BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

APPLICABILITY OF DEEPFAKES IN THE FIELD OF CYBER SECURITY

POUŽITELNOST DEEPFAKES V OBLASTI KYBERNETICKÉ BEZPEČNOSTI

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. ANTON FIRC**
AUTOR PRÁCE

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

BRNO 2021

Brno University of Technology
Faculty of Information Technology

Department of Intelligent Systems (DITS) Academic year 2020/2021

Master's Thesis Specification

Student: **Firc Anton, Bc.**
Programme: Information Technology Field of study: Cybersecurity
Title: **Applicability of Deepfakes in the Field of Cyber Security**
Category: Security
Assignment:

1. Investigate the current state of deepfakes detection in cybersecurity.
2. Get familiar with the currently available techniques of creating various types of deepfakes and verify their technical feasibility, evaluate their availability and complexity.
3. Analyze the various areas that allow deepfakes to be carried out and discuss possible impacts. Define at least 3 attack vectors for deepfake usage and create corresponding syntetic media.
4. Perform experimental implementation of defined attacks using created depfakes and evaluate them. Evaluate quality of relevant detection mechanisms for performed attacks.
5. Discuss the security implications of these types of attacks and suggest possible methods of protection.

Recommended literature:

- Siarohin, Aliaksandr, Stéphane Lathuilire, S. Tulyakov, E. Ricci and N. Sebe. "First Order Motion Model for Image Animation." ArXiv abs/2003.00196 (2019): n. pag.
- Jones, Valencia A: Artificial Intelligence Enabled Deepfake Technology: The Emergence of a New Threat. Utica College, ProQuest Dissertations Publishing, 2020. 27962429.
- Andrei O. J. Kwok & Sharon G. M. Koh (2020) Deepfake: a social construction of technology perspective, Current Issues in Tourism, DOI: 10.1080/13683500.2020.1738357
- Aliaksandr Siarohin, Stéphane Lathuilire, Sergey Tulyakov, Elisa Ricci, Nicu Sebe: First Order Motion Model for Image Animation. https://arxiv.org/abs/2003.00196
- Bateman, Jon. Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios. Report. Carnegie Endowment for International Peace, 2020. I-Ii. Accessed October 30, 2020. doi:10.2307/resrep25783.1.

Requirements for the semestral defence:

- Items 1 to 3.

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

Supervisor: **Malinka Kamil, Mgr., Ph.D.**
Head of Department: Hanáček Petr, doc. Dr. Ing.
Beginning of work: November 1, 2020
Submission deadline: May 19, 2021
Approval date: November 11, 2020

Abstract

Deepfake technology is still on the rise, many techniques and tools for deepfake creation are being developed and publicly released. These techniques are being used for both illicit and legitimate purposes. The illicit usage yields the need for development and continuous improvement of detection tools and techniques as well as educating broad public about the dangers this technology presents. One of the unexplored areas of the illicit usage is using deepfakes to spoof voice authentication. There are mixed opinions on feasibility of deepfake powered attacks on voice biometrics systems providing the voice authentication, and minimal scientific evidence. The aim of this work is to research the current state of readiness of voice biometrics systems to face deepfakes. The executed experiments show that the voice biometrics systems are vulnerable to deepfake powered attacks. As almost all of the publicly available models and tools are tailored to synthesize the English language, one might think that using a different language might mitigate the mentioned vulnerabilities, but as shown in this work, synthesizing speech in any language is not that complicated. Finally measures to mitigate the threats posed by deepfakes are proposed, like using text-dependent verification because it proved to be more resilient against deepfakes.

Abstrakt

Deepfake technológia je v poslednej dobe na vzostupe. Vzniká mnoho techník a nástrojov pre tvorbu deepfake médií a začínajú sa používať ako pre nezákonné tak aj pre prospešné činnosti. Nezákonné použitie vedie k výskumu techník pre detekciu deepfake médií a ich neustálemu zlepšovaniu, takisto ako k potrebe vzdelávať širokú verejnosť o nástrahách, ktoré táto technológia prináša. Jedna z málo preskúmaných oblastí škodlivého použitia je používanie deepfake pre oklamanie systémov hlasovej autentifikácie. Názory spoločnosti na vykonateľnosť takýchto útokov sa líšia, no existuje len málo vedeckých dôkazov. Cieľom tejto práce je preskúmať aktuálnu pripravenosť systémov hlasovej biometrie čeliť deepfake nahrávkam. Vykonané experimenty ukazujú, že systémy hlasovej biometrie sú zraniteľné pomocou deepfake nahrávok. Napriek tomu, že skoro všetky verejne dostupné nástroje a modely sú určené pre syntézu anglického jazyka, v tejto práci ukazujem, že syntéza hlasu v akomkoľvek jazyku nie je veľmi náročná. Nakoniec navrhujem riešenie pre zníženie rizika ktoré deepfake nahrávky predstavujú pre systémy hlasovej biometrie, a to používať overenie hlasu závislé na texte, nakoľko som ukázal, že je odolnejšie proti deepfake nahrávkam.

Keywords

deepfake, cyber security, voice biometrics

Kľúčové slová

deepfake, kybernetická bezpečnosť, hlasová biometria

Reference

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.

Rozšírený abstrakt

Vzostup deepfake technológie pomaly zapríčiňuje aj zvyšovanie ľudského povedomia o tejto technológii. Deepfake média začínajú byť prezentované v televíznom vysielaní a sú následne spomínané aj na sociálnych sieťach. Toto je však len vrchol ľadovca, nakoľko deepfake média majú veľmi široké spektrum oblastí pre uplatnenie, takisto ako veľké množstvo nástrojov pre ich tvorbu.

Nástroje pre tvorbu deepfake médií pokrývajú všetky oblasti od úpravy a generovania videa až po generovanie textu. Takisto sa líši použiteľnosť takýchto nástrojov a znalosti potrebné k ich používaniu. Od najjednoduchších vo forme mobilných aplikácií až po experimentálne implementácie vedeckých publikácií. S vyššou požadovanou výslednou kvalitou úmerne rastú aj nároky na vedmomosti a skúsenosti.

Ľahký prístup k tvorbe médií, ktoré môžu spôsobiť škodu jednotlivcovi alebo skupinám vyžaduje nasadenie techník pre detekciu syntetických médií. K detekcii je možné pristúpiť dvomi spôsobmi: ľudská detekcia a strojová detekcia. Ľudia môžu detekovať syntetické médiá pomocou podrobnejšej analýzy sledovaného videa, obrázku alebo počúvaného hlasu. Artefakty a nekonzistentnosť často prezradí, že sa jedná o syntetické médiá. Vhodnými ukazovateľmi sú napríklad žmurkanie, odrazy v okuliaroch alebo vzhľad vlasov a brady. Pri reči sa zase často prejavuje nezvyčajné, až neľudské tempo a chyby vo výslovnosti. Strojová detekcia znovu využíva artefakty, prípadne je možné monitorovať správanie neurónových sietí a tak detekovať deepfake médiá.

Deepfake médiá sú najčastejšie spájané so škodlivým a často nelegálnym využitiím. V tejto oblasti majú veľa využití pri diskreditovaní osobností, šírení falošných správ, prípadne vydieraní a podvodoch. Takáto zlá povesť potom nedáva veľa priestoru pre využitie deepfake médií tak, aby boli prínosné ľudstvu. Deepfake médiá je možné použiť v zábavnom priemysle, školstve alebo dokonca zdravotníctve.

Na základe zistených znalostí o tvorbe a využití deepfake médií boli navrhnuté tri vektory útokov pre experimentálne vykonanie. Navrhnuté vektory útokov cielia na systémy hlasovej biometrie a overenie rečníka a ich odolnosť a bezpečnosť proti deepfake nahrávkam.

Prvý vektor útoku skúma využiteľnosť syntetického hlasu pre prelomenie zabezpečenia poskytovaného systémami hlasovej biometrie. Výsledky experimentov ukazujú, že systémy hlasovej biometrie bez implementácie detekcie živosti sú ľahko napadnuteľné. Experimenty s rozdielmi medzi overením závislým a nezávislým na texte ukazujú, že overenie závislé na texte poskytuje väčšiu mieru ochrany pred deepfake nahrávkami.

Druhý vektor útoku nadväzuje na prvý a miesto syntézy anglického jazyka preskúmava možnosti vytvorenia modelu syntézy reči pre český jazyk. Nakoľko takmer všetky publikácie, nástroje a datasety pracujú exkluzívne s anglickým jazykom, je vhodné preskúmať, či tento fakt nejako zvyšuje bezpečnosť systémov hlasovej biometrie v inom jazyku. Získané výsledky ukazujú, že vytvorenie takéhoto modelu je technicky uskutočniteňé aj bez rozsiahlejších znalostí z oblasti spracovania a syntézy reči a tvorba deepfake nahrávok nie je limitovaná zvoleným jazykom.

Posledný vektor útoku znovu nadväzuje na predchádzajúce vektory útokov. Nakoľko v drvivej väčšine systém hlasovej biometrie poskytuje výsledok operátorovi call centra, prípadne inej osobe s ktorou je potreba komunikovať, posledný experiment zisťuje či sú deepfake nahrávky schopné oklamať aj človeka. Zo získaných výsledkov je jasné, že ľudia majú problém rozlišovať medzi ozajstnou a syntetickou rečou. Čím starší ľudia sú, tým je táto rozlišovacia schopnosť horšia. Pridanie ľudského faktoru do procesu overenia teda nezvyšuje jeho bezpečnosť.

Táto práca diskutuje o nástojoch pre tvorbu a detekciu deepfake médií. Diskutuje

o možnostiach využitia týchto médií pre škodlivé ale aj pre ľudstvo prínosné účely. Predstavené a vykonané experimenty ukazujú, že deepfake nahrávky predstavujú hrozbu pre systémy hlasovej biometrie a že táto hrozba je aktuálna aj v našich končinách.

Hlavné prínosy tejto práce teda môžu byť sumarizované nasledovne:

- Táto práca dokazuje, že deepfake média predstavujú vážnu hrozbu pre systémy hlasovej biometrie a je teda potrebné zaviesť riešenia ktoré túto hrozbu zmiernia.
- V tejto práci ukazujem, že overenie závislé na texte je robustnejšie proti deepfake nahrávkam než overenie na texte nezávislé.
- Ukazujem, že natrénovanie modelu pre syntézu reči v českom jazyku je vykonateľné aj bez výraznejších znalostí v oblasti spracovania a syntézi reči, z čoho vyplýva, že jazyky iné ako anglický jazyk nepredstavujú vyššiu mieru bezpečia aj keď pre tieto jazyky existuje len malé množstvo predpripravených modelov a datasetov.
- Ukazujem, že ľudia môžu byť deepfake nahrávkami oklamaní takisto ľahko ako systémy hlasovej biometrie.
- Bol publikovaný deepfake dataset obsahujúci anglickú a českú reč[1].

[1] https://drive.google.com/drive/u/3/folders/1vlR-TA7gjKzjYylxzRnA_HzZEyWiLeOk

Applicability of Deepfakes in the Field of Cyber Security

Declaration

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

.....
Anton Firc
May 9, 2021

Acknowledgements

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

I owe a big THANK YOU to my parents, for their endless love, support and advices during my whole life and study.

I want to thank my girlfriend for her support and patience during my study, and finally my friends for their willingness to take part in all of the experiments I came up with.

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Contents

Chapter 1

Introduction

With the emergence of deepfakes, public awareness is finally starting to rise. Deepfakes are starting to be presented in the television broadcast and the word then spreads through social medias. But this is just the tip of an iceberg, deepfakes have many usages, both malicious and beneficial to the society. Up to date, there are many tools used to create deepfakes, and the number of such tools still rises.

The tools used to create deepfakes cover almost any area you can imagine, from manipulating and synthesizing video, to text synthesis. Usability and the knowledge required to use these tools greatly varies. From the most basic tools in form of mobile applications, to the experimental implementations of scientific publications. The higher the final quality, the more knowledge and experience is needed.

The usage of deepfakes and fairly easy access to their creation calls for adequate measures to detect this kind of media. There are two major approaches on how to detect deepfakes: human detection and machine detection. People can spot deepfakes by examining the video, image or speech. Artifacts and inconsistencies disclose deepfakes most of the times. Suitable signs to look for are for example blinking, reflection in glasses or the look of hair. The synthetic speech most of the times has an unnatural tempo or spelling errors. The machine detection again uses the artifact detection, or neural network behavior can be monitored in order to detect deepfake medias.

The most of the attention deepfakes receive is because of their malicious and often illegal usages such as defaming individuals, spreading fake news or extortion and fraud. However, deepfakes can also be used for beneficial purposes, such as in education, entertainment or even healthcare.

The research of areas allowing malicious usage of deepfakes suggested, that the voice biometrics systems might be spoofed by deepfake medias. The voice biometrics systems are most commonly used in call centers of banking institutions or telephony providers that allow their customers to execute actions without actual physical presence. Being able to spoof such a system would allow the attacker to execute actions on behalf of the victim, for example to execute fraudulent wire transfer.

As this area of deepfakes usage remains mostly unexplored, this work aims to research the readiness of currently available voice biometrics systems to face deepfake speech. This process involves multiple factors: the voice biometrics system, call centre operator and the language used to communicate with the operator. To successfully execute such an attack, the deepfake speech has to be able to spoof the voice biometrics system and to converse with the call centre operator in specific language without raising any suspicion that would make the operator end the call. Three attack vectors are presented in this work, that examine

all of the steps needed to perform the mentioned type of attack. The first proposed attack vector examines the ability of deepfakes to spoof voice biometrics systems, the second one examines the speech synthesis in a selected language and the last one examines the ability of deepfakes to spoof humans. Each attack vector contains experiments that examine its technical feasibility and overall usability. During the design of each experiment, research questions were asked, and all of them also answered during the execution of proposed experiments.

The main contributions of this work might be summarized as follows:

- This work proves that deepfakes present a serious threat to the voice authentication systems. That the voice biometrics systems might be easily spoofed, and measures to prevent these kinds of attacks have to be implemented.
- I present that text-dependent verification is more robust than text-independent verification when dealing with deepfakes, by creating custom dataset to compare the security provided by both of the verification types.
- I present that training a text-to-speech model for Czech language is feasible even without extensive knowledge of speech synthesis, which refutes the hypothesis that language different to English is safer, as the majority of text-to-speech tools and datasets is suited for English language exclusively.
- I present that people might be fooled by deepfakes as easily as voice biometrics systems.
- A deepfake dataset containing English and Czech speech was published.

Various tools for creating various types of deepfakes as well as more detailed definition of what a deepfake is are discussed in Chapter 2. Chapter 3 discusses available detection techniques for human as well as machine detection. Both of the malicious and legitimate usages of deepfakes are further discussed in Chapter 4. Chapter 5 proposes attack vectors that incorporate deepfake speech and experiments to evaluate the threats posed by these attack vectors and their technical feasibility. The execution details and results of each proposed experiment are discussed in Chapter 6. Finally Chapter 7 summarizes the findings of this research and proposes techniques to mitigate threats posed by attacks incorporating deepfake media.

Chapter 2

Creating a deepfake

Deepfake is quite a new technology, that is rapidly advancing, it enables user with an ability to generate synthetic media, both audio and visual. Deepfakes are created using AI method called *deep learning* which relies on deep neural networks [3]. These networks need to be trained over a set of training data before they are capable of generating data unseen during the training [32]. There are many opinions circulating through internet and medias on how difficult it is to create a deepfake. This chapter discusses what a deepfake is, reviews the newest and the most interesting available tools and techniques for creating audio and video deepfakes and finally concludes the entry level for deepfake creation.

Most of the discussed tools was tested by me, to evaluate the overall usability and technical feasibility. Getting hands on the tools was necessary in order to gain general knowledge on how difficult it is to create a deepfake.

2.1 What is deepfake

Deepfakes are AI-generated medias depicting made-up events. A slang term „deepfake“ has no agreed-upon technical definition, the word deepfake is a combination of the words ‘deep learning’ and ‘fake’ and primarily relates to content generated by an artificial neural network, a branch of machine learning [3, 21, 32].

As stated in [3], deepfakes are created using AI method called deep learning which relies on deep neural networks. Those networks need to be trained at first. They take training data as input, extract characteristics of targeted person’s voice or look and then create mathematical patterns based on these characteristics. Using patterns obtained in training, these networks can then generate new, synthetic representation of individual’s look or voice.

Best known type of deepfake is face-swap, this technique transposes one person’s face onto others person’s head. Common usage is found within pornographic materials, transposing faces of celebrities onto faces of adult movie actors. Other popular type of deepfake is voice cloning, it copies unique vocal patterns of individual’s voice to recreate or alter individual’s speech [3].

While face-swapping and voice cloning mimic individuals, there are also techniques for creation of entirely fictitious people, sceneries or objects. Those images of nonexistent people, animals, flowers or landscapes are almost unrecognizable from the real ones, artificial intelligence can also produce synthetic text that emulates human-authored texts. In general, deepfakes are a subset of synthetic media [3].

2.2 Technical background

Most of the deepfakes are created using variations or combinations of generative networks and encoder decoder networks [32]. This section describes these networks, how they are created and trained.

2.2.1 Neural networks

All of the facts stated in this section were retrieved from [32]. Neural networks are non-linear models for predicting or generating content based on an input. They consist of layers of neurons, where each layer is connected sequentially via synapses. All synapses have associated weights that collectively define knowledge learnt by the model. Executing network on a n-dimensional input x, is a process known as forward-propagation, where x is propagated through each of the layers and activation function is used to summarize a neuron's output (e.g. Sigmoid or ReLU function).

To be precise, let's define a neural network as stated by Y. Mirsky and W. Lee [32]. Let M be a neural network with total count of layers L. Let $l^{(i)}$ denote the i-th layer of network M, and let $|l^{(i)}|$ denote the neuron count in $l^{(i)}$. The weights connecting $l^{(i)}$ with $l^{(i+1)}$ are denoted as the $|l^{(i)}| - by - |l^{(i+1)}|$ matrix $W^{(i)}$ and $|l^{(i+1)}|$ dimensional bias vector $\vec{b}^{(i)}$. Finally, let θ denote the collection of all parameters as the tuple $\theta = (W, b)$, where W and b are the weights of each layer respectively. Let $a^{(i+1)}$ denote the output (activation) of $|l^{(i)}|$ computed as a result of $f(W^{(i)} * \vec{a}^{(i)} + \vec{b}^{(i)})$ where f is often the Sigmoid or ReLU function. To execute the network on a n-dimensional input x, the forward-propagation process is performed where x is used to activate $l^{(1)}$ which activates $l^{(2)}$ and so on until the activation of final layer $l^{(L)}$ produces m-dimensional output y.

For further use we consider network M a black box and denote its behavior as $M(x) = y$. Training of M in supervised setting is done using a dataset of paired samples in form (x_i, y_i) and loss function $\mathcal{L}$. The loss function generates a signal that is back propagated through M in order to find the errors of each weight. This process is done by an optimization algorithm, such as gradient descent. The function measures an error between the expected output y and the predicted output y'. The result of a training is the function $M(x_i) \approx y_i$ which can be used to make predictions on previously unseen data [32].

Some deepfake networks use a technique called *one-shot* or *few-shot learning*. This enables a pre-trained network to adapt to a new dataset that is similar to the original one. There are two common approaches to use these techniques. One is to pass an information that input belongs to the new dataset to the inner layers of M during the feed-forward process, second one is to perform additional training steps over few samples from the new dataset [32].

2.2.2 Loss function

As Y. Mirsky and W. Lee in [32] state, optimization algorithm, such as gradient descent, needs the loss function to be differentiable in order to update the weights. The loss function is chosen depending on the learning objective. For example, when training a $n-$class classifier, the output of M would be a probability vector $y \in \mathcal{R}^n$. The forward-propagation process will be then used to train M and obtain $y' = M(x)$ that is then used to compute the cross-entropy loss $\mathcal{L}_{CE}$ by comparing the predicted value y with the ground truth y' and afterwards perform a *back-propagation* to update model weights accordingly [32].

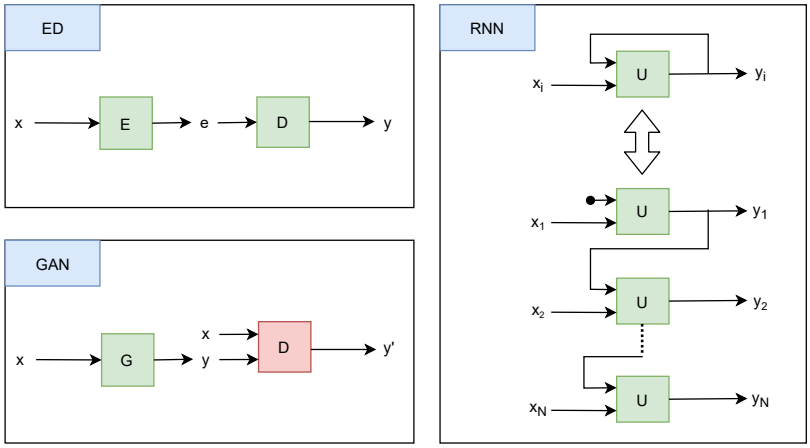

Figure 2.1: Architecture of basic neural networks used to create deepfakes. Green color denotes generator networks and the red color denotes discriminator networks.

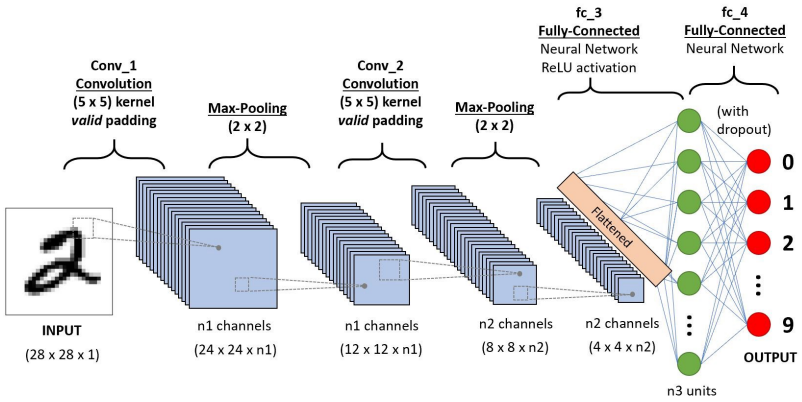

Figure 2.2: Architecture of convolutional neural network. Retrieved from [43].

2.2.3 Generative Adversarial Networks

Combinations or variations of neural networks are used to create deepfakes, this section describes these networks. The architectures of the discussed networks are displayed in Figures 2.1 and 2.2.

Encoder-Decoder network or shortly ED consists of at least two networks, one poses as an encoder and the second one as a decoder [32]. This type of network aims to handle the mapping between highly structured input and output [6]. The encoder is responsible of reading the input data into a continuous space representation, the decoder network then generates output conditioned on the continuous space representation of the input [6]. The function used to calculate the continuous space representation and then output based on this representation is chosen according to the type of the output, for example Boltzmann machine will be used for images segmentation, or a recurrent neural network for natural language modeling [6].

Convolutional neural network or shortly CNN has been designed to specifically work with two-dimensional images [6]. The network consists of multiple convolutional layers and a few fully-connected layers [6]. The convolutional layers form filters that are shifted over the input to form an abstract feature map as the output [32].

Generative adversarial network or shortly GAN is according to Y. Mirsky and W. Lee [32] built using two neural networks that work against each other. The *generator* network learns patters from the training data in order to produce an output that fools the *discriminator* network that learns to distinguish between real and fake samples. The training process ends when the discriminator network is no longer able to tell the difference between generated and real samples. After the training is over the discriminator network is removed, and the generator network is used to generate content [32].

Recurrent neural network or shortly RNN is specialized to handle sequential and variable-length inputs [6, 32]. The network remembers it's hidden internal state between processing part of the input x_i and can use it to process the following part of the input x_{i+1} [32]. The RNNs are mostly used to handle audio and sometimes even video processing [32].

2.2.4 Generalization

According to Y. Mirsky and W. Lee [32], a deepfake network may be designed and trained to work with only a specific set of data, both target and source. Sometimes it might be hard to create a model that is independent of the identity of data, because of correlations learned by the model between data. From the point of generalization we then define three categories as stated in [32]:

- one-to-one: model capable of using a specific entity to drive specific entity
- many-to-one: model capable of using any entity to drive specific entity
- many-to-many: model capable of using any entity to drive any entity

2.2.5 Summary

Technologies used to create deepfakes rapidly emerged in the last few years. As stated by Jon Bateman [3], the deepfake creation methods and tool rely on use of the deep learning. User-friendly software and cheap cloud computing enables technically savvy individuals to create their own synthetic media, sometimes even using a web interface or mobile application. The generated medias vary in quality, the face-swap videos can be unbelievably lifelike, yet most of the times closer observation reveals artifacts, lack of detail or blurred edges. The resulting quality depends vastly on these three factors as stated in [3]:

- choice of algorithm
- amount, quality and variance of training data
- computing power and processing time

2.3 Video deepfakes

Video deepfakes are the most created and discussed type of deepfakes when browsing social media, or searching for deepfake-related topics. Their usage ranges from entertainment, like swapping your face with a celebrity in your favorite movie, to committing crime like using tampered video evidence to extort public figures.

This section discusses some of the available tools for video deepfake creation. For each tool technical background, overall usability and difficulty to use it are discussed. Finally the section contains a summary of the findings.

The discussed tools have been selected because of their availability to broad public as open-source code or free applications, their popularity among the deepfake content creators (DeepFaceLab) and the usage of previously unseen technology and capabilities (First order motion). This section does not provide a complete list of available video deepfake creation tools, rather selects the one I found to be the most interesting and available.

2.3.1 First order motion

The framework proposed by Aliaksandr Siarohin [55] brings a new mean of generating video sequences, where no annotations or prior knowledge about the specific object to animate is needed. Once the training is done over a set of videos from the same category (e.g. faces, human body movement), the method can be applied to any object falling into the same class. This behaviour is achieved by dividing the appearance and motion information by self-supervised formulation. Finally the generator network combines the appearance extracted from source and the motion from the driving video into the resulting video.

The implementation provided by the framework authors is publicly available on GitHub as a CLI application[1]. There are various pretrained models provided for different classes of objects such as faces, body movement or GIFs. The video generation requires previously trained (or pretrained) model, configuration file, driving video and source image to produce source image moving the same as driving video. The usage is fairly simple, although there are some downsides that prevent this framework to be used yet in any other area than research. Both the driving video and source image need to be cropped correctly, which means that the result contains only the targeted object, a head in our case. Example of output using the pretrained dataset of celebrity heads is shown in Figure 2.3. Another noticeable problem is resolution, the resulting video has a resolution of $256x256$px what is considered a little nowadays.

Figure 2.3: Example result of face swap performed by the First order motion tool, using the pretrained celebrity model. Driving video on the left, source image in the middle and result on the right. The frames captured from driving video and result were captured at approximately the same moment. Complete video is available on https://youtu.be/cXFL-P0D4So.

[1] https://github.com/AliaksandrSiarohin/first-order-model

Evaluation

The examples published with the tool are really impressive, however during my experiments I was unable to reproduce that quality. The biggest problem observed with face animations, apart from the right cropping is head movement. If the head remains still and only facial features like lips, eyes or eyebrows move the result is quite convincing. Mostly the rotations of the head caused weird cropping of the face on sides, and movement up or down sometimes caused weird stretching.

This framework presents a very interesting piece of work that has a lots of potential for the future improvement. The ability of generating any type of object is previously unseen and might cause quite a commotion when improved in the future. Unfortunately this framework does not yet provide the functionality wanted for the scope of our research.

2.3.2 DeepFaceLab

DeepFaceLab is an open-source framework for creating face-swap deepfake videos. The framework provides imperative easy-to-use pipeline for all kind of users ranging from beginners with no understanding of deep learning framework to experienced user looking for face-swapping features in their own pipeline [41]. Apart from the framework, DeepFaceLab brings quite a large community of deepfake-interested users. The online forum MrDeepFakes is one of the most famous forums, where the creations, models and general knowledge is being shared [41].

The creation of a face-swap deepfake in outstanding quality is a really demanding process requiring a lot of knowledge and determination, most of the presented results require a lot more details than the ones provided in their research papers, step-by-step annotations of face masks or even 3D models [41]. This framework thus aims to simplify the pipeline of deepfake creation in the most possible way. Another aspect of why to enable majority of people with ability of creating deepfakes is raising awareness, user that has an experience with deepfake creation is more likely to identify such a media circulating social medias or media in general [41].

Evaluation

Unlike the other open-source implementations discussed in this research, DeepFaceLab is very user-oriented [41]. All of the pipeline processes are packaged into CLI scrips, which when run correctly provide simple, yet realistic result. The pipeline follows through all the steps like data preparation, training and final media creation. The only knowledge needed are the parameters to use for the model training, but with many tutorials and even YouTube videos talking about how to work with this framework it is just a matter of hours to get into it.

There are even many YouTube channel publishing face-swap videos created exactly using this framework. The quality of these medias is very realistic, which just support the statement that DeepFaceLab is an simple framework that enables users to create quality deepfakes.

2.3.3 Reface App

Reface App is a mobile app for both iOS and Android smartphones that uses deepfake technology to let users create face swap videos and images [26]. Unfortunately the only

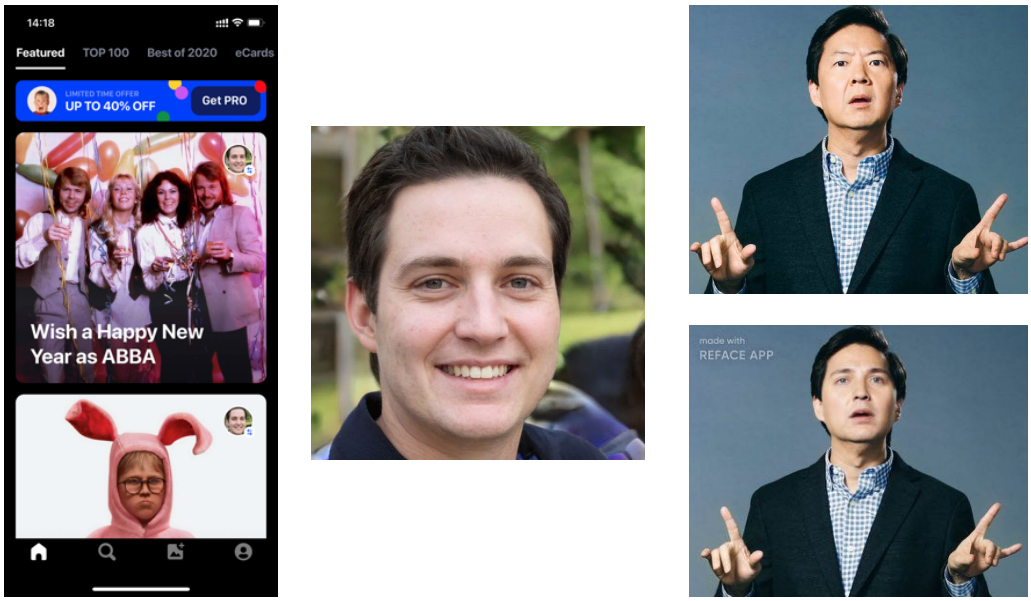

Figure 2.4: The main screen of the Reface App on the left, source image in the middle, and the target image with the result in the top right respectively bottom right corner. *The source image has been generated using the StyleGAN [23] implementation available at: (https://thispersondoesnotexist.com)*

publicly available information on the technology used is that the app is powered by a generative adversarial networks [26].

The app features simple and user friendly UI. The only requirement is having access to a mobile device with supported operation system. The source images can be either selfies taken from within the app, or any picture uploaded from the phone's gallery. User can choose from multiple source images of individuals's faces that he or she has taken or uploaded to be transposed into the selected video, GIF or an image. The destination media library is managed by the application developer and community, thus contains only limited, but quite large set of media. In addition, all of the content is moderated, which means that no illicit content should be created while using this app. The downside of this approach is that there is not as much freedom as we might desire to use this app for experiments, but on the other hand it is a large benefit that keeps the usage of the app purely in the entertainment level and does not create ethical concerns.

Evaluation

Overall, the app is really easy to use. Everyone can create a deepfake in a matter of seconds with just a few clicks. All the computing is done on the server, so the demands on the hardware (smartphone) are minimal. The common workflow includes, uploading or taking a source image, selecting destination media and clicking on the start button. After a few seconds the result shows up that can be then shared or downloaded. Quality of the resulting media is good, but the authenticity lags behind, this means that the result has minimal to none glitches but without knowing the appearance of the person on the source image, you could have a really hard time guessing the identity of person in the deepfake as shown in Figure 2.4.

In conclusion, this application provides a really simple platform for users of all age and technical knowledge to create deepfakes, but the usage is really limited for entertainment purposes.

2.4 Voice deepfakes

There are two main methods for speech synthesis: text-to-speech (TTS) synthesis and voice conversion (VC) [57]. The main difference between these methods is the input data. Text-to-speech synthesis, as the name suggests consumes written text as input and produces synthesized speech that sounds like an particular individual [35]. The voice conversion on the other hand consumes a source voice saying desired phrase and a target voice, and outputs the source phrase spoken by the target voice [35].

This section discusses the available tools for voice deepfake creation from both of the categories, TTS and VC tools. For each tool the technical background, if available, the overall usability and difficulty to use it are discussed. Finally this section contains a summary of the findings.

2.4.1 Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

This section describes a framework for zero-shot voice conversion that requires as less as 5 seconds of reference speech [7]. All information contained in this section has been retrieved from the *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis* [17].

The development of the framework aimed to build a text-to-speech (TTS) system which can generate natural speech for variety of speakers in a data efficient manner. A zero-shot learning setting is used, which enabled the system to synthesize speech in target's voice without updating any model parameters. The main approach is to decouple speaker modeling from speech synthesis. This is achieved by independently training a speaker-discriminative embedding network that captures speaker characteristics, and training a TTS model on a smaller dataset conditioned on the representation learned by the first network. This decoupling enables the networks to train on different data, which reduces the need to obtain high quality training data.

The framework consists of three independently trained neural networks, as illustrated in Figure 2.5: *speaker encoder* that computes dimensional vector from a speech signal, *synthesizer* which predicts a Mel spectrogram conditioned on the speaker embedding vector and WaveNet *vocoder* which converts the Mel spectrogram into time domain waveforms.

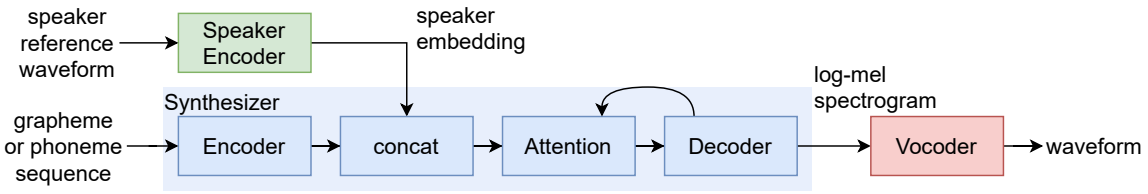

Figure 2.5: Framework architecture. Each of the three colors represents independently trained component. Image retrieved from [17].

Speaker encoder is used to condition the synthesis network. Usage of representation which captures the characteristics of different speakers and the ability to identify these characteristics in short time, independent of phonetic content or background noise is crucial for good generalization. These requirements were satisfied using a speaker-discriminative model trained on text-independent speaker verification task. Even through the encoder network is not directly optimized to learn a representation which captures speaker characteristics relevant to synthesis, training on a speaker verification task provides embeddings which are directly usable to condition the synthesis on speaker identity.

Synthesizer uses attention Tacotron 2[2] architecture to support multiple speakers. An embedding vector or target speaker is concatenated with the synthesizer encoder output at each time step. The training of synthesizer is done using pairs of audio and text transcripts. The text is mapped to a sequence of phonemes which leads to faster convergence and better pronunciation of rare words. The network is trained in a transfer learning configuration, using a pretrained speaker encoder model to extract speaker embeddings. This means the speaker reference signal is the same as the target speech during training.

Vocoder is used to invert synthesized Mel spectrogram provided by the synthesis network into time-domain waveforms. The network is not directly conditioned on the output of the encoder as the Mel spectrogram contains all of the relevant data needed to high quality synthesis of a variety of voices. This allows the multispeaker vocoder to be constructed by training on data from many speakers.

In conclusion, this framework presents a neural-network based system for multispeaker TTS synthesis. The system combines three independently trained parts, a speaker encoder, TTS synthesis network and neural vocoder. Using the learnt knowledge of speaker encoder, the synthesizer is capable of generating high quality speech for speakers seen and also unseen during training. Although, there are some limitations given by the additional difficulty of generating speech for a variety of speakers given just a small amount of data. This results in lower naturalness of generated speech in comparison to single speaker systems. Another limitation is system's inability to transfer accents or inability to completely isolate the speaker voice from the prosody of the reference audio.

2.4.2 Realtime voice cloning

This tool[3] is an implementation of the SV2TTS framework described in Section 2.4.1, thus an text-to-speech synthesis tool [7]. During the completion of this work a new version of this tool including a reworked synthesizer was published. If not stated otherwise, all of the mentions reference version with commit hash: `5425557efe30863267f805851f918124191e0be0`.

The RTVC project was developed to be publicly available as an open-source project [7]. The tool implemented within mentioned project features both graphic user interface (toolbox) and command line interface. The graphic interface does not provide any additional functionality do command line tool, just simplifies the workflow process.

The toolbox lets user to choose utterance audio file to extract embeddings, choose models for encoder, synthesizer and vocoder, write text to be synthesizes using extracted embeddings and buttons to synthesize and vocode given text. The screenshot of the toolbox GUI is shown in Figure 2.6.

The workflow begins with selecting an embedding audio file. Once the file is loaded, the toolbox computes its UMAP projections and draws a Mel spectrogram of loaded audio

[2] https://pytorch.org/hub/nvidia_deeplearningexamples_tacotron2/
[3] https://github.com/CorentinJ/Real-Time-Voice-Cloning

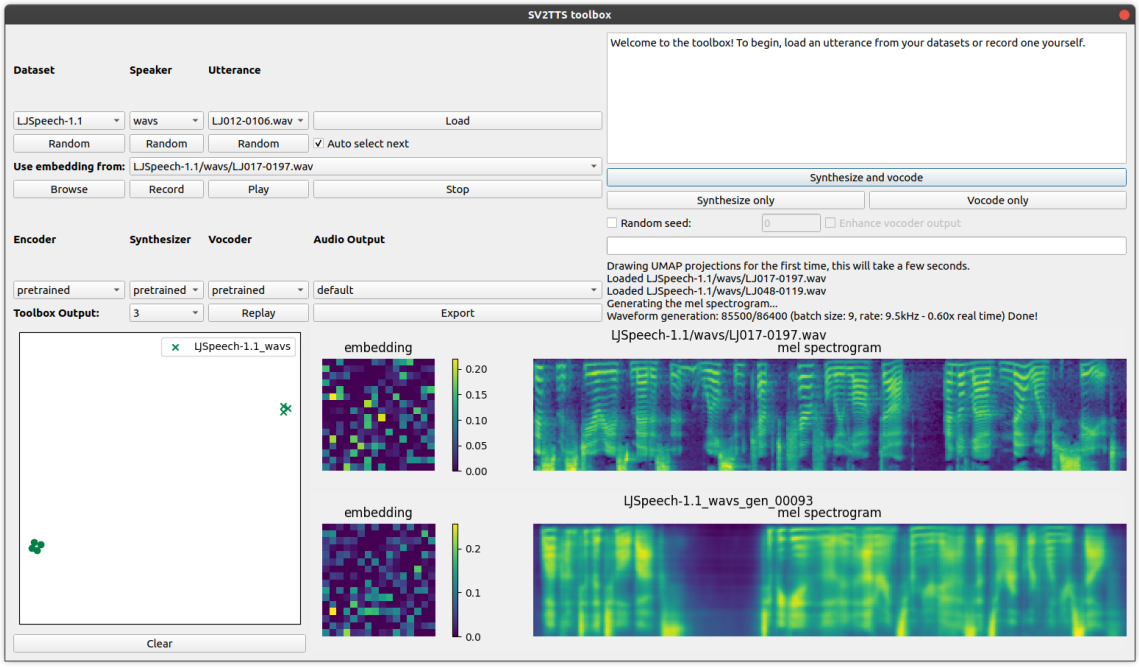

Figure 2.6: Toolbox user interface. Embedding audio file selection and model selection in top left corner. UMAP projection in bottom left corner. Text input and control buttons in top right corner. Extracted (up) and predicted (down) Mel spectrogram in right bottom corner.

file [7]. The Mel spectrogram is drawn just for reference and does not serve any computational purpose [7]. As a next step user enters desired text to be synthesized. Tuning the prosody of synthesized text is done by entering line breaks between parts of text that should be synthesized individually [7]. After the generation process finishes, Mel spectrogram and UMAP projection of resulting audio is drawn and the audio is played.

The command line interface provides similar functionality as the toolbox. It lets user select an embedding audio file, write text to be synthesized and finally synthesize speech.

Apart from speech synthesis, this tool provides a variety of scripts to prepare training data for each part of the framework and also to train each of the parts. These scripts are command-line only and written in python.

In order to run all the scripts I created an environment using conda[4]. The setup consisted of installation of packages from requirements file and PyTorch version 1.2.0. After a brief introduction into the source code I was able to run, and use the command line interface tool to synthesize speech.

The GitHub repository provides pretrained models for all three parts of the framework. The information about these models are provided in Table 2.1. With the pretrained models the usage breaks down to making a python script to run and then synthesizing speech, which is something almost anyone familiar with the command line interface should be able to do.

[4]https://docs.conda.io/en/latest/

	steps	time (days)	GPUs	batch size
Encoder	1560000	20	1	64
Synthesizer	278000	7	4	144
Vocoder	428000	4	1	100

Table 2.1: Information about pretrained models provided within the tool repository. Table contains information about total number of trained steps, the time spent training that amount of steps, number of GPUs (NVIDIA GeForce GTX 1080 Ti) and the batch size used to train each model [8].

Evaluation

The overall synthesized audio quality might be marked as average. The speech is mostly recognizable, the voice is somehow similar to the target speaker‘s voice. Glitches or any imperfections occur mostly as long periods of silence with popping in background and muffled words at the end of the silence as the speech resumes. This occurs mostly in short sentences, 5 to 10 words, the longer ones are mostly fine.

To achieve better results a synthesizer model fine-tuning is recommended by the community. Fine-tuning is a process of training the model for given additional amount of iterations, using only utterances of the target speaker and is further discussed in Section 6.3.

2.4.3 Descript Overdub

As the title on Descript's website states, it is an audio/video editor providing AI features to edit and enhance your recordings [9]. The feature relevant to this research is called Overdub. It is a feature that allows the synthesis of audio in your own voice from typed text, thus an TTS tool. Overdub is capable of generating standalone audio or contextual audio that blends into existing audio [9].

To start using Overdub feature, you have to record training audio. Descript provides a script that is required to be read as mentioned training audio[5]. The minimal required length of the training audio to create an Overdub voice is 10 minutes, but the more the merrier [9]. 30 minutes of recorded audio should provide production-ready audio quality [9]. Once the recording is submitted the model is trained, that takes between 12-24 hours.

Evaluation

To begin with, I decided to record the first part of 10 minutes from the transcript, as it should satisfy the minimal needs for the training and set a baseline for resulting audio quality. Reading the first part took me about 15 minutes, as I am not a native speaker, and the recording also contained some mistakes. The audio quality was average, I used a high-end headset and recorded all the audio in the living room. Some background noise could be heard, but the speech was audible.

After the submission of training audio it took around 12 hours until the voice was ready. The first attempts showed vast similarity to my voice. When listening to generated recordings, I was able to certainly tell that the voice sounds just like mine, even my family members confirmed that.

[5] https://coda.io/@overdub/overdub-scripts

This application provides really visually attractive and also user-friendly UI as shown in Figure 2.7. The application is really simple to use and provides really decent quality of generated speech when provided enough training data. The training text itself is a bit harder to read for non-native English speaking people, but even with mistakes and pauses the resulting quality was very good.

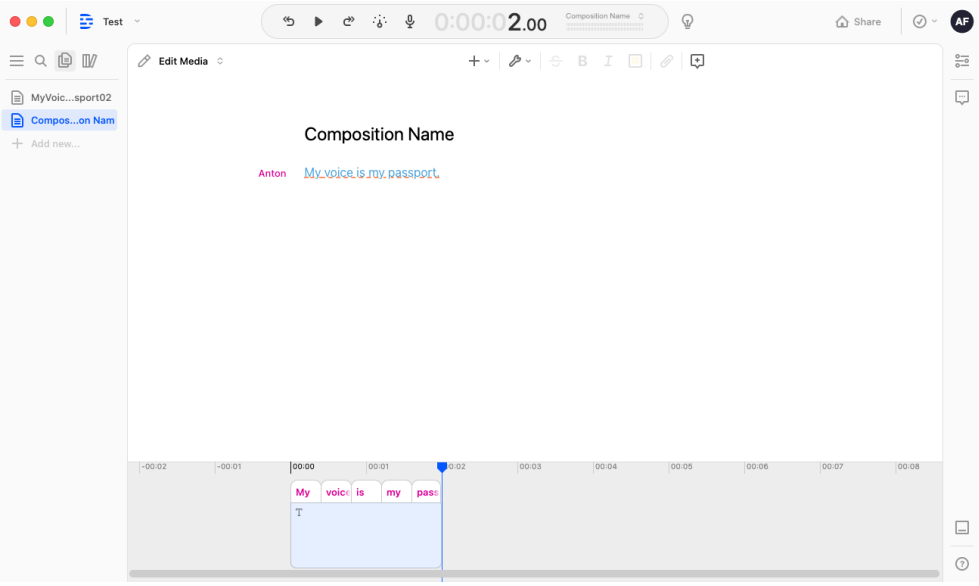

Figure 2.7: Screenshot from the project window of Descript containing written text and chosen speaker to synthesize the text.

2.4.4 Resemble AI

Resemble AI is an online platform for TTS synthesis functionally similar to Overdub. Their web page states that they provide a solution for speech synthesis to be used in podcasts, move narrating, customer care or assistants [51]. API and web interface are both provided for interaction with their system and user-build models [51].

The free plan contains building a custom voice and synthesizing 2000 characters each month. Training the model for synthesis requires at least 50 sentences to be read in the provided interface. Providing training data is also possible by uploading them, but this feature is only available with the paid subscriptions. After reading a sentence, a quality measure appears, and you can re-record the sentence if it was too quiet, there was much background noise or the speech was not distinct as shown in Figure 2.8. After recording the required 50 sentences you can submit the recordings and wait for the model to be trained.

Once your voice is ready you can synthesize any written text. You can add effects to the text in order to modify how the given part will be synthesizes. Provided effects include pause, emphasis, phoneme, prosody, spelling each character, emotion or substitute. The interface for speech generation is shown in Figure 2.9. The quality of generated speech was adequate, even with quite small amount of data the speech sounded like mine, even though the pace was a bit off.

This app is really easy to use, does not require more than a web browser to provide you with with speech generation. The process of training data recording is very user friendly and leads users to submit recordings of high quality.

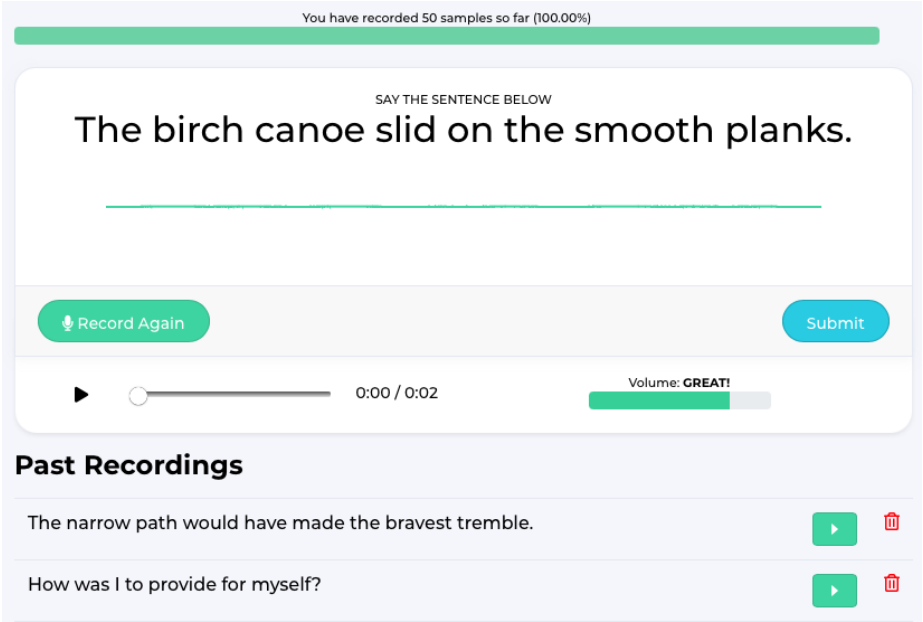

Figure 2.8: Screenshot from the training screen of Resemble web app. The components from top to bottom: progress bar, sentence to be read, control buttons, play again button and quality measure, list of past recordings.

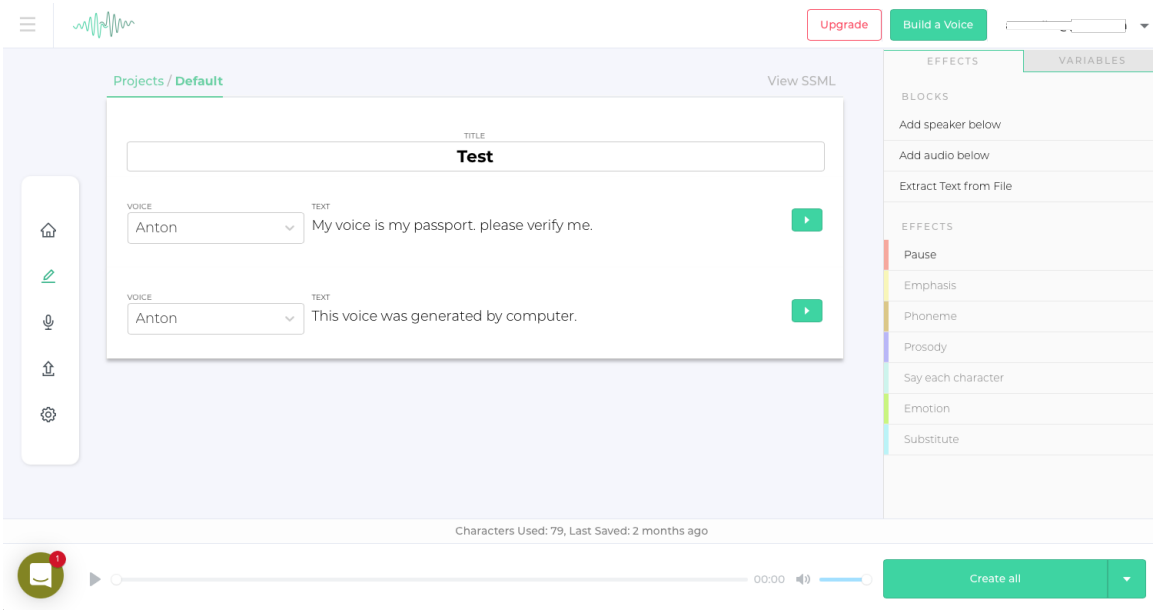

Figure 2.9: Interface for speech synthesis in Resemble AI web app. Menu on the left side, sentences to be generated with speaker selector in the middle, effects menu on the right and „spent characters“ counter on the bottom.

2.4.5 Speech split

The open-source tool *Speech split* belonging to a VC category, is available on GitHub[6] as an implementation of unsupervised speech decomposition framework SpeechFlow proposed by Kaizhi Qian et al. [46].

Speech can be decomposed into four major components: language content, timbre, pitch and rhythm. These components are useful in many speech analysis and generation tasks. The proposed SpeechFlow framework is able to decompose speech into these four components by using three information bottlenecks, which means that this framework is capable to separately transfer on timbre, pitch or rhythm without the knowledge of text labels.

As the Figure 2.10 shows, each of the encoders is responsible for separation of different component. First, the content encoder E_c and rhythm encoder E_r consume speech, while the pitch encoder E_f consumes a normalized pitch contour. Second, the content and pitch encoder perform a random resampling along the time dimension. The resampling operation consists of two steps, firstly the input is divided into segments of random length, secondly all of the segments are randomly squeezed or stretched. This resampling thus might be referred to as an information bottleneck on rhythm. The final outputs of all encoders are called content code Z_c, rhythm code Z_r and pitch code Z_f.

The decoder finally takes all of the speech codes and the speaker identity embedding as inputs and produces a speech spectrogram as output.

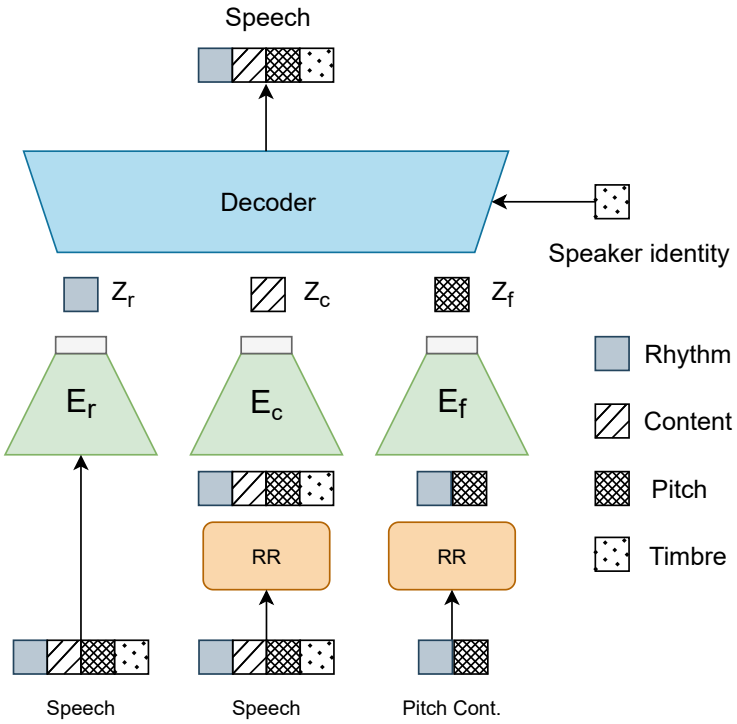

Figure 2.10: SpeechFlow algorithm architecture. Signals are represented as block that denote their components. 'RR' denotes the random resampling operation. The grey block on top of each encoder represents the information bottleneck. The holes in some of the blocks represent a partial information loss. Image and description retrieved from [46].

[6]https://github.com/auspicious3000/SpeechSplit

2.4.6 Cloud services

Apart from commercial tools for content creators or open-source implementations, even big cloud services for speech synthesis exist. Big companies like Google, Amazon, Microsoft or IBM develop their online TTS solutions to be used in any kind of application that might benefit from the usage of speech synthesis.

Almost all of the solutions unfortunately provide TTS synthesis for only a given set of pre-trained voices. While this speech might be suitable for use in smart home assistants, or navigation systems, it cannot be used in any voice spoofing activities as the voice is not that customizable. Only Microsoft with its Text to Speech service [31] offers the ability to completely customise the synthesized voice by providing at least of 30 minutes of audio.

While these solutions offer a text-to-speech synthesis in a very decent quality, the usability for spoofing is limited. The first problem is, that almost all of the available solutions does not have the option to synthesize completely custom voice, just a set of pre-trained voices. The second problem lays within the constant awareness, and most probably even tracking, from the service provider on what you do with their systems. Finally, in contrast to the open-source tools, there is no possibility to perform any modifications to the models or tools.

2.5 Entirely fictitious deepfakes

The previous Sections, 2.3 and 2.4 discussed the tools used to create deepfakes based on a template of human look or voice. Type of deepfake discussed in this section incorporates generation or rather „imagination“ of entirely fictitious people or objects. The presented tools and algorithms are capable of generating synthetic photographs of nonexistent people, animals, flowers, landscapes or emulate human-like text [3].

2.5.1 Image synthesis

As stated by Tero Karras et al. [24], the name *StyleGAN* refers to the style-based GAN architecture that yields state-of-the-art results in data-driven unconditional generative image modelling. StyleGAN as a method for high-resolution image synthesis is currently proved to work reliably on a variety of datasets [24]. The automatically learned and unsupervised separation of high-level attributes, such as pose or identity, and stochastic variation in the generated images, such as freckles or hair enable new, intuitive and scale-specific control of the synthesis [23].

The former StyleGAN method might suffer from the artifact creation, that are mostly invisible to common viewer, but remain in the activations inside of the generator network [24]. These artifacts have been removed in further research [24] and implementation of this research known as *StyleGAN2*. All of the further discussed implementations work with the updated StyleGAN2 method.

There are three major implementations to be found on GitHub:
Original StyleGAN2[7] and it‘s official TensorFlow implementation is provided within a GitHub repository as a CLI application. Compared to tools available for voice deepfake creation, the application is very well documented and the first start required no more than few python libraries to be installed. The image generation is seed-dependent, which means that a number representing seed is needed for each image to be generated. Providing

[7] https://github.com/NVlabs/stylegan2

Figure 2.11: Example of images containing glitches generated using StyleGAN2 and a pretrained network. Image on the left may look normal to an inattentive viewer or in smaller scale, but closer inspection reveals distortion in the background structures such as other cars or trees. The image on the right shows glitches in the main generated object as well as in the background structures.

the same seed repeatedly results in generation of the same image. The provided pretrained networks are capable of generating images ranging from portraits to cars. Generated images look surprisingly real, even though on some of the images there are visible glitches in the background structures as shown on Figure 2.11.

Overall, this tool provides easy to use mean to generate images from chosen area. With the pretrained networks no extensive amount of time or resources has to be provided to achieve decent results.

StyleGAN2 with adaptive discriminator augmentation[8] is an improvement to the original StyleGAN2 method which requires a vast amount of training data, otherwise the discriminator tends to overfit what causes the training to diverge [22]. The modification using adaptive discriminator augmentation (ADA) decreases the need for large datasets with its ability to stabilize training in limited data regimes [22]. Even though this approach improves the result quality with low amount of training data, augmentation still cannot substitute the real data [22].

The official implementation available on GitHub is from the viewpoint of user the same as original StyleGAN described in this section, thus will not be discussed repeatedly.

PyTorch StyleGAN2[9] as the last discussed implementation differs in the used machine learning platform, while it still uses the same method as well as CLI to operate. This implementation seems to be the most user-friendly from all the mentioned StyleGAN implementation within this research. The provided installation script sets up everything needed for the application to run.

The implementation provides a sort of a toolbox that centralizes all the provided functionality like training or image generation. To begin a training only a set of images is required, without performing any transformations or setting parameters.

[8] https://github.com/NVlabs/stylegan2-ada

[9] https://github.com/lucidrains/stylegan2-pytorch

2.5.2 Text synthesis

Generative Pre-trained Transformer or shortly GPT is an open-source AI for text synthesis [47]. GPT shows great performance gains in the area of discriminatively trained models with a generative pre-training of a language model on corpus of unlabeled text followed by discriminative fine-tuning [47].

The latest version labeled GPT-3, in contrast with the preceding one, is released as a commercial product. As the developer (OpenAI) states, this way they can fund the ongoing research, ensure that even smaller companies might benefit from this technology without having large expenses to run it, and finally it makes it easier to respond to any misuse of the technology [39]. The latest open-source version available is GPT-2. Browsing through the examples of generated text shown in [48] leads to an amazement on how real and human-like the generated text feels.

The implementation provides two means of text generation: *unconditional* and *interactive conditional* generation both as a CLI applications. The unconditional generation outputs samples one by one, where each one of them talks about a specific theme. One of the generated texts for example was a Java code for android application, at least the structure and comments in the code. The conditional generation is interactive, it asks for an input and then generates text based on the context of the input. The input text is preferred to be a few sentences in order to match the context, otherwise more of a random content containing keyword from input text is generated.

The implementation does not provide any means of model training, so the usage is reliable on the provided pre-trained models. Fortunately, there are four models available, and their usage requires no more than modifying few lines of code.

Chapter 3

Detecting a deepfake

The emergence of deepfake technology and the threats posed by deepfakes (to be discussed in Section 4.1) lead us to the need for proper detection methods. Both techniques and education for people to distinguish real from fake, as well as algorithms and tools for automatic detection. Being able to differentiate between real and deepfake medium should be the first step in deepfake detection, although humans can make mistakes, and there is even some evidence that human brain might be unable to differentiate between real or fake in some scenarios [36]. This section discusses the techniques and approaches on how people are able to detect a deepfake and algorithms and tools that machines can use to detect deepfakes.

3.1 Human detection

There are many questions about people and deepfake detection, for example: *How do you spot a deepfake?* or *How well can an average person tell the difference between real and deepfake manipulated media?* [13]. A research project carried out in the second half of 2020 [13] aims to answer these questions regarding videos, and brings some recommendations on how to better distinguish between real and AI-manipulated video. The research also brings an online app[1] where anyone can test or practice their skills on deepfake detection. The dataset consists of videos that the machine learning models designed to detect deepfakes classified incorrectly most of the times [13]. All of the manipulations are facial or audio manipulations with many of them being very subtle [13].

Speaking of AI-manipulated media, there are no guaranteed signs to spot a fake, however there are several artifacts that can be looked for:

Facial features

- **Eyes** and their movement. One indicator might be blinking, if the person does not blink or does blink excessively, the second indicator might be the eye movement, for example eyes tend to follow the person that is being talked to [13, 18].
- **Eyebrows** and especially their shadows might reveal a deepfake by wrongly depicting the shadows [13].

[1] https://detectfakes.media.mit.edu

- **Glasses**, specifically the glare might be suspicious, as deepfakes tend to fail in representation of the physics of lightning [13].
- **Facial expressions**, most importantly the unnatural ones such as nose pointing to different direction than the head may signal facial morphing, same as not expressing emotions, a lack of emotion is another indicator that the face might be morphed [18]. In addition, most of the high end deepfakes are facial transformation [13].
- **Hair and facial hair** might not look real, for example frizzy or flyaway hair [13, 18].
- **Skin** contains a few indices at once, the skin color of some parts might not fit the rest of the face or even the body, the aging of skin such as wrinkles might not correspond with the age of hair, eyes or person in general [13, 18]. Even skin features like **moles** and how real they look may provide some information [13].
- **Lips**, concretely the size and color might not match the rest of person‘s face [13].
- **Teeth** and specifically the absence of outlines is another indicator of dealing with a deepfake, as the algorithms are not capable of generating individual teeth [18].

Body features

- **Body position or posture** that might look unnatural or the head and body positioning might be inconsistent [18].
- **Body movement** generated by deepfakes might be distorted or off, mostly when the person moves their head [18].

Voice features

- **Unusual tempo** may make the speech sound not human.
- **Ends of word** might trail off much more than in common human speech [19].
- **Fricatives** (letters like f, s, v and z) are hard to determine from background noise and may sound strange [19].
- **Conversation**, if possible, might reveal the synthesized speech very accurately, as the current technology struggles to converse as effectively as humans [54].

General indices

- **Blurring or misalignment** visible on the edges or visual structures, for example face and head meeting the body may reveal a deepfake medium [18].
- **Inconsistent noise or audio** in the video deepfakes. The authors of video deepfakes mostly spend most to all of the time dealing with the visual side which may result in poor lip syncing, robotic voices, strange pronunciation or even complete audio loss [18].

In addition the tips stated in this section, there are means of using available tools other than deepfake detection engines (discussed in Section 3.2) to help people differentiate between real and fake. Any video editing software capable of slowing down the video

provides better conditions for examining the video, when slowed down the eyes or lips might reveal discrepancies suggesting that the content was manipulated which would be normally invisible [18]. Digital fingerprints and cryptographic algorithms used to prove authenticity of content may be used to detect deepfakes as well as any other forms of media manipulation [18]. Finally even the search engines such as Google might be used for deepfake detection, the reverse image search might be able to find the original content thus to determine if the image or video is real [18].

The stated points should help with identifying deepfakes, although the high quality deepfakes are not easy to spot, but with practice people can build some kind of intuition for identifying what is fake and what is real [13].

3.2 Machine-based detection

As stated by Run Wang et al. [57], machine-based deepfake detectors are are just emerging, which means it will take some time before effective and robust solutions will be dealing with this emerging threat. Some of the former methods for machine-based deepfake detection work with the metadata inspection and byte-by-byte inspection. Latter methods, called bispectral, look for unusual spectral correlations in voices generated by DNNs, these correlations are called *bispectral artifacts*. While the bispectral methods might provide usable results in deepfake detection, it is only a matter of time before the artifacts in deepfake medias will be removed and these methods will be ineffective.

The constant development and improvement in the field of deepfake creation also pushes the development of detection techniques, frameworks and tools.

This section gives an overview on some of the latest findings and improvements in the field of machine-based deepfake detection.

3.2.1 Fake image detection

As the techniques for face-swap creation advance, the detection of these images becomes more and more challenging for forensics models as the pose, facial expression or even lighting can be preserved [38]. This section discussed the approaches to fake image detection as stated by Thanh Thi Nguyen et al. [38].

One of the first methods used the bag of word technique to extract feature set that represents an image and then various classifiers such as support vector machines or random forest to discriminate the real images from the fake ones.

Another technique used incorporates the image preprocessing such as Gaussian blur and Gaussian noise to remove low level high frequency clues of GAN generated images. This way the pixel level statistical similarity between real and fake images is increased and the classifiers are then able to learn more important features that have better generalization capability.

The fake image detection can also be defined as a hypothesis testing problem using a statistical framework. The minimum distance between distributions of real and fake images has to be defined for a particular GAN. The results show that the distance increases when the GAN is less accurate.

One of the latest methods introduced a two-phase deep learning method for fake image detection, where the first phase extracts features based on the common fake feature network (CFFN) and the results are then fed into the second phase which consists of small

convolutional neural network concatenated to the last layer of the CNNF that distinguishes between real and fake images.

3.2.2 Fake video detection

This section discussed the approaches to fake video detection as stated by Thanh Thi Nguyen et al. [38].

Most of the fake image detection methods cannot be used for video detection because of the strong degradation of the frame quality due to video compression. In addition, videos have a temporal characteristics that vary among sets of frames what makes the detection of fake videos more challenging than fake image detection. Two main groups of fake video detection tools can be defined: methods that examine temporal features and methods that examine visual artifacts.

The temporal feature based methods examine the low-level artifacts and their propagation trough frames. Because of the frame-by-frame manipulation of deepfake creation tools, there are inconsistencies across frames that can be exploited for fake video detection.

Another method using the temporal features incorporated the usage of physiological signal, eye blinking, based on an observation that persons in deepfake videos blink a lot less often than in an untampered videos. The detection itself then calculates the blinking rate and determines the realness of the video based on this rate. The experimental results indicate promising accuracy of this method.

The second group are detection **methods based on visual artifacts**. The video is decomposed into frames from which the features are extracted and these features are then fed into a deep or shallow classifier to differentiate between real and fake. Artifact based methods exploit a fact that the deepfake videos are normally created with a limited resolution, which requires an affine face warping approach to match the original faces. This approach creates artifacts, a resolution inconsistency between the warped area and surrounding context. These artifacts can be detected using convolutional neural networks.

3.2.3 Fake voice detection

The framework called *DeepSonar* presented by Run Wang et al. [57] proposes a new method for voice deepfake detection based on the neural network behavior monitoring. All of the information stated in this section is retrieved from.

The framework works above a deep neural network-based speech recognition system by monitoring the activity of neurons in each of the layer when processing an input. It provides a binary classifier that determines whether the voice recording on the input is synthetic or not. This method originates in adversarial attacks detection, where even the minimal changes in the output can be detected by analyzing the neuron behavior [57]. During the training, dataset consisting of real and fake voices was used, and the activity of each neuron in each layer was monitored. The resulting model than includes correlations between the important neurons and important attributes of the input data. Finally, the experiments proved this approach to be robust and effective enough to be potentially usable in the real world environment.

3.2.4 Online detection tools

In the beginning of 2021 an online article showed up claiming that from now on, anyone will be able to detect deepfake videos or images [33]. The article introduces new online platform for deepfake content detection released in January, 2021 by *Sensity*.

The detection platform uses a combination of deep learning algorithms and automated threat intelligence, able to monitor over 500 sources that are well-known for possession of malicious deepfakes [5]. As shown in Figure 3.1, platform features simple UI, that allows for photo and video upload, as well as providing a URL of the photo or video to be checked [5, 33].

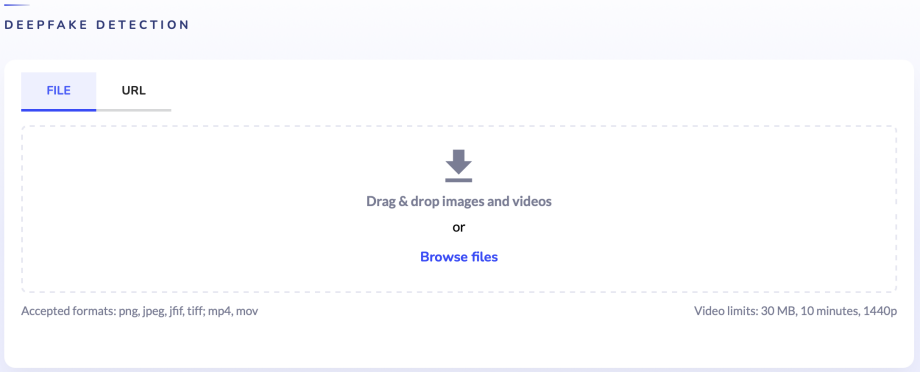

Figure 3.1: UI of the *Sensity* deepfake content detection platform.

As the Figure 3.2 shows, the platform is sometimes even able to detect a model that was used to synthesize given photo or video [5]. The platform was able to detect the GAN-generated face, even when the checked image was taken as a screenshot from this paper. The checked image was the face used as a source image for demonstration of Reface App in Section 2.3.3.

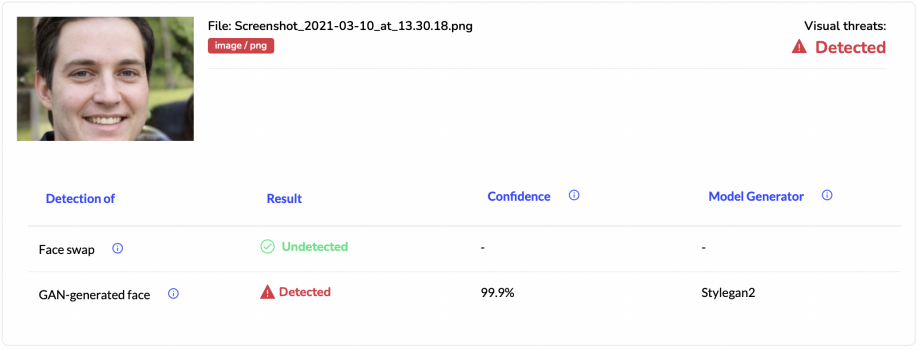

Figure 3.2: Example of detected GAN-generated face.

This platform addresses current security concerns about deepfake medias. The usage spreads from corporate usage when the provided API can for example be used to check attachments in mail communication, to the forensics usage where the investigators might use this platform to verify originality of video or image evidence.

Chapter 4

Deepfakes today – a threat or an asset?

The ability of deepfake technology to generate media depicting events that never happened makes it difficult to differentiate between what is real and fake [3, 21]. Deepfakes do not only pose a threat to individuals, but might also threaten institutions, governments or national security [21].

This chapter poses a literature review that discusses primarily the threats deepfake technology poses to the current digital ecosystem. The first section discusses the impacts deepfakes might have in different areas such as politics, private sectors or social media. The second section is dedicated to discuss the beneficial usages of deepfakes including most of the areas of the ordinary life. The final section contains a short summary on the opinions and facts discovered during the literature review.

4.1 Threats posed by deepfake usage

Even with deepfakes being a relatively new technology, there are various researches, reports and online articles discussing the malicious usage of this technology and the impacts these usages had or might have. The innovation of deepfake technology has been vastly used to defame individuals, create political havoc, spread fake news or threaten stability and democracy globally [21]. New technologies are being developed to detect and suppress deepfakes, however cybercriminals continue to develop more advanced forms of threats to stay ahead [21].

This section discusses the areas and scenarios that deepfakes might have an impact on, and how severe threats are posed by these scenarios. All the scenarios presented in this chapter are feasible because they involve widely available synthetic media technologies [3]. The scenarios described in this section use two main media types: **narrowcast synthetic media** and **broadcast synthetic media** [3]. The broadcast synthetic media are targeted for mass audience, they tend to spread across content-monitored channels like social networks or news reports which provide more opportunities for detection, moderation and fact checking [3]. On the other side, the narrowcast synthetic media transit trough phone lines, emails or personal messaging, which are relatively unmonitored spaces [3].

4.1.1 Government and politics

Governments, politicians or central banks and regulators have battled the spread of rumors numerous times before [3]. Deepfake technology brings a different dimension to creation of those rumors, even with fabricated evidence to support them [3]. Deepfakes might be used in government warfare to spread misinformation that will turn the citizens against their government [21]. Deepfakes also have a potential of discrediting political candidates and opponents reputation, which caused several concerns about the U.S. presidential elections being influenced by spreading misinformation using deepfake medias [21]. Looking back at the U.S. presidential elections, despite the fears of an AI-powered campaign raging through the medias, nothing worth mentioning happened [11, 27]. However, this does not imply that deepfakes are not a threat, just that simpler deceptions like selective editing or outright lies did the trick [27]. One of the most prominent politician‘s deepfakes is a video of Barrack Obama insulting president Donald Trump, even though the video was intended to raise public awareness regarding the threat of deepfake technology [21]. One of the most recent cases of creating deepfakes of politicians, is the video aired during the Christmas time on the UK‘s Channel 4 depicting the Queen of England [49]. The video was again intended to raise awareness about fake news in the digital age [49].

4.1.2 Military

The threat to military posed by deepfake lays within evidence manipulation or discrediting the government. The manipulated evidence might be used to make claims of civilian causalities more plausible during military operations [21]. There is a reported case of manipulating the evidence, where images of a collision of military vehicles was doctored to depict a mangled bicycle [21]. The media then claimed that the soldiers have killed a boy [21]. The deepfake media also bring the fear of discrediting the government with hostile information, in 2019 a video of Gabon‘s president has spread through country setting off massive confusion about his health and thus his abilities as a leader [4, 21].

4.1.3 Justice

Deepfakes have the ability to impact criminal and personal injury by falsifying key evidence [21]. For example a image or video created to convict an innocent target, or an audio recording proving the guilty defendant as innocent. The existence of deepfakes and the threat of potential misuse erodes trust in video evidence and affects its probative value as evidence in court [28]. Even though all of the video or image evidence have to be authenticated before used as evidence, the deepfake techniques are still improving which makes this process very complicated and it is just a matter of time until deepfakes become almost indistinguishable from the real images or videos [28].

4.1.4 Stock market manipulation

Deepfakes can also be used to manipulate the stock markets [21]. The stock markets have suffered from the spread of disinformation long before the invention of deepfake, either by using a scheme called „pump and dump“ in which criminals artificially increase the stock price, or schemes that lower the stock price [3, 21]. The deepfake technology brings a new mean to create statements that might have the power to drop stock value [3, 21]. Two examples of this practice are when media claimed that the former US President Barrack

Obama was injured during an explosion in the White House, or in 2019 when CEO of Tesla, Elon Musk was depicted smoking marijuana [3, 21]. Both of these incidents resulted in lowering the stock value [21]. The opposite can also happen, which means using deepfake technology to fabricate events that will raise the company‘s stock value, for example by depicting celebrities promoting certain product [3].

These scenarios might be really harmful to targeted subjects, for example an especially damaging scenario involves synthesized recording of a highly-ranked officer using sexist language [3]. Even though this recording is purely fake, there are almost no ways to certainly prove that this conversation never happened [3]. Even after disapproving such a recording, it would most likely have a long-term consequences regarding company and personal reputation [3].

The other deepfake-backed form of manipulating stock prices are social botnets as stated by Jon Bateman [3]. With this being a common practice nowadays, deepfakes and deep learning in general can step up the quality of these attacks. Criminals have already started to use AI-generated images as profile pictures to avoid detection of picture reuse. The deep learning could potentially be capable of creating AI-driven synthetic bots that will operate on social networks, using generative models to create and publish posts. The most advanced case of using social botnets presents a scenario with bots posting novel and individualized content, this way they can build and manage their own persona. This process of self-development could run for several months or years with minimal supervision, during this period of time bots will be able to gain organic followers that will increase the impact of their messages and make them far more harder to detect [3]. After triggering by their creator, or operator these bots will start posting about a targeted company using their built persona, this can be used to achieve previously mentioned rise or drop of the company‘s stock price [21].

4.1.5 Social Media and Disinformation

Apart from stock market manipulation introduced in section before, social media have the capability of spreading mass misinformation [3, 21]. Many people view content and information received from their family members or other relatives as reliable without confirmation of it being manipulated [21]. This may lead to an uncontrollable spread of misinformation at a high pace. Considering this mismatch in information provided from the trusted individuals and from the media or other truly reliable sources people might lose the ability to identify what is real or fake or to trust in what they see or hear [21]. The specific scenario impacting financial area might involve spreading fabricated medias to arouse bank runs [3]. For example using a botnet introduced in Section 4.1.4 to spread rumors, or releasing a video or voice recording of bank executive describing liquidity problems [3].

4.1.6 Misinformation and Fake News

The term fake news is nothing that that people are unfamiliar with. Fake news spread through medias for a long time now, becoming a weapon in political sphere [21]. The fake news can be also used as a *clickbait* to manipulate user into clicking on the content that the author benefits from [21]. The deepfake technology brings new means of creating fake news that are more believable and might cause more harm [21].

4.1.7 Extortion

The common cyber extortion scheme consists of criminals claiming to have embarrassing or derogatory information, often of a sexual nature, about the victim and they require payment or different form of ransom in order to not release the information [3].

The extortionists may use deepfakes to generate a convincing blackmail material, for example inserting victim's face into a pornographic image or video to provide a proof of the access to sensitive material [3]. This threat scenario is more than actual, supported with the outcome of a study conducted by Amsterdam-based deep learning company *Deeptrace Labs* in December 2018, which found out that 96% of uploaded deepfake videos had pornographic content and featured primarily women [21]. The extortionists may even proceed to a large-scale attacks with robots harvesting content from social medias, such as personal images and then producing synthetic media and using it to blackmail [3].

The manipulated voice or look has already been used for extortion, Israeli con men impersonated French Foreign Minister over Skype and scammed their targets for over 90M, stating that the funds will be used for secret operations and to pay ransom for the release of hostages [21].

4.1.8 Fraud

The deepfake technology consistently raises a threat of a sensitive personal information breach [21]. Spoofing attacks will be used against biometrics systems to compromise face or voice identification [21]. The spoofed voice or even video can also be used as a more authentic alternative to emails in business email compromise schemes to convince victims to transfer funds or purchase gift cards [3].

4.1.9 Identity theft

According to Jon Bateman's [3] claims, identity theft is one of the most common type of consumer complaint in the United States. The typical case involves a criminal opening a new credit card using personal details of the victim, which can be acquired by breaching into commercial databases. Deepfakes present at least two ways to steal one's identity, the first one involves targeting an individual with intention of personal harm. Criminals will impersonate the victim in a phone call to trick an executive or financial advisor to execute wire transfer. This method can also be used for money laundering, where criminals create bank accounts under false identity using deepfake audio or video. The second method uses stolen identity at a scale, for example using deepfakes in a social engineering campaign resulting in gaining unauthorized access to personal data [3].

4.1.10 Imposter scam

The imposter scam in scenario without the usage of deepfake typically includes making contact by a phone, using spoofed phone number or voice over IP to disguise attacker as a local caller or a specific person as stated by Jon Bateman [3]. Other cases might involve hackers taking control of someone's email or social media account and use it to scam friends or family members. After the establishment of contact between the attacker and victim, attackers threaten with an imminent harm unless amount of money is paid. The reported losses from imposter scams in the U.S. reached amount of 667 million dollars in 2019 [3].

The deepfake-backed scenarios can enhance the believability of imposter, for example by cloning a voice of specific individual that is trusted by the victim like a relative or government official. Voice manipulation is becoming more popular and damaging than video deepfakes since there is no visual information to determine if the call is fake. Some of the scam victims reports claim that a family member's voice was cloned using artificial intelligence, however none of these reports is verified [3].

4.2 Benefits of deepfake usage

The usage of deepfakes is not purely restricted to cause harm, the negative consequences attract so much attention from fact that deepfakes can also have beneficial usage that upholds our society [21, 25]. The impacts can be seen in many areas, but mainly in entertainment, education and healthcare, as discussed in this section [21, 25].

4.2.1 Entertainment

When discussing the area of entertainment, we discover a tight boundary between what is just entertaining, and what is harmful. This fact can be deducted from the information stated in this chapter, where for example fake news might arise from what was meant to be just an entertaining video. Although, social media are not the only place for deepfakes to be used to entertain. The movie industry may benefit from updating the footage instead of re-shooting it, recreation of classical scenes, or creating movies with actors that are no longer alive [21]. The deepfake technology may be also utilized to automatize the dubbing process and delivering the content to many audiences with minimal to none effort [59]. Another example of using deepfakes to entertain and spread awareness is the global malaria awareness campaign that took place in 2019, which aimed to attract diverse audiences with a video created using visual and voice-altering technology [59].

4.2.2 Education

The benefits to education area begin with bringing long-dead historical personalities to life in order to emphasize teaching [21]. Deepfake technology was also used in an exhibition in Dalí Museum, St. Petersburg, USA, where visitors can interact with Salvador Dalí's virtual self and learn about him [21, 25]. Speaking of education, it is crucial to point out the need to educate not only pupils and students, but also broad public about the dangers that deepfakes pose to the public. Although this is not a benefit of deepfake technology to education process, it seems to be quite important in order to lower the threat presented by usage of this technology for example in area of fake news.

4.2.3 Healthcare

The generative ability of deepfake technology may be used to help people in social and healthcare areas, ranging from plastic surgeries to creating data for treatment simulation [21]. Deepfakes can help patients suffering with Alzheimer's disease to interact with a younger faces they might remember [59]. The deepfakes can also be used to recreate amputees limbs, simulate the results of transgender operation, or model the outcome of plastic surgical operation [59]. Oncological or other patients that lost their voice voice because of a disease, might benefit from creation of text-to-speech models that enable them to speak again using their own voice [21]. Generative models might even be used to detect

abnormalities in x-ray images [21]. The AI-based technology could also help to create an virtual population of patients to develop and test various ways to diagnose, monitor and treat diseases with lifelike patients without putting any life to risk in case of failure [21]. The benefit can be found even in medical areas linked to psychology, where bringing a deceased person back to life may help the mourners to say final goodbye [59].

4.2.4 Other areas

There are far more areas that can benefit from the usage of deepfake technology, than the ones mentioned before.

Tourism can benefit from the generation of photorealistic images of scenes that might be used in advertising to attract more visitors and attention to a destination [25].

Fashion industry may accommodate „dressing up“ synthetic models. This way, models do not need to be directly present at the photo shooting, and also their proportions can be easily altered to fit the desired look [59]. This goes hand in hand with the **e-commerce** area, where customers will be enabled to try the outfits on their virtual selves [59].

Advertisement might use the deepfake technology to create advertisements without the need for the chosen actor to be present at the shooting, the desired face might be added to the clip afterward [59]. This also opens up opportunities to shoot advertisement spots with already dead actors or celebrities. While this approach technically seems to be nothing different from creating a deepfake of dead historical person, there might be some ethical concerns.

Gaming is strongly connected with the latest technological trends, and for many years pushes the requirements for personal computers performance sky high. However, we have not seen any outburst in usage of deepfake technology in gaming. Deepfakes can be used to give voice to the characters in games or enable increased telepresence [59]. Synthetic voices are also starting to be used in **smart assistants** to make them sound more natural [59].

4.3 Chapter summary

As stated in this chapter, deepfakes have a vast variety of harmful or illicit usages which causes more misgivings then credence [25]. Deepfake usage finds place in various sectors such as politics, finance, private sector or fake news. The threats presented in various areas present different levels of risk scaling from individuals up to the governments. There is a possibility of deepfakes causing a global damage, but unlikely in the current situation as the synthetic media are better suited to inflict individual harm on targeted individual or company [3]. Also, none of the presented scenarios present a serious threat to the stability of the global financial system or national markets in healthy economies [3]. Even with most of the attention drawn to the malicious usage, there are plenty of ways humanity can benefit from the deepfake usage. With the right education and legal regulations, deepfakes might be able to be viewed in different angle, but before we can profess this statement with certainty, we have to answer yet unanswered questions about the potential impact of deepfake usage [25].

Chapter 5

Experiment design

The previous chapter discussed the threats posed by deepfake technology. This chapter aims to introduce the attack vectors and experiments that will be carried out and analyzed as a part of this research. At first, current situation and public opinions regarding the ability of voice biometrics systems is discussed. The gathered opinions on technical feasibility of spoofing voice biometrics systems are mixed, what leaves us wondering how much effort does it take to spoof such a system.

The proposed attack vectors examine all of the parts of using deepfakes to spoof voice biometrics systems: spoofing voice biometrics system itself, spoofing the human operator and being able to execute this attack in any chosen language. The first attack vector thus examines the technical feasibility of synthesizing speech using both open-source and commercial tools, and the ability of this speech to spoof voice biometrics system. The second attack vector examines the technical feasibility of synthesizing speech in Czech language. The third attack vector finally uses the speech synthesized during our research to examine whether the voice able to spoof voice biometrics system is also able to spoof humans.

Data gathered from the experiments proposed in this section will then be used to evaluate the feasibility of each of the parts that a real-world attack does have to contain.

5.1 Voice biometrics systems and deepfakes

The first attack vector proposes usage of deepfake speech to gain unauthorized access to a system secured by voice biometrics system, such as financial institution call centre. The attack vector composes of three experiments that aim to explore the capabilities of both commercial and open-source text-to-speech (TTS) synthesis tools and usage of synthesized media against voice biometrics systems. At first, current situation regarding voice biometrics systems and their safety against synthetic speech is evaluated by reviewing available sources online. Afterwards a deepfake scenario is presented, where the usage of synthetic media to spoof voice biometrics systems is further discussed based on the available literature discussing first successful documented attempts to spoof voice biometrics system. Finally three sub-experiments are designed that range from exploring the basic principles of using text-to-speech synthesis against voice biometrics systems to a more detailed look on both text-dependent and text-independent verification types and their comparison regarding provided security.

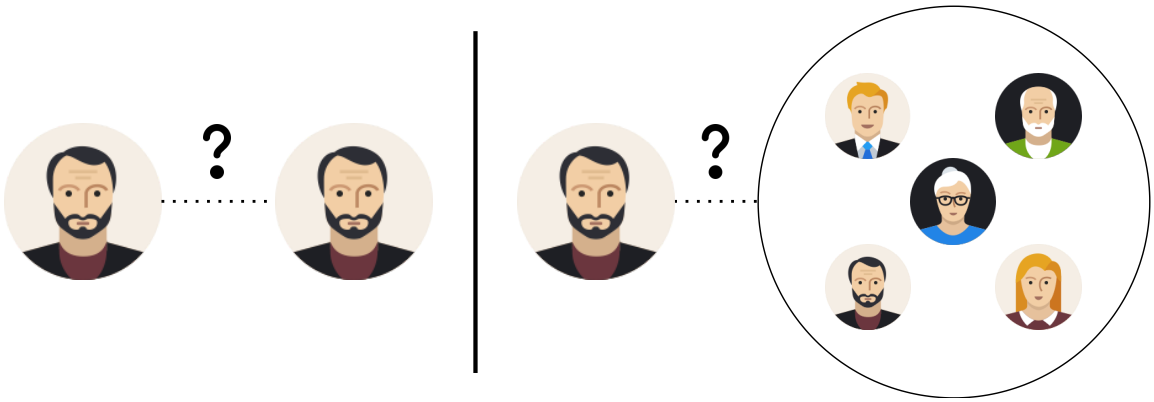

Figure 5.1: The principle of verification process on the left, and the principle of identification process on the right.

5.1.1 Voice biometrics system

This section defines a voice biometrics system the way it is understood in this work, and the metrics used to compare the performance of these systems that are further referenced.

Voice biometrics, or speaker recognition, systems provide a biometric-based security process called speaker authentication [29]. It means, that the system authenticates people on a „*What you are*“ principle, by processing their personal biometrics attribute - voice.

Two essential speaker recognition technologies are *speaker verification* and *speaker identification* [29]. Speaker verification is a process of determining whether a person is the one who she or he claims to be, thus a $1:1$ comparison [29, 30]. Speaker identification is a process of assigning an identity to the voice from a pool of enrolled speakers, thus a $1:N$ comparison [29, 30]. The difference between identification and verification is also shown in Figure 5.1.

Speaker verification can be further divided into two types: *text-dependent* and *text-independent* [29, 30]. Text-dependent verification needs the same passphrase to be spoken in both enrollment and verification phases [30]. This approach might lead to an stronger security, as the phrase might be unknown to an attacker. On the other hand, the text-independent verification has no restrictions on what the speaker says during the enrollment or verification phase and extracts only a similarity score [30].

In conclusion, the voice biometrics system is understood as a black-box, that authenticates a person based on their own voice.

Performance measures

There are several measures that specify the performance (accuracy) of a biometrics systems and are used to compare these systems. To evaluate how accurately a biometrics system is in the scope of verification, many genuine and impostor attempts are made with the system and the achieved matching scores are saved [45]. A genuine attempt is an attempt done by user to match her or his own profile [45]. An impostor attempt is an opposite to the genuine attempt, an attempt done by a user to match someone else‘s profile [45].

Using the saved matching scores, a distribution graph of matching scores by attempt type can be plotted. Both impostor and genuine distributions, are probability density functions, showing how many attempts fall into the given interval of matching score. The

matching score threshold is used to tune the biometrics system: a lower threshold makes the system more convenient, and higher threshold makes the system more secure [45]. Finally a pairs of false reject rate (FRR) or false non-match rate (FNMR) and false accept rate (FAR) or false match rate (FMR) can be calculated by applying a varying matching score threshold [45]. All of the mentioned measures are shown in Figure 5.2.

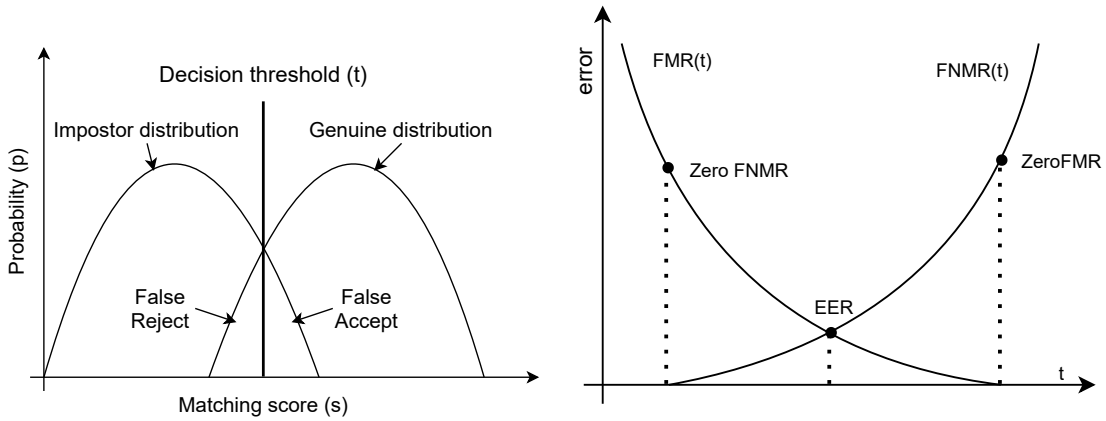

Figure 5.2: Matching scores distribution graph on the left. FMR, FNMR and EER on the right.

The FMR and FNMR measures, can also be used to calculate equal error rate (EER), it is an error value at certain value of matching score threshold, where FMR and FNMR equals [2]. FMR and FNMR rates, along with EER are shown in Figure 5.2.

5.1.2 Attacker model

The attacker is a person with ability to create voice deepfakes and his goal is to gain access into a system secured by voice authentication, such as financial institution call centre. The attacker is in possession of all needed personal information about his victim, all needed details about the common scenario of the voice authentication process, and finally samples of voice belonging to the victim.

The attacker will use all of this information to synthesize speech reproducing the victim‘s speech, and then in the most believable way possible try to access the system secured by voice authentication and use the granted access in his advantage.

The ability to create voice deepfakes can be understood in two main ways. The attacker is either able to collect and prepare enough data and train his own speech synthesis tool that he will later use, or the attacker is able to get unauthorized access into one of the commercial systems and misuse the stored speech synthesis models to generate speech. The second type of the attacker would be less powerful and probable, as only a small portion of people use such commercial systems.

5.1.3 Current situation and public opinion

This section serves a review of opinions that can be found on the internet, on the ability of voice biometrics systems to deal with synthesized speech. A survey performed in December of 2019 by a biometric security startup ID R&D, concluded that the worries about deepfake voice fraud are slowing down the adaptation of voice authentication systems [52]. Around

two thirds of Americans are afraid of their voice being spoofed and used to access their accounts secured by voice authentication [12].

According to the article published by the Guardian [20], the companies behind the voice biometrics technology say there are more than 100 unique physical and behavioral characteristics of each individual. These characteristics include length of the vocal tract, nasal passage, pitch, accent and many more [20]. They are claimed to be as unique to an individual as much as an fingerprint is, and that the systems can even recognize people with cold or sore throat [20]. Nuance communication claimed that even professional voice imitators can't fool their system, while some of the other companies claimed that their voice identification system is even able to spot a difference between identical twins [20].

These statements were (un)fortunately refuted by a BBC reporter when bank let his non-identical twin access his account [20, 56]. As stated before, the bank claimed to use more than one hundred different characteristics of voice to verify an identity [56]. Also, the bank allowed them to get seven attempts of verification wrong, before the eight successful attempt [56]. This huge amount of failed attempts has been pointed as a main problem that weakened the authentication system. Although this incident has taken place in first half of 2017, and since then the technology has advanced rapidly, this incident proves that an attack on voice biometrics using a person with similar voice characteristic was doable, thus it still should be doable, it is just a matter of technology or the right circumstances. This incident also brings a different view on the statements claiming, that these system cannot be fooled by voice imitators or other forms of voice spoofing. As shown in Chapter 2, it is still becoming easier to generate speech spoken by voice of other person, even without any IT education or knowledge. The only thing needed is a recording of target person's voice, which can be nowadays obtained simply by downloading a video from a social network, YouTube or a podcast. This leads to a belief that voice identification should serve as an additional method to verify persons identity, not the main one [20].

The voice biometrics systems are believed to include techniques to detect illicit audio such as replayed or synthetically generated speech by an article posted on TechRadar [42]. Some of the companies like ID R&D even show their liveness detection in action as a promo for their voice biometrics system [14]. A different article published on Techradar denotes the voice authentication systems the strongest and most accurate authentication systems, however also states that this ability might be compromised if the deepfake technology continues to improve [42]. The most recent article discussing this topic published in May of 2020 [16] supports the statement that the voice biometrics systems might be fooled by sufficiently advanced deepfakes. The article even proposes that this ability might lead to an artificial intelligence arms race, with institutions upgrading their authentication systems and criminals improving the quality of the deepfakes to overcome newly implemented measures [16].

5.1.4 Deepfake scenario

The first scientific opinions about spoofing voice authentication using deepfakes shows up in a research published on the Black Hat conference in 2018 [53]. The authors introduced a proof of concept that the voice biometrics systems might be fooled by deepfakes and supported this theory with first real evidence.

Authors of mentioned research demonstrated a deepfake powered attack on voice biometrics system using text-to-speech service called Lyrebird. Lyrebird is an AI research division of Descript that is responsible for development of Descript's Overdub feature men-

tioned in 2.4.3. The quality of synthesized speech was a bit lower, but both Apple Siri and Microsoft Speaker Recognition API authenticated with the generated audio [53].

To demonstrate how a real attack might be performed, authors used the state-of-the-art text-to-speech generation solutions, such as *Tacotron* and *WaveNet*. To generalize the threat model, the text-to-speech models were treated as a black box. The open-source implementations available at GitHub were used although some of them were suffering with the degradation of audio quality. This limitation seems to be just a temporary considering the recent rapid advancements in the text synthesis field.

Authors of this research state that methods introduced in [53] reduce the effort and cost for attacking speech authentication systems to downloading an open-source code, reading some StackOverflow threads, spending few hours transcribing audio and finally running the code on a decent GPU over a bunch of days. In their research, the also proved that the speech authentication, by itself, is vulnerable when the victim‘s audio recordings are publicly available. The voice authentication might be fooled, so it is not recommended to use voice as a primary mean of authentication. The authors also stated that there is no known mechanism to detect this kind of attack.

A different research [37] confirmed that an attack using deepfake speech on voice biometrics systems is doable. However, none of the researches performed these attacks in real-world conditions.

As a standalone security measure, biometric authentication does not provide bulletproof security [10]. Up to date, there have been no reports of spoofing voice biometrics using a deepfake technology [10]. This statement itself raises another question, which is, whether no incident of this kind was ever successfully performed, or the incident(s) occurred, but no report is available to public. Fraudsters nowadays are very intelligent and creative, so the challenge is to keep a step ahead of those people [10].

5.1.5 Experiment design

Given the previous facts, voice biometrics systems have already been fooled just by imitating the speech of another person. Given the content of sources discussed in this section, there are mixed opinions on the technical feasibility and reliability of voice biometrics systems, even with a proof of concept, but the minimum of available evidence, still leave us wondering how much effort would it take to spoof such a system.

The following experiments focus on text-to-speech (TTS) tools, as these tools are currently in later stage of evolution than the voice cloning (VC) tools. The TTS systems are available as commercial solutions like Overdub (Section 2.4.3) or ResembleAI (Section 2.4.4) or as many open-source implementations available on GitHub. On the other hand, the currently available VC tools are just research paper implementations to demonstrate the functionality and running these tools, at least with demonstration data, requires very extensive amount of time, patience and knowledge in area of speech processing. This unfortunately limits the usage of such tools to people dedicated to their development and will take some time before the broader public is able to reliably use them with their own data.

The former idea for execution of these experiments was to cooperate with real-world subjects that use the voice biometrics systems for their customer care solutions, such as call centers. The experiments were meant to be executed in the English language. Using English language instead of mother language has more reasons. Firstly, there are currently no publicly available TTS models for Czech or Slovak language. Also the difficulty of

these languages makes it much more demanding to create a model that could synthesize these languages in an usable quality as will be shown in Section 5.2. Finally, executing the experiments in language that is not so commonly used for voice authentication might bring interesting results. As little to none of these clients are native English speakers, the imperfections or other glitches in the synthetic voice might be overlooked or even amplify the believability of the synthesized speech. Although it is not all sunshine and rainbows as it might sound, the models used to identify these clients are known to be less precise which means there is much more attention paid by the operator that the client is interacting with.

After reaching out to Czech and Slovak subjects with a cooperation proposal, two financial subjects responded with an interest to participate in this research. Unfortunately after the initial discussions about the details of experiment execution and agreement on cooperation, both of the subjects remained silent. This led to a backup plan, where biometric systems I was able to get hands-on were tested instead. I reached out to companies providing voice biometrics solutions with a request to provide me an access to their system for testing the resilience of their system against a created deepfake dataset. The responses ranged from no response at all, to informing me that they are very sorry, but they unfortunately do not provide any kind of licenses for students. After the unsuccessful attempts of getting hands-on as many voice biometrics systems I ended up with following two: Microsoft Speaker Recognition API and the Phonexia Voice Verify demo. I also tried to look for an open-source implementation of any speaker verification system, but the systems did not provide the needed functionality, which is that the system is already functioning without any need for training the model, provides interface for enrollment of any speech recording and uniform results for each verification attempt that can be compared with each other.

The experiments aim to research the technical feasibility of cloning someone‘s voice and the possibilities of using synthetically generated voice of an individual to get access into a system secured by voice biometrics system. Following experiments thus have been proposed:

Exploring basic concepts

The first experiment aims to explore the technical feasibility of following commercial TTS tools: *Overdub, ResembleAI* and an open source-tool *Real-Time-Voice-Cloning*, and provide the basic knowledge on using TTS synthesis to spoof voice biometrics systems. This knowledge will then be used to better specify next experiments and execute them.

The selected commercial tools are currently the most popular ones, when searching for voice deepfakes online. The open-source tool was chosen because of its ability to synthesize specific speech based on a short recording of target speaker, which is a feature that no other TTS tools provide. The synthesized voices will then be compared using Microsoft Speaker Recognition API and the Phonexia Voice Verify demo. As mentioned before, these are the only voice biometrics system that provide wanted functionality and I was able to get access to them.

Firstly, I will evaluate the overall usability of the selected TTS tools, learn to work with them and generate speech. Afterwards, I will acquaint myself with the voice biometrics systems. With the introduction to both of the softwares done, I will continue to generate different utterances and authenticate them in the available voice biometrics systems to see how do the systems behave when deepfake authentication attempts are executed.

This experiment will answer following research questions:

- *How difficult is it to create a synthetic copy (clone) of an individual‘s voice?*

- *How much data is needed to clone an individual's voice?*
- *Are today's voice biometrics systems capable of detecting synthetic voice?*
- *What areas of using TTS synthesis to spoof voice biometrics systems are interesting for further examination?*

Text-to-speech versus text-independent speech verification

The second experiment focuses entirely on the Real-Time-Voice-Cloning (RTVC) tool, and Microsoft Speaker Recognition API. Using the knowledge acquired in the first experiment a deepfake dataset will be created and then evaluated. The aim of this experiment is to extend the scope of results achieved in the previously proposed experiment. The RTVC tool was selected because of its open-source license and ability to synthesize speech based on short recording of target speaker. The open-source license of the RTVC tool also allows for modifications and other tweaks, which makes me believe that a real attacker would use this tool rather than one of the commercial ones. I tried to negotiate an access to the commercial systems to use them during the research, however all of my cooperation request remain unanswered. The Microsoft Speaker Recognition API provides a fast and simple way to calculate matching scores of utterances when compared to enrolled voice profile.

At first, a dataset consisting of 100 speakers will be created. For each speaker, there will be genuine recordings and deepfake recordings generated by fine-tuning the synthesizer model of RTVC tool for 1k and 5k iterations. The recordings synthesized using model fine-tuned for 1k iterations and 5k iterations will be further referenced as deepfake 1k and deepfake 5k respectively. The information about fine-tuning and models used in the RTVC tool can be found in 6.3.1 respectively 2.4.2. The speakers will be selected from the Common Voice Corpus [1], as this dataset is publicly available, provides a wide range of speakers and qualities of utterances.

The dataset will be created following a specific scheme. For each speaker, the longest recording will be selected as an enrolment recording. This recording will then be used to create speaker's voice profile as well as a target speaker recording for RTVC tool for speech synthesis. All of the others recordings will then be used to fine-tune the synthesizer model for 1k and then 5k iterations. At both 1k and 5k fine-tuned iterations, a synthesizer model will be saved for possible further use and 10 recordings will be synthesized. The utterances will be the same for both 1k and 5k synthesized recordings, transcriptions can be found in Appendix B.

After the dataset will be completed, matching scores for genuine and impostor attempts will be calculated as well as matching scores of deepfake recordings for each user. The matching score calculation depends on the type of the recordings used and follows this scheme:

Genuine matching scores will be calculated for each of the speakers by creating a voice profile using enrolment recording, and then calculating matching score for each of the speaker's recordings, even the enrolment one against the created voice profile.

Impostor matching scores will be calculated for each of the speakers by creating a voice profile using enrolment recording, and then calculating matching scores of recordings of other speakers against the created voice profile.

Deepfake matching scores, both 1k and 5k, will be calculated for each speaker by creating a voice profile using enrolment recording, and then calculating matching score of each speaker's deepfake recording against the created voice profile.

To get even better insight on the performance of text-independent verification against deepfakes, I will calculate matching scores for available deepfake datasets from the ASVspoof challenge and voice conversion challenge that are further discussed in Section 6.2. The matching scores calculated from the datasets created for the purpose of training deepfake detection systems and to showcase the abilities of voice conversion tools will provide interesting comparison of results achieved with the RTVC tool.

Each of the matching scores types will then be used to plot user distribution graphs and to calculate the equal error rate (EER). The distribution graphs and EER will be compared and evaluated for each of the matching score types in order to find out how similar the deepfake utterances are to the genuine ones. The matching scores will also be used for further examination and data mining, with a goal to find patterns between speakers, their deepfake recordings and achieved matching scores.

This experiment will answer following research questions:

- *Do the deepfake matching scores distributions differ from the genuine matching scores distribution?*
- *Are there any patterns in matching scores of deepfake recordings based on genuine matching scores?*
- *Are there any other patterns that might lead to detection or improvement of deepfake utterances?*

Resilience of text-dependent vs. text-independent verification

During the execution of the first experiment (Section 5.1.5), a difference in text-dependent and text-independent verification matching scores suggested that the text-dependent verification might be more secure against deepfakes. As the text-dependent verification not only examines the general voice characteristics (independent of the content), but also the way how a specific phrase is spoken, it seems to deliver more security when facing a synthesized speech [29, 30].

To further examine, and possibly confirm this hypothesis, an special dataset containing the phrases accepted by the Microsoft Speaker Recognition API has to be used. As no dataset containing these phrases is currently publicly available, the first step will be to create one. The dataset will consist of the phrases used for text-dependent verification, and then a set of training utterances for the RTVC tool. I have chosen three phrases from the set of available phrases to include phrases of all lengths:

- my name is unknown to you
- my voice is my passport verify me
- I am going to make him an offer he cannot refuse

Each phrase will be recorded 5 times in total, where the first 4 recordings will be used for text-dependent profile enrolment, and the fifth recording will be used as a reference to calculate genuine matching score. The training set will consist of 75 sentences randomly selected from the transcripts of the Common Voice Corpus, and the Harvard Sentences[1].

[1] https://www.cs.columbia.edu/~hgs/audio/harvard.html

With dataset complete, RTVC tool synthesizer models will be fine-tuned for each speaker and each phrase will be synthesized 10 times. Afterwards, matching score calculation for both genuine and deepfake utterances will be done the following way:

A **text-dependent** profile will be created for each speaker, and the genuine and deepfake matching scores for each phrase will be calculated.

A **text-independent** profile will be created, using recordings from the training set. Afterwards, for each speaker the reference utterance and deepfake utterances will be used to calculate the matching scores.

This experiment will answer following research question:

- *Is text-dependent verification harder to breach using deepfakes than text-independent verification?*

All of the experiments proposed in this section are further discussed, executed and finally evaluated in Section 6.3

5.2 Text-to-speech synthesis for Czech language

As the majority of research in the field of text-to-speech synthesis and deepfakes is carried out and presented in English language, almost all of the tools and pretrained models can handle exclusively the English language. Of course there are solutions from big players like Google or Amazon that are capable of synthesizing very naturally sounding speech in Czech language, but the technical background and models are kept private and also there is no possibility to adjust the speaker to meet specifics needs. In order to be able to clone a voice and to use it for voice phishing or an identity theft we need to use a tool that is capable of generating speech of a given speaker with as few training samples as possible.

The second attack vector thus consists of using speech synthesized in Czech language instead of English as in the first attack vector. This experiment thus aims to extend the knowledge of using text-to-speech synthesis to attack systems secured by voice authentication that are being used by Czech speaking people. Text-to-speech synthesis model will be trained for Czech language, to evaluate the amount of data, knowledge, effort and time it takes to train such a model, and finally to discuss whether the achieved quality currently poses a threat to Czech or Slovak speaking people.

Using the model trained to synthesize speech in Czech language, a deepfake dataset similar to the dataset discussed in Section 5.1.5 will be created. The deepfake dataset will then be evaluated to find out how precisely the synthesized speech recalls the original one from the point of view of a voice biometrics system.

The model will be trained for the Real-Time-Voice-Cloning tool (Section 2.4.2) as I gained most experience working with this tool during the research, and it provides the wanted functionality of synthesizing speech of target speaker based on short speech recording. The training will be following the GitHub wiki [8] article about training the tool from scratch for any language and other useful information found in the issues section of the GitHub repository.

This experiment will thus answer following research questions:

- *Is it technically feasible to train TTS model for Czech language from scratch?*
- *Is there enough usable data available for training TTS model for Czech language?*
- *How much time it takes to train TTS model for Czech language?*

- *What is the final quality of synthesized speech?*
- *Do TTS tools currently present a threat to Czech or Slovak speaking people?*

The process of training an TTS model for Czech language, and all the achieved results are further discussed in Section 6.6.

5.3 Speaker similarity survey

The last, third attack vector follows up on the first one. While the first attack vector was directed towards the voice biometrics systems only, this attack vector focuses on the human factor, as in most of the scenarios a voice biometrics system provides authentication results to a call centre operator or other person that the attacker has to communicate with. This means that even if the voice is recognized as genuine by the voice biometrics system, the operator might get suspicious and end the communication. The voice thus needs to be as similar as possible to the original one for both computers and humans. This experiment will examine, whether the voice able to spoof voice biometrics system is also able to spoof a human being.

The evidence collected from online articles and the only published incident report, suggest that human brain, and people in general are not capable of differentiating between real speech and the synthetic one. The article published on BuzzFeed [58] tells a story about how a BuzzFeed reporter used deepfake speech to communicate with his mom, without her noticing. Another evidence is an incident report of fraudsters accomplishing to get 250k USD transferred to their account [21]. Fraudsters impersonated the CEO of the company and manipulated an employee to transfer funds to one of the suppliers using the synthetically generated voice [21].

To test human capability to distinguish between real and spoofed speech a survey will be created, where the respondents will be set into a role of speech verification system. A made up scenario will set the respondents into a role, where their main goal will be to listen to an enrolment utterance and then to three utterances that will represent three independent verification attempts and decide the genuineness of each utterance. The respondents will also be informed, that the voices they will hear are very similar, but some of the verification attempts are made by synthetic or somehow manipulated utterances. Finally, the respondents will be told to be critic and only select the recordings they are completely sure are genuine.

This scenario aims to simulate the process of speaker identification that runs subconsciously in our minds when we hear a voice with no visual information on who we are speaking to, e.g. receiving a phone call. The recordings will be created using models from the deepfake dataset created for the purposes of the experiment discussed in Section 5.1.5. The dataset will be completely created using knowledge and means available to me during the period of research. I decided not to use a dataset prepared for evaluation of biometrics systems or dataset made for demonstration purposes of deepfake tools even though they would provide speech of better quality. This way I am completely aware of the effort and knowledge needed to create the dataset, which means it provides more detailed view on the quality of the created deepfakes by user that uses existing tools rather than implementing them, which better suits assumptive attacker. Ten speakers will be selected in a way to cover the whole range of achieved average matching scores of deepfake recordings from the original English deepfake dataset. This way a link between matching score of a deepfake recordings and believability to humans can be examined. For each speaker, the choice of

recordings type will be the same: genuine recording, deepfake recording generated using model pretrained for additional 1k iterations and additional 5k iterations. The order of recordings will be randomly selected during the creation of the survey.

This experiment will thus answer following research questions:

- *Is there any link between matching scores and believability to human of deepfake utterances?*
- *How many times genuine user was denied access, and how many times impostor was allowed access?*
- *Is the collected data usable to support or refute the human ability of spotting a deepfake speech.*

The survey is expected to be accepting responses for a month, but this period may change based on the amount of received responses and the daily increment of the responses. Further information about the used dataset, survey structure, and results are discussed in Section 6.7.

Chapter 6

Experiment preparation and realisation

This chapter discusses the creation of deepfake media for each experiment, the experiment execution and achieved results. The following sections discuss the experiments in the same order as they were proposed during the design in the Chapter 5.

6.1 Used voice biometrics systems

This section discusses the voice biometrics systems used in the experiments to verify speakers and calculate matching scores of recordings. Genuine utterances are verified using all of the voice biometrics systems to show how do the genuine attempts perform.

6.1.1 Microsoft Speaker Recognition API

The Microsoft Speaker Recognition service provides both speaker verification and speaker identification functionality. After providing training data, an enrollment voice profile is created that can be used to validate speakers identity, or cross-check voice samples against a group of enrolled speakers to see if it matches any profile in the group [30].

The speaker recognition services documentation [30] also states that the speech verification APIs are **not intended** to determine whether the audio is spoken by a real person or is an imitation or recording.

As a first step in experiments with Microsoft speaker recognition and deepfake recordings I decided to try the verification and identification functionality with my real voice to set a baseline of score for genuine attempts. The matching scores of text-dependent verification are shown in Table 6.1 and text-independent in Table 6.2. As the identification is just a differently used text-independent verification, it will not be taken into account.

Recording no.	Attempt 1	Attempt 2	Attempt 3
1	0.90110	0.90110	0.90110
2	0.83630	0.83630	0.83630
3	0.88688	0.88688	0.88688
4	0.82257	0.82257	0.82257
5	0.83245	0.83245	0.83245
6	0.73093	0.73093	0.73093
7	0.78041	0.78041	0.78041
8	0.73779	0.73779	0.73779

Table 6.1: Matching score of genuine recordings using text-dependent verification using phrase „my voice is my passport verify me“. Each recording verified three times independently of the earlier attempts. Recordings number 1 - 4 were used for enrollment, recordings number 5 - 8 were previously unseen for the speaker recognition system.

Recording no.	Attempt 1	Attempt 2	Attempt 3
1	0.86717	0.86717	0.86717
2	0.78694	0.78694	0.78694
3	0.83001	0.83001	0.83001
4	0.80286	0.80286	0.80286
enrol	1.0	1.0	1.0

Table 6.2: Matching score of genuine recordings using text-independent verification. Each recording verified three times independently of the earlier attempts. Recording marked as *enrol* was used as an identity for profile enrollment. Other recordings were previously unseen for the speaker recognition system.

The results indicate that the text-independent verification might provide slightly higher matching scores compared to the text-dependent verification. Data from both of the tables show slight discrepancy between repeated attempts to verify the same recordings, this is even more visible within the results of text-independent verification.

The recording matching scores show, that to reproduce the quality of the genuine recordings with this dataset, the matching scores achieved by deepfake recordings should lay in the interval of $[0.7, 0.95]$. This means that if the created deepfake recordings are stable to reach values from this interval, they mimicked the genuine voice almost completely accurately, and in this scenario will be accepted by the verification system.

6.1.2 Phonexia Voice Verify demo

Phonexia Voice verify is an online service for speaker identification. As stated by the Phonexia website [44], the service provides text-independent and language-independent verification at the same time. To enroll a speaker only 30 seconds of speech is needed, to verify a minimum of 3 seconds of net speech is required.

The presentation pages do not mention that any kind of liveness detection is implemented. Differences between demo version, and production versions surely exist, but no information is available on these differences. However, the demo version represents some

state of development and settings of the system, thus an outdated or improperly configured system should behave the same.

Figure 6.1 shows the complete verification graph of genuine utterance. After the initial oscillation, the verification curve goes right to the top and stays there for the whole verification process. To reproduce the genuine speech deepfakes should not only reproduce the final result but the shape of the verification curve as well.

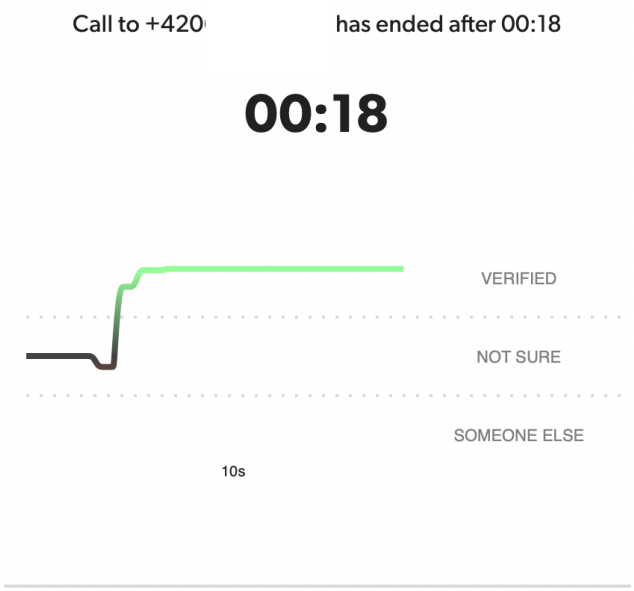

Figure 6.1: Screenshot of a complete verification graph of genuine speech in conditions with no background noise or other form of interference. A small oscillation is visible in the beginning, but afterwards the system shows stable certainty that the voice is genuine.

6.2 Datasets

The datasets that are in scope of this research can be divided into two categories: *deepfake datasets* that contain genuine and corresponding synthetic speech of an individual, and *datasets for speech processing tasks* that contain only genuine speech.

The deepfake datasets are mostly used to train systems capable of differentiating between genuine and deepfake speech, while the datasets for speech processing task are used for various tasks ranging from training speaker authentication systems to speech synthesis systems. Unfortunately no standardised datasets to measure the performance of voice biometrics systems in terms of differentiating between genuine and synthetic speech are available, however the discussed deepfake datasets can be tailored to suit this purpose.

This section discusses publicly available datasets from both of the mentioned categories.

6.2.1 Deepfake datasets

The **ASVspoof challenge** aims to design a solution capable of discriminating between genuine and synthetic speech [60]. For the purpose of development of such a solution, a dataset consisting of synthetic speech created using the state-of-the-art TTS and VC

systems is provided [60]. The most recent dataset currently available was published for the ASVspoof 2019 challenge.

The ASVspoof challenge dataset consists of two parts: logical access and physical access. The logical access refers to an usage of synthetic media over the telephone, while the physical access refers to replaying the synthetic speech which is not only a problem for speaker authentication systems, but also smart objects like home assistants [60].

The **Blizzard challenge** is aimed to measure the advances in text-to-speech synthesis field. A dataset consisting of target speech is published for participants, and the goal is to synthesize speech of the target speaker [63]. Each year a different language to be synthesized may be selected, in 2020 Mandarin and Shanghainese were selected [63].

The original dataset does not contain any synthesized speech, however the published results of the challenge provide enough synthetic data to construct a deepfake dataset. Unfortunately the dataset designed for TTS training is not public and available only for the participants of the challenge. The latest available dataset is from 2013[1].

The **Voice conversion challenge** in contrast to the Blizzard challenge measures advances in the field of voice conversion [61]. A dataset consisting of source and target speaker utterances is published for each challenge. The results are again published, so composing a deepfake dataset is possible.

The presented datasets can be used to evaluate the performance of voice biometrics system to differentiate between genuine and synthetic speech. The ASVspoof challenge and voice conversion challenge datasets provide both genuine and deepfake speech. Considering the purpose for which the datasets were created and their content, the ASVspoof challenge datasets seems to best suit the role of a dataset for testing the resilience of voice biometrics systems against state-of-the-art deepfakes.

6.2.2 Datasets for speech processing tasks

The second category of datasets relevant to the scope of this research, are datasets created for speech processing tasks. In contrast to the deepfake datasets, a wide variety of these datasets exists. The datasets also provide a wide range of speakers, languages and qualities of speech. The following list briefly describes the most popular ones:

- *Common Voice corpus* [1] is a massively-multilingual collection of transcribed speech dedicated for purposes of the automatic speech recognition, but can be used for much more speech related tasks. The Common Voice projects incorporates crowd-sourcing for both data collection and validation. The latest version of dataset, marked 6.1, currently consists of 60 languages and more than 7k hours of validated speech. This makes the Common Voice corpus the largest dataset in terms of number of languages and spoken hours.

- *LJ Speech* [15] dataset contains speech with transcriptions of a single speaker reading passages from 7 non-fiction books in English. The total number of recordings in dataset is more than 13k which is approximately 24 hours of total length.

- *Vox Celeb* [34] is an audio-visual dataset consisting of short clips of speech that were gathered from videos uploaded to YouTube. The dataset consists of more than 7k speakers with more than a million recordings which approximates to more than

[1] https://www.synsig.org/index.php/Blizzard_Challenge_2013

2k hours of total length. All of the recordings are captured „in the wild“, which means there is background chatter, laughter or overlapping speech.

- *LibriSpeech* [40] is a dataset consisting of read English created for training and evaluation of speech recognition systems. The dataset is derived from audiobooks recorded for the LibriVox[2] project.
- *Libri TTS* [62] is a dataset specially designed for text-to-speech use. The dataset is derived from the LibriSpeech dataset, while addressing a number of issues that make the LibriSpeech dataset not ideal for text-to-speech usage.

6.3 Exploring basic concepts

This section describes the execution of experiment proposed in Section 5.1.5. For each tool technical feasibility and overall usability, and an usage of created deepfakes to authenticate with the used voice biometrics systems is discussed.

6.3.1 Basics of deepfake creation

The first part of this experiment, was to get familiar with the used tools and learn how to use them to synthesize deepfake in the best quality possible.

Overdub

As mentioned in Section 2.4.3, I recorded the first part of the training script, which is approximately 10 minutes long, for my speech model training. After submission of the training script, my voice was ready in approximately 7 hours.

Firstly, I verified the quality of synthesized speech by listening to the recordings. To. my ears the synthesized speech sounded very natural and clear, almost the same as the training recording.

After successful hearing quality verification, I put the synthesized recordings to test against the Microsoft Speaker Recognition API. As the Table 6.3 shows, the results for text-dependent verification failed to reproduce the genuine matching scores. I also noticed that the synthesized phrase has different pauses between words than the genuine one. This happened most notably between the words *voice* and *is*. I tried to shorten that pause by concatenating these words, as the last line of the Table shows, and reached significantly higher matching score. The matching scores for text-independent verification on the other hand show great similarity to the genuine ones as shown in Table 6.4.

Recording no.	Synthesized text	Attempt 1	Attempt 2
1	my voice is my passport verify me	0.53939	0.53939
2	my voiceis my passport, verify me	0.64144	0.64144

Table 6.3: Matching scores for text-dependent verification from Microsoft Speaker Recognition API using deepfake recordings synthesized with Overdub tool. Each line contains the text that was synthesized and the matching scores achieved in two attempts.

[2] https://librivox.org

Recording no.	Attempt 1	Attempt 2
1	0.74349	0.74349
2	0.79611	0.79611

Table 6.4: Matching scores for text-independent verification from Microsoft Speaker Recognition API using deepfake recordings synthesized with Overdub tool.

The Microsoft Speaker Recognition API calculated some promising matching scores, mostly for the text-independent verification. As shown in Figure 6.2, the voice authenticated even with the Phonexia Voice Verify demo. The verification result graph is very similar to the genuine one, only the slight drop on the beginning might suggest that deepfake speech instead of genuine speech was used.

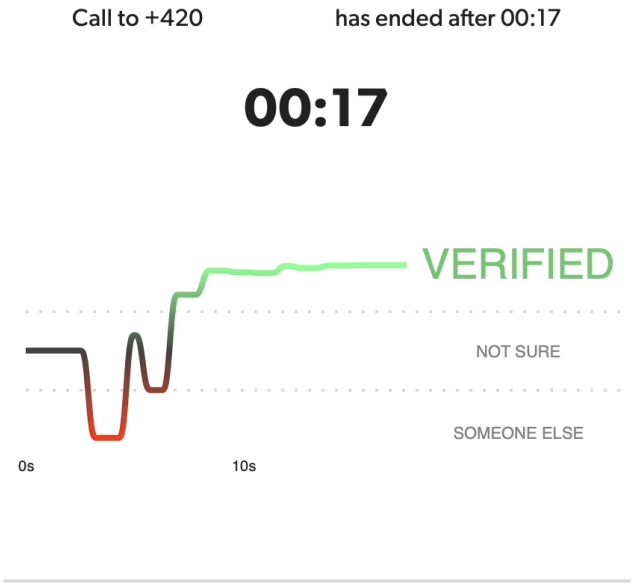

Figure 6.2: Screenshot of complete verification graph of deepfake recording created using Overdub feature of Descript. A big drop is visible right at the beginning, afterwards the course reproduces the genuione verification graph.

The Overdub tool is very simple to use, even an individual with no special IT knowledge is able to synthesize his own speech. As seen, the quality is good enough to spoof the tested voice biometrics systems. However, the usage of this tool is limited to synthesizing own speech, as specific transcript must be read. Of course, there are ways to overcome this problem, either by manipulating the victim to say or read out loud the training script and record it for later use, or breaching into Descript‘s systems and getting a hold of the trained models.

Both of the variants have many factors to fail on, and are moreover theoretical. However that does not imply an attack following one of the scenarios can not happen. It surely is possible, and regarding the achieved results it might cause some damage to the targeted individual.

Resemble AI

The Resemble AI tool, discussed in Section 2.4.4, provides very similar functionality as the previously discussed tool Overdub. To train my speech model, I recorded 150 sentences to approximately match the amount of training data provided to Overdub, to ensure similar conditions to both of the tools. Training the model took exactly 26 minutes.

The verification by ear showed that the synthesized speech is similar to the recordings, however some glitches and unusual pauses could be heard. As Tables 6.5 and 6.6 shows, the Microsoft Speaker Recognition API confirmed that the quality of synthesized speech is lower than the Overdub synthesized speech.

Recording no.	Synthesized text	Attempt 1	Attempt 2
1	my voice is my passport verify me	0.48961	0.48961
2	my voice. is my passport. verify me	0.51071	0.51071
2 (pp)	my voice. is my passport. verify me	0.55974	0.55974

Table 6.5: Matching score of text-dependent verification from Microsoft Speaker Recognition API using deepfake recordings synthesized with Resemble AI tool. First two rows represent an recordings created by synthesizing differently written phrase. The last row shows matching score of recording number 2 that was post processed.

Recording no.	Attempt 1	Attempt 2
1	0.52847	0.53619
2	0.59407	0.60107

Table 6.6: Matching score of text-independent verification from Microsoft Speaker Recognition API using deepfake recordings synthesized with Resemble AI tool.

With the phrases for text-dependent verification, I again noticed that the synthesized speech contains different pauses between words and also the speed at which some words are said differs. To modify the synthesized speech in order to achieve higher matching scores I modified the synthesized text which slightly increased the achieved matching score as the second row of the Table 6.5 shows. To further improve the quality of synthesized speech I decided to shorten the pauses and slow down some parts of the phrase as shown in Figure 6.3, this again led to an increase of achieved matching score as the last row of the Table 6.5 shows.

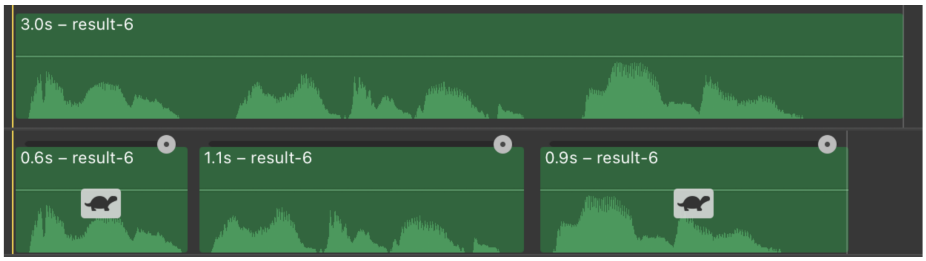

Figure 6.3: Comparison between original synthesized recording and the same recording after post processing. The original recording on the top. Post processed recording with shortened pauses and slightly slowed down first and last segments on bottom.

Finally I tried to verify the synthesized speech with the Phonexia Voice Verify demo. As the Figure 6.4 shows, the speech was verified, even though the course was more shaky than with the genuine recording. The verification graph again shows a significant drop right in the beginning.

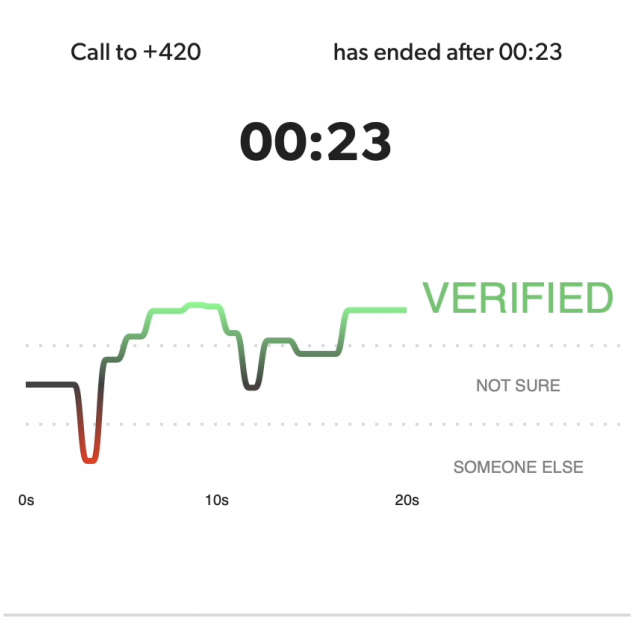

Figure 6.4: Screenshot of complete verification graph of deepfake recording created using ResembleAI tool. Even though the final result states verified, the verification course graph differs from the genuine one.

As mentioned before, the Resemble AI tool is very similar to the Overdub tool, and even an individual with no extensive IT knowledge is able to synthesize his own speech. Again, the training follows predefined sentences, however most of the sentences, if not all, are Harvard sentences[3], so tricking someone into saying them for you might be easier.

Real-Time-Voice-Cloning

The Real-Time-Voice-Cloning tool is an open-source software that provides more customizable functionality, but at the price of more extensive knowledge required to use it.

To begin with, I used the provided pretrained models to get familiar with synthesizing speech without any GUI as in the commercial tools. As the RTVC tool synthesizes speech based on the provided target speaker recording, I decided to experiment with various lengths of this recording. As the Tables 6.7 and 6.8 shows, the achieved matching scores are very low, as well as the verification by listening to synthesized speech showed that there is minimal similarity with the genuine speech.

[3] https://www.cs.columbia.edu/~hgs/audio/harvard.html

Template	Template length	Attempt 1	Attempt 2
Short sentence	3s	0.17202	0.17790
Sentence	12s	0.19861	0.19951
Paragraph	45s	0.16500	0.15664
Long paragraph	5m 12s	0.04032	0.04219

Table 6.7: Matching score of text-dependent verification using phrase „my voice is my passport verify me“ from Microsoft Speaker Recognition API of deepfake recordings created using pretrained models provided within GitHub repository. Each row represents an input of different length used as a template for voice cloning. The approximate length of target speaker recording and matching scores achieved when using the same voice profile as in Table 6.1

Template	Template length	Attempt 1	Attempt 2
Short sentence	3s	0.30780	0.30780
Sentence	12s	0.37237	0.37237
Paragraph	45s	0.47097	0.47097
Long paragraph	5m 12s	0.32789	0.32789

Table 6.8: Matching score of text-independent verification from Microsoft Speaker Recognition API of deepfake recordings created using pretrained models provided within GitHub repository. Each row represents an input of different length used as a template for voice cloning. The approximate length of a target speaker recording and matching scores achieved when using the same voice profile as in Table 6.2

To achieve better results the synthesizer model can be fine-tuned for target speaker [50]. It is recommended as a good mean of achieving better results without any extensive work in the issues section of the GitHub repository[4].

Only the synthesizer model needs to be fine-tuned, as all of the other parts (encoder and vocoder) are speaker independent. The fine-tuning process requires around 0.2 hrs of recordings with transcription [50]. With the data prepared, the fine-tuning process follows the steps of synthesizer training [50]. For this purpose, I prepared a dataset of my own voice with 68 recordings with total length of 12 minutes and 30 seconds what is slightly more than 0.2 hours. The most time exhausting part of this process was gathering the data, and transcribing speech, the processing was done in a matter of minutes, and fine-tuning the synthesizer model for additional 1k steps took around 1 hour using one NVIDIA Tesla T4 GPU. The matching scores of text-dependent verification using this fine-tuned synthesizer model are shown in Table 6.9 and matching scores of text-independent verification using the same synthesizer model are shown in Table 6.10.

[4] https://github.com/CorentinJ/Real-Time-Voice-Cloning

Template	Template length	Attempt 1	Attempt 2
Short sentence	3s	0.53881	0.54017
Sentence	12s	0.41729	0.41839
Paragraph	45s	0.41655	0.41500
Long paragraph	5m 12s	0.41640	0.41611

Table 6.9: Matching score of text-dependent verification using phrase „my voice is my passport verify me“ from Microsoft Speaker Recognition API and deepfake recordings created after a synthesizer model fine-tuning with additional 1k iterations to the pretrained model used in Table 6.7. Each row represents an input of different length used as a template for voice cloning, the approximate length of a target speaker recording and matching scores achieved.

Template	Template length	Attempt 1	Attempt 2
Short sentence	3s	0.53854	0.53854
Sentence	12s	0.61365	0.61365
Paragraph	45s	0.60910	0.60910
Long paragraph	5m 12s	0.55084	0.55084

Table 6.10: Matching score of text-independent verification from Microsoft Speaker Recognition API using deepfake recordings created after a synthesizer model fine-tuning with additional 1k iterations to the pretrained model used in Table 6.8. Each row represents an input of different length used as a template for voice cloning, the approximate length of a target speaker recording and matching scores achieved.

The matching scores achieved after fine-tuning show significant increase, however they still do not reach the genuine matching scores, mainly the text-dependent verification attempts. To further test the capability of fine-tuning for a single speaker, I decided to train the synthesizer model to additional total 5k iterations. Training the additional iterations to the previously fine-tuned model took around 3 hours, using the same GPU as before. As Tables 6.11 and 6.12 show, the matching scores decreased compared to fine-tuning results after 1k iterations. This might be because of the synthesizer model getting over-trained. This behavior is further examined in Section 6.4.

Template	Template length	Attempt 1	Attempt 2
Short sentence	3s	0.47863	0.47863
Sentence	12s	0.59272	0.59272
Paragraph	45s	0.46116	0.46116
Long paragraph	5m 12s	0.39038	0.39038

Table 6.11: Matching score of text-dependent verification using phrase „my voice is my passport verify me“ from Microsoft Speaker Recognition API and deepfake recordings created after a synthesizer model fine-tuning with additional 5k iterations to the pretrained model used in Table 6.7. Each row represents an input of different length used as a template for voice cloning, the approximate length of a target speaker recording and matching scores achieved.

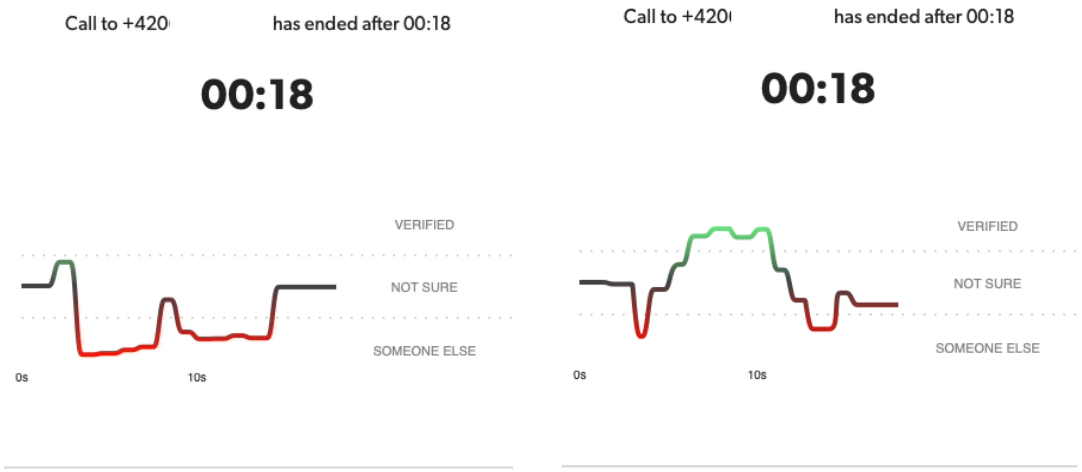

Figure 6.5: Complete result of verification for speech synthesized using synthesizer model fine-tuned for 1k additional iterations on the left and 5k additional iterations to the pretrained synthesizer model on the right.

Template	Template length	Attempt 1	Attempt 2
Short sentence	3s	0.21066	0.21066
Sentence	12s	0.38387	0.38387
Paragraph	45s	0.16417	0.16417
Long paragraph	5m 12s	0.41994	0.41994

Table 6.12: Matching score of text-independent verification from Microsoft Speaker Recognition API using deepfake recordings created after a synthesizer model fine-tuning with additional 5k iterations to the pretrained model used in Table 6.8. Each row represents an input of different length used as a template for voice cloning, the approximate length of a target speaker recording and matching scores achieved.

Finally I tried to authenticate in Phonexia Voice Verify demo using speech synthesized by RTVC tool for both of the fine-tuned synthesizer models. As the Figure 6.5 shows, the lower quality of synthesized speech reflected into the course of verification graphs. But in contrast to matching scores achieved using Microsoft Speaker Recognition API, the speech synthesized using model pretrained for 5k additional iterations reached significantly better results. However, none of the synthesized recordings would have passed the verification in Phonexia Voice Verify demo.

The Real-Time-Voice-Cloning tool provides interesting functionality, but comes with a higher demands for knowledge. As shown, the synthesized speech still lacks quality compared to commercial tools, but the synthesized speech might be usable. More detailed view on how well this tool reproduces speech is provided in Section 6.4.

6.3.2 Experiment conclusion

This experiment discussed speech synthesis using three different tools. The commercial tools synthesized speech of solid quality, that has a potential to get verified by voice biometrics systems as the matching scores and verification graphs show. The quality of speech

synthesized using the open-source tool falls behind the commercial tools, as well as usage requires more knowledge and effort to be made. But again, this limitation is surely just temporary and in a matter of time the usability of open-source tools as well as quality of synthesized speech will approach the current state of used commercial tools.

The experiment results also provide answers to the research questions asked in the design of experiment (Section 5.1.5):

How difficult is it to create a synthetic copy (clone) of an individual‘s voice?

There seems to be more correct answers, depending on what approach is chosen and what is the target quality. Generally speaking of speech synthesis, the process is quite simple, you type desired text, click, and the tool synthesizes speech. However as shown with the RTVC tool, the quality of speech synthesized this way is questionable. If you rather choose to use a commercial tool, the difficult part is to gather the recordings of victim for training, and training the model as well as synthesizing speech remains a matter of clicking some buttons in the prepared GUI.

If you decide to rely solely on the open-source tools, you should be able to train a model of your own and work with it. This requires a lot of knowledge about implementing machine learning tasks and python, as well as understanding of speech processing and synthesis principles.

In summary, a process of creating a synthetic copy of an individual‘s voice is not trivial and requires knowledge and time to get familiar with. The demands on knowledge and effort grow with the desired quality and usability of synthesized speech.

How much data is needed to clone an individual‘s voice?

The needed amount of data depends on how the voice will be cloned. As shown in this experiment, as least as 0.2 hours of speech is enough to fine-tune the model and synthesize speech of decent quality. However, it is still valid that the bigger amount of training data increases the quality of synthesized speech. If the attacker plans to train a model from scratch, much more data would be needed, probably above 20 hours as the pretrained model provided for the RTVC tool was trained using the LJ Speech dataset which contains approximately 24 hours of speech [8].

Are today‘s voice biometrics systems capable of detecting synthetic voice?

Unfortunately this experiment does not provide enough data to answer this question. Only two voice biometrics systems were used during the execution of this experiment, while one of the systems explicitly states that it does not implement any kind of liveness detection, and the other one is a demo preview of the complete system. To answer this question more extended research has to be carried out, using wider range of voice biometrics systems and systems implementing liveness detection. However, this experiment shows that system not implementing liveness detection, or incorrectly setup system might be spoofed by synthetic voice.

What areas of using TTS synthesis to spoof voice biometrics systems are interesting for further examination?

The discrepancy between marching scores calculated using Microsoft Speaker Recognition API and the verification results from Phonexia Voice Verify demo for speech synthesized

using fine-tuned synthesizer models for RTVC tool calls for further examination. To better understand what effects does fine-tuning for 1k and 5k iterations have on the quality of synthesized speech a experiment discussed in Section 5.1.5 was proposed.

The differences in matching scores calculated for text-dependent and text-independent verification using Microsoft Speaker Recognition API suggest that the text-dependent verification is more resilient against the synthetic speech. To further examine this behavior an experiment discussed in Section 5.1.5 was proposed.

6.4 Text-to-speech versus text-independent speech verification

This experiment aims to explore the possibilities of using deepfakes against text-independent verification provided by Microsoft Speaker Recognition API. Three datasets containing genuine and deepfake speech will be used for this purpose, one created by me and two publicly available datasets from the ASVspoof and Voice conversion (VC) challenges. The dataset will be created following the structure proposed in the design of this experiment (Section 5.1.5), thus will consist of recordings for 100 speakers. Creating my own dataset for the scope of this work helps to stay aware of all of the aspects of creating the deepfake recordings and how well do they perform in comparison to genuine recordings and other deepfake recordings from the ASVspoof and VC challenge datasets.

6.4.1 Creating a dataset

The dataset is created as a subset of the Common Voice Corpus [1], the speakers were ordered by count of the recordings, and then first 100 were selected. The longest recording was chosen as enrolment and then for each speaker a synthesizer model was fine-tuned for 1k and 5k iterations. All of the recordings were finally collected into a dataset, where for each speaker a set of genuine and deepfake recordings are provided. The dataset was also published for possible future research[5].

6.4.2 Matching scores distributions and dataset comparison

After the completion of my own Common Voice deepfake dataset I was able to calculate genuine, impostor and deepfake matching scores for each dataset. The methods for calculating each type of matching scores are discussed in the design of this experiment (Section 5.1.5).

As the Figure 6.6 shows, the own dataset created using the RTVC tool performed very well. The deepfake matching scores distributions very much reproduce the shape of genuine matching scores distribution. As the FMR and FNMR plots show, the EER increased significantly when comparing deepfake and genuine plots.

Figure 6.7 shows matching scores distributions for recordings from the dataset created for purposes of the ASVspoof challenge. In comparison to the plots for the Common Voice deepfake dataset, the deepfake matching scores are distributed almost evenly trough the whole range of matching scores. More recordings achieved matching scores below the genuine ones, but on the other hand there were recordings that reached higher matching scores than the genuine ones. The increase in EER is very similar, almost same, to the Common Voice deepfake dataset.

[5] https://drive.google.com/drive/folders/1vlR-TA7gjKzjYylxzRnA_HzZEyWiLeOk?usp=sharing

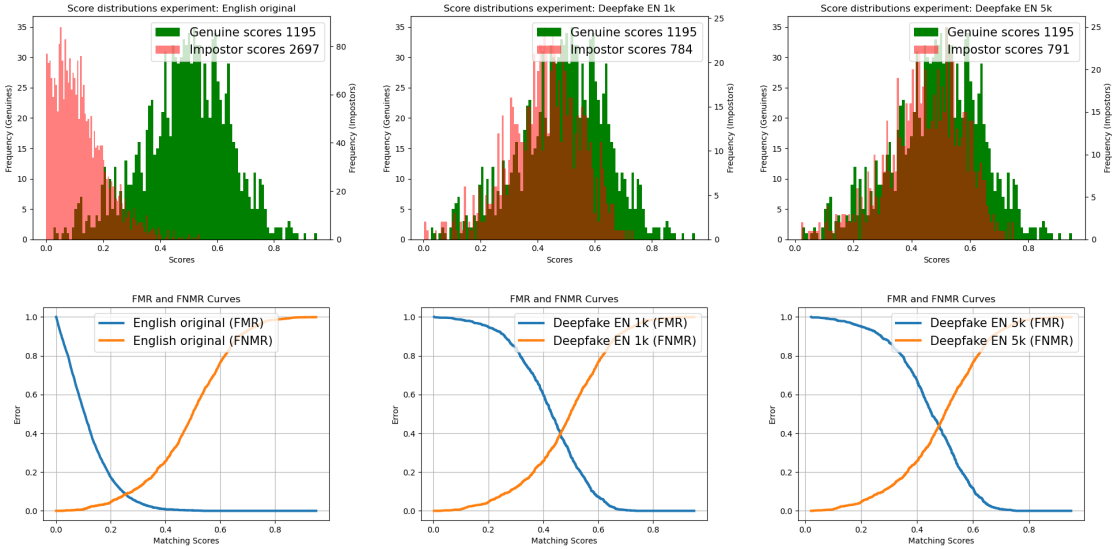

Figure 6.6: Matching scores distribution graphs (top) and FMR / FNMR graphs (bottom) for the Common Voice deepfake dataset. The genuine versus impostor distribution on the left, genuine and deepfake 1k in the middle, and genuine and deepfake 5k on the right.

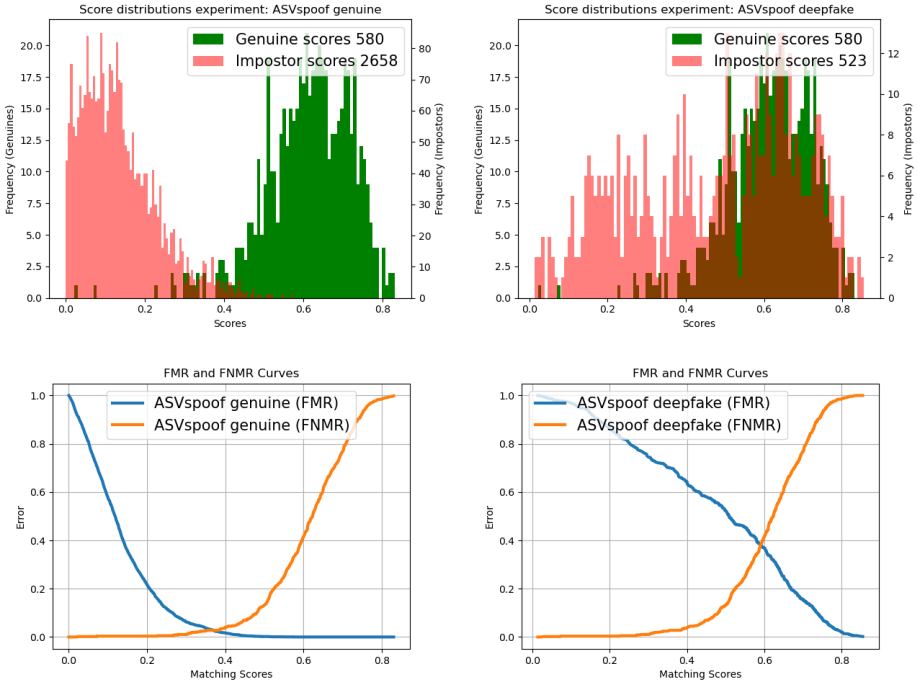

Figure 6.7: Matching scores distribution graphs (top) and FMR / FNMR graphs (bottom) for the ASVspoof challenge dataset. The genuine versus impostor distribution on the left, genuine and deepfake on the right.

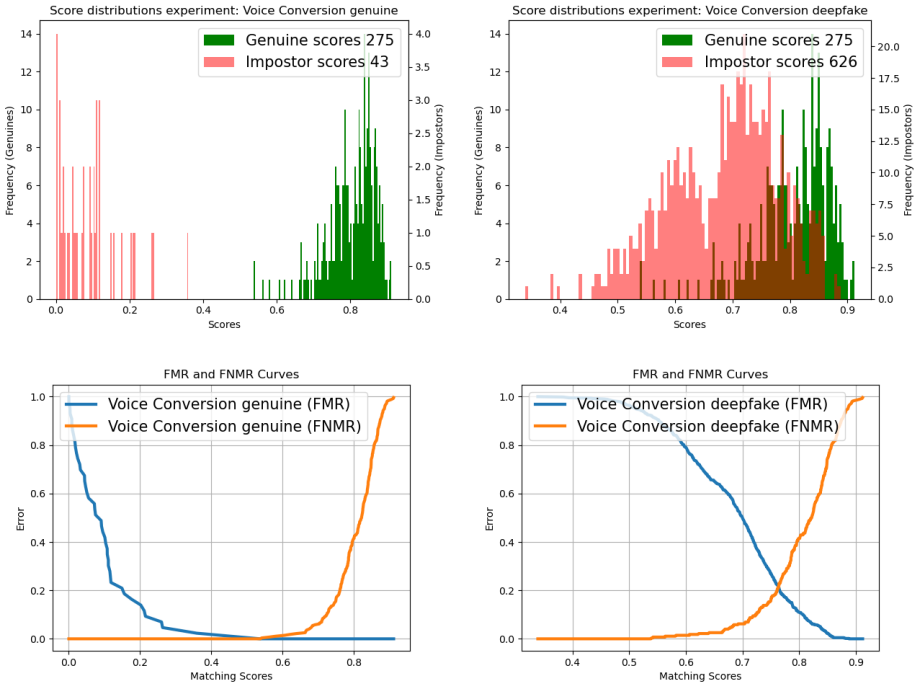

Figure 6.8: Matching scores distribution graphs (top) and FMR / FNMR graphs (bottom) for the Voice Conversion dataset. The genuine versus impostor distribution on the left, genuine and deepfake on the right.

Finally, Figure 6.8 shows matching scores distributions for English recordings from the Voice Conversion challenge dataset. The dataset contains only 4 speakers, so the gap between genuine and impostor distributions is quite large because of the low recordings count. The deepfake matching scores distribution overlaps approximately the lower half of genuine matching scores, but generally lays below. The EER increase is not so significant as with the previous datasets, but is still visible.

The comparison shows that the quality of the created Common Voice deepfake dataset is at the same level as the datasets created by people with more extensive knowledge of deepfake creation. Data gathered from all of the datasets implies that deepfakes indefinitely have the ability to reproduce genuine speech thus to spoof voice biometrics systems. The deepfakes of the highest quality can even reach matching scores higher than the genuine ones.

6.4.3 Mining data from matching scores

To get more extensive knowledge than just the fact that deepfakes are able to reproduce the characteristics of a genuine speech the verification results were used for data mining. Firstly, I decided to look for any link between the genuine matching scores and deepfake matching scores. As Table 6.13 shows, there is a correlation between genuine and deepfake matching scores. Also, there seems to be no link between the count of recordings used to fine-tune the synthesizer model, but the total length of the used recordings has some impact.

MS genuine	MS deepfake 1k	0.73532
MS genuine	MS deepfake 5k	0.75923
MS deepfake 1k	MS deepfake 5k	0.94408
MS deepfake 1k	recordings count	0.01030
MS deepfake 5k	recordings count	-0.01636
MS deepfake 1k	total recordings length	0.19257
MS deepfake 5k	total recordings length	0.20620

Table 6.13: Pairwise correlation table. MS stands for matching score.

Figure 6.9 again shows the correlation between genuine and deepfake matching scores, as well as that the higher the genuine matching score, the bigger the gap between the genuine and deepfake scores. Maximal deviance occurs within the highest interval of $[0.7 - 0.75]$, where the deepfake score drops by 11% in average. This deviance is minimal, and also in other direction, where at the lowest intervals the average deepfake matching scores are higher than the genuine ones.

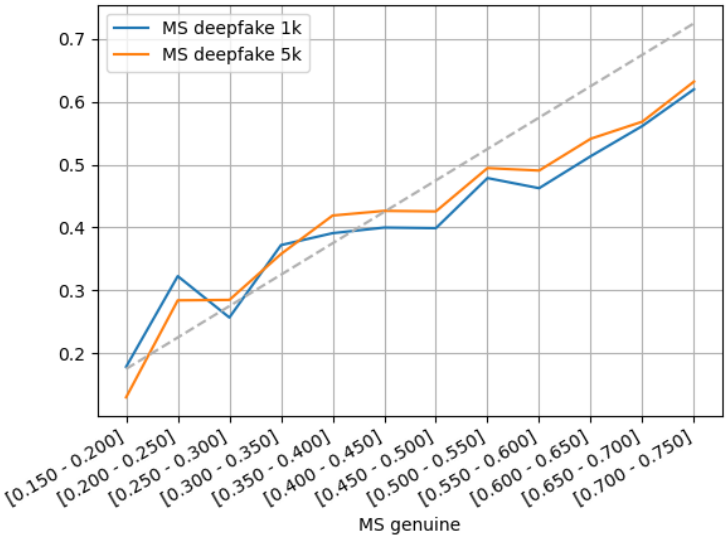

Figure 6.9: Average deepfake matching scores according to the average genuine matching score interval. Gray dashed line connects the middle of each genuine matching score interval with same deepfake matching score value. Lines above this line signalize that the average deepfake matching scores were higher than the genuine ones and vice versa.

Finally I examined the matching scores achieved for each synthesized sentence, Table 6.10 shows that there is minimal difference in average matching scores achieved for each sentence as well as minimal difference in the maximal achieved matching scores. This shows there is minimal to none correlation between the length of synthesized sentence and the achieved matching score as all of the sentences performed approximately the same in both average and maximal matching scores.

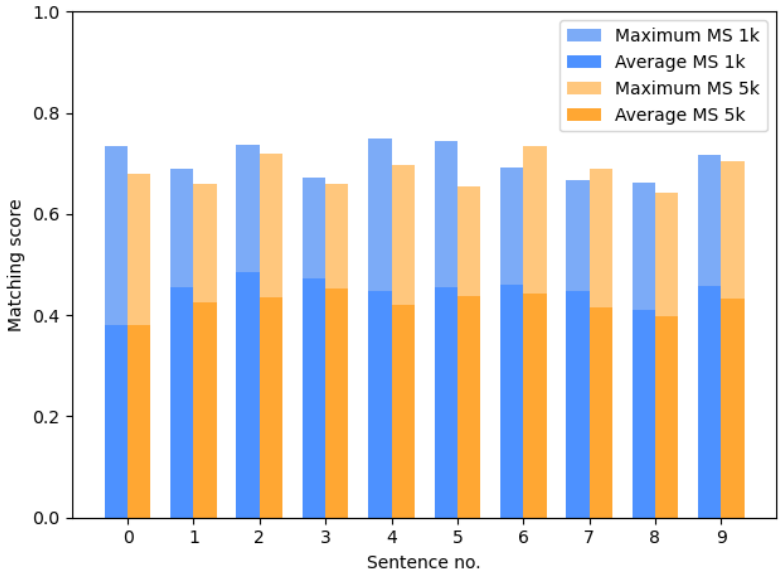

Figure 6.10: Average and maximal achieved matching score for each of the synthesized sentences. Transcriptions to the sentences can be found in Appendix B.

6.4.4 Experiment conclusion

The results of this experiment show that the deepfakes indeed have the ability to reproduce human speech. The created Common Voice deepfake dataset achieved matching scores similar to datasets containing quality deepfakes that are used to showcase voice conversion tools or to train deepfake speech detection systems. The data mining then revealed correlations between the achieved matching scores and properties of used data.

This experiment provided answers to all of the research questions asked during the design of this experiment (Section 5.1.5):

Do the deepfake matching scores distributions differ from the genuine matching scores distribution?

As the matching scores distribution plots show, there is a difference between the genuine and deepfake matching scores. However, there is a large variance between the shape of deepfake distributions from each dataset. The matching scores distribution for the deepfake 5k recordings from the Common Voice deepfake dataset is almost identical to the genuine one. There seems to be no pattern in how do the deepfake distributions look like, but it certainly is possible to reproduce the genuine matching scores distributions.

Are there any patterns in matching scores of deepfake recordings based on genuine matching scores?

As the results show, deepfake matching scores are correlated to the genuine ones. There is some discrepancy as for the speakers with low average genuine matching scores deepfake matching scores were higher, and vice versa for for the speakers with high matching scores, where deepfake matching scores were lower. This implies that the more quality data is used for training, the better the synthesized speech.

Are there any other patterns that might lead to detection or improvement of deepfake utterances?

Despite my expectation that the count of recordings used to train the synthesizer has impact on quality of synthesized speech, the results proved otherwise. However, the quality of synthesized speech depends on the amount of used training data in mean of speech length, not file quantity. The length of synthesized sentences again has no impact on the achieved matching scores. The results of this experiment did not discover any significant difference between genuine and deepfake speech in terms of matching scores that might be used to improve existing detection mechanisms. On the other hand the quality might be improved by gathering data in the best possible quality and quantity.

6.5 Resilience of text-dependent vs. text-independent verification

The differences in matching scores calculated for text-dependent and text-independent verification using Microsoft Speaker Recognition API in the first experiment (Section 5.1.5) suggests that the text-dependent verification might be more resilient against deepfakes.

6.5.1 Creating dataset for text-dependent verification

Microsoft Speaker Recognition API requires a special phrase, from a set of predefined phrases, to be used for text-dependent verification [30]. As mentioned during the design of this experiment (Section 5.1.5), no publicly available dataset consists of these phrases, so one must be created for purpose of this experiment.

I was able to collect recordings for 5 people, 4 male and 1 female. I recorded the voices using a custom created recording tool, that was also published on GitHub[6]. All of the recordings were recorded in same quiet environment, using the internal microphone of 2020 MacBook Pro M1. I decided to record all of the recordings using the same device to have recordings captured in exactly the same conditions, and also to brief every participant about how exactly the recordings of hers or his voice will be used and to provide assistance with the recording process. Unfortunately due to the current pandemic situation with Covid-19 I was able to collect recording only of 5 speakers.

After collecting all of the data, I used the training recordings to fine-tune the synthesizer models of the RTVC tool and synthesize deepfake recordings. The process was the same as when creating the dataset discussed in Section 6.4.

6.5.2 Experiment execution

To evaluate the data, I firstly enrolled a text-dependent profile for each speaker and recorded phrase and collected genuine matching scores as well as deepfake matching scores. As the Table 6.14 shows, there is a significant difference between genuine and deepfake matching scores for text-dependent verification.

[6] https://github.com/AntonFirc/react-voice-recorder

Verification type	Recording type	Avg. Matching score	Standard deviation
Dependent	Genuine	0.84660	0.03549
	Deepfake	0.53636	0.07242

Table 6.14: Average matching scores and standard deviation for text-dependent verification using genuine and deepfake recordings. Matching scores were calculated using Microsoft Speaker Recognition API.

Next, I moved onto text-independent verification, where the profile was enrolled using randomly selected phrase (same for each speaker) from the training set. As the Table 6.15 shows, there is much smaller difference between the matching scores for genuine and deepfake recordings. The genuine matching scores were a bit lower, as well as the deepfake matching scores were higher than with the text-dependent verification.

Verification type	Recording type	Avg. Matching score	Standard deviation
Independent	Genuine	0.67112	0.11360
	Deepfake	0.60283	0.10636

Table 6.15: Average matching scores and standard deviation for text-independent verification using genuine and deepfake recordings. Matching scores were calculated using Microsoft Speaker Recognition API.

The complete comparison is shown in Figure 6.11. The drop in genuine matching scores between text-dependent and text-independent verification as well as the increase in deepfake matching score between these two is visible as mentioned before. The text-independent genuine and deepfake matching scores are much more similar than the text-dependent ones.

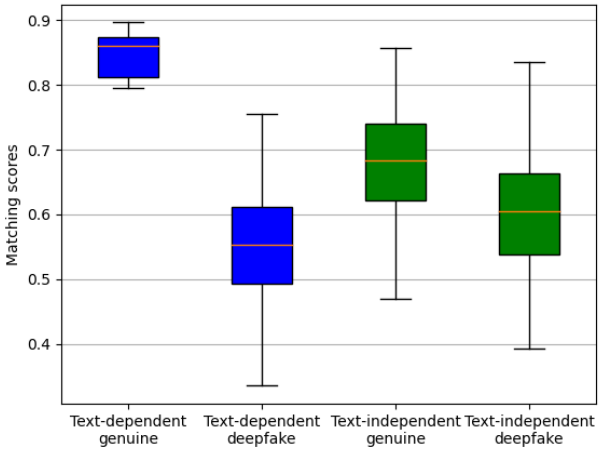

Figure 6.11: Comparison of scores calculated for text-dependent and text-independent verification using the same genuine and deepfake recordings for both verification types.

Figure 6.12 shows the average genuine and deepfake matching scores by each phrase, as well as that there is no correlation between phrase length and calculated matching scores.

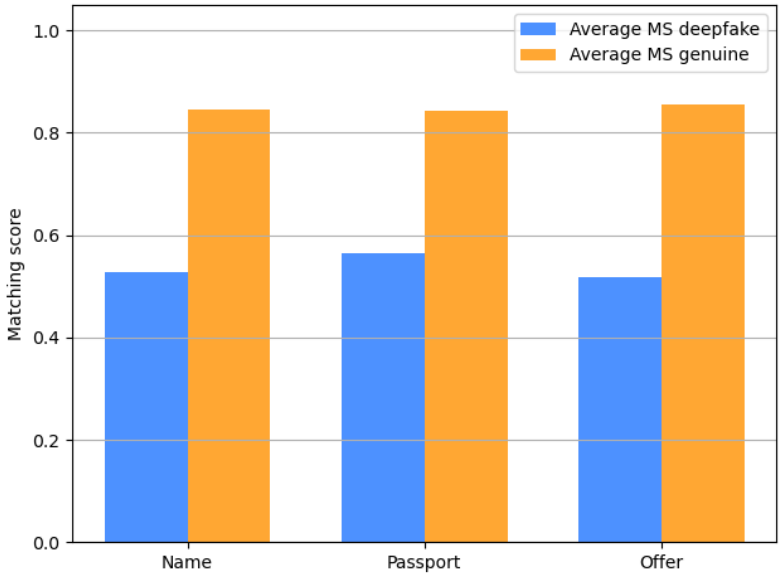

Figure 6.12: Average genuine and deepfake matching score by phrase. Phrases are ordered by length ascending, from left to right.

6.5.3 Experiment conclusion

Even though the used dataset consists only of five distinct speakers, the collected data was enough to show that the difference between text-dependent and text-independent verification when dealing with deepfakes is not a random event. The results show that the text-dependent verification is more resilient to deepfakes than text-independent verification. However to completely prove this hypothesis a more extensive research needs to be carried out, with a much wider spectrum of speakers and voice biometrics systems.

The experiment answered the research question asked during the design of this experiment (Section 5.1.5):

Is text-dependent verification harder to breach using deepfakes than text-independent verification?

As the results have shown, the calculated deepfake matching scores differ vastly from the genuine ones. This difference is vanishing when using text-independent verification. That means, it is much easier to reproduce the matching scores of text-independent verification, what puts text-independent verification into a position of the weaker one when dealing with deepfakes.

6.6 Text-to-speech synthesis for Czech language

As the majority of the research in field of text-to-speech synthesis and deepfakes is carried out and presented in English language, almost all of the tools and pretrained models can handle English language exclusively. Of course, there solutions from big players like Google

or Amazon that are capable of synthesizing very naturally sounding speech in a variety of languages, but the technical background and models are kept private and also there is no possibility to adjust the speaker to meet specific needs of a potential attacker. In order to be able to clone a voice usable in a case of voice phishing or identity theft we need to use a tool that is capable of generating speech of any speaker with as few training samples as possible.

For this purpose the Real Time Voice cloning tool (Section 2.4.2) provides wanted functionality as it can synthesize speech based on a very short recording. This section discusses the execution of experiment proposed in Section 5.2, to train a speech synthesis model for Czech language. Firstly, the used dataset is introduced, then training of each part of the RTVC framework is discussed and finally conclusion on the findings is provided. The training process followed the GitHub wiki [8] on training the tool from scratch for any language.

6.6.1 Used dataset

The most important part of training a TTS model is a quality dataset. The major problem is lack of usable data in Czech language. This problem was partially solved by the *Common Voice* initiative by Mozilla. Their latest *Common Voice Corpus* 6.1 contains 40 hours of speech in Czech language with transcriptions, of which 36 hours are validated, which means that the clip was reviewed by community and the recording matches its transcription in an adequate quality [1].

The 302 speakers and 36 hours of transcribed audio do not provide enough data for the encoder model training, but fortunately there were datasets containing recordings of other languages similar to Czech or Slovak language. The complete list of languages used during the training and approximated lengths and speaker count is shown in Table 6.16.

Language	Total length (hr)	Validated length (hr)	Speaker count
Czech	45	36	302
Polish	129	108	1914
Russian	130	111	860
Slovenian	7	5	36
Ukrainian	43	30	342
Total	**354**	**290**	**3454**

Table 6.16: Table showing languages used during RTVC tool training, total length available, validated length and approximate speaker count after cleansing and grouping the data by speaker hash.

6.6.2 Training encoder model

For the encoder training, recordings have to be divided into folders, where each folder represents one speaker. The RTVC tool itself does not provide support for the Common Voice dataset, in terms of data parsing, however the discussions in the GitHub repository issues section contains threads with all the info and codes needed to transform this dataset

to usable form[7] as well as recommendations on the encoder model settings for training on language different to English[8].

After the extra step of data preparation, and addition of methods to preprocess the Common Voice dataset, training follows the basic routine as described in the repository wiki page [8].

Figure 6.13 shows the progress of encoder training from the beginning to the end after more than 32k iterations, what took around one week including data preparation tasks using one NVIDIA Tesla T4 GPU. The training was finished after the reported loss stopped to change. The converged loss function as well as tight clusters seen in UMAP projections of speakers show that the encoder training process was successful.

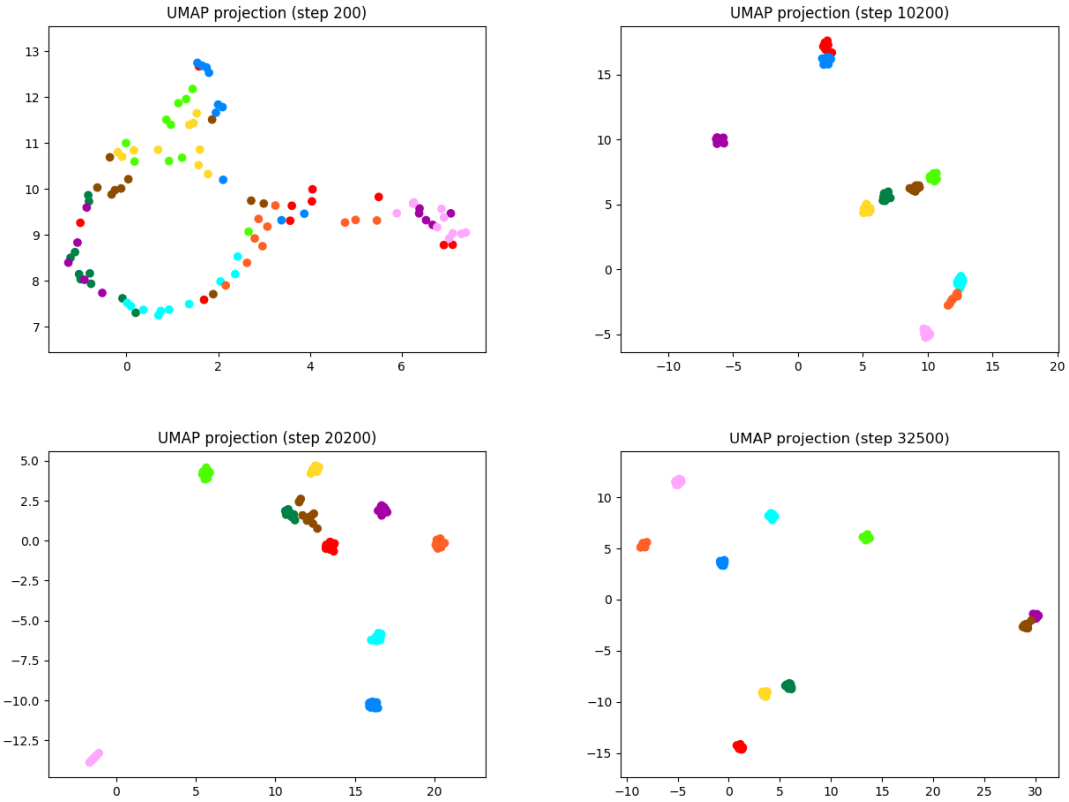

Figure 6.13: Evolution of UMAP projections of speaker clustering generated during the training. Each projection contains different speakers, thus there is no connection between these projections. Fast convergence can be observed in the early stages of the training, and slower convergence during the later iterations of the training. The last projection is the closest one to the final step of training. As the encoder part of the RTVC tool extracts features for speakers, it has to be able to differentiate between speakers, thus speakers should form tight clusters.

[7] https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/458
[8] https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/126

6.6.3 Training synthesizer model

The synthesizer training process again required additional data preparation. Apart from grouping recordings by speaker, each recording needs a transcription provided. The transcriptions in Common Voice dataset are provided in a CSV file, and they are linked to the recordings by the recording file name. The synthesizer preprocessing tool of the RTVC tool needs the transcript to be present in a text file with the same name as corresponding recording. This dataset modification was achieved by a slight modification of the data preparation scripts from the encoder training process.

The performance of synthesizer training process can be measured in two ways: the loss that is calculated in each iteration, and the loss calculated during the evaluation phase. Both of the loss function indicate the difference between the predicted and ground-truth frame of the mel-spectrogram. However, the prediction in the training is based on the ground-truth data, while in the evaluation phase whole mel-spectrogram is predicted from the beginning. This difference in prediction makes the loss reported during the training way better than the model really is. This difference is shown in Figure 6.14.

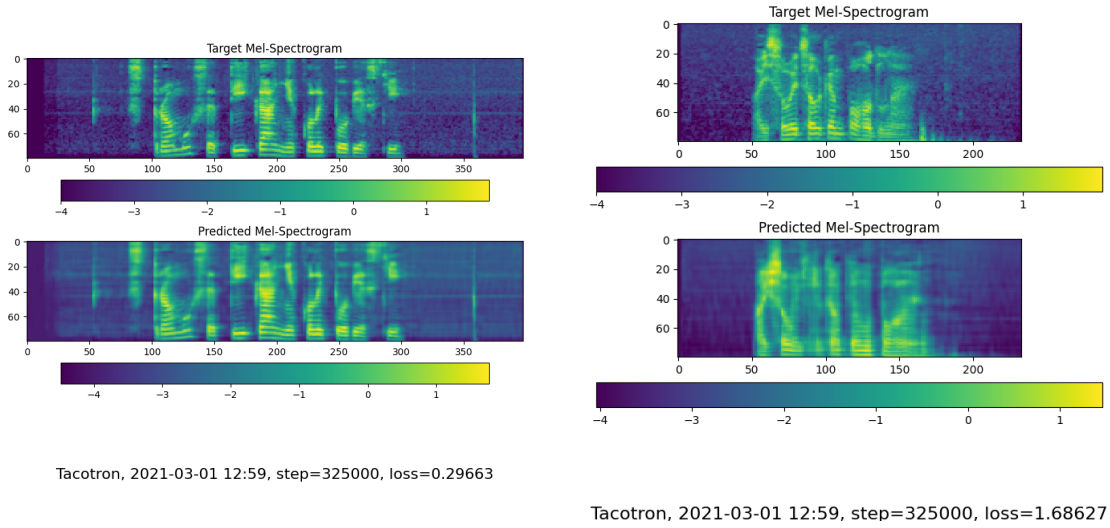

Figure 6.14: Difference in the mel-spectrograms and loss calculated during the training (left) and the evaluation phase (right). The prediction of next frame during training is based on the ground-truth data, while the prediction during the evaluation phase is based completely on the previously predicted frames.

The discussions in the issues section and showcase of pretrained models[9] suggest that 300k iterations should be just enough to train sufficient synthesizer model. Following this advice I stopped the synthesizer model training at 325k iterations. As Figure 6.15 shows the loss calculated during the evaluation phases was decreasing at the beginning, but at a certain point it started to grow. I experimented with different model settings as well as silence removal int he input data. Surprisingly, I achieved the best results using recordings without silence trimming. The whole process of synthesizer training took around 14 days using one NVIDIA Tesla T4 GPU.

[9] https://blue-fish.github.io/experiments/RTVC-5.html

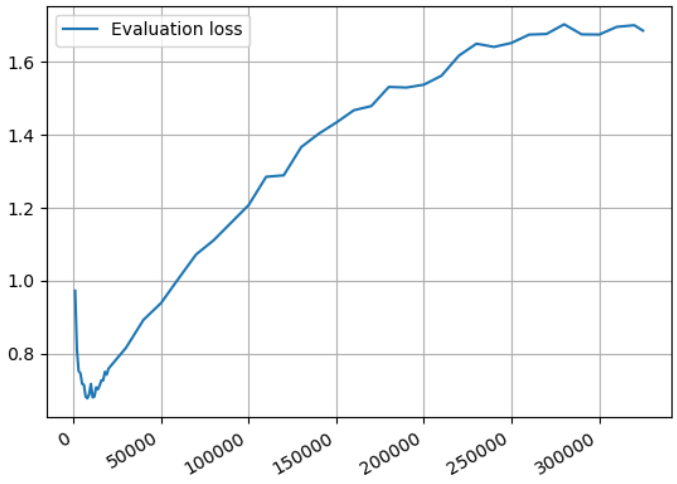

Figure 6.15: Loss calculated during the evaluation phase of synthesizer model training.

6.6.4 Training vocoder model

Vocoder is the only language-independent part of the RTVC tool [7]. As the vocoder training process is time consuming, as well as the impact of the vocoder model on the reproductive ability is minimal, I decided to use the pretrained one without any further modifications.

6.6.5 Results

The loss reported from the evaluation phases during the synthesizer training indicated that the synthesized speech will be of dubious quality. The first synthesized samples were a bit surprising, as the speech came out better than expected. I expected the voice to be very deep, and the speech to contain a lot of pauses and other glitches. However, the synthesized speech matched the voice of the target speaker quite accurately, but the words were mostly mumbled. From the clear parts it was fairly easy to tell that the phrase is spoken in Czech language.

The Figure 6.16 shows that the synthesized speech, or rather voice because of the mumbled sections, quite fairly reproduces the target speakers voices. The matching score distributions of deepfake and genuine recordings mostly overlap, and the EER vastly increased.

Because of the mumbled sections of the speech, the synthetic speech was immediately distinguishable from the genuine one for humans. However, the verification results show that the voice biometrics systems, at least the Microsoft Speaker Recognition API, analyze the voice characteristics and no the speech as a whole. If such a system was used to provide authentication without any human factor in the authentication process, it might represent a considerable security risk. The text-dependent verification might again be more resilient, as the spoken phrase has to be examined. Unfortunately this statement still remains a hypothesis, as I was unable to test any voice biometrics system that provides the text-dependent verification for Czech language.

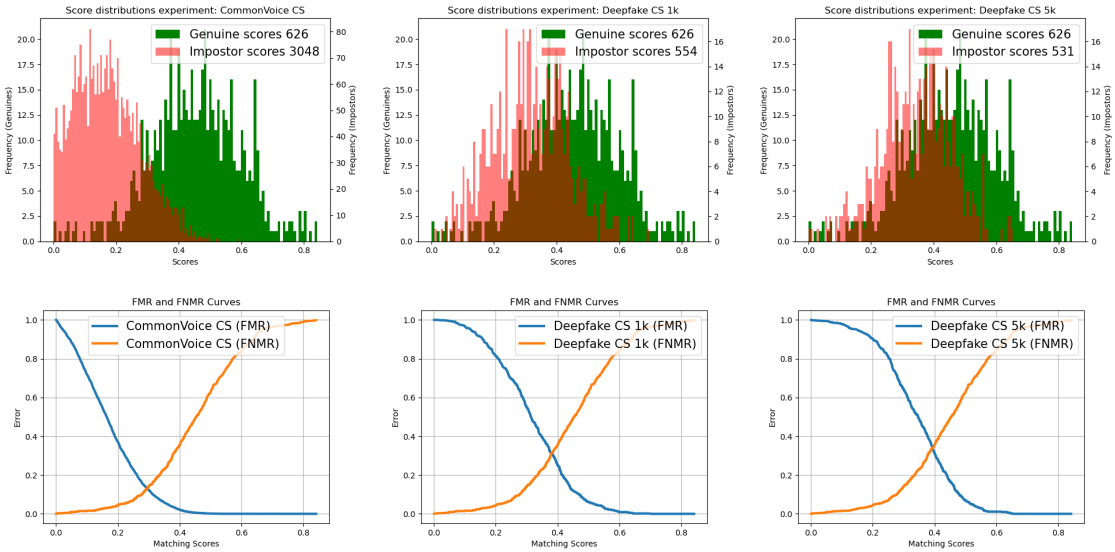

Figure 6.16: Matching score distribution graphs on top. From left to right a genuine/impostor distribution, deepfake 1k and genuine distribution and deepfake 5k and genuine distribution. The deepfake and genuine matching scores distributions are mostly overlapping, which shows that the synthesized voices are similar to the genuine ones. FMR and FNMR curves on bottom, each plot corresponds to a distributions plotted above. A large increase in EER can be seen when comparing genuine/impostor distribution with the deepfake distributions. Distributions and FMR and FNMR curves were plotted using matching scores calculated by Microsoft Speaker Recognition API.

6.6.6 Experiment conclusion

This experiment discussed the training of a text-to-speech synthesis tool for a Czech language from scratch. As shown, gathering enough data even in Czech language is now quite simple as a lot of datasets with speech and transcriptions exist and datasets such as Common Voice provide even very large variety of languages. The process of training the synthesizer model is the most demanding part of the whole process, and has major impact on final quality. The process itself requires more knowledge than just synthesizing speech and also much more time to go trough the trial and error process. All research questions asked during the design of this experiment (Section 5.2) can be answered:

Is it technically feasible to train TTS model for Czech language from scratch?

As shown in this experiment, training the TTS synthesis model for Czech language is definitely technically feasible. The training for any language might be generalized to obtaining data of selected language from public datasets, preprocessing the audio such as noise and silence removal and finally training the TTS model. Training the model from scratch does not require any different knowledge that is needed to fine-tune such a model, only more extensive amount of time is needed.

Is there enough usable data available for training TTS model for Czech language?

The experiment has shown that there is already enough processed data, suitable for TTS related tasks, publicly available for Czech language. Also, the number of datasets for Czech language as well as other languages will surely grow overtime. Apart from already made datasets, social medias, YouTube or podcast provide huge number of recordings in Czech language. This audio can be transcribed using some of the online services to create a custom dataset. This option requires more time and effort, to correctly split the sentences and provide corresponding transcripts. However with the actual state of the technology it is easily achievable task.

How much time it takes to train TTS model for Czech language?

The time needed to train TTS model depends on many variables, for example: the technical knowledge of an individual, the computational resources available or the quality of data. The knowledge of an individual with the computational resources can be pointed out as the major factors. If an individual knows how does the TTS system work, and is able to use it, she or he is able to prepare the data and train model in relatively short amount of time. This time then depends on the available computational resources. The data quality also plays a role. If the data is of a low quality, more time has to be spent processing it into a usable form. To provide an exact answer, the process of training the TTS model described in this section, computational time and time spent working with the tool other than learning, took around 20 days, while the most time-depleting task was the synthesizer model training that took almost 14 days.

What is the final quality of synthesized speech?

The quality of synthesized speech might be viewed from two points of view. One point of view is the ability of synthesized speech to fool a human, the other point of view is the ability to fool voice biometrics system. Considering the first point of view, ability to fool human, the synthesized speech is of a low quality. Anyone would be able to tell that there is something wrong with the speech even without any awareness about deepfakes and speech synthesis as the words are mumbled and inaudible sometimes. From the second point of view, ability to spoof voice biometrics systems, the synthesized speech has better quality. As the matching scores distribution plots and FMR and FNMR plots have shown, the synthesized speech can get verified by a voice biometrics system as genuine speech.

Do TTS tools currently present a threat to Czech or Slovak speaking people?

The findings of this experiment show that the current state of text-to-speech synthesis definitely present a threat to languages other than English, such as Czech or Slovak. Even though the threat is not yet as acute as with the English language, with the right knowledge and data synthesis of Czech language in decent quality is possible. This threat will rise as the TTS technology will advance and become usable to wider range of users as well as more models and datasets will be published.

6.7 Speaker similarity survey

The voice biometrics system might not be the only factor that is involved in the authentication process. An attack vector proposed in Section 5.3 uses the synthesized speech to communicate with a call centre operator and initiate unauthorized action on behalf of the victim. In order to successfully perform such an attack vector it is required not only to fool the biometrics system, but the operator as well. To further examine whether the synthesized speech used in experiments discussed before is able to spoof humans an survey was created.

As stated during the experiment design, the survey puts respondents into a role of voice biometrics system. The survey consists of 10 speakers, and for each speaker there are three recordings: one genuine and two deepfake. The recordings were selected from the dataset created during the experiments discussed in Section 6.4, to be completely aware of all of the aspects of creating these recordings. Survey was published in two versions, one with instruction in Czech language and the second one in English language, the content was exactly the same.

6.7.1 Survey findings overview

Survey was published at the beginning of March 2021 on the Google Forms platform. During the period of one month exactly 100 responses were collected while the Czech : English response ratio is 1 : 2.

To get general insight on how well the deepfake recordings mimicked the target speakers false accept and false reject rates were calculated. As the Table 6.17 show, both FAR and FRR are quite high. Almost every third deepfake attempt was successful, as well as more than one third of genuine attempts were rejected. As expected, these rates were more favorable for the English respondents, as the better knowledge of the language helps to spot the synthesized speech. To examine whether there is a link between the matching scores of recordings and the false accept rate, matching scores for the used recordings were calculated and finally a correlation between the matching scores and accept rates. As the Table 6.18 shows, the deepfake matching scores are correlated with both of the accept rates.

	both	English	Czech
FAR (%)	30.67	27.29	36.57
FRR (%)	39.06	37.04	42.57

Table 6.17: FAR and FRR calculated from the survey results.

	Genuine	Deepfake 1k	Deepfake 5k
FAR	-0.26135	0.45173	0.38530
FRR	-0.22226	0.44963	0.44963

Table 6.18: Correlation between FAR, FRR and utterance matching scores. Correlation is calculated from FAR, FRR and matching scores of recordings for each speaker.

Next, I decided to examine the relation between age and sex, and the false accept and reject rates. As the Figure 6.17 shows, the error rates are correlated with age, as well

as they are influenced by sex of the respondent. For each sex and age interval there is a significant difference in the error rates, the only exception is the false reject rate that is almost similar for the age interval of 51 y.o. and older.

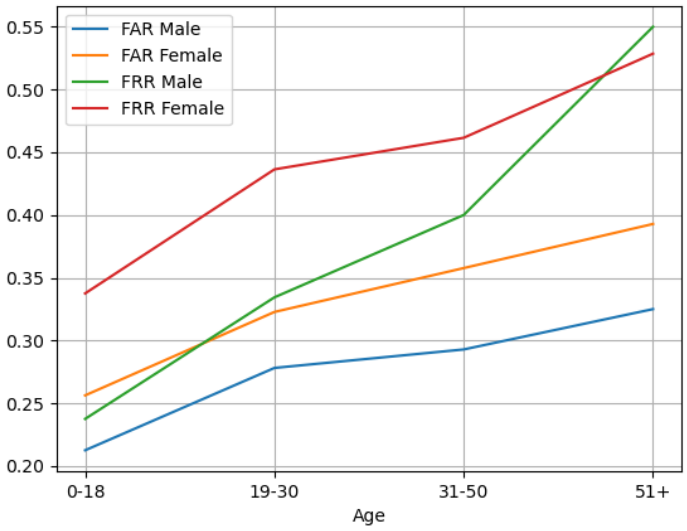

Figure 6.17: FAR and FRR values difference according to sex of the respondent.

Speaker number 5 achieved the highest FAR with more than 83%, this was caused by quality deepfake recordings along with a low quality genuine recording. On the other hand, speaker number 3 achieved the lowest FAR just below 19%, this was most likely caused by high matching score of the genuine recording and low matching scores (below 45%) for the deepfake recordings. These results imply that the respondents gained information for the decision process also from the comparison of the recordings. I believe that the false accept rates for each speaker would increase if each verification attempt was really independent, there would be no other attempts to compare with, for the person entrusted to decide the genuineness of verification attempts.

6.7.2 Experiment conclusion

The last experiment examined the believability of deepfakes synthesized during this research to humans. Relations between the matching scores of the recordings, age and sex of the respondents and the final results were examined. The results show that humans have no special ability to distinguish between real and synthesized speech if the synthesized speech is of a decent quality.

All of the research question asked during the design of experiment (Section 5.3) can be answered:

Is there any link between matching scores and believability to human of deepfake utterances?

The collected responses show that there is not negligible correlation between the matching scores calculated for both genuine and deepfake recordings. High genuine matching score

lowered the FAR and FRR rates, while high deepfake matching score increased these rates. This correlation implies that higher the matching score, the more naturally and believably the recording sounds to a human.

How many times genuine user was denied access, and how many times impostor was allowed access?

As the results show, the error rates of false accept and false reject were high. Approximately every third deepfake recording was accepted as genuine, while the rate of rejecting genuine recordings was even higher where more than 40% of the genuine recordings were denied.

Is the collected data usable to support or refute the human ability of spotting a deepfake speech.

The collected data show that it is hard for people to distinguish between real and synthesized speech. The speech of course has to be of a decent quality, but as the results show the denial of deepfake recordings and acceptance of the genuine might be considered as random. The collected results definitely support the hypothesis that humans cannot distinguish fake speech from real.

Chapter 7

Conclusion

The fact that deepfake technology is just on the rise and it has already caused commotion in the medias and common life, leads us to belief that mankind should keep an eye on it and use it wisely otherwise great harm can be caused. There are many available tools and techniques for deepfake creation ranging from the most simple ones presented as smartphone applications to the more complicated ones state-of-the-art techniques with implementation, so called *papers with code*.

The statement that everybody can create deepfakes now is in general true, however there are some obstacles. Swapping your face with celebrity in sequences of your favourite movie using your smartphone is really an activity almost anyone can do, but the final quality and usability of deepfake of this kind is none. The more advanced tools require more knowledge and produce more believable results on almost any kind of input data, i.e. speech, video, text and much more.

Most of the focus is oriented towards video deepfakes and face swapping, even communities of people creating this type of deepfakes exist. On the other hand, least of the focus is oriented towards fictitious image generation (Section 2.5.1) or text generation (Section 2.5.2), as the usability in the cybersecurity field is low as well as public attractivity of this kind of media. Voice deepfakes find their place in the middle, more tools, even commercial ones exist, but majority of these systems is used for text-to-speech synthesis with application for example in navigation systems or smart assistants.

Deepfakes can be used in many ways, both illicit and legitimate. The illicit usage of deepfakes has almost all of the attention dragged to itself. These usages range from fabricating false events or defaming individuals to identity theft or extortion. On the other hand, there are areas that might profit from the deepfake usage such as education, healthcare or entertainment.

The illicit usages yield the need for detection techniques and systems. There are two ways of detecting a deepfake, one is doing it personally and identifying it based on the visual or audio properties, while the second one relies purely on the machines and the implemented detection engines. The quality of deepfakes nowadays is still not perfect, so looking for artifacts in the video or facial features like glasses, hair or beard might reveal inconsistencies typical for deepfakes that may lead to detection.

This research more closely examines the threats that voice deepfakes present to voice authentication. The results of proposed and executed experiments show that deepfakes present a serious threat to voice biometrics systems as well as to people.

The results of experiments discussed in Section 6.3 show that systems that do not implement any kind of liveness detection are easily spoofed. Unfortunately no system

implementing liveness detection was tested, thus no conclusion on whether the liveness detection improves the provided security of authentication system may be done.

The majority of text-to-speech related tools and datasets is intended for the English language exclusively. This may may lead to an idea that using voice authentication in language different to English provides increased security. To examine this hypothesis I trained my own TTS model for Czech language (Section 6.6). As the results of this experiment shown, training own TTS model for selected language is possible even with no extensive knowledge of speech synthesis. This means that using voice authentication in language different to English does not imply safety from deepfakes.

The last executed experiment (Section 6.7) has shown that adding human factor to the authentication process does not ensure improvement in security. The performed survey has shown that in average every third deepfake verification attempt was accepted by humans, while in average every second genuine verification attempt was rejected. These results imply that there is little to none difference on how do people process the deepfake and genuine speech, which means they can be easily spoofed.

Considering all three parts of a real-world attack on voice biometrics systems using deepfake speech, the easiest one is to synthesize speech that is able to spoof voice biometrics systems. Synthesis of speech in language different to English in a decent quality is the most demanding part. The speech has to be able to fool humans and also to converse with them. The tested voice biometrics systems authenticated even the recordings that contained mumbled or inaudible speech, that any human would instantly mark as suspicious. The combination of these parts thus increases the complexity of such an attack to the level of the most demanding part, synthesizing speech in selected language that is believable to humans.

A mean to mitigate the threat posed by deepfakes to voice authentication was discovered during this work. As shown in Section 6.5, text-dependent verification presents more secure way of voice verification. Even though the experiment was executed with a small sample of speakers, it has proven that this is not a random event. To further explore the differences in security provided by text-dependent and text-independent verification, much broader spectrum of speakers is needed.

A deepfake dataset containing Czech and English speech as a subset of the Common Voice Corpus was published[1].

To further extens the scope of this research, more speech synthesis and voice conversion tools have to be tested as well as more voice biometrics systems. The attack vectors should be evaluated in a real-world conditions that exactly simulate the conditions at which the attacks might be carried out.

In summary, this work has shown that deepfakes do present a serious threat to voice authentication in any language. None of the tested voice biometrics systems was able to detect usage of deepfake speech and accepted it. People as well did not perform better in distinguishing between real and fake speech. Fortunately a mean to mitigate the threat deepfakes present to voice biometrics systems was discovered, by using text-dependent verification over the text-independent. Finally, areas allowing future research were presented.

[1] https://drive.google.com/drive/folders/1vlR-TA7gjKzjYylxzRnA_HzZEyWiLeOk?usp=sharing

Bibliography

[1] Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J. et al. Common Voice: A Massively-Multilingual Speech Corpus. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, p. 4218–4222. ISBN 979-10-95546-34-4. Available at: https://www.aclweb.org/anthology/2020.lrec-1.520.

[2] Bahaa Eldin, A. M. A medium resolution fingerprint matching system. *Ain Shams Engineering Journal*. 2013, vol. 4, no. 3, p. 393–408. DOI: https://doi.org/10.1016/j.asej.2012.10.001. ISSN 2090-4479. Available at: https://www.sciencedirect.com/science/article/pii/S2090447912000895.

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

[4] Breland, A. *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink* [online]. 2019 [cit. 2020-12-26]. Available at: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

[5] Cavalli, F. *How to detect a deepfake online* [online]. 2021 [cit. 2021-3-10]. Available at: https://sensity.ai/how-to-detect-a-deepfake/.

[6] Cho, K., Courville, A. and Bengio, Y. Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks. *IEEE Transactions on Multimedia*. 2015, vol. 17, no. 11, p. 1875–1886. DOI: 10.1109/TMM.2015.2477044.

[7] Corentin, J. *Real-time Voice Cloning*. Liège, Belgique, 2019. Master thesis. Université de Liège, Liège, Belgique. Available at: https://matheo.uliege.be/handle/2268.2/6801?locale=en.

[8] Corentin, J. *Real-time Voice Cloning - Wiki page*. GitHub, 2019. Available at: https://github.com/CorentinJ/Real-Time-Voice-Cloning/wiki.

[9] Descript. *Descript webpage* [online]. 2020 [cit. 2020-11-30]. Available at: https://www.descript.com.

[10] Dobbie, S. Challenge of biometric security for banks. *Biometric Technology Today*. 2020, vol. 2020, no. 3, p. 5 – 7. DOI: https://doi.org/10.1016/S0969-4765(20)30037-0. ISSN 0969-4765. Available at: http://www.sciencedirect.com/science/article/pii/S0969476520300370.

[11] Gary Grossman, E. Deepfakes may not have upended the 2020 U.S. election, but their day is coming. [online]. 2020, [cit. 2021-1-5]. Available at: https://venturebeat.com/2020/11/01/deepfakes-may-not-have-upended-the-2020-u-s-election-but-their-day-is-coming/.

[12] Ghosh, S. *Are You Confident About Distinguishing Between a Computer-Generated Voice and Human Voice?* [online]. 2019 [cit. 2020-12-20]. Available at: https://aithority.com/ait-featured-posts/are-you-confident-about-distinguishing-between-a-computer-generated-voice-and-human-voice/.

[13] Groh, M. *Detect DeepFakes: How to counteract misinformation created by AI* [online]. 2020 [cit. 2021-1-3]. Available at: https://www.media.mit.edu/projects/detect-fakes/overview/.

[14] ID R&D. *Combat Voice Spoofing Attacks* [online]. 2021 [cit. 2021-4-15]. Available at: https://www.idrnd.ai/voice-anti-spoofing/.

[15] Ito, K. and Johnson, L. *The LJ Speech Dataset* [online]. 2017 [cit. 2021-4-2]. Available at: https://keithito.com/LJ-Speech-Dataset/.

[16] Jefferson, E. *Are voice biometrics the new passwords?* [online]. 2020 [cit. 2020-12-26]. Available at: https://www.raconteur.net/technology/cybersecurity/voice-biometrics/.

[17] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J. et al. *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis*. 2019.

[18] Johansen, A. G. *How to spot deepfake videos — 15 signs to watch for* [online]. 2020 [cit. 2021-1-3]. Available at: https://us.norton.com/internetsecurity-emerging-threats-how-to-spot-deepfakes.html.

[19] Johnson, D. *Audio Deepfakes: Can Anyone Tell If They're Fake?* [online]. 2020 [cit. 2021-2-3]. Available at: https://www.howtogeek.com/682865/audio-deepfakes-can-anyone-tell-if-they-are-fake/.

[20] Jones, R. *Voice recognition: is it really as secure as it sounds?* [online]. 2018 [cit. 2020-12-20]. Available at: https://www.theguardian.com/money/2018/sep/22/voice-recognition-is-it-really-as-secure-as-it-sounds.

[21] Jones, V. A. *Artificial Intelligence Enabled - Deepfake technology The Emerge of a New Threat*. 2020. Master thesis. Utica College.

[22] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. et al. *Training Generative Adversarial Networks with Limited Data*. 2020.

[23] Karras, T., Laine, S. and Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *CoRR*. 2018, abs/1812.04948. Available at: http://arxiv.org/abs/1812.04948.

[24] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. et al. Analyzing and Improving the Image Quality of StyleGAN. *CoRR*. 2019, abs/1912.04958. Available at: http://arxiv.org/abs/1912.04958.

[25] Kwok, A. O. J. and Koh, S. G. M. Deepfake: a social construction of technology perspective. *Current Issues in Tourism*. Routledge. 2020, vol. 0, no. 0, p. 1–5. DOI: 10.1080/13683500.2020.1738357. Available at: https://doi.org/10.1080/13683500.2020.1738357.

[26] Lomas, N. *Are You Confident About Distinguishing Between a Computer-Generated Voice and Human Voice?* [online]. 2020 [cit. 2020-12-31]. Available at: https://techcrunch.com/2020/08/17/deepfake-video-app-reface-is-just-getting-started-on-shapeshifting-selfie-culture.

[27] Mak, T. and Temple Raston, D. Where Are The Deepfakes In This Presidential Election? [online]. 2020, [cit. 2021-1-5]. Available at: https://www.npr.org/2020/10/01/918223033/where-are-the-deepfakes-in-this-presidential-election?t=1609838586916.

[28] Maras, M.-H. and Alexandrou, A. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*. 2019, vol. 23, no. 3, p. 255–262. DOI: 10.1177/1365712718807226. Available at: https://doi.org/10.1177/1365712718807226.

[29] Markowitz, J. A. Designing for Speaker. In: Dahl, D., ed. *Practical Spoken Dialog Systems*. Dordrecht: Springer Netherlands, 2004, p. 123–139. ISBN 978-1-4020-2676-8. Available at: https://doi.org/10.1007/978-1-4020-2676-8_7.

[30] Microsoft. *About the Speech SDK* [online]. 2020 [cit. 2021-2-3]. Available at: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview#speaker-verification.

[31] Microsoft. *Text to Speech* [online]. 2021 [cit. 2021-5-3]. Available at: https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/.

[32] Mirsky, Y. and Lee, W. *The Creation and Detection of Deepfakes: A Survey*. 2020.

[33] Mitro, M. *Deepfake videá a fotky už odhalí každý. Zverejnili unikátny nástroj, vyskúšali sme ho* [online]. 2021 [cit. 2021-3-10]. Available at: https://fontech.startitup.sk/deepfake-videa-a-fotky-uz-odhali-kazdy-zverejnili-zaujimavy-nastroj-vyskusali-sme-ho/.

[34] Nagrani, A., Chung, J. S., Xie, W. and Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*. 2020, vol. 60, p. 101027. DOI: https://doi.org/10.1016/j.csl.2019.101027. ISSN 0885-2308. Available at: https://www.sciencedirect.com/science/article/pii/S0885230819302712.

[35] National Academies of Sciences, Engineering, and Medicine. *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop*. Washington, DC: The National Academies Press, 2019. ISBN 978-0-309-49450-2. Available at: https://www.nap.edu/catalog/25488/implications-of-artificial-intelligence-for-cybersecurity-proceedings-of-a-workshop.

[36] Neupane, A., Saxena, N., Hirshfield, L. and Bratt, S. E. *The Crux of Voice (In)Security: A Brain Study of Speaker Legitimacy Detection*. 2019. Available at:

https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_08-3_Neupane_paper.pdf.

[37] Nguyen, Q. *Penetration tests of verification system.* Brno, 2020. Master thesis. Brno University of Technology, Faculty of Information Technology. Available at: https://www.vutbr.cz/en/students/final-thesis/detail/129892.

[38] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T. and Nahavandi, S. *Deep Learning for Deepfakes Creation and Detection: A Survey.* 2020.

[39] OpenAI. *OpenAI API* [online]. 2020 [cit. 2021-1-5]. Available at: https://openai.com/blog/openai-api/.

[40] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2015, p. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[41] Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S. et al. *DeepFaceLab: A simple, flexible and extensible face swapping framework.* 2020.

[42] Petersen, J. *Combating deepfakes with voice biometric technology* [online]. 2019 [cit. 2020-12-23]. Available at: https://www.techradar.com/news/combating-deepfakes-with-voice-biometric-technology.

[43] Phani8. *What is the Convolutional Neural Network Architecture?* [online]. 2020 [cit. 2021-5-1]. Available at: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/.

[44] Phonexia. *Phonexia Voice Verify* [online]. 2021 [cit. 2021-4-15]. Available at: https://www.phonexia.com/en/product/voice-verify/.

[45] Precise Biometrics AB. *Understanding biometric performance evaluation* [online]. 2014 [cit. 2021-02-15]. Available at: https://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation-QR.pdf.

[46] Qian, K., Zhang, Y., Chang, S., Cox, D. and Hasegawa Johnson, M. *Unsupervised Speech Decomposition via Triple Information Bottleneck.* 2020.

[47] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. *Improving language understanding by generative pre-training.* 2018.

[48] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. et al. Language models are unsupervised multitask learners. *OpenAI blog.* 2019, vol. 1, no. 8, p. 9.

[49] Rahim, Z. *'Deepfake' Queen delivers alternative Christmas speech, in warning about misinformation* [online]. 2020 [cit. 2020-12-28]. Available at: https://edition.cnn.com/2020/12/25/uk/deepfake-queen-speech-christmas-intl-gbr/index.html.

[50] *Single speaker fine-tuning process and results* [Real-Time-Voice-Cloning GitHub]. 2020. [Online]. Available at: https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/437.

[51] Resemble AI. *Resemble AI webpage* [online]. 2020 [cit. 2020-11-30]. Available at: https://www.resemble.ai.

[52] Schwartz, E. H. *Deepfake Security Concerns Are Limiting Voice ID Adoption: Survey* [online]. 2019 [cit. 2020-12-20]. Available at: https://voicebot.ai/2019/12/19/deepfake-security-concerns-are-limiting-voice-id-adoption-survey/.

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

[54] Sharkey, L. *Deepfake Audio Is The Beyond Sneaky Scam Trend You Need To Get Clued Up On ASAP* [online]. 2019 [cit. 2021-2-3]. Available at: https://www.bustle.com/p/how-to-spot-deepfake-audio-fraud-because-this-technology-may-be-more-widespread-than-you-think-18684814.

[55] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E. and Sebe, N. First Order Motion Model for Image Animation. In: *Conference on Neural Information Processing Systems (NeurIPS)*. December 2019.

[56] Simmonss, D. *BBC fools HSBC voice recognition security system* [online]. 2017 [cit. 2020-12-20]. Available at: https://www.bbc.com/news/technology-39965545.

[57] Wang, R., Juefei Xu, F., Huang, Y., Guo, Q., Xie, X. et al. *DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices*. 2020.

[58] Warzel, C. *I Used AI To Clone My Voice And Trick My Mom Into Thinking It Was Me* [online]. 2018 [cit. 2020-12-26]. Available at: https://www.buzzfeednews.com/article/charliewarzel/i-used-ai-to-clone-my-voice-and-trick-my-mom-into-thinking.

[59] 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.

[60] Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X. et al. *ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database*. 2019. Available at: https://doi.org/10.7488/ds/2555.

[61] Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K. et al. *Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —*. 2020. Available at: https://www.isca-speech.org/archive/VCC_BC_2020/pdfs/VCC2020_paper_13.pdf.

[62] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J. et al. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. *CoRR*. 2019, abs/1904.02882. Available at: http://arxiv.org/abs/1904.02882.

[63] Zhou, X., Ling, Z.-H. and King, S. *The Blizzard Challenge 2020* [online]. 2020. Available at: http://www.festvox.org/blizzard/bc2020/BC20_zhou_ling_king.pdf.

Appendix A

Contents of the included storage media

Root folder contains following directories:

- `dep_indep_dataset`: dataset used in text-dependent vs. text-independent verification experiment discussed in Section 6.5
- `experiments_data`: data used in basic concepts experiment discussed in Section 6.3, and link to the deepfake datasets from Sections 6.4 and 6.6
- `latex_src`: source codes for this work
- `survey_results`: raw survey results (Section 6.7)

Appendix B

Synthesized sentences

Following sentences were synthesized to create an English deepfake dataset discussed in Sections 5.1.5 and 6.4:

1. My voice is my passport verify me.
2. October arrived, spreading a damp chill over the grounds and into the castle.
3. Hello. Yes, I would like to inform myself on the topic of spoofed voice and the security implications.
4. In some cases specific syllables and particular words are consistently represented by specific syllables.
5. The black cat has reflexes, agility and stamina of an olympic level acrobat.
6. The grape is still popular in North Africa, Algeria, Morocco and Tunisia.
7. Eight minutes later she went to general quarters and enemy bodies were reported.
8. Elisabeth attended university of Chicago Laboratory schools.
9. Hale docking occurs in one of two ways.
10. Debate may also end if no senator wishes to make any further remarks.

Following sentences were synthesized to create a Czech deepfake dataset discussed in Sections 5.2 and 6.6:

1. Tato pláž byla oceněna modrou vlajkou.
2. Domníval jsem se, že to bude představovat tisícové náklady.
3. Každoročně se schází vrcholný výkonný orgán tvořený hlavami států nebo předsedy vlád členských států.
4. Tento typ strukturální podpory je příležitostí, jak dosáhnout našich cílů.
5. Kvůli nedostatku přesnosti se moderní porodnictví termínu vyhýbá.
6. Z mezinárodního hlediska se používá vždy mezera.

7. Musíme i nadále podporovat naše zemědělce při modernizaci jejich podniků.
8. Menšími úpravami prošla také karoserie.
9. Kanada patří mezi nejdůležitější partnery, které Evropská unie má.
10. Hráč si může udělat i své vlastní návštěvníky.

Appendix C

Survey answer key and links

The answer key to the speaker similarity survey (Sections 5.3 and 6.7) and links to corresponding recordings.

Speaker no.	Utterance A	Utterance B	Utterance C
1	genuine	deepfake 1k	deepfake 5k
2	deepfake 5k	genuine	deepfake 1k
3	genuine	deepfake 1k	deepfake 5k
4	deepfake 1k	deepfake 5k	genuine
5	deepfake 1k	genuine	deepfake 5k
6	deepfake 5k	deepfake 1k	genuine
7	deepfake 5k	deepfake 1k	genuine
8	genuine	deepfake 1k	deepfake 5k
9	genuine	deepfake 1k	deepfake 5k
10	deepfake 1k	genuine	deepfake 5k

- Speaker 1 : https://youtu.be/n6zttgkIass
- Speaker 2 : https://youtu.be/FQHoFeNZtvc
- Speaker 3 : https://youtu.be/495SM7WmAvM
- Speaker 4 : https://youtu.be/hsgVcSDOht8
- Speaker 5 : https://youtu.be/dFJGTOiR8QE
- Speaker 6 : https://youtu.be/OWWeXxFTIUE
- Speaker 7 : https://youtu.be/tMwbX_mJ_CQ
- Speaker 8 : https://youtu.be/sGS_h8LracQ
- Speaker 9 : https://youtu.be/LxH1wpLYps8
- Speaker 10 : https://youtu.be/XtZ83viyS-I