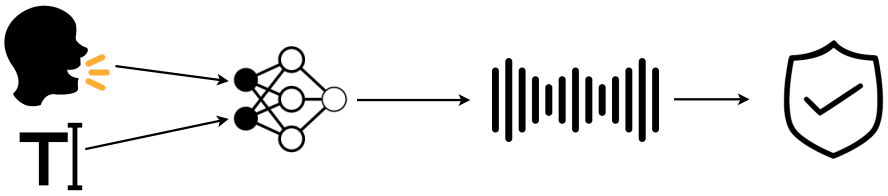
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2021
http://excel.fit.vutbr.cz

Deepfakes – really fake?

Bc. Anton Firc*



Abstract
Deepfake technology is on the rise, many techniques and tools for deepfake creation are being developed and publicly released. These techniques and tools are being used for both illicit and legitimate purposes. 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 or tools are tailored to synthesize the English language, one might think that using a different language might mitigate mentioned vulnerabilities, but as shown in this work, synthesizing speech in any language is not that complicated. Finally measures to mitigate the threat posed by deepfakes are proposed, like using text-dependent verification because it proved to be more resilient against deepfakes.

Keywords: deepfake — cybersecurity — voice biometrics

Supplementary Material: Synthesized speech samples — Deepfake dataset — Raw survey results

*xfirca00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

The emergence of deepfakes is progressively increasing the public awareness. As the deepfakes begin to appear in the television broadcast and the word spreads through the social media, broad public slowly starts to realise what threats the synthetic medias might pose.

The slang term *deepfake* has no agreed-upon technical definition, it is just a combination of words 'deep learning' and 'fake', and primarily relates to an content generated by an artificial neural network [1]. As the name suggests, deepfakes are created using an AI method called deep learning that utilizes deep neural networks [1]. The networks first have to be trained, the input data may be of various formats e.g. images, videos or speech and the network then extract characteristics of the input data and create a mathematical pattern based on the training data [1].

This generative ability of neural networks has a lot of potential to be used in both malicious and beneficial ways. The malicious usage finds place in various areas like politics, finance, private sector or fake news [1]. The inflicted harm ranges from making a fool out of an individual to an identity theft or subverting the government [1]. Fortunately, the threats posed by deepfakes seem to be more individual oriented, and should not cause a global damage [1]. The beneficial usages do not get as much attention as the malicious usages, but that does not mean there are none. The beneficial usages may find place in entertainment, education or even healthcare [2]. The movie creators can use deepfake technology to bring long dead actors to movies, or synthetic speech may bring back voice to

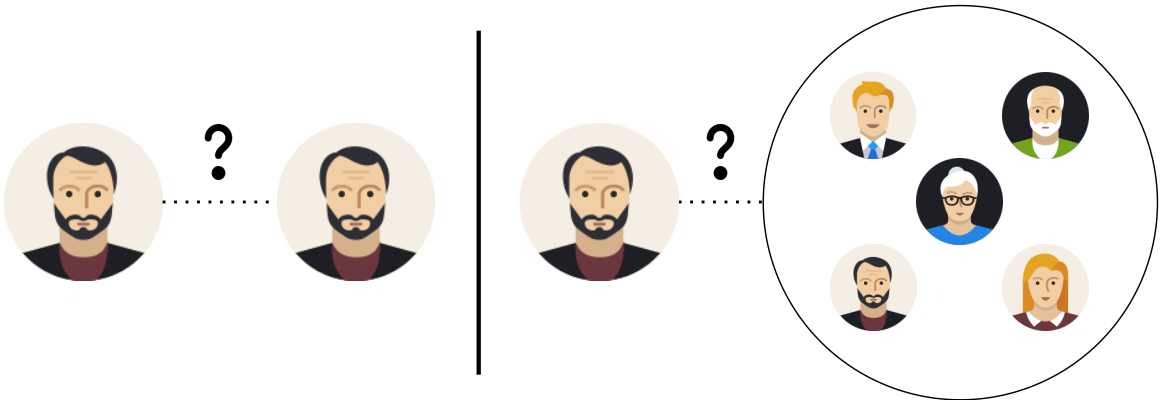


Figure 1. The principle of verification process on the left, and the principle of identification process on the right.

oncology patients that lost it [2].

While the most popular type of deepfakes are the video deepfakes, the most harmful type seems to be the voice deepfakes, as you cannot visually confirm that you speak with a person whose voice you hear [1]. Regarding the voice deepfakes, there is an interesting area of voice authentication, that lacks any scientific evidence on the resilience of voice authentication systems against deepfakes.

In this work, I prove that the deepfake speech is a threat to the voice authentication, that the text-dependent verification is more secure than the text-independent verification and examine the feasibility of training own text-to-speech synthesis model for Czech language.

2. Synthetic speech, a threat or an asset?

There are different opinions on the security of voice authentication systems that can be found on the internet. There are researches dedicated to this topic, however they show just a proof-of-concept instead of testing in real-world conditions. In this section I will define a voice biometrics system and the performance measures that will be later used to evaluate the realness of deepfakes, review the public opinion on security of voice authentication against deepfakes and present the deepfake scenarios of attacking a voice biometrics system.

2.1 Voice biometrics system

The voice biometrics, or speaker recognition, system provides a biometric-based security process known as speaker authentication [3]. Freely translated, it means that the system authenticates its users based on *what they are*, by processing their unique voice characteristics. Two technologies exist: speaker verification and speaker identification [3]. As the Figure 1 shows, verification is a process of determining whether a person is who she or he claims to be, while identification is a process of determining an identity of a person from a pool of known identities [3, 4]. Speaker verification further divides to text-dependent and text-independent [4]. Text-dependent verification needs the same phrase to be spoken during the enrolment and the verification phase, the text-independent verification on the other hand has no restriction on the spoken content [4].

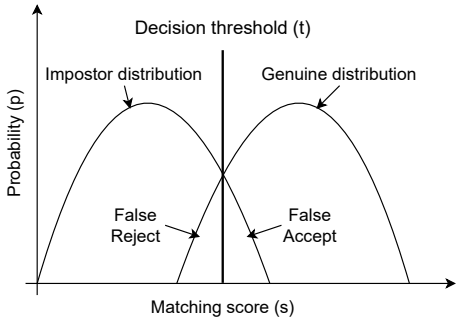


Figure 2. User matching scores distribution by attempt type.

Shortly, voice biometrics system is a black-box that authenticates a person based on their own voice.

2.1.1 Performance measures

The performance of a biometrics system might be measured and compared using numerous measures. The verification performance evaluation is done by performing many genuine and impostor attempts while the matching scores are saved [5]. A genuine attempt is performed by an user to match his own profile, an impostor attempt is performed by an user to match someone else‘s profile [5].

Saved matching scores can be used to plot user matching scores distribution (see Figure 2), and to calculate error rates. The most commonly used error rates are false reject rate (FRR) or false non-match rate (FNMR) and false accept rate (FAR) or false match rate (FMR) by applying a varying matching score threshold [5]. As shown in Figure 3 FMR and FNMR are then used to calculate the equal error rate (EER), which is a point where the FMR and FNMR equal [5].

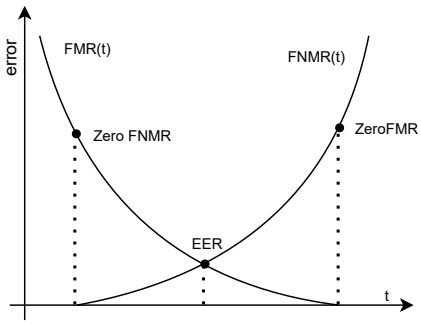


Figure 3. FMR, FNMR and EER.

2.2 Current situation and public opinion

The public opinion regarding security of voice authentication when facing deepfake recordings is mixed, and the lack of any incident reports leaves us wondering, where the truth really is. There is already evidence supporting the ability of deepfakes to fool humans. Article published on BuzzFeed [6] tells a story of how a reporter used deepfake speech to communicate with his mom, and the he first, and only, documented incident reports a theft of 250k USD using deepfake speech [2].

The companies developing voice biometrics systems mostly claim that their systems are impenetrable [2]. This statement was already disproved by a reported incident, when BBC reporter let his twin access his account [2]. Even though this happened back in 2017, and the technology has taken a massive leap forward since then, it implies that under the right circumstances spoofing voice biometrics system is possible. This possibility was as well confirmed by an article published in the beginning of 2020 [7].

2.3 Deepfake scenario

The first real talk about spoofing voice authentication using deepfakes shows up in a research published on the Black Hat conference in 2018 [8]. The authors introduced a proof of concept that the voice biometrics systems might be spoofed by deepfakes and supported this theory with first real evidence.

A text-to-speech system was used to create deepfake speech, and this speech was then used against Apple Siri and Microsoft Speaker Recognition API. Both of the systems authenticated the deepfake speech. There were some limitations regarding quality of synthesized speech, but this was believed to be just temporary limitation as the deepfake technology still advances. In conclusion, authors stated that there is no known mechanism to detect this kind of attack.

Now, almost three years after publishing the mentioned research, the topic is still more than actual, as well as no further evidence or reports were published since then. This is why the topic of this work is important, because it measures the advances in both deepfake and voice authentication areas.

3. (Deep)faking voice authentication

This section closely explores the possibilities of using deepfakes to spoof voice authentication. The attacker model is proposed, and experiments evaluating the overall technical feasibility, creating deepfake dataset and bulk testing and finally comparing text-dependent versus text-independent verification in terms of security against deepfakes.

3.1 Voice deepfakes

Currently, there are two main approaches for creating deepfake voices: text-to-speech (TTS) synthesis and voice conversion (VC). As the name suggests, TTS synthesis consumes written text as input and produces synthesized speech, VC on the other hand consumes a source voice saying desired phrase and a target voice, and produces a source phrase spoken by the target voice [9]. The VC tools are currently just research paper implementations, and their usage for people not involved in their development is still limited to running them with demo data for showcase, if even that is possible as the implementations are very fragile and require an extensive amount of time and knowledge to run them. The TTS tools on the other hand already feature commercial implementations, thus they are available to a wide range of users.

The TTS tools used for the purpose of this research will be treated as a black-box. Two commercial tools: *Overdub* and *Resemble AI*, and one opens-source *Real-Time-Voice-Cloning* will be used.

3.2 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 bank call centre. The attacker is in possession of all needed personal information about his victim, all needed details about the authentication process, and finally samples of voice belonging to the victim.

The attacker will use all of this information to synthesize speech similar to the victim‘s, and then in the most believable way possible try to access the secured system 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.

3.3 Victim model

A victim is any person that uses her‘s or his voice to authenticate into any kind of system.

3.4 Introduction to TTS and voice biometrics

To begin with, I explored the basic concepts and capabilities of selected TTS tools: *Overdub*, *Resemble AI*, *Real-Time-Voice-Cloning* (RTVC) when creating deepfakes, and try to use the synthesized speech to spoof two voice biometrics systems I was able to get my hands on: *Microsoft Speaker Recognition API* and *Phonexia Voice verify demo*. Unfortunately, I was unable to get access to any other voice biometrics systems, as all of the companies I reached out to either did not response, or responded that they do not provide any kind of ”student license”.

To gain the initial information, I tried to verify my own voice against my profile. As the Table 1 shows, the MS Speaker Recognition API returns matching scores for genuine attempts from interval $[0.70, 0.95]$, and there seems to be no significant difference between text-dependent and text-independent matching scores.

As a next step, I created deepfakes using both commercial tools, and the RTVC tool using the pre-trained models provided with the implementation. As Table 2 shows, the RTVC tool reached very low matching scores, while both of the commercial tools almost reached the genuine text-dependent matching scores, and reached the text-independent matching scores.

RTVC tool improves quality of speech when synthesizer model is fine-tuned for target speaker [10]. Fine-tuning is done by training the synthesizer model for at least additional 200 iterations using at least 0.2 hours of target speech [10]. This process greatly increases the quality of synthesized speech, as shown in Table 3.

After reaching solid results with the Microsoft Speaker Recognition API, I moved on to the Phonexia Voice Verify demo. The system uses an ISP provider, and the verification is done through a phone call, so to perform the tests, I decided to play synthesized audio from a notebook speaker to a phone placed nearby. As shown in Table 4, the commercial tools were able to synthesize speech that passed the verification process, the RTVC tool fell a bit behind with verified and rejected parts of the recording resulting with a non-verified recording.

Table 1. Achieved average matching scores for each type of verification using Microsoft Speaker Recognition API.

Verification type	Matching score
Dependent	0.83815
Independent	0.82174

Table 2. Best achieved matching scores for each TTS tool for each type of verification using Microsoft Speaker Recognition API.

Verification type	Tool	Matching score
Dependent	RTVC	0.19861
	Overdub	0.64144
	Resemble AI	0.55970
Independent	RTVC	0.47097
	Overdub	0.79611
	Resemble AI	0.60146

Table 3. Best achieved matching scores of TTS tool after fine-tuning the synthesizer model for both types of verification using Microsoft Speaker Recognition API.

Verification type	Tool	Matching score
Dependent	RTVC	0.59272
Independent		0.62365

Table 4. Results of verification in Phonexia Voice Verify demo using the best recording synthesized with each of the used tools.

Tool	Verified
Overdub	yes
Resemble AI	yes
RTVC	no

3.5 Creating deepfake dataset

Using the RTVC tool, I created a deepfake dataset consisting of 100 speakers. I decided to use this tool because of its open-source license and ability to synthesize speech conditioned by short target speaker utterance, despite the fact it achieved the lowest scores in previous experiment. For each speaker there are genuine recordings, deepfake recordings, and an enrolment utterance that was used as a sample to synthesize deepfakes and to enroll speakers voice profile. Each deepfake recordings is a one sentence, the sentences are same for all speakers. All of the sentences were taken from the Common Voice Corpus transcripts. To show contrast, I took the ASVSpoof 2019 challenge dataset [11], and calculated the same metrics as for the created dataset. As the Figure 4 shows, the distribution of deepfake matching scores matches the genuine one quite precisely, and EER vastly increased comparing the original and deepfake one.

In contrast to the ASVspoof dataset, the deepfake matching scores of the created dataset almost exactly

Table 5. Average matching scores and standard deviation for both text-dependent and text-independent verification using genuine and deepfake recordings.

Verification type	Recording type	Avg. Matching score	Standard deviation
Dependent	Genuine	0.84660	0.03549
	Deepfake	0.53636	0.07242
Independent	Genuine	0.67112	0.11360
	Deepfake	0.60283	0.10636

reproduce the genuine ones. If viewing the matching scores distributions, there is no significant difference between genuine and deepfake speech.

Finally, this was possible using the tool that performed worst in previous experiment, which suggests that the other (commercial) tools should perform minimally the same.

3.6 To depend or not to depend on the text?

As results of the first experiments (see 3.4) suggest, the text-dependent verification seems to be more resilient against deepfakes. As no dataset containing needed phrases is currently publicly available, I created my own dataset for this purpose. The dataset consist of phrases used for text-dependent verification and training data to fine-tune the RTVC synthesizer model. Up to date, the dataset consists of 5 speakers, 4 male and 1 female.

As Table 5 shows, the difference between genuine and deepfake average matching score for text-dependent verification is very significant. On the other hand, the average matching scores for text-independent verification, decreased for genuine attempts and increased for deepfake attempts, which makes the difference between genuine and deepfake approximately four times smaller.

The results definitely support the hypothesis, that the text-dependent verification is more resilient against deepfakes. The average matching scores and standard deviation show the significant difference between genuine and deepfake recordings when using text-dependent verification, and also that this difference almost vanishes when using text-independent verification.

To further confirm this hypothesis a testing with bigger spectrum of speakers has to be done, as well as using more voice biometrics systems. However the achieved results show that this behavior was not a random event, and is worth to research further.

4. Fooling people as well as machines?

The second experiment follows up on the idea of spoofing voice authentication and initiating a fraudulent action for example in a bank. In most of the cases, after being authenticated by the voice biometrics system, there is a human operator you have to talk to to perform the desired actions. If the operator gains any suspicion, she or he will most likely end the conversation. This means, that the deepfake recording must be able to spoof the voice biometrics system and the operator at the same time.

To examine whether the created deepfakes might be accepted by humans as genuine recordings, I created a survey, where the respondents are be set into a role of voice biometrics system. A made up scenario sets respondents into a role, where their main goal is to decide whether each attempt is genuine or not. The respondents are also told to be careful and select only the recordings they are completely sure are genuine.

The survey consists of 10 speakers from the deepfake dataset created during the first experiment based on their average deepfake matching scores to range from the lowest to the highest. Second criteria was to balance sex distribution, so 5 female and 5 male speakers were selected.

For each speaker, I took the models fine-tuned for 1k and 5k iterations and synthesized sentences randomly chosen from the Common Voice dataset transcripts. The choice of deepfake recordings aims to investigate any link between the amount of steps-fine tuned and believability to humans. The enrolment utterance was chosen to be the same as the template recording used when creating the deepfake dataset. Finally, the lowest quality utterance was selected as genuine, to simulate the worst case scenario of genuine attempt and better blend in with lower quality deepfakes.

For each speaker I combined the utterances in random order, while preserving the count of one genuine and two deepfake (1k and 5k) recordings. I also created two version of the survey, one with instructions in Czech language and second in English language.

Both of the surveys were released in the beginning of March 2021, and since then almost 100 responses were collected. The ratio of Czech to English responses was 1 : 2, and the female to male ratio was surprisingly exactly 1 : 1.

The responses were evaluated the same way as performance of biometrics systems is evaluated. Each utterance represents an verification attempt, genuine or impostor. False accept rate and false match rate were calculated to evaluate how precise the respondents were in distinguishing between genuine and deepfake

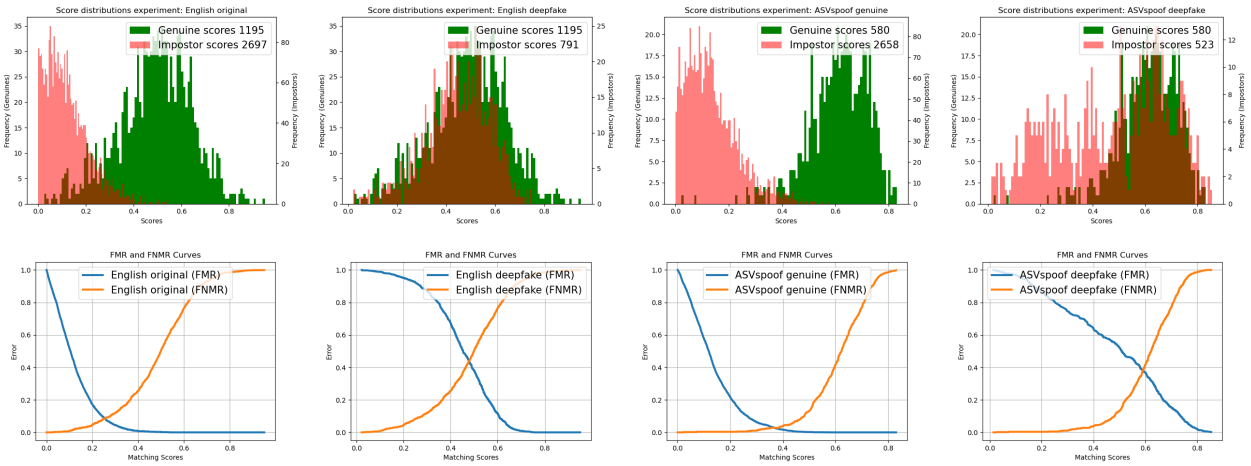


Figure 4. Matching scores distribution graphs (top) and FMR / FNMR graphs (bottom). The left two plots represent the created deepfake dataset. The two right plots represent the ASVSpoof 2019 challenge dataset [11]. For both datasets, left plots show genuine vs. impostor matching scores comparison, while the right plots show deepfake vs. genuine matching scores comparison. Distributions were produced by plotting all calculated genuine/impostor and genuine/deepfake matching scores calculated using the Microsoft Speaker Recognition API.

Table 6. FAR and FRR calculated from the survey results.

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

Table 7. Correlation between FAR, FRR and utterance matching scores. Correlation is calculated using FAR, FRR and matching scores of recordings for each speaker.

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

utterances. Table 6 shows results of both surveys together, and also each one separately.

As Table 7 shows, there seems to be hint of correlation between the matching scores and believability for humans. The genuine matching score tends to decrease the false accept and false reject rate, as well as the deepfake matching scores tend to increase the false accept and false reject rate for each speaker.

The results show that approximately one of the three deepfake authentication attempts was successful. The quality of the genuine utterance had large impact on the false accept rate, as the speaker number 5 with strangely sounding genuine utterance reached FAR of 83.33%.

The achieved results show, that voice deepfakes have the ability to fool humans, and that there is some correlation between calculated matching scores and the accept or reject rates.

5. Do deepfakes speak Czech?

The third experiments evaluates the usability of deepfakes in Czech or Slovak Republic. As no publicly available tools or models provide the ability to create deepfakes of any voice in any language, an attacker would have to create his own model for this purpose.

This experiment thus aims to explore the technical feasibility of training own TTS synthesis model for Czech language, and concluding whether using Czech language is more secure then English regarding voice authentication. For this purpose I decided to train all needed models for the Real-Time-Voice-Cloning tool in order to synthesize speech in Czech language. The training followed instructions on GitHub wiki and tips found in the issues section.

As training dataset, I have chosen the Common Voice Corpus [13], because it is the only dataset publicly available containing Czech speech with transcriptions. The dataset is dedicated to help with the speech recognition and synthesis tasks, thus should provide suitable data to train on.

5.1 Real-Time-Voice-Cloning

Real-Time-Voice-Cloning is a open-source implementation of a TTS framework originally proposed in [14]. As shown in Figure 5 the pipeline consists of three parts: *encoder, synthesizer* and *vocoder* [14, 12]. The encoder conditions the synthesis network on a reference speech signal from the desired target speaker [14]. The synthesizer predicts a mel-spectrogram of synthesized speech [14]. The vocoder finally transforms the mel-spectrogram to speech [14].

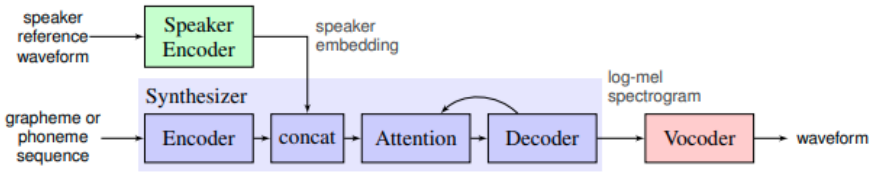


Figure 5. Overview of framework implemented within the RTVC tool. Retrieved from [12]

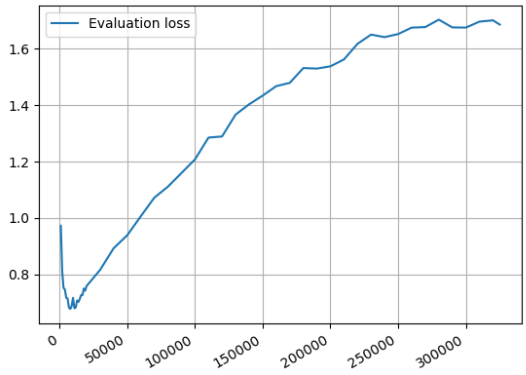


Figure 6. Loss calculated during the evaluation phase of RTVC synthesizer model training.

5.2 Training RTVC for Czech language

To train the encoder, a large number of speakers is needed. To maximize speaker count, I decided to use all available Slavic languages: Polish, Russian, Slovenian and Ukrainian. Final count of distinct speakers was 3454 with 290 hours total length of recordings. I proceeded to train the encoder model for 5 days using the nVidia Tesla T4 GPU, until the training reported adequate performance.

To train the synthesizer model, only one language must be used, so only the Czech subset was used. While experimenting, I found out that recordings with no silence removal produced the best results during training. According to the discussions in the issues section and some showcased models trained using RTVC tool, around 300k iterations should provide enough knowledge to synthesize quality speech. Following this fact, I trained the synthesizer model for total of 328k iterations. Unfortunately the training did not go so well as with the encoder, even in the best trials the loss slowly increased as shown in Figure 6. The time needed to train the 328k iterations and prepare the data was 14 days using the nVidia Tesla T4 GPU.

The vocoder model is the only language-independent part of the RTVC tool, and because the training takes extensive amount of time, I decided to use the already pretrained one, available with the tool [12].

5.3 Results

The synthesized speech came out better than expected because of the increasing synthesizer training loss. Most sections of the synthesized speech are unclear and the words are often inaudible, however during the more quality sections of synthesized speech, it is really obvious that the synthesized language is Czech.

To get better insight on the similarity of synthesized speech to the original one, I created deepfake dataset with the same structure and method as in Section 3.5. As the Figure 7 shows, the genuine and deepfake distribution overlap, and the deepfake recordings increased the EER almost two times compared to the impostor recordings.

In conclusion, the synthesized speech shows similarity to the original speakers, despite lower quality. If the recordings were used only against the voice biometrics system to gain access, there is a chance for success. However, any human being will surely spot that something‘s wrong. Considering, that the RTVC tool was treated as a black-box and the imperfections in synthesizer training, the results are quite alarming, and show that the usage of deepfakes is not limited to the English language, however the Czech language can still be viewed as more secure.

6. Conclusions

The goal of this work was to analyze the threat that deepfakes present to the voice authentication, and to propose measures to mitigate that threat.

This work shows that deepfakes do pose a serious threat to voice biometrics systems and also people. As was shown, the authentication systems that do not implement any kind of liveness detection are easily spoofed. Adding a human factor into the authentication process also does not ensure improvement in security. The creation of a deepfake capable of spoofing a voice biometrics system or human might be generalized to cloning a GitHub repository, learning to work with it, collecting voice samples of the victim, transcribing them and finally synthesizing speech.

Even though most of the models and tools are suited for English language, I have shown that training a text-to-speech model for different language is possible even with no extensive knowledge of speech

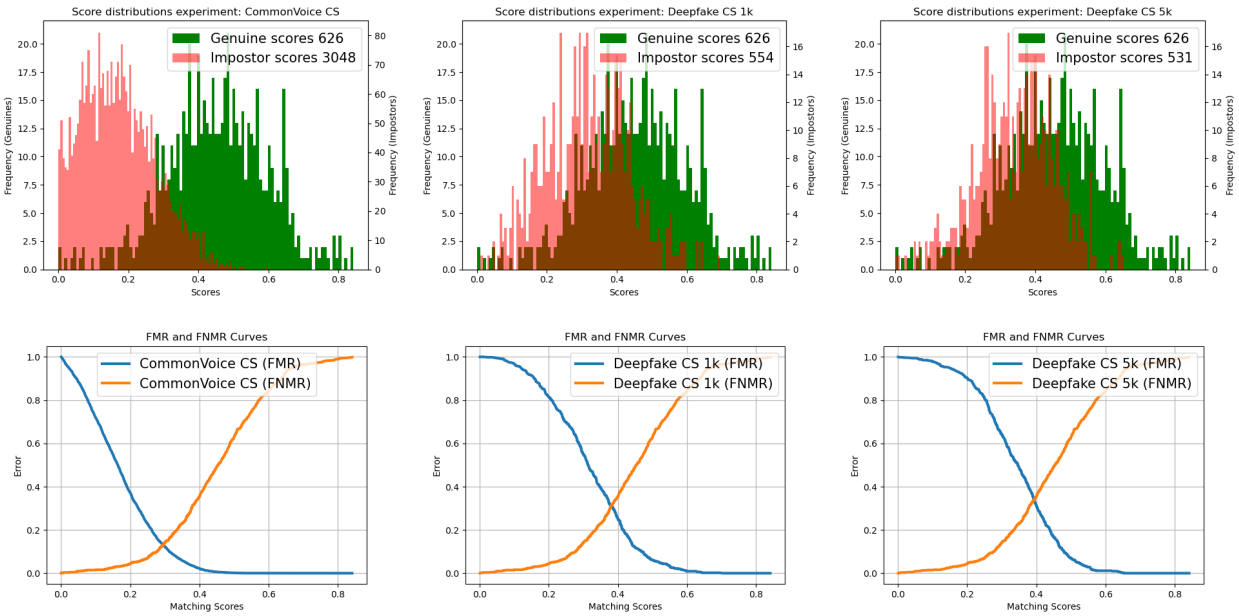


Figure 7. Matching scores distribution graphs (top) and FMR / FNMR graphs (bottom) for the Czech deepfake dataset. The genuine versus impostor distribution on the left, genuine and deepfake 1k in the middle, and genuine and deepfake 5k on the right. Distributions were produced by plotting all calculated genuine/impostor and genuine/deepfake matching scores calculated using the Microsoft Speaker Recognition API.

synthesis. This means that using voice authentication in language different than English does not imply safety from deepfakes.

This work shows that using text-dependent verification should mitigate the threat posed by deepfakes. As none of the tested voice biometrics systems implemented liveness detection, it is necessary to test such a system and conclude whether the liveness detection might mitigate the threat posed by deepfakes even more.

In summary, deepfakes are fake, they are synthetically constructed, however as I demonstrated in this work, people and voice biometrics systems reckon them as real.

Acknowledgements

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

References

[1] Jon Bateman. Deepfakes and synthetic media in the financial system: Assessing threat scenarios. Technical report, Carnegie Endowment for International Peace, 2020.

[2] Valencia A. Jones. Artificial intelligence enabled - deepfake technology the emerge of a new threat. Master thesis, Utica College, 2020.

[3] Judith A. Markowitz. *Designing for Speaker*, pages 123–139. Springer Netherlands, Dordrecht, 2004.

[4] Microsoft. About the speech sdk. online, 2020. https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview#speaker-verification.

[5] Precise Biometrics AB. Understanding biometric performance evaluation. online, 2014. https://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation-QR.pdf.

[6] Charlie Warzel. I used ai to clone my voice and trick my mom into thinking it was me. online, 2018. https://www.buzzfeednews.com/article/charliewarzel/i-used-ai-to-clone-my-voice-and-trick-my-mom-into-thinking.

[7] Ed Jefferson. Are voice biometrics the new passwords? online, 2020. https://www.raconteur.net/technology/cybersecurity/voice-biometrics/.

[8] John Seymour and Azeem Aqil. Your voice is my passport. online, 2018.

https://www.blackhat.com/us-18/briefings/schedule/#your-voice-is-my-passport-11395.

[9] National Academies of Sciences, Engineering, and Medicine. *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop*. The National Academies Press, Washington, DC, 2019.

[10] Single speaker fine-tuning process and results. online, 2020. https://github.com/CorentinJ/Real-Time-Voice-Cloning/issues/437.

[11] Junichi Yamagishi, Massimiliano Todisco, Md Sahidullah, Héctor Delgado, Xin Wang, Nicolas Evans, Tomi Kinnunen, Kong Aik Lee, Ville Vestman, and Andreas. Nautsch. Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database, 2019.

[12] Jemine Corentin. Real-time voice cloning. Master thesis, Université de Liège, Liège, Belgique, 2019.

[13] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[14] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis, 2019.