáček [12], the authors point out that one such attack happened

[4] https://www.youtube.com/watch?v=oXeAWdHP0uY

in 2019 when a fraudster using deepfakes created a transaction of almost $250,000. The CEO of an energy company thought he was talking to his boss on the phone and when the caller asked him for an urgent transfer of this money the victim did not hesitate and sent the money believing he was completing a task from his boss. There are many cases like this today. The same article says that vishing was reported by 69% of companies in 2021, a big increase from 2020 when 54% of companies reported it. Spoofing is also very often associated with this scam, giving the scam much more credibility. For example, the fraudster can call the victim from the real phone number of the person they are playing. Spoofing can also be used with the phone number of a bank, the police, etc. In this way, the attackers try to exert authority on the victim, who is then more likely to disclose the required information in fear.

Politics has not escaped deepfakes either. Especially in the form of misinformation, when, primarily in the pre-election period, it is basically expected that a video will be circulated that tries to badmouth a politician and somehow influence the elections. Such a disinformation video may, for example, include a politician taking a bribe, using a racial epithet, admitting complicity in the crime and so on. This distribution is called a disinformation campaign and is used to manipulate public opinion in any area of interest. They are often used to spread propaganda [47]. For example, in the US in May 2019, a video of Speaker of the House Nancy Pelosi, shown in Figure 2.4, appeared in which she speaks as if she is unintelligible as if she is drunk [34]. Although this video is presented by the media as a deepfake, the method by which it is produced puts it in the category of cheap fakes [10]. In the end it is irrelevant whether it is a deepfake or a cheap fake, the impact on public opinion is very much the same, regardless of quality. This fraudulent video was even shared by Donald Trump himself, still president at the time, on his Twitter account. Political videos, while manipulative and usually quite credible, tend to be quickly debunked. Given the reach they have, it's in a lot of people's interest to debunk these attacks, and they are succeeding. Another article [8] on the same topic, quotes the University of California Berkeley professor Hany Farid as saying "What if somebody creates a video of President Trump saying, 'I've launched nuclear weapons against Iran, or North Korea, or Russia?' We don't have hours or days to figure out if it's real or not." This could be one of the blackest uses of deepfake technology, where, as Hany Farid says [8], we won't have time to figure out if it's fake or not. Fortunately, nothing like this has happened to this day.

Figure 2.4: Cheap fake of Nancy Pelosi [34].

This technology is and probably will be one of the main information weapons of war. Such videos are appearing, for example, in the current war in Ukraine [44]. Three weeks after the war began, in March 2022, a deepfake video of the Ukrainian President calling for

the laying down of arms and surrender appeared on Ukrainian national news. This video was also debunked very soon.

The topic of deepfakes in pornography is the most widespread on the Internet and information about it is easier to obtain, and although it is given the most space in this section it is not exactly the biggest problem and all the scenarios mentioned here deserve equal attention.

2.3 People's approach to deepfakes

In August 2022, iProov released a study [19], asking people what they thought about deepfakes. The study surveyed 16,000 people from eight countries. They asked questions like „Which of the following worries you most about how deepfakes could be used against you?“ (see Figure 2.5), where the most common answer was identity theft, and to their bank and other accounts. Another question asked, „Which of the following statements do you agree with most about deepfakes?“ (see Figure 2.6), where people most often voted that it makes us lose trust in things on the internet.

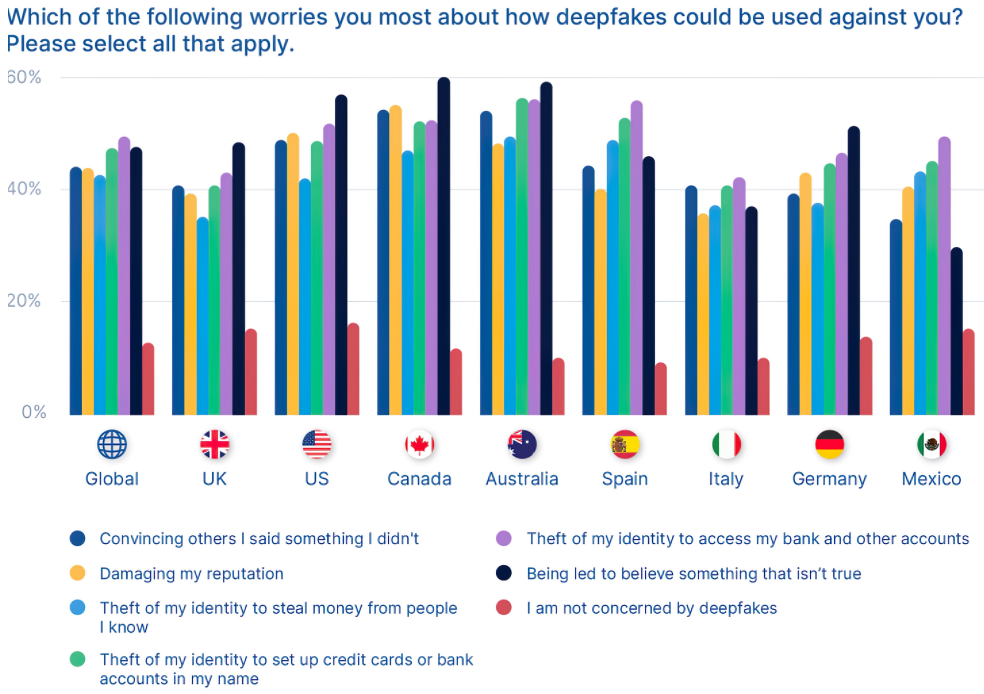

Figure 2.5: Result of a survey about how people see deepfakes [19].

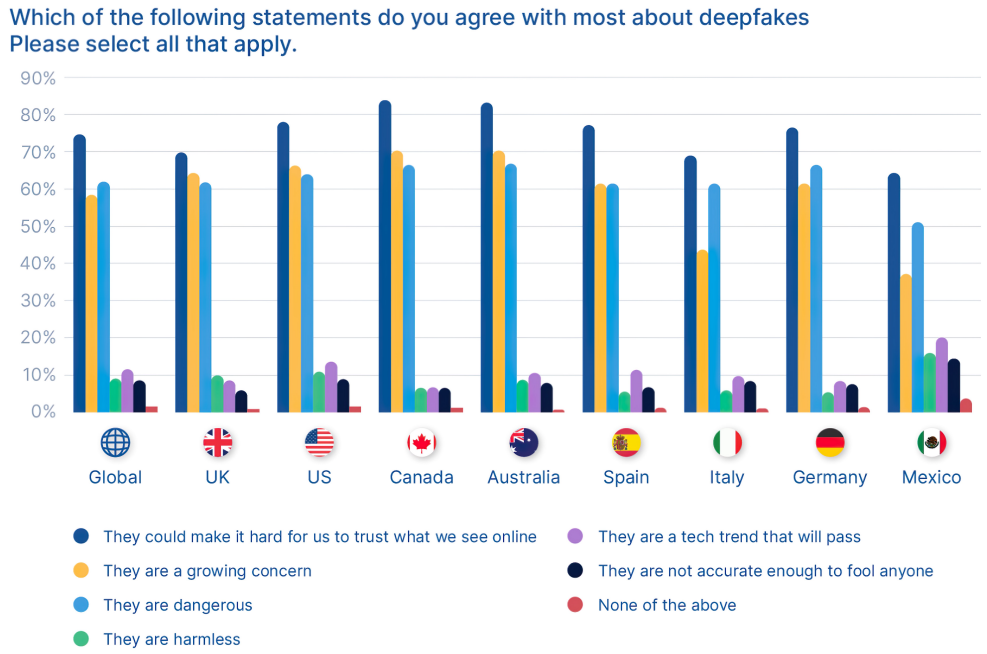

Figure 2.6: Result of a survey about how people see deepfakes [19].

But that people's interest in deepfakes is growing is evident, for example, in the analytics of Google searches[5]. Here in Figure 2.7 that plots interest in searches for the specific word „deepfake“ on a scale of 0-100 relative to searches in a given period. We can see that the interest is increasing, with the highest values in the last months.

Figure 2.7: The relative interest in the search term „deepfake“ on a scale of 0-100, based on the specified period from January 2017 to March 2023.

[5] https://trends.google.com

2.4 Creation of voice deepfakes

Two main methods dominate the creation of voice deepfakes today, namely **Text-To-Speech (TTS)** and **Voice Conversion (VC)**. TTS uses text as input data, while VC uses direct recording as input data, which is further processed. Both methods are described below in Sections 2.4.3 and 2.4.4.

Their common goal is to create a waveform with all the characteristics of the native speaker's speech in which the sound is produced. The output should sound intelligible and natural, i.e., all the words of the input text should be spoken clearly and sound intelligible and natural to the native listener. They should therefore correctly use all the phonetic features of the languages in question. Another important aspect is the sound quality itself, which should not contain noise or any speech artefacts. Ideally, the method should be applicable in all world languages. [17, 46]

2.4.1 Few-shot training strategy

Deep learning approaches previously used mainly supervised learning and required huge datasets of training data. This is unrealistic in some domains such as healthcare or robot motion planning, either due to the unavailability of data or the very high cost of obtaining it. As a solution to this problem, semi-supervised and unsupervised learning for deep architectures started to be used. This is called K-shot learning and involves deep learning approaches that can learn efficiently from only a small handful of examples by experimenting with the network architecture, improvising learning algorithms and exploiting the nature of data. Extreme cases are then **One-shot** or **Zero-shot**, which use only one or no training example. This is where the term **Few-shot** came from. These approaches bridge the gap in the knowledge of the learner by borrowing knowledge (learning from training data of related tasks and reusing it), creating knowledge by generalizing (generalizing approaches those abstract and learning representations), recalling (already) learned knowledge with memory and attention (cognitive abilities added to models allow them to store knowledge and retrieve it again from memory when doing another task of the same type) and acquiring/generating data that helps in creating new knowledge (creating synthetic data by extending data or learning rich representations). [22]

Approaches that allow training without large amounts of data are often used in the creation of synthetic speech, where often only a few seconds of target speech is needed and the model is able to produce a very convincing result.

For synthetic speech generation, along with few-shot approaches, we also use **fine-tuning**, which allows us to build interfaces over such pre-trained models for further usage. Fine-tuning takes advantage of labelled application data to train additional application-specific parameters, whereby the pre-trained data is frozen or only minimally modified [21].

2.4.2 Generative Adversarial Networks

Generative Adversarial Networks (GAN) are a framework designed by Goodfellow et al. [14] to remove the difficulty of approximating intractable probabilistic computations that arise in, for example, maximum likelihood estimation and related strategies. The authors proposed a framework that contrasts generative models with discriminative models. The two models face each other like two players of a minimax game. The generative model G captures the distribution of the data and the discriminative model D estimates the probability of the sample coming from the trained data and not from model G. Both models are trained

together and provide each other with data through which they learn. Training model G consists of maximizing the probability of error of model D. In the paper, the authors mention an example of police officers (model D) and counterfeiters (model G), where both learn to produce and detect counterfeits until they are indistinguishable from real money. The workflow of both models is shown in Figure 2.8. GAN is used by many models to create synthetic speech as a vocoder, i.e. to generate the resulting waveform.

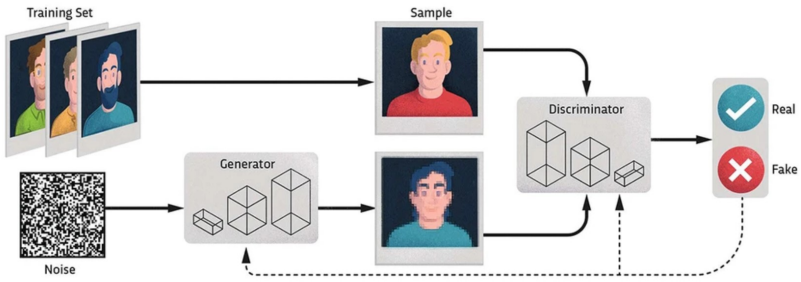

Figure 2.8: Work of Generative Adversarial Networks [37].

2.4.3 Text to speech

Text-to-speech (TTS) synthesis is able to generate a waveform based on the input text. Creating these models is an individual process, so we chose the particular model described by Jia et al. [20], which is a good example of the principles being used. In this paper, the authors created a multispeaker text-to-speech synthesis, which is not common for TTS systems, TTS models are usually trained for one specific output speaker. The synthesis uses a Zero-shot approach supported by fine-tuning. Their model consists of three independently trained neural networks namely a **recurrent speaker encoder**, a sequence-to-sequence **synthesizer** and a **vocoder**, shown in Figure 2.9.

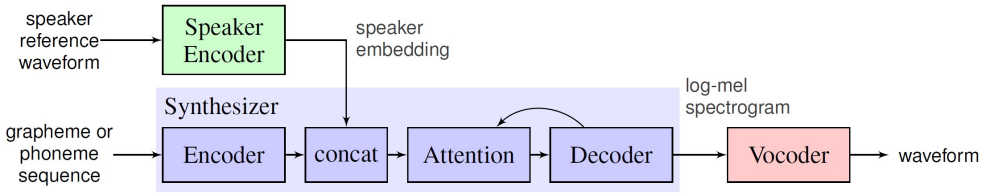

Figure 2.9: Text-to-speech synthesis diagram [20].

Speaker encoder is used to characterize the target speaker based on his/her reference speech signal. Good generalization then results from using a representation that can identify the characteristics of different speakers independently of its phonetic content and background noise and using only a short adaptation signal. The network is trained such that the embeddings of utterances, numeric vectors containing representations of objects in high-dimensional space, of the same speaker have a cosine similarity high and the embeddings of different speakers are distant in the embedding space.

The input to the **synthesizer** is the text analysis and mapping of text to phoneme sequences, which is done by the grapheme-to-phoneme conversion described by Gundle and

Chavan [17]. In this conversion, the grapheme form of the input text is converted to a phoneme form for speech synthesis. This synthesis is then the exact pronunciation of each word in the input sentence. The synthesizer is then trained on pairs of this converted text and the audio of the target speaker obtained from the speaker encoder. The output of the synthesizer can be displayed on spectrograms as shown in the paper.

As a **vocoder**, the authors used the WaveNet network [38]. This network can invert the synthesized spectrograms into time-domain waveforms. Since the synthetic network already predicts all the necessary information for high-quality output sound, it is now very easy to build a vocoder for different voices with only simple training.

2.4.4 Voice conversion

Voice conversion attempts to combine the non-linguistic information of the target speaker's waveforms with the content information of the source speaker. Non-linguistic information is information such as speaker identity, accent or pronunciation. As with text-to-speech, the creation of each model is individual, so we again chose a representative model, this time described by Chou, Yeh and Lee [9].

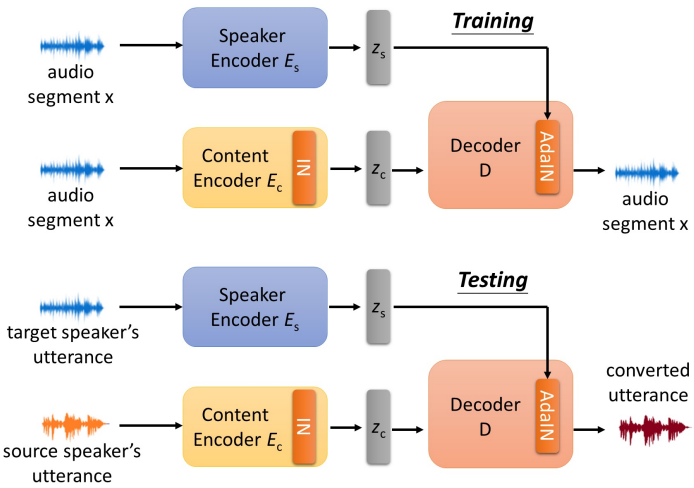

Figure 2.10: Voice conversion synthesis diagram [9].

The authors focus on the fact that speech signals carry both static and linguistic information. Static information is invariant throughout the speech, it is information such as speaker, acoustic state, etc. On the other hand, linguistic information changes every few frames. This allows them to split speech into speaker representation and content representation, this is achieved by **speaker encoder** (a model trained to encode static speaker information into speaker representation), **content encoder** (a model trained to encode linguistic speaker information into content representation) and **decoder** (performs the synthesis of both representations together). The involvement of the components is shown in Figure 2.10. This architecture makes the model capable of factorization, allowing it to perform One-shot voice conversion. For conversion, the model needs only one utterance from the source speaker and one from the target speaker, then the model extracts the representations of both utterances and finally combines them using a decoder. The model handles the conversion without any speaker labelling.

However, there are also models that can handle both approaches, such as **YourTTS** [7] with an open-source project and demos on GitHub[6]. Multilingual, multi-speaker and Zero-shot TTS model, which due to its architecture can also be used for voice conversion. This is made possible by the Zero-shot approach and the fact that the encoder has no information about the identity of the speaker. Thus, the distribution predicted by the encoder is forced to be speaker-independent and YourTTS is able to convert voices using the model's **Posterior Encoder** [24], a neural network used to learn a compressive representation of the dataset for downstream tasks, **the decoder** and **the HiFi-GAN Generator** [25], a generator based on the aforementioned GAN 2.4.2. Conditioning the embedding of external speakers allowed the authors to mimic the voices of unknown speakers in a Zero-shot voice conversion setting. The model requires less than a minute of speech to fine-tune for a speaker with very different characteristics.

2.4.5 Online tools

There are plenty of online tools that allow you to create voice deepfakes for free. The vast majority of them are TTS systems that convert text into pre-trained models, such as *ElevenLabs*[7] or *FakeYou*[8]. Web application *Speechify*[9] allows you to create multi-character conversations using the same creation principle and adjust parameters such as pitch, volume and more for individual speeches. *Resemble.ai*[10] can do a similar thing, where you can also create/train your own voices if you buy the Premium Package. With VC systems it's similar there are a lot of them, but most of them only let you use your voice as a reference and not as a target, they have pre-trained models for target voices. You can supposedly create your own voice to use as a target voice in *Voice.ai*[11]. It is an application that is required to install. In the app, you can create your model using any of the sound files. The files are approved first and it takes time to get them approved. Once approved, a model is created, however, the quality was not very high on first listen, compared to other tools. The application is more useful for voice-overs created by voice instead of text and the live voice changer, which can be used for amusement due to the usable pre-trained models in good quality. It is also important to say that the app has a problem with processing reference recordings and most of the conversion attempts have failed due to input data processing failures.

It should be added that even though the web applications that can be used online for free are not the best in terms of quality and flexibility in creating deepfakes. There are plenty of open-source projects for creating high-quality synthetic records, which we have also used in this experiment. For a person experienced in IT, it should not be difficult to create a sufficiently high-quality deepfake. Similarly, paid programs bring higher quality and flexibility of creation.

[6] https://github.com/Edresson/YourTTS
[7] https://beta.elevenlabs.io
[8] https://fakeyou.com
[9] https://voiceover.speechify.com
[10] https://app.resemble.ai
[11] https://voice.ai

Chapter 3

Related work

There is already related work assessing the human ability to detect deepfakes in various ways. Based on relevance and differences in approach, We have selected a few of them and discussed their methodologies, results and benefits. At the end of the chapter, we provide a summary of all the works and compare them with our experiment.

3.1 Voice

The research that deals with the detection of voice deepfakes by humans is described in the scientific article by Müller, Pizzi and Williams [36]. In this research, the authors focused on the ratio of the success rate of deepfakes detection by humans and artificial intelligence.

The experiment compared human and machine detection capabilities, using a game-based challenge in which the respondent always played a recording and then determined whether it was fake or real. The web interface is shown in Figure 3.1. They made the same decision with machine learning models trained only on the training data of the dataset they used. For the experiment, the authors used the ASVspoof 2019 dataset, a dataset created for the ASVspoof 2019 Challenge, which aims to test Automatic Speaker Verification (ASV) systems resistant to spoofing attacks. In addition, the respondent was asked about their age, their level of IT experience, and whether they were a native English speaker.

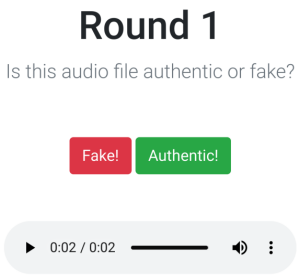

Figure 3.1: The web interface as presented to the survey participants [36].

Through the experiment, the authors found that the human ability to recognize deepfake and real recordings reaches 80%. The results were compared to the two trained models, with the first model, based on naive AI [6], solving tasks using a set of trivial features, such as the presence or absence of silence in particular audio. Its success rate for detecting a fake was 95%. The second model, based on realistic AI, achieved similar results to humans.

Further, the experiment found that recordings created using TTS fooled humans much more than voice-conversion or waveform concatenation systems. Especially one TTS attack made trouble for AI detectors, but not for humans. The authors believe it could be because it was using GAN as the waveform generator.

Other interesting results are that, native speakers handled recognition better than non-native speakers, while the level of IT experience did not affect performance, and that people's ability to recognize deepfakes decreases with increasing age. It is also interesting to note that people learned very quickly and as the article says after the first ten rounds the success rate improved from 67% to 80%, but promptly stabilised at those levels and did not improve.

The research website is still active after its evaluation, and anyone can test their skills for themselves at https://deepfake-demo.aisec.fraunhofer.de.

In 2019, ID R&D commissioned a study [32] to find out the attitudes of people in the US towards voice deepfakes. The research found that 36% of respondents were confident they could detect a computer-generated voice and only 30% of respondents were not sure they could tell the difference. The research also says that 66% of US adults are worried about their identity being stolen by using fraudulent copies of their voice to gain access to their accounts.

3.2 Image

One of the studies that looked at the ability of humans to detect a deepfake image was conducted by Groh et al. [16]. In this experiment, the authors use a proprietary technology called Deep Angel, which is a technology that can remove objects from an image and replace their pixels with new ones that create a background that could be the same without the object.

The experiment results say that the average correct detection rate of manipulated media is 86%. And some of the manipulated images were detected in more than 90% of the cases. However, the paper has a link[1] to the source images used in the experiment, and the quality of many of the manipulated images is not good at all. Although there are some nice deepfakes at first look, in many of them the object is just replaced by a grey square or the object was blurred and there was a clearly detectable smudge in the image. We had generate ten random numbers of images and for these images tried to determine the exact location of the removed object, which we got right for nine of them. So the results of the work could be fundamentally affected by poor good quality media.

Interestingly, however, the experiment yielded similar statistics in human learning as the voice deepfakes experiment [36]. The authors found that the average success rate of correctly identifying the first image is 78%, and it gradually increases with the number of identified images, reaching 88% after the first ten, and then begins to stabilize.

Another similar experiment is described in the scientific paper by Rössler et al. [42]. Here, the authors used a custom dataset consisting of images cut from videos such as newscasts, etc. The dataset thus contains real photos of faces and their deepfakes. The authors showed these images to the respondents for a limited time, randomly 2, 4 or 6 seconds, and then asked them whether the image was fake or genuine.

The results found that deepfake detection accuracy is dependent on the overall quality of deepfake media. Lower quality leads to a decreasing accuracy rate, which averages

[1] https://github.com/mattgroh/human-detection-machine-manipulated-media-data-code

68.69% for raw videos, 66.57% for high-quality videos, and 58.73% for low-quality videos. The responses in the study are also classified using the deepfake generation method and the graphs presented in the study clearly show that some methods are more effective than others. For example, the fake detection rate of the *NeuralTextures* method does not exceed 40% in either quality, while the *DeepFakes* method has the highest detection rate in all qualities.

The next very interesting study that deals with image deepfakes is an article by Godage et al. [13]. This study looks at the success rate of people detecting a special kind of deepfake called "Face morphing".

For clarity, Face morphing is a method that combines two or more images of different human faces to create one new face with features from both of them [13]. Such images pose the biggest threat to airport border controls. Some countries do not take professional photos of people when creating a passport but only want to bring in a printed photo and they will then re-digitize it. This opens the door to morphed facial images. If a passport is created with such a photo it can be used by both people whose faces were used to create it [12, 13]. This can be a huge problem when someone wanted by the police is travelling with an innocent person's passport and a photo that looks undeniably like him, as shown in Figure 3.2.

For detecting morphed images there are two approaches both fall under morphing attack detection (MAD). One is Single-Image MAD (S-MAD) which works only with the image without any reference to compare against. The other is Differential MAD (D-MAD) which has an image reference, for example from an Automated Border Control (ABC) gate, which takes a photo and then compares it to a passport photo. Border controls are carried out in two ways, one by the ABC gates and the other by the people, the passport controllers. The authors of the study put people in the position of deciding whether it is fraud or not. They conducted the study on both D-MAD and S-MAD and for the experiment, they prepared a HOMID dataset with a subset of HOMID-D samples for D-MAD, which contains studio photos, morphs and they chose the ABC gate photos as trustworthy images. And HOMID-S for S-MAD with morphs and studio photos. Both subsets had printed-scanned photos modified by digital postprocessing.

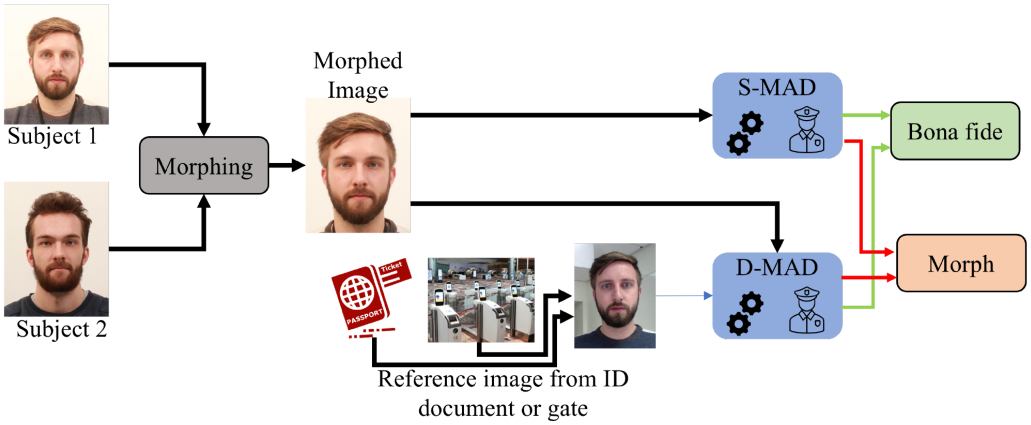

Figure 3.2: Pipeline of Single-Image Morphing Attack Detection (S-MAD) and Differential MAD (D-MAD) [13].

The study was conducted for each method by over 400 respondents (469 for D-MAD and 410 for S-MAD), primarily government employees in the field such as Border Guards, Case handlers, Face comparison experts or ID Experts from over 40 states. Respondents were asked about their training and experience and then the iHOPE platform, shown in Figure 3.3, developed for the experiment let them into the D-MAD experiment, here they always saw a trusted live capture and an image for which they had to determine if it was genuine or a morph, they had to make 100 such decisions. A similar scenario took place in the S-MAD experiment, but here they did not have a photo to compare and thus had a single photo to decide on as in D-MAD. In this experiment, they also tagged 100 pictures. In both experiments, they could zoom in on the picture and pause the experiment at any time and resume later.

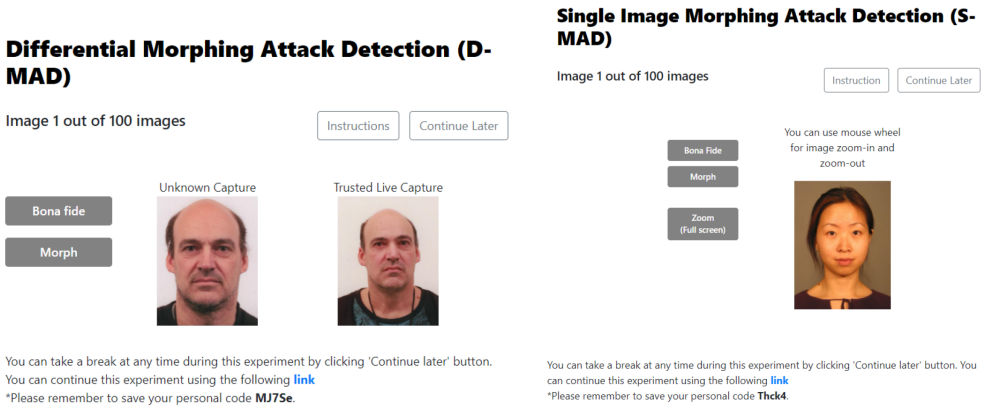

Figure 3.3: Sample of the iHOPE platform used for both experiments [13].

The results say that in the experiment for D-MAD, 11 observers had a success rate higher than 90%, the best even 99.5%. The overall average morph attack detection rate is 64.1%. For the S-MAD experiment, the average accuracy is lower at 58.98%. The study also reports that they did not find a significant difference depending on training when split by years of training to less or more than 1 year and less or more than 5 years. It also describes the success rate by occupation where Face comparison experts perform best in both experiments with an average accuracy of 72.56% for D-MAD and 64.63% for S-MAD.

3.3 Video

An example of research that has looked at human detection of deepfakes in videos is described by Korshunov and Marcel [26]. The experiment used data from the Facebook dataset, which according to this article was the largest and most recent dataset at the time. Another advantage of this is that the videos are ranked and divided into five groups based on how easy it is to spot their visual artefacts. It is therefore possible to determine success rates according to the quality of the recordings in question. The scoring used technologies and methods to ensure that respondents watched the entire video and did not skip answers. They also conducted a screen brightness test and showed respondents a close-up of their face next to the video, so that it would have comparable conditions to a machine that only checks the face area. Before the experiment, each respondent was presented with several examples of test videos of different categories of fake and real videos. Then 40 videos were played to him and he chose the answers “Fake”, “Real”, and “I don’t know”.

The experiment found that only 24.5% of people marked quality deepfake videos as fake, even though they knew they were looking for fakes. Further, 71.1% of people labelled the easily detected deepfakes as fakes, and 82.2% were able to label the real recordings as genuine. The graphs then show that the correct labelling rate decreases with better quality, and quite rapidly, typically by 10-15% for each higher quality group.

The second survey is described in the paper by Groh et al. [15]. Here the authors perform two experiments. In the first one, respondents answer a two-alternative forced choice about whether the video is real or fake. In the second, they present the video to the respondent and ask him to answer what percentage he thinks is a deepfake, then give him another chance to modify his answer after showing him the model's prediction. This is how they evaluate the influence of the respondent's decision.

The deepfake detection rate for the first experiment ranges from 83% to less than 50% for different of the selected videos. The authors do not state the reason for this reduced success rate for some videos. Interestingly, unlike previous experiments, in this one, the authors did not observe any evidence of improvement or learning by the respondents. It is also described that the longer respondents took to make decisions the lower their accuracy was, or that responses were 5.6% less accurate for inverted than for upright videos.

In the second experiment, participants always responded on a scale with percentages ranging from 51 to 100% on one side for real and on the other side for fake, with a threshold of 50% indicating a tie. The detection model made the same decision and showed its prediction to the respondents, which influenced them so much that they changed their answer in 24% of the trials, half of which crossed the 50% threshold, i.e., changed their decision from more likely real to more likely fake and vice versa. The influence alone is said to have increased the average accuracy of the determination by 10.4% for the 40 videos that the model identified correctly. And for the 10 videos that the model identified incorrectly, the accuracy worsened by an average of 2.7%. Further, the authors found that the presence of two people in a video leads to a 7.6% decrease in accuracy or that the presence of a flickering face increases accuracy by 24.2%. The success rate of crowd wisdom was a few percentage points higher and more accurate, rising from 66% to 74% for recruited respondents and from 69% to 80% for non-recruited respondents.

Video deepfakes are also addressed by Tahir et al. [45]. Specifically, the ability of people to distinguish video deepfakes from real ones in two different experiments. The authors use three datasets of deepfake videos (Celeb-DF [29], FaceForensics++ [42] and DeepFaceLab [41]) and genuine videos.

In the first experiment, the respondents were played four random videos and after watching all of them, they were shown a screen to determine which videos were fake and which were not. In this, when the respondent looked at the monitor the authors took a picture of his eyes and analyzed which part he was looking at. They then plotted this data in a heatmap, shown in Figure 3.4.

The results then claim that the participants detected 88% of the genuine videos, and the success rate of detecting fakes ranged between 21%-58%, depending on the dataset. According to the graphs, respondents had the lowest success rate for Celeb-DF and then DeepfaceLab which, the paper claims, proves that background and hair are important in classification. FaceForensics++ had the highest detection success rate, and for what the article describes as lower-quality fakes, people focused primarily on the eyes, nose, and other facial features. Overall, people reportedly focused most often on the background, eyes, forehead, lips, cheeks, and expressions. The results showed no correlation between success rates and age, gender or education.

Figure 3.4: Visual representation of where algorithms and humans focus on when attempting to detect DeepFakes. The leftmost image is the real image, the second from the left is the heatmap from the „Ensemble of CNNs“ technique cited by the paper, the second from the right is the heatmap from the self-reported fake-looking features in the baseline survey, and the rightmost is the eye gaze data from the first experiment [45].

The second experiment focused primarily on learning efficiency. They, therefore, selected participants who were at the literacy end of the spectrum and had limited exposure to the latest technology. They divided the group into two groups, the control group (no training) and the treatment group (underwent training). Both groups played four videos in the initial test and then showed all of them again at once to let them choose which were real and which were fake, similar to the first experiment. The treatment group was then thoroughly trained, reminded of the important features of each video, and told why these features were characteristic of deepfake. Finally, both groups were given a final test, the same as the first but with different videos and ratios of real and fake videos to avoid bias. The course of the experiment is illustrated in Figure 3.5.

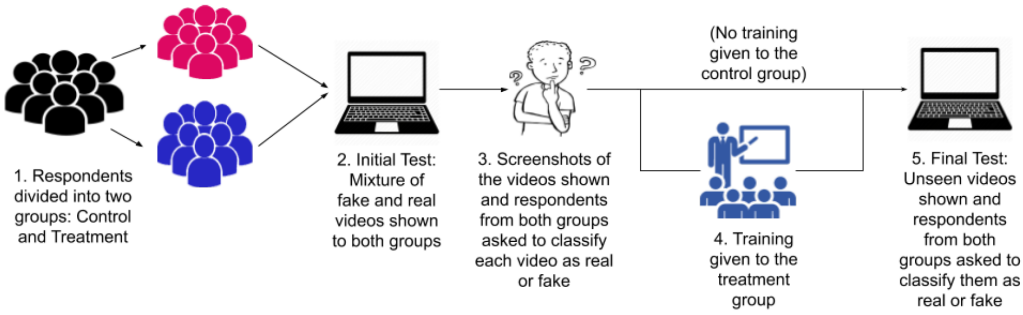

Figure 3.5: Flowchart describing the steps in the specialized training survey [45].

Before training, both groups were able to detect deepfake with about the same accuracy, exactly 55% and 58% (treatment and control group). However, after the final test, the trained group improved a lot and increased the accuracy of detecting fake videos to 88%, and the untrained group had an accuracy of 57% in the second round, with no shift. The treatment group had the largest shift on the simpler deepfakes but also performed much better than the control group on the more complex ones.

The research is also the only human detection research to describe the design of a training procedure, which was also used to train a treatment group. During the training, the participants were presented with examples of deepfakes from both the easy and difficult ends

of the detection spectrum. In addition, during each training session, they were presented with a real video highlighting the differences for comparison but also the most important and common significant features of the deception videos. The authors tried to create content that emphasized the importance of rationality and objective content analysis. Each video had detailed instructions with highlighted points and custom analysis strategies that one should pay crucial attention to. The authors created these strategies based on the results of the detection algorithms and their own results, such as the heatmaps mentioned above. The main features that were presented to the participants included discolouration of the skin, random flickering in the face, blurred spots on different parts of the face, uneven or extra eyebrows, and clumsy lip movements. Overall, the authors state that one should focus primarily on the face, based on the fact that respondents who focused more on the face in the first experiment had a higher success rate than those who focused on the body. This training method clearly had a significant effect on the group of participants, based on their improvement of the aforementioned 33%.

3.4 Overall

Based on the results of these studies, we cannot say whether or not people can recognize deepfakes in any of the media. Detection accuracy ranges from roughly 20%-90% depending on the experience of the respondents, previous training, and most importantly the quality of the deepfakes. From the results, we can notice that the success rate is reduced by either better quality deepfakes or worse quality of the medium itself. What some research also tells us is that training can be a major turning point. For some, it was just based on recognizing the first ten images/recordings of Groh et al. [16] and Müller et al. [36]. And with the Tahir et al. [45] research we can see a huge improvement based on properly targeted training. Also interesting is the combination of detection by deep learning models and human observers, which increases their success rate [15].

Also in all experiments, the participants knew they were looking for the deepfake and therefore targeted it. This is where our research differs very fundamentally from others.

Chapter 4

Experiment design

The main part of this work is an experiment in which we decided to test the human ability to detect a voice deepfake, or at least to notice something strange that might evoke a sense of danger in a person, to which they would respond with increased attention in conversation.

When designing the experiment, we were inspired by Matyáš et al. [33] experiment, we took into account the information given to us by related work and the experience of people in the team who helped with the work. The fundamental difference we want to distinguish from other related work is that the respondents will not know that they are supposed to detect deepfakes. Trying to create a scenario of a real attack where we change a real voice that the victims know and don't find suspicious to a deepfake version and try to see if they react somehow or not.

The experiment was conducted in the Czech Republic and therefore all communication was in the Czech. This is also related to the production of deepfake voice in the Czech language. Despite the fact that the models for non-English languages are not as good as the English ones, their listening quality is still high.

4.1 Research questions

For the whole experiment, we have identified a few main research questions, which are:

1. **Are humans able to identify deepfake recording during casual conversation?**
   We are interested in whether people can notice during a conversation that they have received a computer-generated recording and how they respond to it.

2. **Are humans able to detect a deepfake recording among genuine ones?**
   We wanted to see if people were able to retrospectively identify which of the messages in the conversation was a deepfake recording. Also if they will only mark one or mark multiple ones for their confidence even if they are authentic.

3. **What is people's awareness of deepfake technology?**
   Given that victim knowledge of deepfakes is critical to detecting these scams, we were interested in how many had heard of the technology, or were actively interested in it, and what is their experience with deepfakes.

4.2 Respondents

Respondents are randomly approached by people who meet the requirement of being native speakers of the Czech language and over 18 years of age.

4.3 Course of the experiment

First, when approaching the respondents, we provide basic information about the experiment using a prepared questionnaire. For example, how time-consuming the experiment can be for them, what they need and when they will be contacted. We then asked them for a phone number to contact them via WhatsApp and asked them to prepare the application itself.

We contacted all participants in turn. We started the conversation by introducing, presenting the pre-prepared cover story, explaining the rules of the experiment, explaining the rules of the game we are going to play and telling the respondents that whenever they find anything wrong they should report it to us. This is important for our experiment because we need them to report any concerns to us. It is also important for us to get them used to my voice and to listen to it. We then gradually send them voice messages in which are made statements about the game. They listened to these statements and also replied to me with voice messages as well. In this way, we sent five sets, and one was a pre-prepared deepfake set. As shown in Figure 4.1 describing the course of the experiment.

If the respondent recognized that they had received a deepfake, we would refer them directly to the questionnaire. Otherwise, after all five sets have been exhausted, we sent the respondent a link to the questionnaire to complete. At the end of the questionnaire, we added a few links so that respondents can look up other interesting facts about deepfakes and learn more about them.

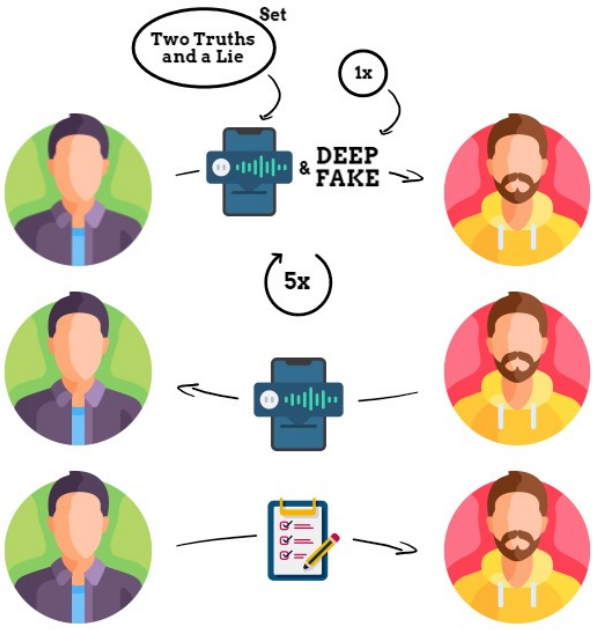

Figure 4.1: Flowchart describing the course of the experiment. (www.freepik.com)

4.4 Coverstory

Our primary goal is to get enough voice recordings to the respondent without telling them what we are up to or making them suspicious in any way. So in order to create a real attack scenario, we need to hide the whole experiment under a prepared cover story.

We told participants in the experiment that we were testing the usability of voice messages. Set the whole thing in a narrative in which we position voice messages as the future dominant form of communication. We said that this is why we need to test their usability, and why we will only need to communicate with them using voice messages during the experiment. Then explained the content of our communication, as described below, and asked them not to be afraid to reply by voice message whenever they received a message from me, whether at work, on public transport or in a shop. We hoped that this would lend credibility to our cover story, and moreover, it would ideally model the real situation in which the victim of an attack might find themselves, namely a situation in which they are not focused on an incoming call because they are distracted by their surroundings and needs to deal with the call as quickly as possible.

The real meaning of the experiment and the revelation of the cover story comes only in the second part of the survey when users learn about deepfakes.

4.5 Content of communication

In order to give the respondent the feeling that we're testing the usability of voice messages, we need to somehow create a short and direct conversation line. We found it ideal to play a game with them. The content of the recordings is therefore designed for the game *Two truths, one lie*.

We chose this game mainly because we think we are able to keep the attention of the respondents. By having to determine which statement is a lie, respondents have to listen to the set multiple times, which increases the chance that they may notice something odd in the recordings. And it offers them a clear answer so that they don't have to think about it for a long time and can answer within a few minutes. We believe this interaction supported the cover story about testing voice message usability.

We picked a topic country in the world and three interesting facts about each country, then edited one each time to create a lie.

The difficulty of uncovering the lie is irrelevant to the experiment, after all, information from Wikipedia should be enough. We just need to take into account not to lose the respondents' attention with too primitive examples and make it as interesting as possible for them.

For example, here is one set from the game that we used for an experiment. The others are listed in Appendix A.

India

1. *India is the second most populous country in the world.*
2. *The capital is Mumbai.*
3. *The most widespread religion is Hinduism.*

The lie and the correct version of it is 2. The capital is New Delhi.

4.6 Creating a deepfake

For this experiment, We created only one deepfake recording, which we tried to improve to perfection by post-processing. The quality of the deepfake is crucial to us in this, and a bad deepfake could fundamentally affect the results. The created deepfake must be in English, that's one of the reasons why we lean towards voice conversion technology. Models trained for voice conversion are able to work for several different users. Instead, text-to-speech models are mostly trained with one target speaker's voice, an exception is the multispeaker text-to-speech tool by Jia et al. [20]. Only a few really good models have been trained in the Czech language. Voice conversion allows us to use an already trained model and just give it data that it processes and generates a deepfake for us.

There are several tools that are easily traceable on the internet and most of them are open-source projects:

- **Adaptive Voice Conversion**[1], is an implementation by Chou, Yeh and Lee [9] which is already discussed in connection with the creation of VC deepfakes in Section 2.4.
- **YourTTS** [7], which although the name doesn't match, is a tool that supports both TTS and VC technology. The only one mentioned is not open source, but has a publicly available demo on Google Colab[2].
- The other tools are **FragmentVC** [30] and **StarGANv2-VC** [28].

In the end, we chose the YourTTS tool [7], mainly due to the access to the tool via a demo on Google Colab and the fact that we possess a version with a trained model in the Czech language. The creation of the recording itself was then easy. All we had to do was compile and provide the target speaker recordings and the source recording. We created several such recordings with multiple source recordings and then selected the most accurate one. However, even that one had errors, so we decided to go with post-processing.

First, we had to remove the noise that was created during creation, which we did use the *Noise Reducer*[3] tool. Next, it was necessary to smooth out the frayed phonemes, which we solved using *Audacity*[4] by cutting out the part of the recording where the phonemes resonated. We also adjusted the pitch of the voice in *Audacity*. And after a test run added artificial brown noise, the same as we play into voice messages, to achieve the same quality of both recorded inputs and resolve the differences described in Section 4.9.

4.7 Deepfake evaluation

We were trying to find a tool to evaluate the deepfake recording itself so we could describe its quality. Since we could not find any tool that could do this, and since we only needed this information for indicative evaluation, we decided to use an evaluation inspired by the **Mean Opinion Score (MOS)** subjective listening test method described by Loizou [31]. Unfortunately, we could not find any scorecard suitable for this type of evaluation. These tables are mainly based on the Quality of Experience and Quality of Call ratings, which in my opinion are quite different from the deepfake processing quality ratings. Furthermore,

[1] https://github.com/jjery2243542/adaptive_voice_conversion
[2] https://github.com/Edresson/YourTTS
[3] https://noisereducer.media.io/speech-enhancement
[4] https://www.audacityteam.org

the MOS evaluation takes place in two phases, the first is a training phase where subjects are played a poor-quality recording and a good-quality recording. In the second, evaluation phase, respondents rate the test recording according to the previous ones. Even in this case, we are not able to replicate the given metric exactly, but despite this, we decided to at least use the evaluation patterns and played the recording during our **Security@FIT** research group meeting to people who work with deepfakes regularly and are familiar with them. We, therefore, consider them to be experts on the subject and we know that they know deepfake recordings in all qualities. We asked them to rate the quality of the recording on a scale of 1 to 5, with 1 being poor and 5 being excellent. All of them gave the recording an average rating, i.e. 3, and therefore the recording qualitatively corresponds to the rating „Fair“.

4.8 Creating a survey

To create the questionnaires we used www.limesurvey.org.

When creating the survey, it was important to keep in mind that our experiment is hidden under a cover story. It was necessary to determine the correct sequence of questions so that the questions could not influence those yet to follow. We knew that we had to get to the questions that were most important to us slowly, and at the same time, the survey follows after the experiment and therefore must not be time-consuming. Therefore, we finally chose the following option. The survey is divided into six groups of questions:

1. **Respondent profile.**
   Here we asked for basic information about the respondent such as age, gender, the field in which he/she works and a phone number which allows us to check the validity of the answers linked to the experiment.

2. **Usability**
   In order not to start asking the respondent about deepfakes right away, we decided to select one question on the usability of voice messages that might be useful for evaluation.

3. **Recordings**
   In the next area, we asked the respondent about the recordings and if they found anything strange or unnatural about them. Alternatively, what it was. This is one of the most important questions in this research.

4. **Deepfakes**
   At this moment we told the respondent what deepfakes are and asked whether they have ever encountered them and possibly in what context. The next question asked how confident they are that they would recognise a deepfake. This question responds to research that looked at how confident Americans are that they can recognize a computer-generated voice masquerading as a human voice [32].

5. **Real experiment**
   At this point, we explained our full experiment, reveal the cover story, and admit that we sent the deepfake in our conversation. We checked if the respondent was able to recognize which set are deepfakes when they already know they received at least one.

6. **Conclusion**
   In the last section, we revealed which recording was not authentic and determined whether the respondent was surprised by the quality of the voice deepfakes and whether he is more or less confident that he can recognize a deepfake after this experience and after revealing the true meaning of the experiment.

At the end of the survey, we recommend links where respondents can learn more about deepfakes. A detailed overview of all the questions can be found in Appendix B.

4.9 Test experiment

Before we started with the actual experiment we did test runs where we ran the experiment exactly as we did afterwards with real respondents. After the test we contacted them and they gave us feedback.

This helped to improve the lyrics, tweak the deepfake recording, and uncover potential issues that might affect the experiment. One such problem is the actual transmission of the recordings. This is because the WhatsApp app recognizes voice messages recorded at a given moment through a microphone, to which it puts a photo of the sender and recordings sent from a computer or phone, to which it puts a default picture, as shown in Figure 4.2. This is actually a simple but very interesting security feature that an attacker has to deal with. By sending voice messages to the respondent in the experiment within a fairly short time span, all of roughly the same quality, and with no significant distractions, and with similar volume levels, the true voice recordings are quite accurate, high quality, and similar to each other. Deepfake, on the other hand, we have to play from speakers and record their output, which has slightly different parameters in volume, noise and so on. For these reasons, we decided to add artificial noise both to the deepfake recording and to play it with the actual voice messages. As is described in Section 4.6.

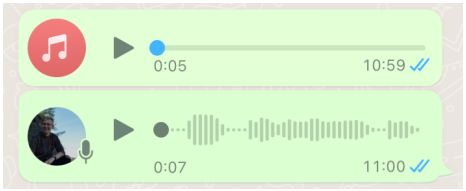

Figure 4.2: The difference between sending a recording and recording a voice message. The top message is a pre-recorded message and the bottom is a message spoken live into a microphone.

Chapter 5

Experiment

Following the design of the experiment, we then performed the actual run. Thirty-one respondents completed the experiment and finally completed the questionnaire. We analyzed, evaluated and described the results in this chapter.

5.1 Reaching respondents

The experiment itself began by reaching out to people. Respondents were contacted mainly via the Internet, former colleagues, classmates, and similar. We did not offer anyone anything for completing the experiment, and all respondents participated in the experiment of their own volition to help us. In the end, we managed to reach more than 130 people who opened the sign-up form, but unfortunately, only a few of them signed up for the experiment.

5.2 Profile of the surveyed group

Thirty-one people signed up for the experiment and we asked each of them about their gender, age, and field of work. All of these parameters are quite related to the environment in which we personally operate because we collected the respondents ourselves and approached mainly people from our own community. All respondents also fulfilled the conditions set by us, i.e. all were over 18 years of age and all were native speakers of the Czech language.

In terms of gender, 71% of respondents were male and 29% were female, shown in Figure 5.1.

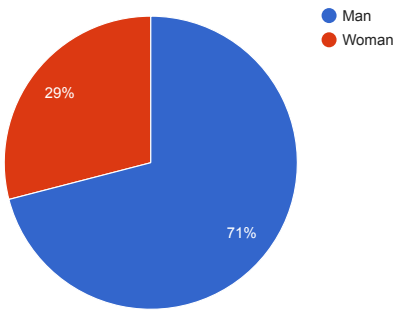

Figure 5.1: Gender of respondents who participated in the experiment

The age of the respondents ranges from 18 to 46, but **80% of the values are less or equal to 23** and the average age is about 22.39 years. When we compare age with gender, the average age of females was 23.78 and males was 21.82. The age range of respondents including gender is shown in Figure 5.2.

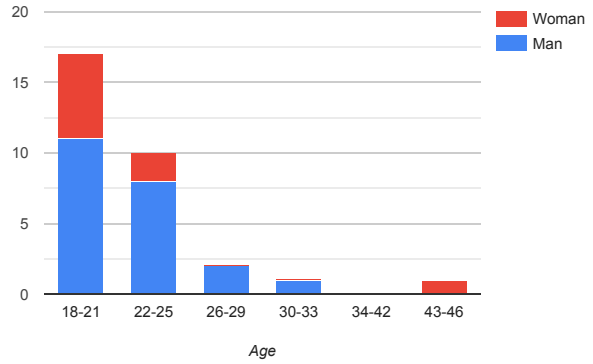

Figure 5.2: Age of respondents with a look at the gender ratio in five age groups

In focus on the field of work, IT has the highest representation with 41.9% of respondents, the next common field is teaching with 19.4%, law and healthcare with 6.5% and other fields like machinery, marketing, military, art etc. All shown in Figure 5.3.

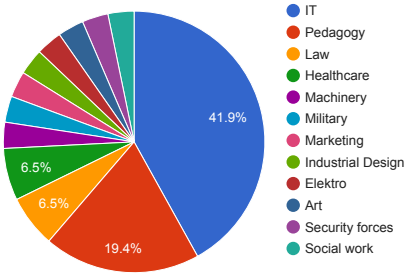

Figure 5.3: Proportions of fields in which respondents work

5.3 Course of the experiment

The whole experiment went quite smoothly and without any major problems. However, a few minor problems showed up.

One of them was that a lot of the people interviewed didn't have WhatsApp installed, people use Messenger, Instagram etc., and WhatsApp is not the most used communication platform here in the Czech Republic. We believe this, along with the fact that we asked respondents for their phone numbers as personal information, was the main reason why more people did not sign up for the experiment.

The second major problem was the quality of the recordings. As described in Section 4.9 and 4.6, we played artificial noise in the background of the recordings. And although when we played the recordings back (on iPhone 11) the noise was minimal and we could understand everything without any problems, a lot of people got back to me saying that the quality of the recordings was really bad mainly because of the noise. We suppose it depends on the device on which the respondent listened to the recordings, some devices maybe can

reduce the noise others can't. Poor quality and noise was also the most common thing that respondents identified as odd about the conversation, 13 people did in the questionnaire.

On the other hand, the feedback from the respondents was that they enjoyed the experiment and learned new things, they found it interesting and liked the topic of deepfakes, and a lot of them told me after the experiment that they didn't expect or think that there could be a fraudulent recording or something like that in the conversation. We even talked to a few of them, especially those in the IT industry, about the creation of deepfakes, how it works, in what situations deepfakes become dangerous, etc.

In terms of time, We did the experiment for about 3 weeks, with most of the time taken up by finding people. Then some people went through the experiment in 15 minutes and some took 4 days, it was a casual conversation so apart from reminding ourselves we didn't push anyone to answer and left it up to them. No one complained about the time commitment or anything like that.

5.4 Results

Due to the small number of respondents and the predominantly verbal responses, we evaluated the experiment's results manually, without any evaluation scripts or similar. As described in Section 5.2, 31 respondents participated in this experiment.

5.4.1 Research questions

1. **Are humans able to identify deepfake recording during casual conversation?**
   The first research question was evaluated based on the participants' reactions to the deepfake message during the conversation, whether they said something specific, whether they indicated that something was suspicious, etc. And also on the basis of the responses from the Voice messages question group, shown in Appendix B. When evaluating, we were mainly guided by the verbal responses and categorised them into different groups. We formed these groups based on the responses and there were four, nothing noticed, irrelevant comment, lower quality, and deepfake sign. As a deepfake sign, we labelled responses that contained a description of something specific to the voice deepfake.

   No one reacted to the deepfake at all during the conversation. One respondent even asked to repeat this set, yet he continued on and answered the question as the others did without noticing.

   Only one respondent mentioned anything specific about deepfakes to the question "Did you find anything unnatural about the voice messages you received?", mostly the worse quality mentioned by 13 people, and there were stray comments such as "Too much geography" or respondents said they do not notice anything unnatural. So there was only one respondent who said that he found it suspicious that the recording was artificial, although he did not give anything more specific, we consider this answer as a revelation of the experiment. Without asking anyone or looking up this statistic, 9 people, almost one-third, mentioned to me either after the experiment or in their text responses to the questionnaire that the possibility of a fraudulent recording had not occurred to them during the interview, and they focused primarily on the content and the correct answer, stating that they considered the lower quality to be normal.

2. **Are people able to detect a deepfake recording among genuine ones?**
We evaluated this research question based only on responses to the Real Experiment group question, described in Appendix B, answered when participants knew they had received at least one deepfake during the conversation and were asked to identify which voice message contained it. We evaluated whether the respondent detected or not according to the set he/she marked. Here we have two statistics, namely whether they labelled the correct set and whether they only labelled the correct set, i.e., did not label any other set and correctly identified the deepfake and real messages. Furthermore, we only investigated the reasons for marking the deepfake set, we did not include the reasons for marking real messages in this evaluation. We evaluated and categorized the responses. For the answers, we created three groups that the respondent used to justify his/her choice, namely lower quality compared to the real recordings (lower sound quality, less intelligibility and so on), the message was simply different (different voice, different sound system, quieter, different intonation and so on) and deepfake sign as in the first research question, i.e. answers that contained a description of something specific to the voice deepfake. We categorised some justifications into two categories if the justification was primarily based on either reduced quality or differences from other reports, but also contained some deepfake sign. Some participants omitted explanations.

   The deepfake set was identified by 96.8% of the respondents, but 13.3% of them also identified other possibilities, so we can say that **83.9% of all respondents correctly identified only the deepfake set**. Thus, only one person did not identify the correct deepfake set. For all respondents who did not mark only the correct set, we can say that in general, they did not reveal the correct set.

   When justifying this choice, the most common reason, 54.8% of respondents, indicated the correct set because it is simply different. The second reason was lower quality compared to real recordings, mentioned by 29% of respondents. And the last one is the identification of some artefacts that deepfakes are characterized by, 22.6% of respondents.

   All those who did not mark only the correct set said at the beginning of the questionnaire that they found the low quality suspicious. Interestingly, 4 of the 5 were from the IT industry, which may be a coincidence given that almost half of the people are from the IT industry, but from their explanations is clear that they tended to approach it differently than the others. As it said most people were marking the deepfake set because it was different from the others, whether because of lower quality or other aspects. People in the IT industry were much more cautious about it, they justified a lot of it by monotone pronunciation, steps in intonation, or pronunciation of word endings. Obviously, they know the features of deepfake technology, even after talking to them we learned that maybe they only know the TTS method and therefore assumed pronunciation without overspeaking etc. Overall, they were more cautious and emphasized a bit different aspects than just that the recording sounds different. This is not necessarily a bad thing, this subject is quite dangerous and needs to be approached very carefully, and they may have known that, but it doesn't change the fact that the people who took the easy way out were more successful.

3. **What is people's awareness of deepfake technology?**
   The evaluation of the last research question is based on the answers from the first two questions of the Deepfakes questionnaire set described in Appendix B. Responses to the text-based questions were again classified into the groups described below.

   The response rate for the question "Have you ever heard of deepfakes?" is shown in Figure 5.4. Here, respondents had a choice of three options, 16.1% of respondents answered "I've never heard of them", "I've heard of them before" was 64.5% and "I'm actively interested in them" which was marked by 19.4%.

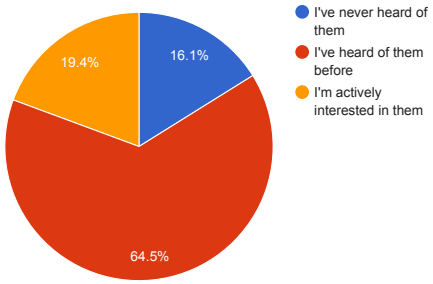

Figure 5.4: Proportion of deepfake knowledge groups

If we look at the success rate of guessing right only the deepfake set depending on whether the respondents have heard of deepfakes or not, we find that those who have never heard of deepfakes always answered correctly, but people who are actively interested in deepfakes only got it right four out of six times, which I think supports the previous reflections about the attitude of people from IT backgrounds.

Where they heard about deepfakes is variable, but can still be classified into several groups. More than a quarter of people, 25.8% to be exact, said that they heard about deepfakes on social media, mainly in some informative videos, articles, etc. One respondent said they had encountered deepfake videos of politicians on TikTok. Consistently, 19.4% of people wrote that they simply heard about them on the internet, nothing more specific, or that they simply heard about them and did not specify where, or tried to create them themselves which were mainly people who are in the IT environment. The smallest group with 16.1% of respondents are those who have never heard of deepfakes. These results are shown in Figure 5.5.

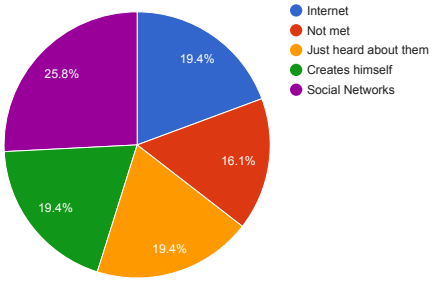

Figure 5.5: Percentage of groups from where people know deepfake

Based on these results, 83.9% of the participants have at least heard of deepfakes and this is mainly from social media and informative videos.

5.4.2 View of deepfakes

After revealing the experiment at the moment when the respondents already knew if they had correctly detected the deepfake set or not and knew which set was fraudulent so they could replay it, we asked them if they were surprised by the quality of today's voice deepfake in the Czech language. 74.2% of the respondents said they were surprised and 25.8% said they were not surprised, as shown in Figure 5.6. Most of those who said they were not surprised by the quality were people who are at least marginally involved in the IT industry.

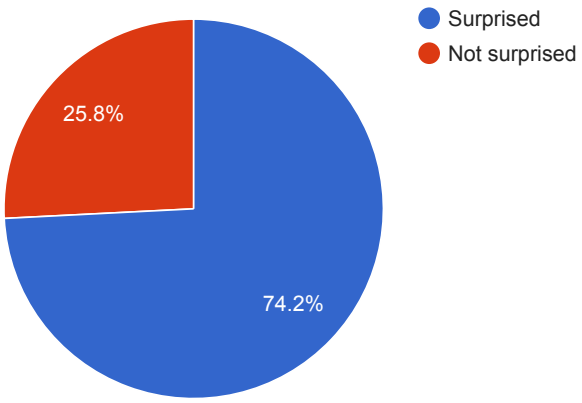

Figure 5.6: Chart of respondents' answers to the question, if they were surprised by the quality of today's voice deepfakes.

Whether or not the interviewees were surprised by the quality had no effect on whether they increased or decreased the value of how confident they were that the voice deepfake would be detected after the experiment. We asked respondents at the beginning of the questionnaire before the experiment was revealed if they believed that voice deepfake would reveal, and we asked them the same question at the end of the questionnaire if the experience had changed their minds in any way. The answers are shown in Figures 5.7 and 5.8.

Seventeen people, or 55%, changed their initial opinion, 16 of whom said they were more confident than at the beginning, and only one lowered his view. A total of 52% of respondents increased this value, 45% did not change it, and only 3% decreased it. All those who increased the value correctly detected the fake set, the only one who decreased also guessed, and none of those who did not guess right changed their answer.

The respondents were given the opportunity to rate their convictions on a scale of 1-5. The overall average was initially 2.29 and eventually 2.94, so overall the average increased by 0.65 points. Those who increased the value raised it by an average of 1.31 points. We divided the scale into three groups, less than 3 means the respondent was not confident, equal to 3 means they are unsure and greater than 3 means they are confident. So in the beginning, only 2 people, 6.5%, were confident, while by the end 9 people, 29%, were confident. The unsure group changed from 29% to 35.5% of the respondents and the not confident group decreased from the beginning to the end of the experiment from 64.5% to 35.5% of the people.

Compared to the article [32] that asked similar questions about how confident people are in recognizing the computer-generated voice, the results are quite similar, to the results from the end of the survey. The article states that 36% of people are confident they are able to detect deepfake, our numbers were 6.5% and finally 29%. And it goes on to say that

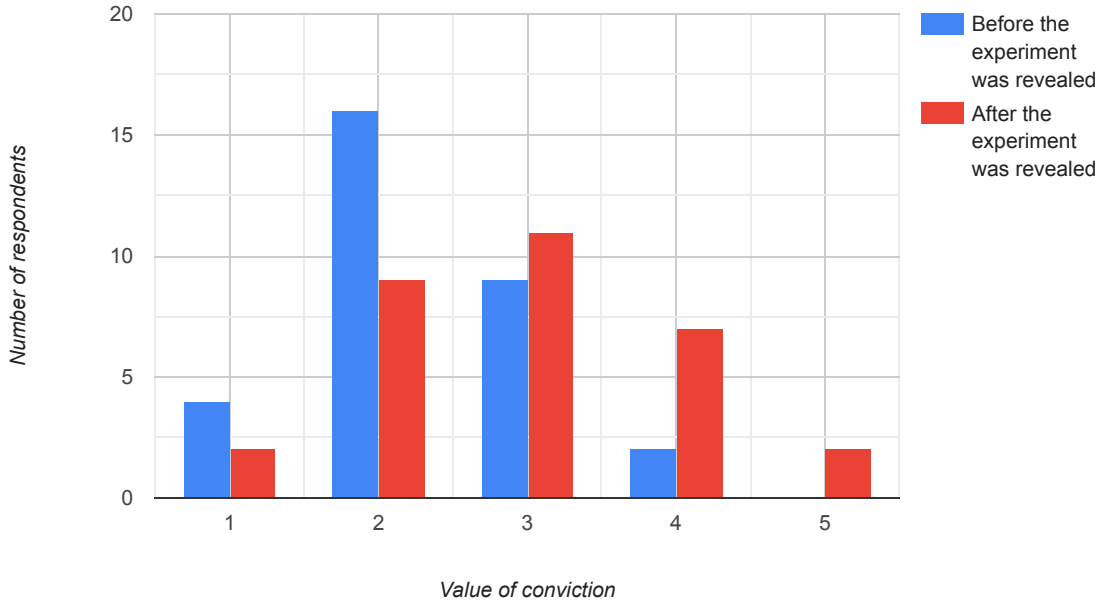

Figure 5.7: Graph of responses to the question of **how confident respondents are in detecting a deepfake**, quantified by a number of respondents.

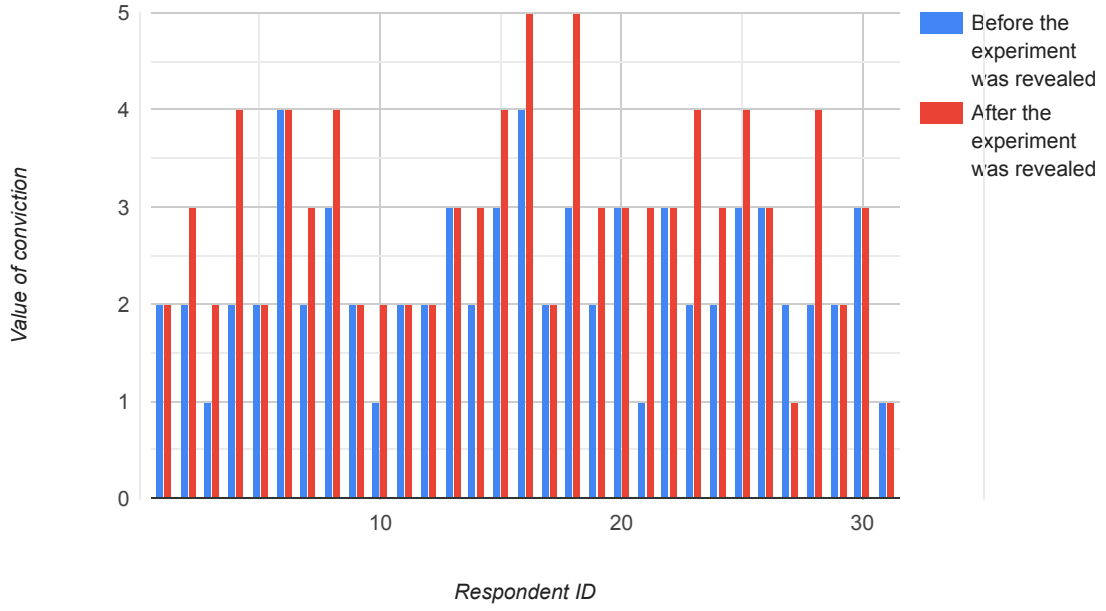

Figure 5.8: Graph of responses to the question of **how confident respondents are in detecting a deepfake** by respondents and their answers.

30% of people were not confident, our numbers were 64.5% and then 35.5%. This article does not state whether the respondents had a choice on a scale like in this research or if they only chose from, say, three options, nor does the article state whether the respondents had any experience or not. Therefore, we would venture to say that when comparing our results from the end of the questionnaire to the results of the article, our results are quite similar.

As is written, the overall average has risen by 0.65 points. If we split the respondents into men and women, or by whether they are in IT or not, the results would only differ by hundredths of a point. However, it is more interesting to look at the correlation with age, by dividing the respondents into three dominant age groups (17 people aged 18-21, 10 people aged 22-25 and 4 people aged 26 and older), we get quite different numbers. For the youngest group, the average increases from 2.18 to 3.06, an increase of 0.88 points. For the middle group, it changes from 2.6 to 3.1, or 0.5 points. And of the oldest group, no one changed their answer and at the end of the questionnaire, the average remains the same as at the beginning, namely 2.0. If we wanted to express the correlation and calculate it for the average age of the groups and the average score difference of the groups we would get the rounded value of the *Pearson correlation coefficient* -0.9736, which means that the older the respondents were, the difference in values decreased. Personally, we would say that the older ones approach this more cautiously and although the respondents were mostly young people and even the oldest group has an average age of 33 years so we think that the confidence of the older ones in technology is less than the younger ones and this attitude breaks down at an early age.

5.4.3 Impact of using voice messages

We also asked respondents how often they use voice messages and their responses are shown in Figure 5.9. We found no correlation with other values in these answers. People who answered that they use voice messages frequently, those who answered 4 or 5, had different knowledge of deepfakes, values of how confident they are with deepfake detection, the success of detection, and different justifications. People who answered that they do not use voicemails at all, thus answering 1, were similar. Here, however, we can say that all of them detected the deepfake set, but we don't think this has any deeper relevance given that most of the time their justification was that the recording was simply different from the others. If we wanted to look at how often people use voicemails as seen in the graph below, most people don't use them much.

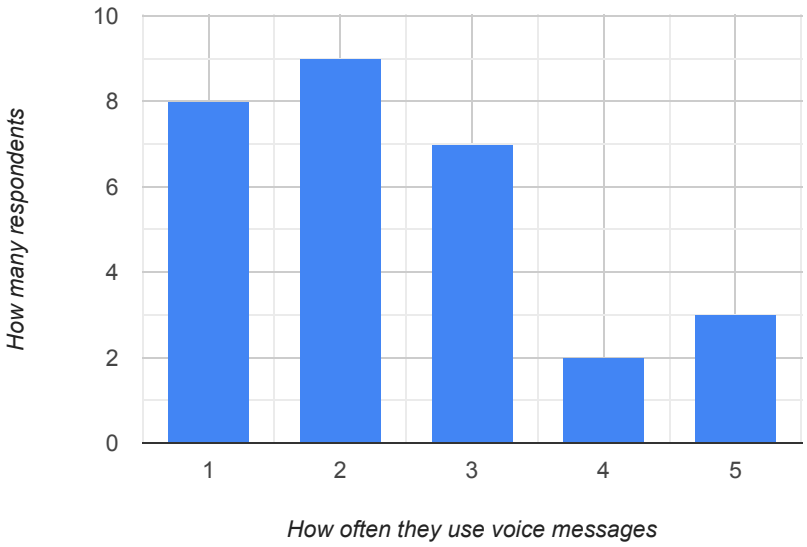

Figure 5.9: A graph of how often respondents use voice messages, where 5 means very often and 1 means never.

5.5 Experiment conclusion

The results of the experiment gave us a lot of information. They answered the research questions we identified in Section 4.1, and equally provided information on which we are able to assess the human ability to recognize synthetic speech, the main goal of this work. First, the answers to the research questions:

1. **Are humans able to identify deepfake recording during casual conversation?**
   None of the respondents noticed anything during the interview, or at least none of them found it strange enough to react to it, and only one of them described the features of deepfake technologies in his answer. Thus, we can say that only one of the thirty-one participants revealed the real intention of the experiment, and only after we asked what they found strange. A lot of participants even said that they focused mainly on the content and tried to answer the questions, ignoring the degraded quality or similar things because they thought it was a common technological shortcoming.

2. **Are people able to detect a deepfake recording among genuine ones?**
   The deepfake set was correctly identified by 96.8% of respondents and only this correct set was identified by 83.9% of all respondents, the rest either identified a completely different set or identified more than one of them. Most respondents justified their choice by saying that the message was clearly different from the others or of lower quality, and that it was very easy to spot on second listen once they knew they were looking for deepfake. Which interestingly goes against the results of the first research question, so participants did not notice anything special at all on the first listen when they were focusing on the content, but on the second listen when they were looking for a deepfake, they usually noticed the difference clearly. The results also reveal that most of those who did not clearly identify the deepfake set were people from IT backgrounds. Their reasoning suggests that they were familiar with and focused on the features typical of deepfake and therefore often identified multiple possibilities, including genuine recordings. Or, for example, that they were familiar with the TTS method and did not anticipate better speech fluency or overspeaking. So even though they were more cautious, knowing the methods didn't help them in detection and people who took the easy way out were more successful.

3. **What is people's awareness of deepfake technology?**
   Deepfakes were encountered or at least heard about by 83.9% of respondents. Of those, 76.9% have only heard about them, seen a video or read an article, and 23.1% are actively interested in them, these were only IT people. And so 16.1% of all respondents had never heard of them. The main place where participants were introduced to deepfakes was social networks, especially informative videos and the like. A few respondents even created deepfakes themselves.

Interestingly, after the experiment, more than half of the people increased the value they were confident of detecting a deepfake. The younger part of the respondents increased the value, while the older ones kept the value the same despite correctly detecting a deepfake. And exactly 74.2% of respondents said they were surprised by the quality of today's deepfakes.

From the results of the experiment, we can conclude that the human ability to recognize deepfakes is not at such a level that we can trust it, it is insufficient, and it is not easy for humans to recognize even obvious deepfakes in ordinary conversation.

Related work, described in Chapter 3, has evaluated human ability with often more than a 60% success rate. The success rate of deepfake detection in our scenario is 3.2% and this result is quite different. Of course, it is important to say that our approach is fundamentally different from those of others and perhaps that is why our results are so different. A similar scenario to the other work is offered by the second research question, where the success rate is 96.8%.

Chapter 6

Suggestions

An article by Westerlund [47] quotes computer scientist Hao Li [1] as saying "*This is developing more rapidly than I thought. Soon, it's going to get to the point where there is no way that we can actually detect [deepfakes] anymore, so we have to look at other types of solutions.*"

Looking at the results of our work and the work of others, we have to admit that we need to adapt to artificial intelligence faster than we expected.

6.1 People training suggestions

There are some interesting methods such as those presented by Tahir et al. [45], who educated respondents by illustrating deepfake videos along with highlighted points and analysis strategies. The training in this research led to significant improvement as is described in Section 3.3. This method could easily be transferred to audio media, it would require analyzing and defining the specific artefacts of audio deepfakes recordings and, as in research [45], educating people about these possible markers of synthesized sound using specific examples. But here, we think, we come to the problem of speech media. Videos travelling through the internet today are of high quality. On the other hand, media transmitted by sound, such as phone calls or voice messages, can be significantly screwed up in quality just because of the transmission or recording technique and can show significant signs of deepfake even normally. This would lead us to focus our training on a narrower set of artefacts indicative of synthetic voice, such as equal length of spaces between words, robotic voice, etc. And even so, many people may consider these features to be just reduced sound quality, as they were in our experiment.

The interesting thing our experiment found is that people only checked the content on the first listen, ignoring the sound artefacts, and no one detected the deepfake. But on the second listen, they focused on the sound artefacts and most of them detected the deepfake. This leads to a possible method of identification that could be called *Double-listening prevention*, where the first listening focuses on the content as it is natural and the second listening focuses only on the audio and if something unnatural comes up, the user should verify the content as described in the next paragraph.

Another option is to train for verification and caution because, as research shows, deepfakes deceive us already and their believability will increase. Even our research has shown that knowledge of deepfakes or even creation techniques may not ensure a higher detection success rate, described in Section 5.4. An article published by FBI [4] mentions the SIFT

method i.e. Stop, Investigate the source, Find trusted coverage, and Trace the original content when consuming information online. This method can lead to uncovering disinformation, exposing genuine media and debunking hoaxes, and people who adopt it will be much better protected against deepfakes than those who focus on their signs and rely on their own senses. The article also advises against considering an online persona legitimate just because it has a video, image, or audio on social media. Use multi-factor authentication on all systems. And also confirming any sensitive information through some other communication channel. An example of such a situation, inspired by the attack described in Section 2.2, would be for each company to set up rules for double authentication of money movement requests. For example, when someone requests a money transfer over the phone, it is a good idea to email the person to confirm the transaction. In the case of spoofing, the attacker only "calls" from a given number in one direction, if you terminate the call and call back and get the previous call confirmed, you can prevent the attack. Even this simple validation of requirements can be very effective against such fraud. The solution could therefore be to educate the public about securing their communications, verifying sources and developing a neutral attitude towards information and media appearing on the internet, especially those that disseminate polarising content.

Another approach to detecting misinformation could be initial distrust, we often see recommendations to detect if the information is false. What if we consider each piece of information to be false first and then verify its authenticity. Just like when a stranger on the street shows people a picture depicting something shocking, most people would probably not believe it and would only accept the information after verifying it. On the internet it often works exactly the other way around, people believe it first and then look for a reason to disprove the information.

Tools for detecting deepfakes can also be a good way to detect them. As demonstrated by Groh et al. [15] the opinion of a detection tool increases the chance of detecting fraudulent media. Sadly, even though there are plenty of articles on deepfake detection, plenty of models and so on, if one wants to find a non-commercial tool that can be used online to verify a media isn't easy. We find tool `www.deepware.ai` which can tell a deepfake video from a real one, but it only accepts videos from Youtube, Facebook or Twitter.

The most ideal would be to combine all these options to create a place, an educational platform, a web app or something similar to offer people:

- Demonstrations of deepfake technologies, abuse cases, their vulnerabilities, or how to defend against them.
- Interactive training in detecting any kind of synthetic media.
- Teaching caution about information and how to verify it.
- Summary of the detection tools, their introduction and instructions for use.
- Recommendations and links to sites to help people affected by deepfakes technologies, such as the aforementioned `www.napisnam.cz` operating in the Czech Republic.

Publicly accessible web applications, where people could freely view tutorials, try to work with deepfakes and generally learn about the technology, could increase the public's resilience to these scams.

6.2 Suggestions for spreading awareness of deepfakes

My survey showed that more than 83% of respondents have heard of deepfakes before, especially on the internet and in videos that discuss the issue. Today, content creators have a huge reach on the internet and this can be seen for example in advertising campaigns, which often take place primarily on social media. There are plenty of creators out there dealing with these topics, and even today they are succeeding in informing the public with at least marginal information about the dangers of deepfakes. Using the reach that these people have could be the key to spreading this information to the general public and it's up to us, the people in the industry, to get this information out to them.

However, it is important to note that deepfakes may not only be the spreading of misinformation through public social networks but also attacks companies and similar institutions. People in this field need to be trained in the technologies they may face and shown how to deal with them. The platform suggested in the previous section would be sufficient for such training, with examples that people can go through and become familiar with the problem.

It is probably not necessary to conduct training in the form of a face-to-face meeting with a professional. Certainly, there would be better interaction, and people could ask questions, but there would not be a significant difference in the amount of information imparted that would equal the cost of such training versus a well-structured guide on a platform.

Chapter 7

Conclusions

The aim of this work was to evaluate the human ability to recognize synthetic speech. We have structured this task into three main questions that should be answered, as described in Section 4.1. To answer them, we designed and executed the experiment described in Chapter 4.

The results of the experiment (Section 5.4) show that during the conversation, none of the thirty-one respondents reacted in any way to the fraudulent deepfake message. After evaluating the questionnaire to see if they noticed anything, we agreed that only one person noticed and described something specific to deepfakes. On the other hand, when identifying a deepfake message, 96.8% of respondents correctly identified it. And 83.9% of respondents correctly marked only the deepfake message and not the other genuine ones. Thus, the results show that although the deepfake message was clearly identifiable among others, no one reacted in any way to it.

The experiment also found that 83.9% of respondents had at least heard of deepfakes, primarily from social media, educational videos or simply stumbling across them online. The survey also asked respondents how confident they were that they would detect a deepfake. We asked them this question twice, at the beginning and at the end of the questionnaire, and had them rate their view on a scale of 1 to 5. The average was 2.29 at the beginning and after the experiment was revealed, at the end of the questionnaire, the average increased to 2.94. The value increased mainly with younger respondents.

At the end of the thesis (Chapter 6), are present suggestions for training people in recognizing deepfakes and ways to spread awareness of them. These suggestions are built on articles that discuss this topic and on the results of the experiment we made. Their main part is a content design of a fully accessible platform for public familiarization with deepfake technologies, synthetic media detection training and information verification methods in general.

The experiment could be improved by focusing on improving the quality of the deepfake voice and solving it so that background noise is not necessary. This noise is heard differently in different devices, as the respondents told us, which is an uncontrollable fact on our part. This problem could be further investigated, for example, in terms of the impact of the end-user device to deepfake recognition. Work that could follow would be an implementation of the educational platform described in Chapter 6, proving the effectiveness of the proposed *Double-listening prevention* from the same chapter, or finding the effect of age and field of operation on knowledge about deepfake technology as well as other artificial intelligence technologies.

Work has shown that the human ability to recognize voice deepfakes is not at a level we can trust. It is very difficult for people to distinguish between real and fake voices, especially if they are not expecting them. The human ability to detect deepfakes is probably largely influenced by the fact that people don't think about the voice they are listening to, are used to poor-quality audio conversations, and focus primarily on the content of the message. However, it is very easy to hide deception under poor quality.

Bibliography

[1] *Hao Li*. Available at: https://www.hao-li.com.

[2] *Karen Hao*. Available at: https://www.karendhao.com.

[3] *Increasing Threats of Deepfake Identities*. Department of Homeland Security, 2019. Available at: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

[4] *Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations*. March 2021. Publisher: FBI. Available at: https://www.aha.org/system/files/media/file/2021/03/fbi-tlp-white-pin-malicious-actors-almost-certainly-will-leverage-synthetic-content-for-cyber-and-foreign-influence-operations-3-10-21.pdf.

[5] Ajder, H., Patrini, G., Cavalli, F. and Cullen, L. *The State of Deepfakes: Landscape, Threats, and Impact*. www.deeptracelabs.com, september 2019. Available at: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[6] Barak, T., Avidan, Y. and Loewenstein, Y. *Naive Artificial Intelligence*. arXiv, september 2020. ArXiv:2009.02185 [cs]. Available at: http://arxiv.org/abs/2009.02185.

[7] Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E. et al. *YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone*. arXiv, february 2022. ArXiv:2112.02418 [cs, eess]. Available at: http://arxiv.org/abs/2112.02418.

[8] Cbsnews.Com. *Doctored Nancy Pelosi video highlights threat of „deepfake“ tech*. May 2019. Available at: https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/.

[9] Chou, J.-c., Yeh, C.-c. and Lee, H.-y. *One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization*. arXiv, august 2019. ArXiv:1904.05742 [cs, eess, stat]. Available at: http://arxiv.org/abs/1904.05742.

[10] Donovan, J. and Paris, B. Beware the Cheapfakes. *Slate*. june 2019. ISSN 1091-2339. Available at: https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html.

[11] Ellery, S. *Fake photos of Pope Francis in a puffer jacket go viral, highlighting the power and peril of AI*. March 2023. Available at: https://www.cbsnews.com/news/pope-francis-puffer-jacket-fake-photos-deepfake-power-peril-of-ai/.

[12] 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*. Elsevier BV. april 2023, vol. 9, no. 4, p. e15090. DOI: 10.1016/j.heliyon.2023.e15090. Available at: https://doi.org/10.1016/j.heliyon.2023.e15090.

[13] Godage, S. R., Lovasdaly, F., Venkatesh, S., Raja, K., Ramachandra, R. et al. Analyzing Human Observer Ability in Morphing Attack Detection -Where Do We Stand? *IEEE Transactions on Technology and Society*. Institute of Electrical and Electronics Engineers (IEEE). 2023, p. 1–1. DOI: 10.1109/tts.2022.3231450. Available at: https://doi.org/10.1109/tts.2022.3231450.

[14] Goodfellow, I. J., Pouget Abadie, J., Mirza, M., Xu, B., Warde Farley, D. et al. *Generative Adversarial Networks*. arXiv, june 2014. ArXiv:1406.2661 [cs, stat]. Available at: http://arxiv.org/abs/1406.2661.

[15] Groh, M., Epstein, Z., Firestone, C. and Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*. 2022, vol. 119, no. 1, p. e2110013119. DOI: 10.1073/pnas.2110013119. Available at: https://www.pnas.org/doi/abs/10.1073/pnas.2110013119.

[16] Groh, M., Epstein, Z., Obradovich, N., Cebrian, M. and Rahwan, I. Human detection of machine-manipulated media. *Communications of the ACM*. october 2021, vol. 64, no. 10, p. 40–47. DOI: 10.1145/3445972. ISSN 0001-0782, 1557-7317. Available at: https://dl.acm.org/doi/10.1145/3445972.

[17] Gundle, P. A. and Chavan, R. *Survey on Text to Speech Synthesis Models and Methods*. www.ijser.org, 2016. Available at: https://www.ijser.org/researchpaper/Survey-on-Text-to-Speech-Synthesis-Models-and-Methods.pdf.

[18] Hukkelås, H., Mester, R. and Lindseth, F. *DeepPrivacy: A Generative Adversarial Network for Face Anonymization*. arXiv, september 2019. ArXiv:1909.04538 [cs]. Available at: http://arxiv.org/abs/1909.04538.

[19] iProov. *Protect Against Deepfakes*. August 2022. Available at: https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection.

[20] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J. et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa Bianchi, N. et al., ed. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018, vol. 31. Available at: https://proceedings.neurips.cc/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf.

[21] Jurafsky, D. and Martin, J. H. Fine-Tuning and Masked Language Models. In: *Speech and Language Processing*. July 2023. Available at: https://web.stanford.edu/~jurafsky/slp3/11.pdf.

[22] Kadam, S. and Vaidya, V. Review and Analysis of Zero, One and Few Shot Learning Approaches. In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, April 2019, p. 100–112. DOI:

10.1007/978-3-030-16657-1_10. Available at: https://doi.org/10.1007/978-3-030-16657-1_10.

[23] Kamal, M. and Newman, W. J. Revenge Pornography: Mental Health Implications and Related Legislation. *Journal of the American Academy of Psychiatry and the Law Online*. Journal of the American Academy of Psychiatry and the Law Online. 2016, vol. 44, no. 3, p. 359–367. ISSN 1093-6793. Available at: https://jaapl.org/content/44/3/359.

[24] Kim, J., Kong, J. and Son, J. *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. arXiv, june 2021. ArXiv:2106.06103 [cs, eess]. Available at: http://arxiv.org/abs/2106.06103.

[25] Kong, J., Kim, J. and Bae, J. *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. arXiv, october 2020. ArXiv:2010.05646 [cs, eess]. Available at: http://arxiv.org/abs/2010.05646.

[26] Korshunov, P. and Marcel, S. *Deepfake detection: humans vs. machines*. arXiv, september 2020. ArXiv:2009.03155 [cs, eess]. Available at: http://arxiv.org/abs/2009.03155.

[27] LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. *Nature*. Springer Science and Business Media LLC. may 2015, vol. 521, no. 7553, p. 436–444. DOI: 10.1038/nature14539. Available at: https://doi.org/10.1038/nature14539.

[28] Li, Y. A., Zare, A. and Mesgarani, N. *StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion*. arXiv, july 2021. ArXiv:2107.10394 [cs, eess]. Available at: http://arxiv.org/abs/2107.10394.

[29] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S. *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics*. arXiv, march 2020. ArXiv:1909.12962 [cs, eess]. Available at: http://arxiv.org/abs/1909.12962.

[30] Lin, Y. Y., Chien, C.-M., Lin, J.-H., Lee, H.-y. and Lee, L.-s. *FragmentVC: Any-to-Any Voice Conversion by End-to-End Extracting and Fusing Fine-Grained Voice Fragments With Attention*. arXiv, may 2021. ArXiv:2010.14150 [cs, eess]. Available at: http://arxiv.org/abs/2010.14150.

[31] Loizou, P. C. Speech quality assessment. *Multimedia analysis, processing and communications*. Springer. 2011, p. 623–654.

[32] Martin, K. *New ID R&D research finds over 1 in 3 Americans confident they could detect a computer-generated voice pretending to be a human voice*. Aug 2020. Available at: https://www.idrnd.ai/voice-deepfake-survey/.

[33] Matyas, V., Krhovjak, J., Kumpost, M. and Cvrcek, D. Authorizing Card Payments with PINs. *Computer*. march 2008, vol. 41, p. 64 – 68. DOI: 10.1109/MC.2008.40.

[34] Mervosh, S. Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump. *The New York Times*. may 2019. ISSN 0362-4331. Available at: https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html.

[35] Mirsky, Y. and Lee, W. The Creation and Detection of Deepfakes. *ACM Computing Surveys*. Association for Computing Machinery (ACM). january 2021, vol. 54, no. 1, p. 1–41. DOI: 10.1145/3425780. Available at: https://doi.org/10.1145/3425780.

[36] Müller, N. M., Pizzi, K. and Williams, J. Human Perception of Audio Deepfakes. In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. October 2022, p. 85–91. DOI: 10.1145/3552466.3556531. ArXiv:2107.09667 [cs, eess]. Available at: http://arxiv.org/abs/2107.09667.

[37] Nasar, B. F., T, S. and Lason, E. R. Deepfake Detection in Media Files - Audios, Images and Videos. In: *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, December 2020. DOI: 10.1109/raics51191.2020.9332516. Available at: https://doi.org/10.1109/raics51191.2020.9332516.

[38] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O. et al. *WaveNet: A Generative Model for Raw Audio*. arXiv, september 2016. ArXiv:1609.03499 [cs]. Available at: http://arxiv.org/abs/1609.03499.

[39] Paris, B. and Donovan, J. *Deepfakes and Cheap Fakes*. September 2019. Available at: https://datasociety.net/library/deepfakes-and-cheap-fakes/.

[40] Patel, M., Gupta, A., Tanwar, S. and Obaidat, M. S. Trans-DF: A Transfer Learning-based end-to-end Deepfake Detector. In: *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, October 2020. DOI: 10.1109/iccca49541.2020.9250803. Available at: https://doi.org/10.1109/iccca49541.2020.9250803.

[41] Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S. et al. *DeepFaceLab: Integrated, flexible and extensible face-swapping framework*. arXiv, june 2021. ArXiv:2005.05535 [cs, eess]. Available at: http://arxiv.org/abs/2005.05535.

[42] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. et al. *FaceForensics++: Learning to Detect Manipulated Facial Images*. arXiv, august 2019. ArXiv:1901.08971 [cs]. Available at: http://arxiv.org/abs/1901.08971.

[43] Scott, D. *Deepfake Porn Nearly Ruined My Life*. February 2020. Available at: https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/.

[44] Szoldra, P. *Deep fakes are Russia's newest 'weapon of war'*. Available at: https://www.theruck.news/p/deep-fakes-are-russias-newest-weapon.

[45] Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M. et al. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, May 2021. DOI: 10.1145/3411764.3445699. Available at: https://doi.org/10.1145/3411764.3445699.

[46] Thombre, S., Bhattacharyya, P. and Jyothi, P. *Survey: Text-to-Speech Synthesis*. Indian Institute of Technology, Bombay, 2022. Available at: https://www.cfilt.iitb.ac.in/resources/surveys/2022/shyam-tts-survey-29jun22.pdf.

[47] 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: timreview.ca/article/1282.

Appendix A

Sets of game facts

Sets of questions used to content the conversation in the game *Two Truths One Lie*, translated from the Czech language. As a deepfake set, we had pre-prepared a fourth set, facts about Australia.

1. **set - Spain** (LIE: 1. The full name is the Kingdom of Spain.)
    1. The full name is Spanish Monarchy.
    2. Spain is the fourth largest country on the European continent.
    3. The painter Pablo Picasso was born in Spain.
2. **set - India** (LIE: 2. The capital is New Delhi.)
    1. India is the most populous country in the world.
    2. The capital is Mumbai.
    3. Hinduism is the most widely spread religion.
3. **set - Argentina** (LIE: 3. The national sport is pato.)
    1. The official language here is Spanish.
    2. Argentina is the eighth largest country in the world.
    3. The national sport is football.
4. **set - Australia** (LIE: 3. The largest city is Sydney.)
    1. It lies near the Pacific Ocean.
    2. The capital is Canberra.
    3. The largest city is Melbourne.
5. **set - Japan** (LIE: 1. The most common religion is Shinto.)
    1. The most widespread religion here is Buddhism.
    2. The head of Japan is the emperor.
    3. Japan is made up of almost 7,000 islands.

Appendix B

Survey questions

List of questions from the questionnaire and info provided by question groups in the same order.

1. **Respondent profile**
   - Phone number used for identification (Short Free Text)
   - Age (Short Free Text)
   - Gender (Gender)
   - The field in which you operate (Short Free Text)
2. **Usability**
   - How often do you use voice messages in your everyday communication? (5 Point Choice)
3. **Voice messages**
   - Did you find anything unnatural about the voice messages you received? (Yes/No)
   - If yes, what did you find strange? (Long free text)
4. **Deepfakes**
   Deepfakes are an artificial intelligence technology that can create fake videos, photos and voices that look like the real thing. These media are then almost indistinguishable from reality and very difficult to detect. Nowadays, they are increasingly being used to commit online fraud, spread hoaxes and steal identities.
   - Have you ever heard of deepfakes? (List (Radio))
   - If so, describe your experience. (Long free text)
   - How confident are you that you could detect a voice deepfake? (5 Point Choice)
5. **Real experiment**
   We come to the real goal of this experiment. Its aim was to highlight the quality of voice deepfakes and to test whether people are able to detect deepfakes. In the course of the experiment, I sent you five sets of facts to play with. However, one set was not narrated by me, it was a deepfake, it was narrated by someone else and only dressed up by my voice.

- If you go through the conversation again, which sets would you say can be deepfake? Please don't forget to add why. (Multiple choice with comments)

6. **Conclusion**
   The Deepfake recording was the fourth set, the facts about Australia.

   - Were you surprised at what level today's voice deepfakes can be? (Yes/No)
   - How confident are you that you would know deepfake after this experience? (5 Point Choice)