BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

IMPACT OF SPEAKER CHARACTERISTICS ON THE PERFORMANCE OF DEEPFAKE SPEECH DETECTORS

VLIV OSOBNÍCH CHARAKTERISTIK MLUVČÍHO NA VÝKON DETEKTORŮ DEEPFAKE HLASU

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **KAREL SRNA**
AUTOR PRÁCE

SUPERVISOR **Ing. VOJTĚCH STANĚK**
VEDOUCÍ PRÁCE

BRNO 2025

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

165219

Institut: Department of Intelligent Systems (DITS)
Student: **Srna Karel**
Programme: Information Technology
Title: **Impact of Speaker Characteristics on the Performance of Deepfake Speech Detectors**
Category: Security
Academic year: 2024/25

Assignment:

1. Familiarise yourself with the topic of speech deepfakes, tools for their creation, and existing datasets.
2. Design the structure and content of a deepfake dataset that evaluates the impact of various speaker characteristics (e.g., age, language, gender, minimum of 4 characteristics) on the performance of deepfake detectors.
3. Create the designed dataset. As a basis, you can use existing datasets with the necessary annotations (if such datasets exist).
4. Using the created dataset, evaluate the impact of the selected characteristics on the performance of at least three deepfake detectors with different architectures. Standard experimental evaluation procedure is expected (experimental design, hypothesis formulation, statistical analysis, interpretation of results, etc.)
5. Analyse and conveniently present the measured results and the impact of individual characteristics.

Literature:

- FIRC Anton, MALINKA Kamil and HANÁČEK Petr. Deepfakes as a threat to a speaker and facial recognition: an overview of tools and attack vectors. *Heliyon*, vol. 9, no. 4, 2023, pp. 1-33. ISSN 2405-8440
- X. Liu *et al.*, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507-2522, 2023, doi: 10.1109/TASLP.2023.3285283.
- Wang, Xin, et al. "ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale." *arXiv preprint arXiv:2408.08739* (2024). https://arxiv.org/abs/2408.08739

Requirements for the semestral defence:
Items 1 - 3 (only a limited size prototype is required for 3)

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

Supervisor: **Staněk Vojtěch, Ing.**
Head of Department: Kočí Radek, Ing., Ph.D.
Beginning of work: 1.11.2024
Submission deadline: 14.5.2025
Approval date: 31.10.2024

Abstract

This thesis is focused on the issue of voice deepfakes, generated audio recordings sounding like uttered by real people. The motivation behind this work is to investigate whether speaker demographics influence the performance of deepfake speech detection systems. At the beginning, it is explained what deepfakes are, along with the main types of speech synthesis. Then follows a review of existing datasets for real and fake speech. Their weaknesses are pointed out, and a new dataset is proposed featuring speakers' annotations like gender, language, age, and education. The dataset contains 237,250 utterances with a total duration of more than 500 hours, compiled from 50 speakers in five languages. It is used for evaluating multiple speech detection systems. The obtained results are analyzed in detail – the detectors' performance is presented for each of the speaker's characteristics, and it is statistically evaluated whether there exists some bias between demographic subgroups. The results show that modern detectors are not unbiased and special care should be taken for training and deploying fair and non-discriminative deepfake speech detectors.

Abstrakt

Tato bakalářská práce se zaměřuje na problematiku hlasových deepfakes, generovaných zvukových nahrávek, které znějí, jako by je namluvili skuteční lidé. Cílem této práce je zjistit, zda demografické údaje mluvčího ovlivňují výkonnost systémů pro detekci deepfake řeči. V úvodu práce je představena technologie deepfake a taktéž klasifikace hlavních typů syntézy řeči. Práce se dále zabývá přehledem existujících datových sad pro skutečnou a falešnou řeč. Je poukázáno na jejich slabiny a je navržena nová datová sada obsahující anotace mluvčích, jako je pohlaví, jazyk, věk a úroveň vzdělání. Navržená datová sada obsahuje 237 250 nahrávek s celkovou délkou trvání více než 500 hodin, sestavených z řeči 50 mluvčích v pěti jazycích. Dataset je použit k vyhodnocení několika systémů pro detekci řeči. Získané výsledky jsou podrobně analyzovány – je uvedena výkonnost detektorů pro každou z charakteristik mluvčího a je statisticky vyhodnoceno, zda existuje nějaké zkreslení mezi demografickými podskupinami. Výsledky ukazují, že moderní detektory nejsou nezaujaté a že je třeba věnovat zvláštní pozornost trénování a nasazení spravedlivých a nediskriminačních detektorů deepfake řeči.

Keywords

Deepfake, Synthetic speech, Dataset, Bias & fairness, Deepfake Detection

Klíčová slova

Deepfake, Syntetická řeč, Datová sada, Diskriminace & bias, Detekce deepfake

Reference

SRNA, Karel. *Impact of Speaker Characteristics on the Performance of Deepfake Speech Detectors*. Brno, 2025. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Vojtěch Staněk

Rozšířený abstrakt

Tato bakalářská práce se zabývá problematikou hlasových deepfakes, neboli syntetických zvukových nahrávek imitujících reálné lidské hlasy.

V úvodu je představena motivace ke zkoumání této oblasti, zejména s ohledem na rostoucí počty případů zneužití těchto nástrojů. Práce stručně vysvětluje základní principy deepfake technologií a popisuje hlavní přístupy k syntéze řeči, jako je generování pomocí textu nebo audia (text-to-speech, voice conversion), a také techniky detekce syntetick7ch nahrávek.

Další část sestává z přehledu stávajících datových sad skutečné i falešné řeči. Tyto datasety jsou analyzovány z hlediska jejich rozsahu a pokrytí různorodých charakteristik mluvčích. Je poukázáno na nedostatečné zastoupení různých demografických skupin v těchto datasetech a na absenci detailních anotací, které by umožnily hlubší analýzu detekční výkonnosti podle těchto charakteristik.

Na základě zjištěných nedostatků je navržen nový dataset, jehož cílem je umožnit evaluaci detekčních systémů napříč různými skupinami mluvčích. Jsou popsány požadavky kladené na nový dataset, včetně volby charakteristik, jako je pohlaví, jazyk, věk, etnická příslušnost a úroveň vzdělání. Následuje podrobný popis procesu kompilace dat, zahrnující výběr reálných nahrávek, jejich syntézu pomocí několika moderních nástrojů, a jejich finální zpracování do standardizovaného formátu.

Datová sada, která je rozdělena do 237 125 nahrávek, je sestavena za pomoci čtyř nástrojů syntézy řeči, konkrétně XTTSv2, F5-TTS, DDDM-VC a Open Voice v2, a obsahuje řeč padesáti mluvčích v pěti jazycích s celkovým trváním přesahující 500 hodin. Klíčovou vlastností datasetu je přítomnost detailních metadat, což umožňuje analyzovat jednotlivé demografické skupiny. Zdrojem reálné řeči je dataset VoxPopuli [77], který vychází ze záznamů z Evropského parlamentu.

V experimentální části práce jsou otestovány nástroje pro detekci deepfake řeči. Výsledky detekce jsou analyzovány nejen z hlediska celkové úspěšnosti, ale zejména s důrazem na rozdíly mezi skupinami mluvčích. Tyto rozdíly jsou kvantitativně hodnoceny statistickými testy a metrikou pro měření chybovosti. Výstupy jsou diskutovány z hlediska možného zkreslení detekčních systémů.

Závěr práce shrnuje hlavní poznatky a upozorňuje na rizika spojená s nedostatečně reprezentativními daty při vývoji a evaluaci nástrojů pro detekci deepfakes. Byl zjištěn nízký, v praxi téměř zanedbatelný, bias mezi muži a ženami. Mezi jazyky je situace obtížnější a většina detektorů měla nižší výkonnost při detekci německých nahrávek. Při zkoumání vlivu věku řečníka byla odhalena menší přesnost vůči mluvčím staršího věku. Statisticky významný rozdíl v rozdělení skóre detektorů byl potvrzen i v případě míry vzdělání.

Navržený dataset představuje krok směrem k spravedlivější a spolehlivější detekci syntetické řeči napříč různými demografickými skupinami.

Impact of Speaker Characteristics on the Performance of Deepfake Speech Detectors

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Mr. Ing. Vojtěch Staněk. I have listed all the literary sources, publications, and other sources that were used during the preparation of this thesis. I acknowledge the usage of generative artificial intelligence for stylistic improvements and language refinement during the writing of this thesis.

. . . .
Karel Srna
May 14, 2025

Acknowledgements

I would like to thank my supervisor, Ing. Vojtěch Staněk, for his guidance, support, and endless patience with my countless questions. Thanks also go to Bc. Michal Luner for his help and advice with fine-tuning synthesis tools. I would like to thank my family and friends for supporting me throughout my studies. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Contents

Glossary

ASR Automatic Speaker Recognition. 14

ASV Automatic Speaker Verification. 17

CNN Convolutional Neural Network. 8, 21

CQCC Constant-Q Cepstral Coefficient. 20

CQT Constant-Q Transform. 20

CSV Comma Separated Values. 26, 38

DDDM Decoupled Denoising Diffusion Model. 10

DiF Diffusion Transformer. 9

DL Deep Learning. 20

EER Equal Error Rate. 22, 31–38

FNR False Negative Rate. 22

FPR False Positive Rate. 22

GAN Generative Adversarial Networks. 8, 9

GMM Gaussian Mixture Model. 21

GNN Graph Neural Network. 21

GPT Generative Pre-trained Transformer. 9

HMM Hidden Markov Models. 7

LCNN Light Convolutional Neural Network. 30

LFCC Linear Frequency Cepstral Coefficients. 20

MEPs Members of European Parliament. 24

MFCC Mel-Frequency Cepstral Coefficients. 20

MHFA Multi-Head Factorized Attentive Pooling. 31

ML Machine Learning. 20

PSTN Public Switched Telephone Network. 17

RNN Recurrent Neural Network. 8

SOTA State Of The Art. 8

SPSS Statistical Parametric Speech Synthesis. 7

STFT Short-Time Fourier Transform. 20

SVM Support Vector Machine. 21

TNR True Negative Rate. 22

TPR True Positive Rate. 22

TTS Text To Speech. 4, 6–10, 12, 17–20, 23

VAE Variational Autoencoder. 9

VC Voice Conversion. 4, 6, 9, 17, 19, 26

VCC Voice Conversion Challenge. 18

VoIP Voice over IP. 17

VQ-VAE Vector Quantised-Variational AutoEncoder. 8

Chapter 1

Introduction

Deepfake technology has changed various spheres of human interest. Educational courses are made by animating historical figures, such as Agatha Christie [29]. It is employed to raise awareness about serious topics like the campaign against Malaria [8]. Usage is also found in medicine, for instance, educating patients or training medical students [5]. And the list can go on.

However, with great power comes great responsibility, and deepfakes can be misused as well. Spreading misinformation, creating non-consensual explicit content, impersonating public figures in scams, and committing fraud are all situations in which this technology is used maliciously. For example, a fake video was released during the Russo-Ukrainian War with President Zelenskyy talking about surrender [75].

To deal with the possible dangers of deepfakes, it is crucial to develop new strategies to face these challenges. One of them is creating new deepfake datasets, which help researchers build better and more accurate tools for spotting fake content. By training detectors on a wide range of synthetic samples, experts can improve their models' generalization capabilities, ultimately reducing the harmful impacts of deepfakes on society.

This thesis focuses on creating a new voice deepfake dataset to study whether certain speaker characteristics, such as age, gender, language, and education level, affect how well deepfake detectors work. If it turns out that some groups of speakers are harder to analyze accurately, it could mean that these people are more vulnerable to deepfake attacks. Understanding these potential biases is important for improving the reliability of detection systems and ensuring that all users are equally protected. In addition, the AI Act[1] of the European Union requires that high-risk applications, such as biometric systems, under which deepfake speech detection falls, must be fair and unbiased, further emphasizing the importance of this research.

Firstly, the current state of the research is presented, including a description of the deepfake creation techniques. Then, available genuine speech corpora are reviewed, including the ones derived from hearings of political institutions. An overview of existing deepfake datasets follows, along with a description of the performance metrics used for evaluating deepfake detectors.

The main result of this thesis is a dataset compiled using Text To Speech (TTS) and Voice Conversion (VC) tools.

[1] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

It contains more than 500 hours of speech in 237,250 utterances across 50 speakers distributed in five languages. A crucial feature of the created dataset is the availability of detailed metadata, allowing disaggregated analysis for each demographic group and possible bias. The utterances are annotated with a transcript and speaker-related information like language, gender, age, level of education, and ethnicity.

The dataset is used to evaluate multiple available deepfake speech detectors to asses their performance. The detectors' outputs are analyzed with regard to bias and the influence of the examined speaker characteristics. Formal statistical testing is employed to provide a solid foundation for drawing experimental conclusions.

Chapter 2

Current state of research

This chapter aims to answer foundational questions regarding the scope of this thesis, which is voice deepfakes. The important questions are: What is a deepfake, and what is a voice deepfake in particular? What are the types of voice deepfakes? What is the difference between TTS and VC speech generation approach?

2.1 What is a deepfake?

The term deepfake is derived from „Deep learning“ and „Fake,“ and it refers to synthetic media generated by deep learning algorithms. These technologies are able to create highly realistic audio, video, or images that are difficult to distinguish from real content [15].

This new technology can be used in various fields of multimedia processing, namely for entertainment, movie production, online live broadcasting, and privacy protection [48]. On the other hand, it can be misused, for example, to spoof biometric security systems, to commit fraud, and spread misinformation [15]. Another malicious usage of this technology can be found in areas like spreading propaganda, fomenting political discord and hate, or even harassing and blackmailing people [52].

Understanding voice deepfakes is essential for creating relevant and usable datasets needed for training deepfake detectors. Using outdated methods contributes less, as detection systems are already trained to recognize them [64]. Employing modern tools ensures the generation of more convincing deepfakes capable of evading modern detectors.

2.2 Text-to-speech

Text-to-speech is a deepfake creation technique that takes written text as input, and the output is a synthetic voice reading the input text [68]. The main goal of this approach is to automate the speech generation [53], which can be leveraged, for example, in smart home assistants or navigational systems [15].

An introduction to the problem is presented here, along with a description of selected models. Portions of this content, unless explicitly stated, are drawn from Tan et al. [68].

Brief history of text-to-speech synthesis

Early speech synthesis explored the idea of creating artificial articulatory organs to mimic human speech production, a method known as articulatory synthesis. However, the results fell short of expectations, producing speech that lacked a natural, human-like quality. Consequently, this approach was abandoned in favor of its successor, formant synthesis and concatenative synthesis.

The principle of formant synthesis is to mimic the formant frequencies in human speech. These are a concentration of acoustic energy at a specific frequency, created by the shape of the vocal tract during speech. It helps distinguish different vowel sounds by emphasizing certain sound levels. [23]

Concatenative synthesis generates speech by combining small prerecorded units such as phonemes, biphonemes, and triphonemes. While effective, this method requires significant memory storage and often produces unnatural-sounding speech, as timbre (voice colour) and prosody (intonation, rhythm, speaking style) can be diminished during concatenation.

Statistical Parametric Speech Synthesis (SPSS) addresses the drawbacks of concatenative synthesis by introducing three modules: a text analysis module, an acoustic model, and a vocoder. Firstly, in the text analysis part, the text is processed by extracting linguistic features. Linguistic features are properties of speech that capture essential elements of language, like pronunciation and prosody. It helps the model to produce more natural-sounding speech. From these, the acoustic model, usually implemented by Hidden Markov Models (HMM), extracts acoustic features, which are then fed to the vocoder, processing them into the final waveform. The SPSS was considered a state-of-the-art speech synthesis approach for a long time. Nevertheless, it could not compete with the rise of Neural speech synthesis.

Neural speech synthesis

Neural network-based speech synthesis is a modern approach to generating speech. These systems are composed of two main components: an acoustic model (also called encoder) and a vocoder, which are adopted from the SPSS. The TTS pipeline can be seen in the Figure 2.1. This chapter will dive deeper into the individual parts of the system and will provide a description of some specific models.

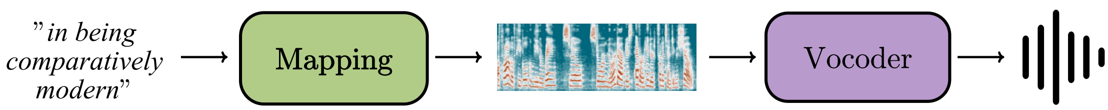

Figure 2.1: Standard neural TTS pipeline, where text is mapped to acoustic features (spectrogram in this case) by an acoustic model. Then the vocoder transforms the input into the final waveform. Taken from Frank and Schönherr [16].

Acoustic models (encoders)

Acoustic models take linguistic features (characters or phonemes) as input and convert them to acoustic features, e.g., or spectrograms. *Mel-spectrograms* are commonly used, as they leverage a logarithmic-like mel scale that is better aligned to human perception of sound [64]. Acoustic features are then passed to a vocoder model. This area's main types of architecture are based on Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), or transformer architectures [7].

Vocoders

Vocoders generate waveforms from acoustic features and can be classified into several distinct categories based on their underlying architectures. Vocoders based on Generative Adversarial Networks (GAN) are one of the most widely used models. In general, a GAN consists of a generator and a discriminator. The word „adversarial“ in this context means that these two modules compete against each other. The generator is trained to maximize the probability of the discriminator making a mistake, while the discriminator's task is to correctly distinguish whether the output comes from the generator or from the training data. The main goal of GAN is to produce audio as realistic as possible. [20]

End-to-end models

End-to-end models do not need manual extraction of context-dependent linguistic features from the text. Instead, they learn these features directly from it using a text encoder. This encoder transforms input text into character embeddings, which are then converted into a representation capturing the necessary context [45]. Fully end-to-end models directly convert characters or phonemes into a waveform. This approach brings the following advantages:

1. Less human annotation and feature development is needed.
2. Error propagation can be avoided compared to models consisting of stages.
3. A model is trained jointly at once, which leads to reduced cost of training, development, and deployment.

This also comes with big challenges in the model training to overcome, mostly due to the differences in format and length between text and speech waveforms.

Tacotron

Tacotron is an end-to-end neural acoustic model that converts text into speech, and it was one of the first widely used models. It takes character inputs, processes them into spectrogram frames, and reconstructs waveforms using the Griffin-Lim algorithm. The model consists of three main components: an encoder, an attention-based decoder, and a post-processing network. The encoder transforms input text into internal representations using a pre-net and the CBHG module (Convolution Bank, Highway Network, Bidirectional GRU), which reduces overfitting. The decoder employs an attention mechanism [63], which helps the model to focus on the important aspects to generate compact seq2seq [65] targets, which are later converted into waveforms. While Tacotron uses Griffin-Lim for simplicity [81], Tacotron 2 improves output quality by incorporating a WaveNet vocoder [60].

XTTS-v2

XTTS-v2[1] is a multilingual, zero-shot TTS model, achieving State Of The Art (SOTA) results in 17 languages. It utilizes Vector Quantised-Variational AutoEncoder (VQ-VAE),

[1] https://huggingface.co/coqui/XTTS-v2

which creates lower-level discrete token representations from the input speech. The token representations are then fed to a Generative Pre-trained Transformer (GPT) model, together with an input text, which predicts the resulting series of tokens. Finally, from this, the resulting waveform is generated using the Hifi-GAN vocoder. [7]

F5-TTS

F5-TTS stands for Fairytaler Fakes Fluent and Faithful speech with Flow matching. It is a fully non-autoregressive zero-shot TTS system, which relies on flow matching with Diffusion Transformer (DiF). The text input is padded with filler tokens to match the length of the input speech. Then, the denoising is performed for speech generation. [10]

FastSpeech

FastSpeech offers significant improvements over Tacotron by addressing its slow inference speed and lack of robustness. Unlike Tacotron, which generates mel-spectrograms sequentially, FastSpeech uses a feed-forward Transformer-based approach to produce mel-spectrograms in parallel, significantly accelerating the process. [55]

WaveNet

WaveNet is a deep neural network designed to generate raw audio waveforms. The model operates in a fully probabilistic and autoregressive manner, meaning that it predicts each audio sample's probability distribution based on all the preceding samples. [44]

HiFi-GAN

HiFi-GAN is a vocoder model that uses the power of GAN. The generator takes spectrograms as input and generates audio waveforms. It has two discriminators work together to improve the quality of the generated audio. Multi-Period Discriminator focuses on short, equally spaced segments. However, because it only looks at these separate pieces, it might miss the overall flow of the sound. The Multi-Scale Discriminator complements it by analyzing the entire audio sequence at different resolutions. [20]

2.3 Voice Conversion

Parts of this section are taken from *Overview of Voice Conversion Methods Based on Deep Learning, by Walczyna and Piotrowski* [76], if not stated otherwise. Voice conversion is a process of seamlessly transferring the speaker's identity to another speaker while preserving the content of the speech. The algorithms are used for speech analysis, speaker classification, and vocoding. The state-of-the-art technology uses deep neural networks. The current voice conversion procedure consists of two parts: one that analyzes and decomposes the voice and extracts its features, and another that maps and combines the extracted elements via reconstructions using a vocoder. This can be seen in the Figure 2.2. The stages of encoder, decoder, and vocoder are the same as with TTS, but in the case of VC, there are two more stages that come before the three mentioned: speaker identity extraction and linguistic content extraction.

A Variational Autoencoder (VAE) is a neural network commonly used for this purpose. It consists of two parts that work together but are built separately: an encoder and a decoder. The encoder's task is to take input data and turn it into an abstraction called a „latent representation“, or embedding, that captures the most important information for the task. The decoder then uses this representation to try to recreate the original data. It is used for both feature extraction and data generation. [28]

Afterward, it merges the target speaker's identity embedding and the source speaker's content embedding to deliver the desired sentence.

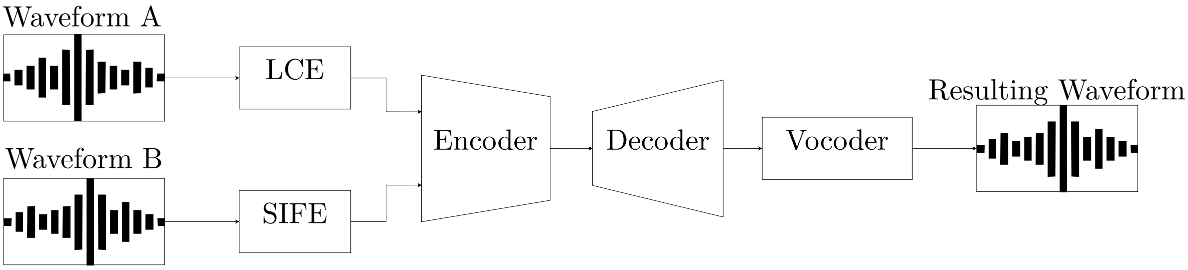

Figure 2.2: Voice conversion pipeline with modules for linguistic content extraction (LCE) and speaker identity feature extraction (SIFE), encoder, decoder, and vocoder, adapted from Walczyna and Piotrowski [76].

Recently, an increasing number of zero-shot[2] voice conversion models have been developed. This aims to change the speaker's characteristics without the need for extended training.

Usage can be found in tasks like expressive speech synthesis, speech enhancement, or pronunciation correction [12]. Other fields of potential application could be entertainment, medicine, teaching, and the military industry. Specific use cases could be the creation of new voices for virtual characters in video games and animated movies, and the ability to synthesize speech for people with speech impairments.

Open Voice v2

Open Voice is able to perform a zero-shot cross-lingual voice cloning, even for languages not included in the training set. It uses a base speaker VITS TTS model and a tone color converter module to embody the reference voice into the generated voice. The resulting waveform is generated by the HiFi-Gan vocoder. [51]

DDDM-VC

DDDM-VC is a voice conversion system that is capable of cross-lingual zero-shot generation, even for unseen languages. The Decoupled Denoising Diffusion Model (DDDM) training process consists of gradually adding noise to a source input until it is just pure noise. During generation, the process is reversed: the model starts from random noise and progressively removes the noise to generate a new output. The term decoupled refers to the fact that the denoising is performed for speech attributes like content of the speech, timbre, and prosody, allowing control of different aspects of the speech. [11]

[2] The model is trained on a large amount of data, so only a short reference audio is needed for calibration to a certain speaker.

FreeVC

The FreeVC framework introduces a non-autoregressive, fully unsupervised voice conversion model, enabling high-quality speaker transformation without requiring paired data or text labels. It leverages self-supervised learning for robust content and speaker disentanglement. [33]

Chapter 3

Existing genuine speech datasets

An overview of existing genuine speech datasets is presented here, which play a crucial role as the foundation for many deepfake databases. Examining these datasets is essential when creating a new one, as it helps identify whether an existing resource can be utilized for the current purpose or if there are gaps that need to be addressed. Special attention will be paid to corpora from political institutions, which seem to be a good source of speech for the purpose of this thesis, as meetings from these institutions are often transcribed, and the demographic information about the speakers can usually be retrieved from publicly available sources. This is particularly useful for generating speech via TTS. Another advantage is the presence of different people, who are publicly known, and therefore, it enables the gathering of additional speaker characteristics.

CSTR VCTK

Speech corpus containing speech of 110 English speakers with various accents reading newspaper articles. Each speaker reads about 400 sentences, making it a total of 44,000 utterances with a duration of 44 hours. [85]

Common Voice Corpus

Large, open-source corpus with 131 languages, with more than 22,000 validated hours and more than 33,151 recorded hours, includes speaker characteristics like age, gender, and accent. [39]

LibriSpeech

Corpus with approximately 1,000 hours of read English by a total of 2,484 speakers, of which 1,201 are female, and 1,283 are male, derived from read audiobooks from the LibriVox project. LibriSpeech is widely used for the training of acoustic models. [47]

LibriTTS

Multi-speaker English corpus of about 585 hours of read English designed for TTS research and derived from the same materials as LibriSpeech. The speech data comes from 2,456 speakers. [88]

VoxForge

Large open-source collection of transcribed speech recordings. The English part has approximately 120 hours of speech from more than 2,900 speakers. [74]

Multilingual Librispeech

Multilingual speech corpus derived from LibriVox audiobooks consisting of eight languages – English, German, Dutch, Spanish, French, Italian, Portuguese, and Polish. Published in the year 2020. In total, it consists of about 44,500 hours of English and about 6,000 hours of other languages. [50]

AISHELL

AISHELL is a family of Mandarin speech corpora. AISHELL-1 was published in 2017 and includes a speech from 400 people with different accent areas [6].

AISHELL-3, released in 2020, contains a total of 88,035 utterances by 218 speakers. Their characteristics, such as gender, age group, and native accents, are provided in the corpus as well as transcripts at the Chinese character level. [61]

AISHELL-4 is recorded using a circular microphone array for speech processing in conference scenarios. The aim is to provide a dataset suitable for multi-speaker processing. [18]

Multilingual TEDx

Multilingual speech corpora with eight languages and transcripts. The languages are Spanish, French, Portuguese, Italian, Russian, Greek, Arabic, and German. Transcripts are aligned with the audio. The total length of the audio is 765 hours, distributed across the languages. Available metadata includes the source language, title, speaker, duration of the audio, keywords, and a short description of the talk. [57]

DAPS

Corpus with speech from 20 speakers (10 male and 10 female), where each speaker reads five texts from public domain stories. [40]

The EMIME Bilingual Database

This database consists of speech recordings from 42 speakers, evenly divided between males and females. It includes data in English, Finnish, and German, with the Finnish and German speakers being bilingual. This dataset is valuable for research on bilingual speech and language processing. [82]

LJSpeech

The LJSpeech dataset is a public domain speech resource containing 13,100 short audio clips of a single speaker reading excerpts from seven non-fiction books. Each clip includes a transcription and ranges from 1 to 10 seconds in length, with a total duration of approximately 24 hours. [24]

Arctic

Speech data from 24 non-native English speakers who are native speakers of these languages: Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. Two male and two female speakers are present in the dataset for each language. [92]

VoxCeleb

The VoxCeleb is a large-scale audio-visual speaker recognition dataset curated from open-source media like YouTube. It contains over one million utterances from more than 6,000 speakers, making it significantly larger than other publicly available Automatic Speaker Recognition (ASR) datasets. Males account for 61% of the total, while females make up 39%. The majority of the speakers are from the following countries: the USA, the UK, Germany, India, and France. Transcripts are not available. [41]

Parliamentary Corpora

Parliamentary corpora were identified as a potential source of genuine speech data for this research. They offer notable advantages, such as large amounts of speech content for individual speakers. This is particularly beneficial for training voice deepfake creation models, where extensive data from a single individual helps enhance the model's ability to learn speaker-specific traits. Furthermore, parliamentary corpora typically include transcripts and speaker identification, which can be utilized to derive additional speaker-related features.

When utilizing this source to create new datasets, it is crucial to consider its limitations. These collections often lack diversity, as the individuals tend to share similar characteristics, such as age and gender. For instance, in the current Czech Chamber of Deputies, women make up only 25% of the participants, which is a historical record, and the average age is nearly 50 years[1]. Additionally, the minimum age to be elected in the Czech Republic is 21, further limiting the demographic range. Although the situation may vary in other countries, these constraints highlight that parliamentary corpora do not provide the most balanced or diverse source of genuine speech, as the contributors often come from a relatively narrow spectrum. As such, it is important to consider alternative sources for constructing a more inclusive dataset.

This section offers an overview of speech datasets available from political institutions, with information sourced partly from the *Parliamentary Corpora* resource family on the CLARIN website[2].

Norwegian Parliamentary Speech Corpus

Speech from 267 speakers makes up to 140 hours in total and is taken from recordings of meetings in the Norwegian parliament.[3]

[1] https://csu.gov.cz/zeny-a-muzi-v-rozhodovacich-procesech?pocet=10&start=0&skupiny=30&razeni=-datumVydani

[2] https://www.clarin.eu/resource-families/parliamentary-corpora

[3] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/

Finnish Parliament ASR Corpus

Database extracted from the Finnish parliament plenary session transcripts and videos, consisting of more than 3,000 hours of speech gathered by 449 speakers. Transcripts are included together with speaker annotations. [73]

Parczech 4.0

Recordings compiled from Chamber of Deputies' meetings. In total, 196,185 utterances have a duration of 4,590 hours [30]. The previous version contains speech from 486 speakers, implying that the number of speakers should be even higher in the newer version [31]. The metadata that can be obtained is, for example, speaker identities, gender, web page links, affiliations committees, and political groups [31].

ParlaSpeech-HR

Croatian parliamentary corpus containing transcribed speech from 309 speakers, divided into 381,849 utterances and annotated with the speaker's name, party, gender, age, status, and role. [37]

VoxPopuli

The VoxPopuli is a large-scale multilingual corpus with 1,800 hours of transcribed speeches in 15 languages and aligned oral interpretations into 15 target languages, totaling 17,300 hours of annotated data. Data is sourced from European Parliament event recordings. [77]

Other genuine speech corpora

This section discusses sources that are not being used for model training, nor are derived from recordings from political institutions. The aim is to find databases that would fill the gap in the parliamentary corpora. Overall, it is about younger speakers or individuals with some notable characteristics worth further research.

DIALEKT

Czech corpus containing recordings from 291 speakers, mostly older than 60 years, from all over the Czech Republic, distributed into 43,628 utterances. A notable annotation, as the name suggests, is the individual's dialect. [19]

Estonian Dialect Corpus

Speech from 381 speakers, annotated with gender, age, and dialect. The average age of the speaker is around 78 years [34].

aGender

German speech corpus with recordings of people speaking over public telephone lines. The data is annotated with gender and age, which range from 7 to 80 years. While no additional annotations are provided, it is worth considering whether the use of telephone recordings is a limitation or a feature. The telephone channel's compression and noise could introduce artifacts that might impact the accuracy of voice deepfake detectors. On the other hand, it may enhance robustness by simulating real-world scenarios. The corpus could also be useful with its wide age range, including children.[4]

Corpus RVG1_CLARIN

Recordings of 500 Germans from different dialect regions in Germany. The metadata includes information about the speaker's age and gender.[4]

Voice Onset Time in Cochlear Implant Patients and Vowel Production in Cochlear Implant Patients

Two corpora, each containing German speech from 48 patients with cochlear implants and 48 control group speakers. The annotations consist of transcription, age, and gender.[4]

Gamli: Icelandic Oral History Corpus

The dataset comprises 210 unique speakers (90 women and 120 men) and includes annotations for age and gender, with most participants being over 60 years old (94.8%) and an average age of 77 years [42].

JASMIN Speech Corpus

Recordings of the Dutch language with speakers of ages ranging from 7 to 60 years old. The speaker selection was based on variables such as region of origin, nativeness, dialect region, age, gender, and Dutch proficiency level for non-native speakers. [13]

Fair-Speech dataset

Focused is given on performance across different demographic groups. Transcribed speech from 593 Americans is divided into about 26,500 utterances. Provided metadata includes information about age, gender, ethnicity, geographic variation, and whether the participants consider themselves native English speakers. [72]

AfriSpeech-200

Around 200 hours of transcribed Pan-African English speech in 67,577 clips from 2,463 unique speakers across 120 Indigenous accents from 13 countries is included in this dataset. [43]

TalkBank

TalkBank[5] claims to be the largest open-access repository for spoken language data. Audio data can be found from speakers with various conditions, such as patients with dementia and aphasia. There is also a subset for studying child language development in 42 languages. While TalkBank is a valuable resource for genuine and transcribed speech, especially from speakers with diverse characteristics, the reviewed corpora were small in size, and a significant amount of data preprocessing would be needed.

[4] https://clarin.phonetik.uni-muenchen.de/BASRepository/
[5] https://ca.talkbank.org/

Chapter 4

Existing deepfake datasets

Studying existing voice deepfake datasets is crucial not only to avoid repeating past mistakes – such as leaving silence in the utterances, as seen in the case of ASVspoof 2019, which can mislead the detectors to focus more on silence rather than speech [90] – but also to identify gaps in the current data, improve the quality of new datasets, and ensure they reflect diverse speaker characteristics for better generalization.

An overview of already existing voice deepfake datasets is presented. The focus is on how many speakers, languages, utterances, voice deepfake creation tools, and core features, that distinguish the dataset from the others.

ASVspoof 2019 LA

The ASVspoof initiative and challenge series started to foster the development of countermeasures to protect Automatic Speaker Verification (ASV) systems against spoofing attacks [35]. The database is derived from the VCTK corpus, which is only English. The whole dataset is divided into three parts: training, development, and evaluation. The total number of speakers for all three subsets is 78, of which 33 are male and 45 are female. Spoofed utterances are generated by one of two voice conversion and four speech synthesis algorithms. The evaluation subset's fake data is created by variants of the algorithms used in the training and development datasets. The number of clips for both training and development parts is 5,128 genuine and 45,096 fake. Speaker characteristics that can be obtained from the dataset are only gender, which can have three values: male, female, and gender independent [84]. The number of spoofing algorithms is 17 in total, with six seen and 11 unseen during training [80].

ASVspoof 2021 LA

The speech in this dataset, both bona fide and spoofed, is transmitted across real telephony systems, like Voice over IP (VoIP) and Public Switched Telephone Network (PSTN), to achieve more realistic results instead of using clean audio. The purpose is to reduce the gap between laboratory conditions and those expected in real-life scenarios. Speech source data is taken from the previous run of the challenge and, therefore, also from the VCTK corpus and English only. For the faked speech, one of 13 different VC, TTS, or hybrid algorithms is used. A total number of utterances is 16,492 bona fide and 148,148 fake ones. The dataset consists of speech from 37 females and 30 males. [36]

ASVspoof 2021 DF

The dataset is divided into two splits, progress, and evaluation, the same as with the LA part. Both parts combined include 87 female speakers, 73 male speakers, 20,637 bona fide utterances, and 572,616 spoofed utterances. The speech source is partly from the ASVspoof 2019 LA evaluation set and another part comes from Voice Conversion Challenge (VCC) databases from years 2018 and 2020. The VCC databases are derived from DAPS and EMIME corpora. The spoofed utterances were created by more than 100 algorithms, taken from contributions to ASVspoof 2019 or VCC databases. [36]

ASVspoof 5

The database of ASVspoof 5 is based on the MLS dataset [50]. It contains data from around 2,300 male and 2,400 female speakers. The generated speech comes from more than 30 systems. It is exceptionally large with more than one million high quality utterances. Furthermore, various codecs are utilized and it incorporates adversarial attacks on deepfake detectors for the first time in the history of the challenge. [79]

TIMIT-TTS

An audio-visual deepfake database that can be used as a standalone audio dataset. Only TTS technique is used from 12 different spectrogram generating systems and 2 vocoders. The total number of speakers in this corpus is 37, and the language is English. The dataset has four variants depending on the sound quality, each having 19 780 utterances. [58]

ADD 2022

The speech for this dataset's subsets training and dev was acquired from 40 males and 40 females from Mandarin Chinese corpus AISHELL-3 [61]. The number of utterances is 5,319 genuine ones and 45,367 fake ones. [86]

CFAD

Chinese dataset for deepfake audio detection with fake utterances generated by twelve mainstream speech synthesis techniques. Real speech is sourced from six corpora, including AISHELL-1 [6] and AISHELL-3 [61]. The number of speakers is 1,023 real ones and 279 fake ones, which produced 347,400 clips. [38]

FoR

Fake or Real, known as FoR, is an English deepfake dataset with more than 198,000 utterances, and it was published in 2019. It includes state-of-the-art speech synthesis systems from that time. The generated speech was collected with seven TTS systems, which were learned to produce 33 voices. For a collection of genuine speech, Arctic [92], LJSpeech [24] and VoxForge [74] datasets were used. Apart from that, educational videos like TED talks were utilized, adding speeches from 140 individuals and 3,720 clips in total. [54]

WaveFake

Dataset based on English LJSpeech [24] and Japanese corpus JSUT [62], making it bilingual. Each of the corpora contains speech from one female speaker. Therefore, WaveFake includes speech from two speakers. Spoofed utterances are acquired from six systems. The total number of clips is 117,985. [16]

SynSpeechDDB

Speech corpus based on VCTK [85] and AISHELL-1 [6]. From these corpora, 14,400 genuine utterances were derived in either English or Chinese language, and 127,890 fake speech clips were synthesized using 16 tools. Both TTS and VC systems are included. [91]

Chapter 5

Deepfake detection

The increasing number of audio deepfake tools has drawn the attention of many researchers, with Machine Learning (ML) and Deep Learning (DL) methods being developed to detect them [1].

The detection methods do not focus on a specific kind of voice deepfake, according to the literature. The focus is given on generalization across different speech synthesis technologies [15]. Two main architectural approaches can be found: pipeline and end-to-end solution. [87]

This section will discuss the features used for detection and their types, together with the classification approaches. Understanding how the audio deepfake detectors work can be beneficial for the right selection of the detectors needed for the experiments. Parts of this and the following two subsections were taken from a survey by Yi et al. [87], if not stated otherwise.

Features

The features can be divided into two groups: handcrafted and features extracted by deep neural networks. A large number of these characteristics exist. Therefore, only a few will be mentioned for illustration purposes.

Power spectrum features are the most studied type in the area of fake audio detection. The features are part of the short-term spectral features, which are computed by applying Short-Time Fourier Transform (STFT) on a speech signal. Some of the most popular spectral features include Linear Frequency Cepstral Coefficients (LFCC) [46], Mel-Frequency Cepstral Coefficients (MFCC) [21]. Another category is long-term spectral features, which are better at capturing long-range information from speech signals compared to the previously mentioned short-term spectral features, which capture the information in a frame-by-frame fashion. Features based on Constant-Q Transform (CQT) can be mentioned as a part of this category, with a concrete example being Constant-Q Cepstral Coefficient (CQCC) [32].

The handcrafted features are good at representing information. However, they can suffer biases. The deep features extracted by models based on deep neural networks try to fill the gap. Spectrograms [17] can be used as features in deepfake detection and not only for speech synthesis, as discussed in Section 2.2, about TTS. Another way to extract features is by utilizing a deep neural network feature extractor. These extracted features are generally called embeddings, and as an example, can be mentioned i-vectors or x-vectors. Yet, the best-performing detection systems to date use a novel approach: embeddings extracted by a self-supervised learning, as is apparent from the results of the latest ASVspoof challenge [78]. Using this approach, a model is trained on unlabeled data, and then it is fine-tuned on labeled data, leading to less annotated data needed for training [4]. Wav2vec 2.0 [4] is one of the most popular examples of models from this class. This model combines multi-layer CNN with a Transformer [71] network.

Recently, deep neural networks have achieved strong performance by integrating feature extraction and classification in an end-to-end manner, optimizing both directly on raw speech waveforms [78].

Classification

Many of the classic machine learning approaches were previously used for classification, with the most widely used being Support Vector Machine (SVM) [9] and Gaussian Mixture Model (GMM) [84]. The trend of today's classification lies in deep learning methods, which outperform the previously mentioned methods. For instance, the classifiers can be based on CNN [83], deep residual network (ResNet) [69], or Graph Neural Network (GNN) [66].

5.1 Deepfake detection metrics

To evaluate the performance of deepfake detectors, it is essential to establish appropriate metrics. Metrics of detector performance are introduced in this section. The parts of this text were taken from Altuncu et al. [2].

Confusion matrix

The confusion matrix is a tool used to evaluate the performance of a detector by categorizing its results into four distinct sections. These sections are based on the true class of the input and the predicted class from the detector. The matrix helps to compare the actual outcomes with the predicted results, allowing for a detailed assessment of the detector's accuracy and errors.

The four classes in the confusion matrix are as follows:

- **True Positive (TP)**: Refers to instances where the detector accurately identifies a positive input as positive, meaning both the actual and predicted classes are positive.
- **False Positive (FP)**: Represents cases where the detector mistakenly classifies a negative input as positive, with the actual class being negative but the prediction being positive.
- **False Negative (FN)**: Describes situations where the detector incorrectly categorizes a positive input as negative, where the actual class is positive, but the prediction is negative.
- **True Negative (TN)**: Refers to instances where the detector correctly identifies a negative input as negative, with both the actual and predicted classes being negative.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 5.1: Confusion Matrix

Classification metrics derived from the confusion matrix

Considering the predicted positive samples, two metrics can be defined: True Positive Rate (TPR) and False Positive Rate (FPR). In the confusion matrix, TPR is the fraction of predicted positive samples in the first row, as defined in Equation (5.1), and FPR is the fraction of predicted positive samples in the second row, as shown in Equation (5.2). Similarly, two additional metrics focusing on negative predicted results can be defined: True Negative Rate (TNR), given in Equation (5.3), and False Negative Rate (FNR), defined in Equation (5.4).

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$FPR = \frac{FP}{FP + TN} \tag{5.2}$$

$$TNR = \frac{TN}{TN + FP} \tag{5.3}$$

$$FNR = \frac{FN}{FN + TP} \tag{5.4}$$

Equal Error Rate (EER)

Equal Error Rate (EER) measures a detection system's performance by focusing on two key error rates: the FPR and the FNR. These rates typically conflict, meaning reducing one often increases the other. The EER identifies the threshold where these error rates are equal, offering a single metric to assess a system's overall balance.

Thus EER is simple and easy to interpret, its limitation lies in its equal weighting of FPR and FNR. This makes it unsuitable for applications where one type of error – such as missing a critical deepfake FNR – is far more costly than false alarms FPR.

A major advantage of EER is that it is independent of the decision threshold, which is useful for comparing models without relying on a specific operating point. Additionally, it gives insight into how well-separated the classes are in the feature space – the lower the EER, the better the separation, and therefore, the better the model's performance.

Chapter 6

Dataset design

The previous chapter gave insight into the current state of research, deepfake synthesis and detection, and available databases. The research has shown that other voice deepfake datasets usually do not contain speaker characteristics except for gender and spoken language. Therefore, this thesis aims to fill this gap by proposing a new one.

In particular, it explores the question: How does the effectiveness of voice deepfake detection systems vary based on specific speaker characteristics, and how might these factors introduce biases or impact detection performance? Addressing this question requires a dataset design that comprehensively accounts for diverse speaker attributes.

6.1 Requirements

Developing a new dataset from scratch would be very time-consuming. Therefore, using one of the existing corpora and enhancing it with deepfake utterances is a preferable way of creating the database. However, there are certain limitations that need to be taken into account:

- The utterances have to be transcribed. This is crucial for creating TTS content.
- Rich annotations should be included to enable detailed analysis.
- The speakers should be paired with the recordings annotated with additional characteristics, or it should be possible to obtain them.
- The dataset should be public with a license for research purposes.
- English should be included, as it can be used as a reference point, given that most of the other voice deepfake datasets are mostly English.
- The presence of more languages is important to observe the impact of the language on the deepfake detectors.

For the purpose of this thesis, the VoxPopuli [77] dataset was selected, as it meets all the specified conditions. Another advantage of this dataset worth mentioning is that the number of languages can be scaled up to 16. This opens the door for further investigation in this domain.

6.2 Speaker characteristics

The characteristics that are already part of the VoxPopuli are gender and language. The annotations include speaker identification, from which the level of education and ethnic group can be obtained, and also the birth date, which, combined with the date of the recording, enables computing the age of the speaker. This metadata, which were not included in the original dataset, had to be manually researched and verified using publicly available online sources. Following is a list of the studied characteristics.

- Language: To capture linguistic variety, five languages were selected based on the number of Members of European Parliament (MEPs) representing each corresponding country in the dataset:
  - English
  - French
  - German
  - Spanish
  - Czech

  This selection rule was introduced in order to achieve the most diverse set of speakers. Czech is an exception. Even though it has less MEPs, it is not present in most of the available speech deepfake datasets. Including Czech provides an insight into how these models perform with less-seen languages. It also aligns with the possibility of using the Czech subsection for other research purposes on the faculty.

- Age: For illustration purposes, the average age of MEPs was 53 years in 2014 [59].

- Gender: To achieve gender balance, the dataset will include an equal representation of male and female speakers.

- Ethnic background: The United Kingdom and France have the highest proportion of members from racialized minorities, whereas the Czech Republic, with significantly fewer individuals from such groups, serves as a control group in this context [14].

- Level of education: No relevant statistical data were found during the design phase. Yet, it is likely that members of the European Parliament are predominantly university-educated, which may limit variability in this characteristic. To address this, incorporating additional datasets could enhance diversity and enable more comprehensive analysis.

By incorporating a balanced and varied selection of speakers across these attributes, the dataset aims to provide a robust foundation for analyzing the impact of different speaker characteristics on the performance of voice deepfake detection systems. However, the selection is limited by the speakers present in the European Parliament.

Chapter 7

Dataset compilation

The previous chapter discussed the thoughts behind the dataset design and high-level overview, while this chapter focuses on the practical aspects of the creation process. Firstly, the selection of the speakers is discussed, along with the synthesis tools used and the processing pipeline. Finally, the size and the distribution of speaker characteristics of the resulting dataset are reviewed. This chapter is also a possible source of information for anyone who wants to reproduce this dataset or enhance it with more languages or speakers. The dataset was designed with this in mind, so it is relatively easy to reproduce or extend.

Speaker selection

The criteria on which the speakers in the dataset were selected play an essential role in the overall attributes of the dataset. The main one was the number of real utterances to achieve the most amount of speech possible. The other constraint was a roughly balanced age distribution and not having speakers with similar birth years present in the dataset. Speakers with an ethnic background of some ethnic minority were prioritized. The education level was not taken into account during this phase, and it was only annotated afterwards.

It is worth noting that the speakers from the English subset are not native speakers, and they had the fewest real utterances available – 25, to be specific. The number of utterances from other speakers had to be equalized to this number in order to maintain an equal distribution of real speech across all languages.

Used Synthesis Tools And Fine-tuning

The tools used for the dataset compilation were XTTSv2 [7], F5-TTS [10], Open Voice v2 [51], and DDDM-VC [11].

The XTTSv2 supported all languages included in the dataset for zero-shot, multi-speaker synthesis. Both DDDM-VC and OpenVoice v2 promised cross-lingual, multi-speaker synthesis. The only tool that required additional fine-tuning was the F5-TTS, as the base model[1] trained on the Emilia dataset [22] did not support Czech and Spanish. Czech speech was sourced from the Parczech [31] dataset, specifically *parczech-3.0-asr-train-2021*[2]. Speech data for Spanish and rest of the languages was sourced from the VoxPopuli train split. Since splits in the VoxPopuli dataset are not speaker-disjoint, utterances from

[1] https://github.com/SWivid/F5-TTS

[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3631?show=full

speakers included in the final dataset were removed from the fine-tuning data. This ensures that there will not be any additional bias, caused by the synthesis tool synthesising speech of a seen speaker. The length of the audio in each language-specific fine-tuning set was matched with the 47-hour long Czech dataset. Scripts written by Michal Luner, available from his GitHub repository[3], were utilized for preparing the audio for the fine-tuning. Models across all languages were each trained for approximately 12 hours (73 epochs) using a batch size of 4,096 samples.

7.1 Pipeline

The first necessary step was to obtain the source audio files. Firstly, raw data from the VoxPopuli [77] dataset were downloaded. These were then cut into small bits of audio with a duration of just a few seconds, called segments, using the scripts from the repository[4]. The VoxPopuli is divided into three splits: train, dev, and test. To avoid potential bias, only the dev and test splits were considered for sourcing speech. The train split was excluded because the synthesis tools may have seen this data during training, which could possibly improve their performance.

The dataset was compiled in a pipeline of several steps:

Select speakers and gather annotations

- Identify target speakers and manually collect metadata (e.g., education, ethnicity, birth year).

Create protocol files for each speaker

- Compile a Comma Separated Values (CSV) protocol of the speaker's real utterances.
- Create the fake utterance protocol by subtracting real utterances from the merged `dev` and `test` splits.
- This prevents synthesis systems from generating speech using the same speaker as both source and target.
- Use the same protocol structure for all synthesis tools.

Select reference audio for synthesis

- Choose the longest utterance between 10–15 seconds from the speaker's real data.
- For F5-TTS, append 1 second of silence at the end. Using this approach, the output was more consistent with the source text, with less mispronunciation, in comparison to the reference audio without the silence at the end. This practice was adopted from the inference tips[5].

Convert audio formats

- Convert reference audio (originally in OGG format) to WAV for compatibility with all VC tools.

[3] https://github.com/Mike327327/F5-TTS/
[4] https://github.com/facebookresearch/voxpopuli/tree/main
[5] https://github.com/Mike327327/F5-TTS/blob/main/src/f5_tts/infer/README.md

- Convert all recordings in the `dev` and `test` splits similarly.

Annotate fake utterances

- Use the reference audio's recording year and the speaker's birth year to calculate the age of the speaker at the time of synthesis.
- Accept that this estimation may be off by one year, which is not considered significant for mature speakers.

Apply character limit for synthesis

- Note character length constraints in XTTSv2, which vary per language, as can be seen from Table 7.1.
- Use the smallest safe limit, 180 characters, for all utterances to ensure consistency.
- Apply this limit to normalized text.
- Use raw transcripts for synthesis input, as they yield more natural-sounding speech.

Post-process all utterances

- Equalize the number of synthesized utterances per speaker-tool pair to 1,180 (based on the smallest available split).
- Resample all audio files to 16 kHz and trim leading and trailing silence.
- For real utterances, equalize to 25 per speaker and apply the same resampling and silence-trimming.

Table 7.1: Character limit for input text for the XTTSv2.

Language	Character limit
English	250
German	253
French	273
Spanish	239
Czech	**186**

7.2 Dataset properties

The dataset consists of 50 speakers in total, with 10 per language. Each language split includes five male and five female speakers. Fake utterances were generated using four different speech synthesis tools. Each combination of speaker and tool from the two hundred combinations contains 1,180 utterances, resulting in 236,000 fake utterances. In addition, there are 25 bonafide utterances for each speaker. Finally, the total dataset size is 237,250 utterances. This can be seen in the Table 7.2 for better clarity.

The resulting dataset is balanced across gender and all five languages. Additionally, it features annotation with age, ethnicity, and level of education. The distribution of the utterances can be seen from Figures 7.1, 7.2, and 7.3.

Table 7.2: Dataset Composition Overview

Language	Deepfake utterances	Genuine utterances	M+F speakers	Tools
English	47,200	250	5 + 5	4
German	47,200	250	5 + 5	4
French	47,200	250	5 + 5	4
Spanish	47,200	250	5 + 5	4
Czech	47,200	250	5 + 5	4
Total	236,000	1,250	25 + 25	4

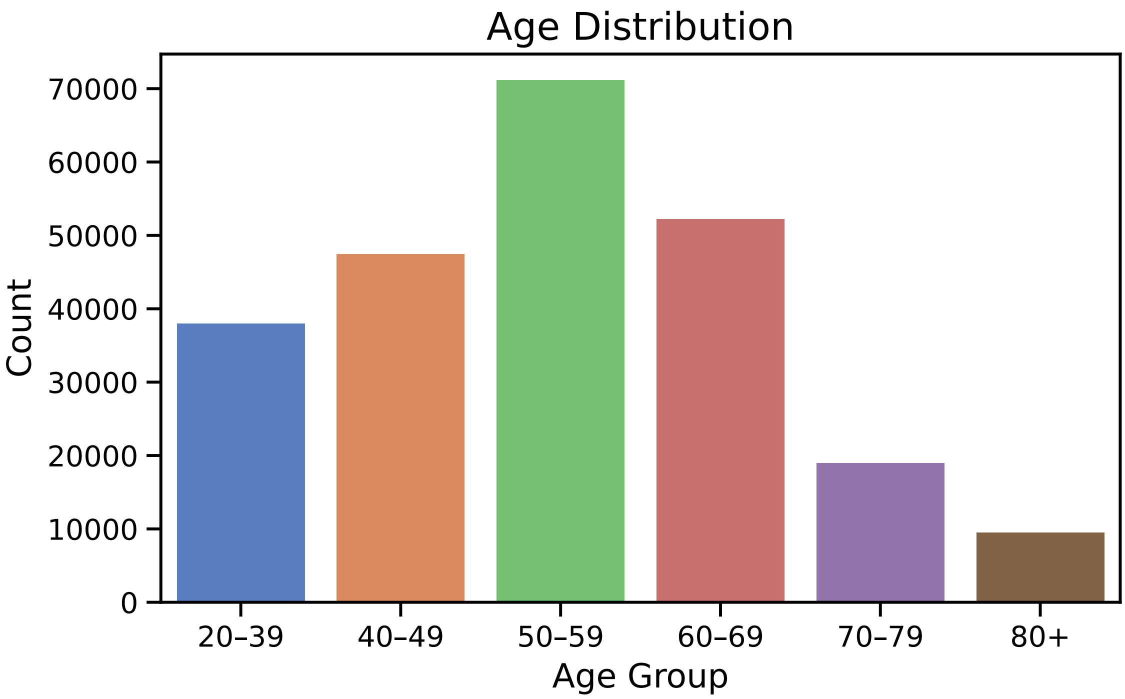

Figure 7.1: Age distribution of recordings across defined age bins.

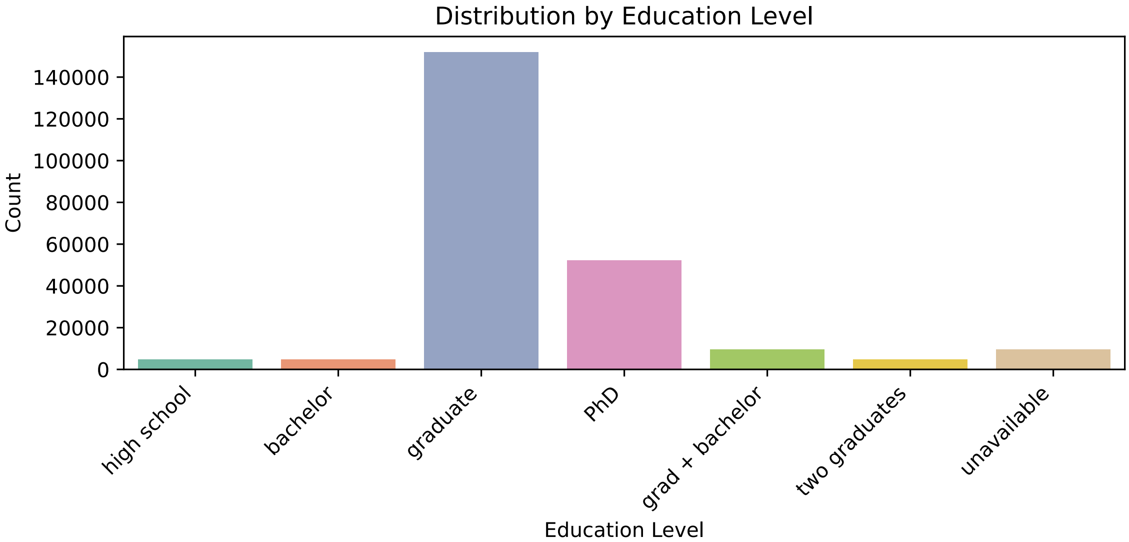

Figure 7.2: Distribution of recordings across education levels.

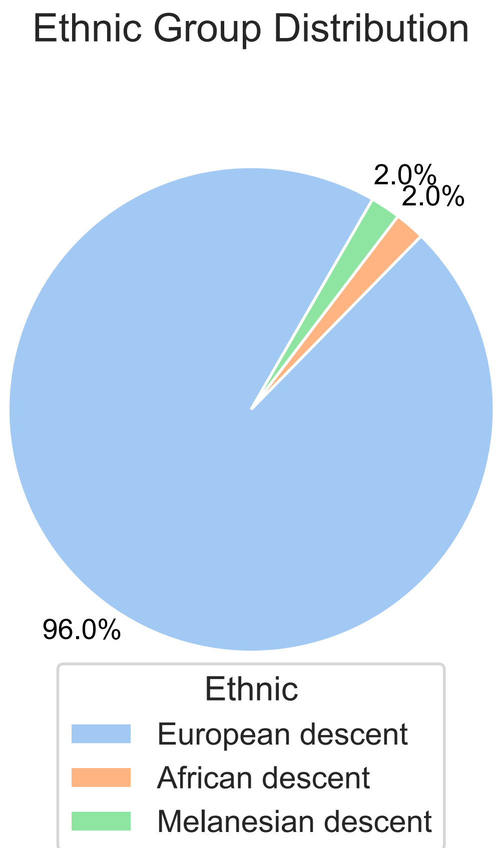

Figure 7.3: Ethnicity distribution of recordings.

Chapter 8

Experiments

The previous chapter presented the proposed dataset. In this chapter, it is put to the test by a total of nine detectors described in Section 8.1. Each annotated speaker characteristic is evaluated in its own section. The impact of ethnicity is not evaluated in this thesis. However, it is included here because the dataset is easily extendable, and the annotation is prepared for anyone who wants to continue this research.

8.1 Used detection tools

The detectors used for the experiments come from the AADA framework [26] and from the Informed methods for deepfake speech detection [64] repository[1], provided by the supervisor of this thesis. The detectors were trained on the ASVspoof2019 LA [84] dataset, and some of them were trained on the ASVspoof 5 [78]. In such a case, it is explicitly written in the name of the tool.

ADAA detectors

The detectors in this section are of an older date, compared to the ones in the following section, but are still used to date [27]. Their feature is that they are very lightweight, and they have low inference time.

- Light Convolutional Neural Network (LCNN) is designed specifically for tasks that require efficient and fast computation, and is also frequently used for real-time processing.
- Rawnet3, as the name suggests, ingests raw audio as input, which does not require any preprocessing. This is a very useful feature when it comes to detecting deepfakes [67].
- Mesonet was originally intended to detect spoofed video but can be used for audio detection as well. One of MesoNet's main strengths is its capacity to process acoustic sequences of different lengths, allowing it to detect subtle artifacts in altered or synthetic audio.
- SpecRNet's name is derived from the term spectrogram, as it processes audio data represented in the time-frequency domain.

[1] https://github.com/vojteskas/Informed-methods-for-deepfake-speech-detection

Informed methods for deepfake speech detectors

The detectors from this section have in common that they all use a pretrained large-scale self-supervised, cross-lingual feature extracting model XLS-R-300M [3]. The models differ in the pooling method used, which helps summarize the rich embeddings extracted from the model into a more compact and usable form.

The first used method is Multi-Head Factorized Attentive Pooling (MHFA) [49], which uses multiple attention heads to group and focus on different parts of the speech signal, capturing more detailed information.

AASIST [25] applies attention mechanisms to graph-structured representations of audio features. This means it treats feature embeddings as nodes in a graph and learns which connections between them are most important. By doing this, it can model complex relationships in the data and focus on the most relevant parts for detecting deepfakes.

SLS [89] is a robust technique that aggregates important features from the layers of the pretrained XLS-R model using adaptive weighting and pooling to obtain a single feature vector used for final classification.

8.2 Impact of gender on detection performance

The goal of this experiment is to determine whether the gender of the speaker influences the performance of voice deepfake detectors. The detectors from the previous Section 8.1 evaluated each utterance from the dataset with a score, a prediction, whether the utterance is bonafide or fake. Their score distributions are statistically evaluated using the Mann-Whitney U test, which is a nonparametric test of the null hypothesis that compares whether two populations have the same distribution. The results of the test are also compared with the EER for each tool. These hypotheses are considered:

H_0 : The detector's score distributions are the same for male and female speakers.

H_1 : The detector's score distributions differ between male and female speakers.

Results

The statistical test was computed separately for the distribution between real and fake utterances. P-values of the test were low (<0.01) with four exceptions, where the difference was not confirmed to be statistically significant. The cases are for real utterances with Rawnet3, SpecRNet, MHFA trained on ASVspoof 5, and SLS tools. Figure 8.1 and Table 8.1 show EER for each detector and gender. For most of the detection systems, it is true that they show better performance with utterances from female speakers, indicated by lower EERs. The results have shown the presence of a bias, which can be seen from the difference in EER and from graph. Even though the difference in score distribution is statistically significant, on the example of the AASIST, trained on the ASVspoof2019 LA, can be seen that the EER varies only slightly, suggesting a limited impact in real-life scenarios.

8.3 Generalization abilities across languages

Given that the detectors were trained only on English datasets, the subject of this experiment is to evaluate the impact of the language on detection abilities. Whether the

Table 8.1: EER (%) across detectors and genders

Training data	Detector	Total	Gender Male	Gender Female
AS19 [80]	LCNN [27]	37.70	36.72	36.63
	RawNet3 [27]	44.21	44.34	43.55
	MesoNet [27]	49.72	47.48	52.04
	SpecRNet [27]	46.40	47.10	45.97
	AASIST [25]	7.56	7.35	6.80
	MHFA [56]	11.38	15.38	11.70
AS5 [78]	AASIST [25]	14.78	15.92	13.73
	MHFA [56]	13.32	11.39	11.79
	SLS [89]	15.37	15.87	14.02

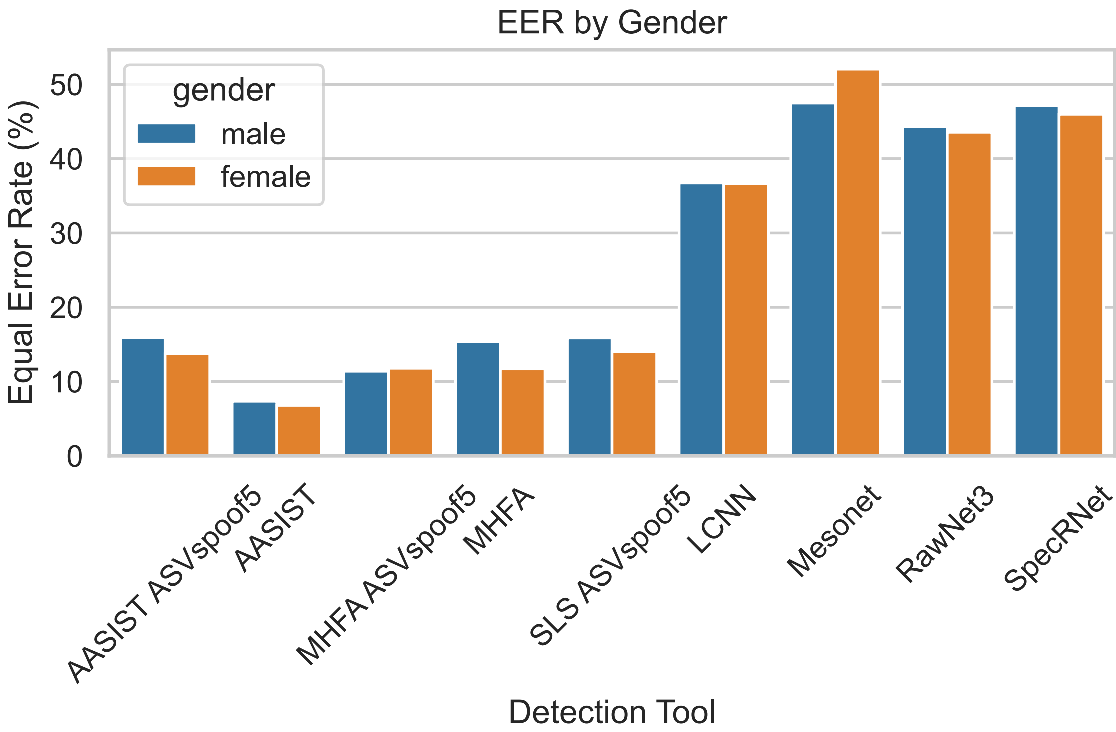

Figure 8.1: EER across gender for all tested detectors.

difference in score distributions across languages is statistically significant was resolved by computing the Kruskal-Wallis H test, which is a nonparametric statistical test that extends the Mann-Whitney U and is able to compare more than just two groups of utterances. The hypotheses are defined as follows:

$$H_0 : \text{The detector's score distributions are the same across languages.}$$
$$H_1 : \text{The detector's score distributions differ across languages.}$$

Results

EER was computed for each detector and language. The results can be found in the Table 8.2, and for better visualization, there is also a bar plot in the Figure 8.2. The statistics were computed separately for real and fake utterances in order to achieve better insight. It is apparent that the p-value is lower than 0.01 for all detectors and fake utterances, and that the null hypothesis H_0 is rejected for all detectors. On the other hand, the hypothesis H_0 is accepted for seven out of nine detectors in the scope of real recordings. It is suspected that the reason behind this phenomenon originates from a slightly different quality of produced deepfakes in different languages, as well as a small sample size of the bonafide recordings.

Table 8.2: EER (%) across detectors and languages

Training data	**Detector**	**Total**	**Language**				
			CS	**DE**	**EN**	**ES**	**FR**
AS19 [80]	LCNN [27]	37.70	30.17	42.88	41.41	36.80	35.67
	RawNet3 [27]	44.21	43.20	50.00	43.27	45.15	41.49
	MesoNet [27]	49.72	52.43	46.00	46.89	52.80	49.35
	SpecRNet [27]	46.40	45.33	50.07	43.79	47.59	42.80
	AASIST [25]	7.56	5.30	9.99	7.20	6.41	7.14
	MHFA [56]	11.38	10.41	15.05	13.20	11.00	15.20
AS5 [78]	AASIST [25]	14.78	13.07	18.48	10.95	14.13	13.50
	MHFA [56]	13.32	13.01	12.76	11.52	7.99	8.63
	SLS [89]	15.37	12.40	18.40	14.95	12.80	12.59

A notable fact can be seen in the Figure 8.2 for the role of English. Intuitively, it would be expected that English would have lower EER from the other languages, due to the fact that all of the detectors were trained only using English. But that is not the case here. English is comparable to other languages, with German as an exception, which represented the biggest challenge for almost all detectors. This is quite surprising because German and English are both members of the Germanic branch of the Indo-European language group, i.e., can be considered similar. This brings up the question of why German stands out. Maybe the synthesis tools are better at generating German speech. Another explanation could be that German is known for its long, concatenated words, which might pose a challenge to detection tools.

To sum things up, the statistics have shown present bias in the score distribution across languages, which can be observed from both the data in Table 8.2 and Figure 8.2. Many tools had the biggest problems with detecting German speech, and English is comparable with other languages, quite unexpectedly.

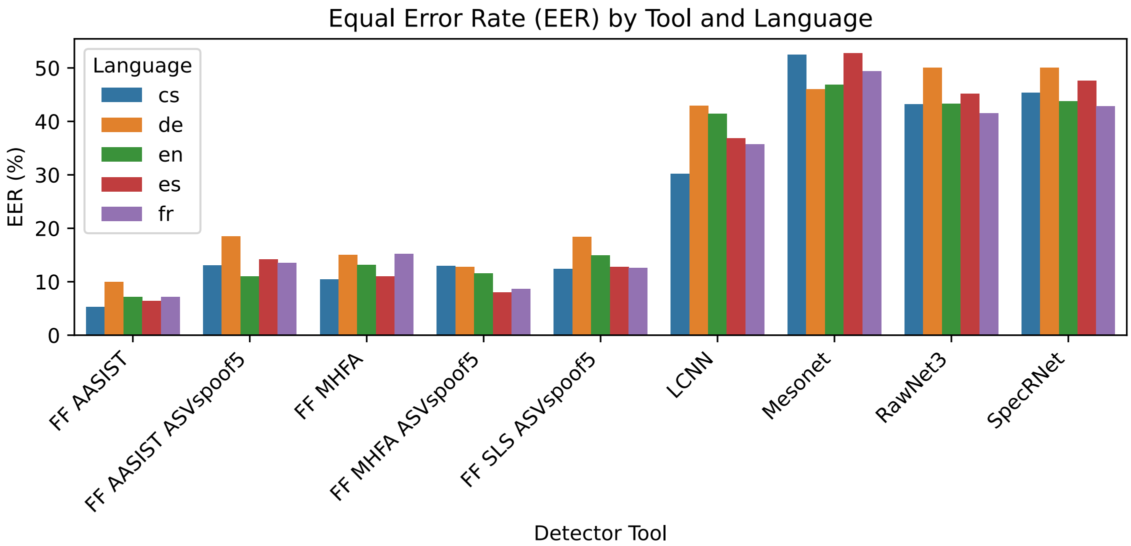

Figure 8.2: EER for each tool and language.

8.4 Age

The evaluation of the impact of age consists of the same steps as described in the previous section. The whole age range was divided into bins by ten years, resulting in six groups: 20-39, 40-49, 50-59, 60-69, 70-79, and 80+ years. Again, EER is computed for each detection tool and evaluated group, together with the Kruskal-Wallis H test for real and fake utterances. The considered hypotheses are these:

H_0 : The detector's score distributions are the same across the age of the speakers.

H_1 : The detector's score distributions differ across the age of the speakers.

Results

From the statistical results, it is apparent that the null hypothesis was rejected in all cases, except one. The score distribution of RawNet3 score across age groups amongst real recordings differs without statistical significance, with a p-value of 0.02, which is the only case of accepting the null hypothesis H_0. In Figure 8.3, it can be seen that the age group of 80+ years causes the most problems to the majority of detection tools. There can also be observed a trend that the performance gets worse with the age of the speaker. The results of the ADAA detectors are worse compared to the rest of the field.

Figure 8.3, based on the data from Table 8.3, reveals a clear trend, that detection performance generally declines with increasing speaker age. The 80+ age group is the most challenging for most tools. This group shows the highest error rates. The ADAA detectors perform particularly poorly in comparison with the rest of the field. A statistical test

Table 8.3: EER (%) across detectors and age groups

Training data	Detector	Total	Age group 20-39	40-49	50-59	60-69	70-79	80+
AS19 [80]	LCNN [27]	37.70	33.98	31.45	32.77	42.95	49.39	50.00
	RawNet3 [27]	44.21	45.95	44.86	45.03	45.56	43.56	40.00
	MesoNet [27]	49.72	44.77	52.84	53.25	45.68	40.32	62.36
	SpecRNet [27]	46.40	50.74	47.88	41.64	45.63	48.13	55.02
	AASIST [25]	7.56	5.62	6.85	7.36	7.70	9.00	10.83
	MHFA [56]	11.38	13.37	12.45	13.03	14.87	13.96	18.00
AS5 [78]	AASIST [25]	14.78	14.92	12.46	14.10	16.77	14.33	14.00
	MHFA [56]	13.32	8.12	10.67	11.64	10.85	11.43	16.39
	SLS [89]	15.37	11.44	13.63	15.26	15.18	17.78	18.00

confirmed the presence of bias, indicating that spoofed speech from older speakers is less likely to be correctly identified.

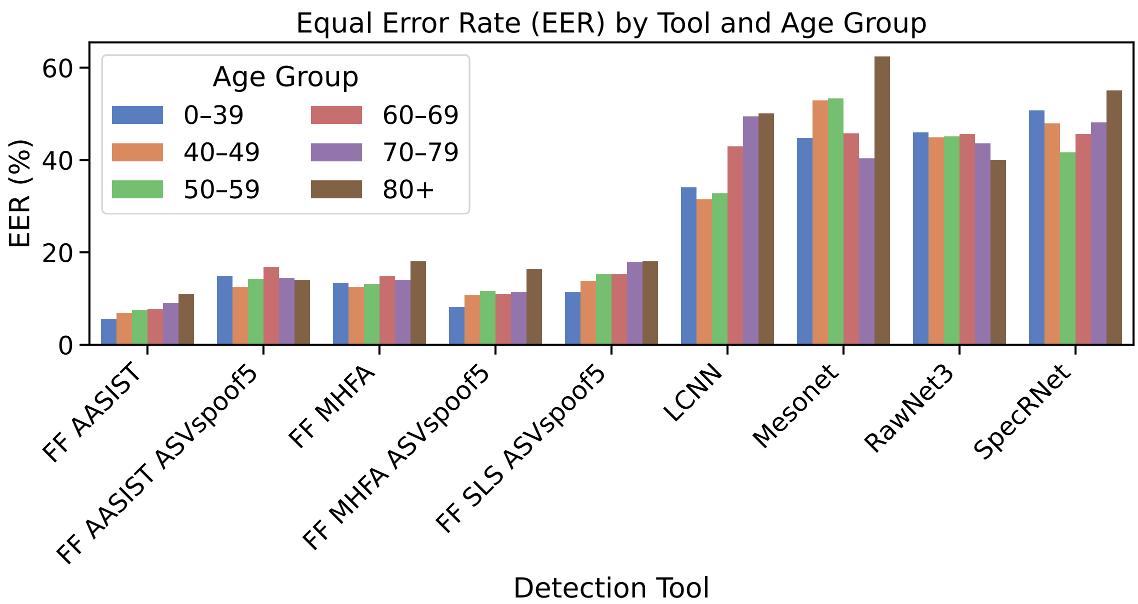

Figure 8.3: Score distribution across age groups for all detectors.

8.5 Education

Impact of education is also computed with the same methods as in the two previous sections: EER and Kruskal-Wallis H test are employed. The hypotheses are these:

H_0 : The detector's score distributions are the same across the education of the speakers.

H_1 : The detector's score distributions differ across the education of the speakers.

Results

The results show that the distribution differs with statistical significance for all of the fake utterances, rejecting the null hypothesis H_0 and accepting the hypothesis H_1. With the real recordings, the null hypothesis H_0 is accepted for three detection systems, namely RawNet3, SpecRNet, and MHFA trained on ASVspoof 5, but rejected for the others.

A closer look at the EER distribution in Figure 8.4 shows that the detectors struggled the most with the speech of the single speaker who had a high-school level of education, while the only speaker with a bachelor's degree proved to be the easiest to detect. Regarding that there is just one speaker per education category, it does not provide a firm ground for any meaningful conclusions, because it could be just a coincidence as well.

Table 8.4: EER (%) across detectors and level of education groups

Training data	Detector	Total	Education group						
			H-S	Bc.	Grad	PhD	Grad + Bc.	2 Grads	NA
AS19 [80]	LCNN [27]	37.70	42.58	40.00	37.63	36.73	45.31	36.00	30.00
	RawNet3 [27]	44.21	39.09	48.00	43.50	46.18	48.00	40.00	34.00
	MesoNet [27]	49.72	56.00	44.00	51.75	45.40	50.00	52.00	45.73
	SpecRNet [27]	46.40	48.00	53.03	46.19	45.09	52.00	55.38	38.14
	AASIST [25]	7.56	12.00	4.00	7.00	8.36	10.01	4.00	9.05
	MHFA [56]	11.38	16.00	11.31	13.51	12.14	11.99	16.00	15.03
AS5 [78]	AASIST [25]	14.78	24.00	6.67	14.63	14.18	15.17	16.00	18.00
	MHFA [56]	13.32	16.00	3.29	11.76	11.24	14.33	9.10	11.58
	SLS [89]	15.37	22.86	10.30	14.75	16.06	18.00	9.56	15.12

Data from the Table 8.4 is visualized in the Figure 8.4. For the majority of the tools, the EER is the highest for the speaker with a high school level of education, while it is lowest for the speaker with a bachelor's degree.

8.6 Discussion

A big majority of the executed experiments suggest that there is a statistically significant difference in the score distribution across various speaker characteristics. This could mean that there is a bias present in the detection ability of the studied tools, which could be mitigated, or at least significantly reduced, by deploying novel training datasets. Generally speaking, the performance of the legacy detectors from ADAA is significantly worse compared to the newer ones. A positive finding is that for a non-negligible number of detectors and speaker characteristics, the difference between score distributions is statistically insignificant for genuine utterances. This is most probably caused by a smaller sample size.

Released deepfake speech detectors should pay special attention to language bias, especially if deployed in multilingual scenarios. One way to help mitigate the language bias could be developing new datasets in less-represented languages (or in the case of this thesis, German, which caused the biggest trouble for detectors). Using these datasets, the detectors could be better trained and prepared for the unseen scenarios in practical deployment.

Another observed fact, which could raise some concerns, is that most of the detectors are less accurate towards elderly people aged 80+ years. Older people are often the target of various scams because of their gullibility. The potential attack vector could look like this: an attacker could exploit this vulnerability by mimicking the voice of an elderly person's

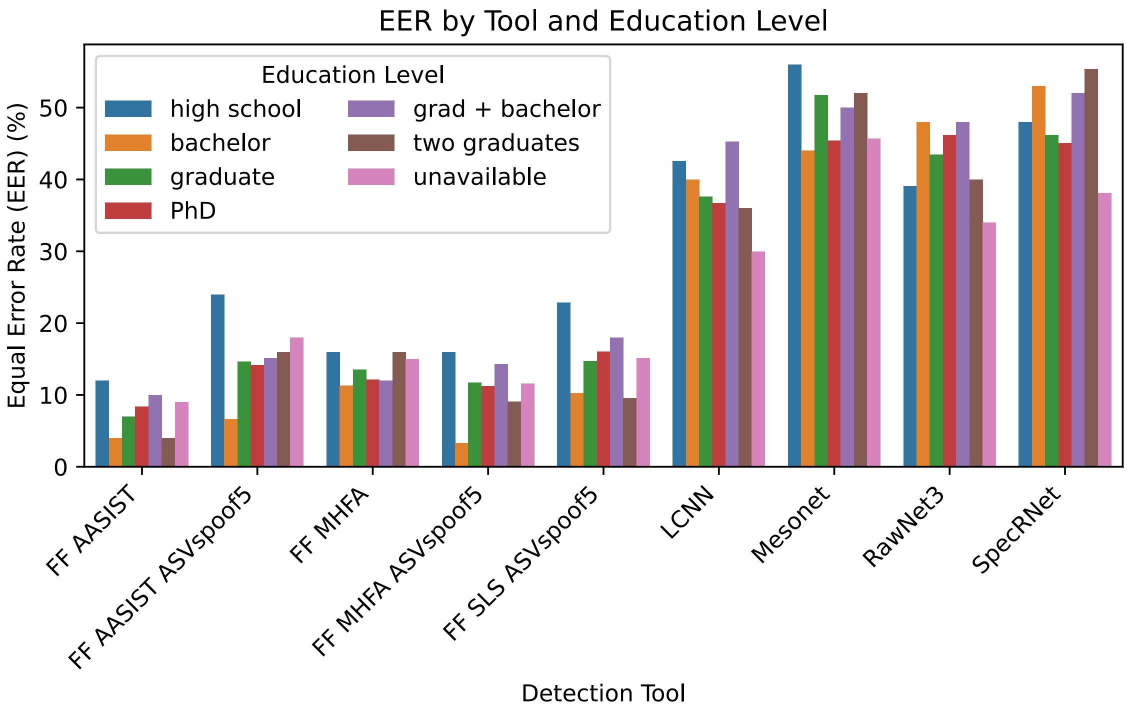

Figure 8.4: EER across all detection tools and levels of education.

spouse, pretending to be in a financial or health emergency, and requesting a transfer of money to a specified bank account. With the bias present in the detection tools, it would be difficult to prove the misuse of the deepfake technology. Including a more diverse representation of elderly people in training data could enhance the detectors' performance, helping protect this digitally vulnerable demographic group.

Chapter 9

Conclusions

The aim of this thesis was to study the impact of speaker characteristics on the performance of voice deepfake detectors. In the theoretical part, the topic is introduced to the reader by describing the types of voice deepfakes and a conceptual overview of how these tools work. The detection tools are discussed as well, and special attention is paid to already existing datasets, both genuine speech corpora as well as deepfake speech datasets.

Then, the design of the new dataset is presented, together with the description of its compilation process and the features of the new dataset. The research on the existing datasets has shown that the best candidate for creating the proposed dataset is the VoxPopuli [77], and also that there is a plethora of other datasets that could be used for creating more deepfake datasets for studying different speaker characteristics.

The created dataset consists of speech from 50 speakers in five languages: Czech, English, French, German, and Spanish. The total duration is more than 500 hours, distributed in 237,250 utterances. Four synthesis tools are employed, namely XTTSv2, F5-TTS, DDDM-VC, Open Voice v2. It features detailed protocols in CSV format. Annotations of the dataset include speaker information, such as gender, age, language, ethnicity, and education level. Each utterance also has its transcription. The dataset can be easily extended by other transcribed languages in the VoxPopuli dataset.

Finally, the dataset is used to measure the performance of nine detection tools. The evaluation of the results is done by comparing EERs and statistical tests, which have shown a statistically significant difference in the detectors' performance across various speaker characteristics. This might be a sign that the detectors under test are biased. The first of the tested characteristics was gender, and it was shown that real male utterances are more likely to be falsely labeled as deepfakes and vice versa. The statistical tests have also confirmed bias between languages, and many detectors had the biggest problem with detecting German utterances. The third tested characteristic was age. It was uncovered that the speech of older people is harder to detect. Last, but not least evaluated characteristic was education level, where the difference in score distribution was statistically significant.

The problem with this dataset is that there is not enough variety of ethnicity, which makes it impossible to do any relevant experiments in this area, yet. Future work could lead in this direction by obtaining more diverse utterances from some other datasets. Another interesting idea is to create a well-rounded dataset that would be balanced in all of its aspects. A lack that could be filled is also studying if there is any present bias towards children and young people, who are not included in this dataset.

In summary, this bachelor's thesis identified a gap in the fairness evaluation of deepfake detection systems. A novel multilingual dataset was designed and compiled, featuring rich speaker annotations. Experiments conducted on various detectors and demographic groups have confirmed the presence of bias in the performance of these tools. The findings accent the need for developing and evaluating detection tools with a greater focus on equality.

Bibliography

[1] Almutairi, Z. and Elgibreen, H. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms*, 2022, vol. 15, no. 5. ISSN 1999-4893. Available at: https://www.mdpi.com/1999-4893/15/5/155.

[2] Altuncu, E.; Franqueira, V. N. L. and Li, S. Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data*. Frontiers Media SA, september 2024, vol. 7. ISSN 2624-909X. Available at: http://dx.doi.org/10.3389/fdata.2024.1400024.

[3] Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q. et al. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. 2021. Available at: https://arxiv.org/abs/2111.09296.

[4] Baevski, A.; Zhou, H.; Mohamed, A. and Auli, M. *Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. Available at: https://arxiv.org/abs/2006.11477.

[5] Birchwood University. *Exploring the Bright Side of Deepfake Technology*. 2024. Available at: https://www.birchwoodu.org/exploring-the-bright-side-of-deepfake-technology/.

[6] Bu, H.; Du, J.; Na, X.; Wu, B. and Zheng, H. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In: *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. 2017, p. 1–5.

[7] Casanova, E.; Davis, K.; Gölge, E.; Göknar, G.; Gulea, I. et al. *XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model*. 2024. Available at: https://arxiv.org/abs/2406.04904.

[8] Center, W. Positive Use Cases for Deepfakes, 2025. Available at: https://www.wilsoncenter.org/article/positive-use-cases-deepfakes. Accessed: 2025-05-08.

[9] Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* New York, NY, USA: Association for Computing Machinery, may 2011, vol. 2, no. 3. ISSN 2157-6904. Available at: https://doi.org/10.1145/1961189.1961199.

[10] Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C. et al. *F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching*. 2024. Available at: https://arxiv.org/abs/2410.06885.

[11] Choi, H.-Y.; Lee, S.-H. and Lee, S.-W. *DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion*. 2023. Available at: https://arxiv.org/abs/2305.15816.

[12] Chou, J. chieh; Yeh, C. chieh and Lee, H. yi. *One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization*. 2019. Available at: https://arxiv.org/abs/1904.05742.

[13] Cucchiarini, C. and Van hamme, H. The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People. In: Spyns, P. and Odijk, J., ed. *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, p. 43–59. ISBN 978-3-642-30910-6. Available at: https://doi.org/10.1007/978-3-642-30910-6_3.

[14] European Network Against Racism (ENAR). *Diversity in the European Parliament 2024*. 2024. Available at: https://www.enar-eu.org/wp-content/uploads/ENAR-Diversity-in-the-European-Parliament-2024.pdf. Accessed: 2025-01-12.

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

[16] Frank, J. and Schönherr, L. *WaveFake: A Data Set to Facilitate Audio Deepfake Detection*. 2021. Available at: https://arxiv.org/abs/2111.02813.

[17] Fu, Q.; Teng, Z.; White, J.; Powell, M. and Schmidt, D. C. *FastAudio: A Learnable Audio Front-End for Spoof Speech Detection*. 2021. Available at: https://arxiv.org/abs/2109.02774.

[18] Fu, Y.; Cheng, L.; Lv, S.; Jv, Y.; Kong, Y. et al. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario. In: *Proceedings of Interspeech*. 2021. Available at: https://arxiv.org/abs/2104.03603.

[19] Goláňová, H.; Waclawičová, M. and Lukeš, D. *DIALEKT: nářeční korpus, verze 2 z 23. 12. 2021*. Ústav Českého národního korpusu FF UK, Praha, 2021. Available at: http://www.korpus.cz.

[20] Goodfellow, I. J.; Pouget Abadie, J.; Mirza, M.; Xu, B.; Warde Farley, D. et al. *Generative Adversarial Networks*. 2014. Available at: https://arxiv.org/abs/1406.2661.

[21] Hamza, A.; Javed, A. R. R.; Iqbal, F.; Kryvinska, N.; Almadhor, A. S. et al. Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access*, 2022, vol. 10, p. 134018–134028.

[22] He, H.; Shang, Z.; Wang, C.; Li, X.; Gu, Y. et al. *Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation*. 2024. Available at: https://arxiv.org/abs/2407.05361.

[23] Hong Kong, T. E. U. of. *Formants of Vowels* https://corpus.eduhk.hk/english_pronunciation/index.php/2-2-formants-of-vowels/. 2025. Accessed: 2025-05-08.

[24] Ito, K. and Johnson, L. *The LJ Speech Dataset* https://keithito.com/LJ-Speech-Dataset/. 2017. Accessed: 2025-01-05.

[25] Jung, J. weon; Heo, H.-S.; Tak, H.; Shim, H. jin; Chung, J. S. et al. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In: *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2022.

[26] Kawa, P.; Plata, M. and Syga, P. Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection. In: *Interspeech 2022*. ISCA, September 2022, p. 4023–4027. Available at: http://dx.doi.org/10.21437/Interspeech.2022-10078.

[27] Kawa, P.; Plata, M. and Syga, P. Defense Against Adversarial Attacks on Audio DeepFake Detection. In: *Interspeech 2023*. 2023, p. 5276–5280. ISSN 2958-1796.

[28] Kingma, D. P. and Welling, M. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*. Now Publishers, 2019, vol. 12, no. 4, p. 307–392. ISSN 1935-8245. Available at: http://dx.doi.org/10.1561/2200000056.

[29] Knight, L. BBC harnesses AI to create writing classes given by Agatha Christie. *The Guardian*, 2025. Available at: https://www.theguardian.com/books/2025/apr/30/bbc-harnesses-ai-to-create-writing-classes-given-by-agatha-christie. Accessed: 2025-05-08.

[30] Kopp, M. *ParCzech 4.0*. LINDAT/CLARIAH-CZ digital library, 2024. Available at: http://hdl.handle.net/11234/1-5360.

[31] Kopp, M.; Stankov, V.; Krůza, J. O.; Straňák, P. and Bojar, O. ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata. In: *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 293–304. ISBN 978-3-030-83526-2. Available at: https://doi.org/10.1007/978-3-030-83527-9_25.

[32] Lai, C.-I.; Chen, N.; Villalba, J. and Dehak, N. *ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual neTworks*. 2019. Available at: https://arxiv.org/abs/1904.01120.

[33] li, J.; tu, W. and xiao, L. *FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion*. 2022. Available at: https://arxiv.org/abs/2210.15418.

[34] Lindström, L.; Todesk, T. and Pilvik, M.-L. *Eesti murrete korpus*. 2022. Available at: https://datadoi.ee/handle/33/492.

[35] Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H. et al. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* IEEE Press, june 2023, vol. 31, p. 2507–2522. ISSN 2329-9290. Available at: https://doi.org/10.1109/TASLP.2023.3285283.

[36] Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H. et al. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Institute of Electrical and Electronics Engineers (IEEE), 2023, vol. 31, p. 2507–2522. ISSN 2329-9304. Available at: http://dx.doi.org/10.1109/TASLP.2023.3285283.

[37] Ljubešić, N.; Koržinek, D.; Rupnik, P. and Jazbec, I.-P. ParlaSpeech-HR – a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus. In: Fišer, D.; Eskevich, M.; Lenardič, J. and Jong, F. de, ed. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, p. 111–116. Available at: https://aclanthology.org/2022.parlaclarin-1.16/.

[38] Ma, H.; Yi, J.; Wang, C.; Yan, X.; Tao, J. et al. *CFAD: A Chinese Dataset for Fake Audio Detection*. 2023. Available at: https://arxiv.org/abs/2207.12308.

[39] Mozilla Foundation. *Mozilla Common Voice Datasets*. 2025. Available at: https://commonvoice.mozilla.org/en/datasets. Accessed: 2025-01-04.

[40] Mysore, G. J. Can we Automatically Transform Speech Recorded on Common Consumer Devices in Real-World Environments into Professional Production Quality Speech?—A Dataset, Insights, and Challenges. *IEEE Signal Processing Letters*, 2015, vol. 22, no. 8, p. 1006–1010.

[41] Nagrani, A.; Chung, J. S.; Xie, W. and Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*. Elsevier, 2019.

[42] O'Brien, L.; Ingimundarson, F. Á.; Steingrímsson, S.; Þorsteinsdóttir, R.; Dagsson, T. et al. *Gamli: Icelandic Oral History Corpus (2023-06-05)*. 2023. Available at: http://hdl.handle.net/20.500.12537/315. CLARIN-IS.

[43] Olatunji, T.; Afonja, T.; Yadavalli, A.; Emezue, C. C.; Singh, S. et al. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, december 2023, vol. 11, p. 1669–1685. ISSN 2307-387X. Available at: https://doi.org/10.1162/tacl_a_00627.

[44] Oord, A. van den; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O. et al. *WaveNet: A Generative Model for Raw Audio*. 2016. Available at: https://arxiv.org/abs/1609.03499.

[45] Pamisetty, G. and Murty, K. S. R. *Prosody-TTS: An end-to-end speech synthesis system with prosody control*. 2021. Available at: https://arxiv.org/abs/2110.02854.

[46] Pan, J.; Nie, S.; Zhang, H.; He, S.; Zhang, K. et al. Speaker recognition-assisted robust audio deepfake detection. In: *Interspeech 2022*. 2022, p. 4202–4206. ISSN 2958-1796.

[47] Panayotov, V.; Chen, G.; Povey, D. and Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, p. 5206–5210. Available at: https://api.semanticscholar.org/CorpusID:2191379.

[48] Pei, G.; Zhang, J.; Hu, M.; Zhang, Z.; Wang, C. et al. *Deepfake Generation and Detection: A Benchmark and Survey*. 2024. Available at: https://arxiv.org/abs/2403.17881.

[49] Peng, J.; Plchot, O.; Stafylakis, T.; Mosner, L.; Burget, L. et al. *An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification*. 2022. Available at: https://arxiv.org/abs/2210.01273.

[50] Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G. and Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. *ArXiv*, 2020, abs/2012.03411.

[51] Qin, Z.; Zhao, W.; Yu, X. and Sun, X. *OpenVoice: Versatile Instant Voice Cloning*. 2024. Available at: https://arxiv.org/abs/2312.01479.

[52] Rana, M.; Nobi, M.; Murali, B. and Sung, A. Deepfake Detection: A Systematic Literature Review. *IEEE Access*, january 2022, vol. 10, p. 1–1.

[53] Rashad, M. Z.; El Bakry, H. M.; Isma'il, I. R. and Mastorakis, N. An overview of text-to-speech synthesis techniques. In: *Proceedings of the 4th International Conference on Communications and Information Technology*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2010, p. 84–89. CIT'10. ISBN 9789604742073.

[54] Reimao, R. and Tzerpos, V. FoR: A Dataset for Synthetic Speech Detection. In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019, p. 1–10.

[55] Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S. et al. *FastSpeech: Fast, Robust and Controllable Text to Speech*. 2019. Available at: https://arxiv.org/abs/1905.09263.

[56] Rohdin, J.; Zhang, L.; Plchot, O.; Staněk, V.; Mihola, D. et al. BUT systems and analyses for the ASVspoof 5 Challenge. In: *The Automatic Speaker Verification Spoofing Countermeasures Workshop*. 2024, p. 24–31.

[57] Salesky, E.; Wiesner, M.; Bremerman, J.; Cattoni, R.; Negri, M. et al. Multilingual TEDx Corpus for Speech Recognition and Translation. In: *Proceedings of Interspeech*. 2021.

[58] Salvi, D.; Hosler, B.; Bestagini, P.; Stamm, M. C. and Tubaro, S. TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access*, 2023, vol. 11, p. 50851–50866. ISSN 2169-3536.

[59] Service, E. P. R. *Age of MEPs by Member State*. 2024. Available at: https://epthinktank.eu/2024/06/04/european-parliament-facts-and-figures/4-age-of-meps-by-ms-3/. Accessed: 2025-01-12.

[60] Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N. et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. Available at: https://arxiv.org/abs/1712.05884.

[61] Shi, Y.; Bu, H.; Xu, X.; Zhang, S. and Li, M. AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines. In: *Proceedings of Interspeech*. 2020. Available at: https://arxiv.org/abs/2010.11567.

[62] Sonobe, R.; Takamichi, S. and Saruwatari, H. *JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis*. 2017. Available at: https://arxiv.org/abs/1711.00354.

[63] Soydaner, D. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*. Springer Science and Business Media LLC, may 2022, vol. 34, no. 16, p. 13371–13385. ISSN 1433-3058. Available at: http://dx.doi.org/10.1007/s00521-022-07366-3.

[64] Staněk, V. *Differential-based Deepfake Speech Detection*. Brno, 2023. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Firc, I. A.

[65] Sutskever, I.; Vinyals, O. and Le, Q. V. *Sequence to Sequence Learning with Neural Networks*. 2014. Available at: https://arxiv.org/abs/1409.3215.

[66] Tak, H.; Jung, J. weon; Patino, J.; Todisco, M. and Evans, N. *Graph Attention Networks for Anti-Spoofing*. 2021. Available at: https://arxiv.org/abs/2104.03654.

[67] Tak, H.; Patino, J.; Todisco, M.; Nautsch, A.; Evans, N. et al. *End-to-end anti-spoofing with RawNet2*. 2021. Available at: https://arxiv.org/abs/2011.01108.

[68] Tan, X.; Qin, T.; Soong, F. and Liu, T.-Y. *A Survey on Neural Speech Synthesis*. 2021. Available at: https://arxiv.org/abs/2106.15561.

[69] Tomilov, A.; Svishchev, A.; Volkova, M.; Chirkovskiy, A.; Kondratev, A. et al. STC Antispoofing Systems for the ASVspoof2021 Challenge. In: *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*. 2021, p. 61–67.

[70] Trnovská, E. *Multilingual Voice Deepfake Dataset*. Brno, 2024. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Available at: https://theses.cz/id/a9kvdl/. Supervisor: Mgr. Kamil Malinka, Ph.D.

[71] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. et al. *Attention Is All You Need*. 2023. Available at: https://arxiv.org/abs/1706.03762.

[72] Veliche, I.-E.; Huang, Z.; Kochaniyan, V. A.; Peng, F.; Kalinli, O. et al. *Towards measuring fairness in speech recognition: Fair-Speech dataset*. 2024. Available at: https://arxiv.org/abs/2408.12734.

[73] Virkkunen, A.; Rouhe, A.; Phan, N. and Kurimo, M. Finnish parliament ASR corpus: Analysis, benchmarks and statistics. *Language Resources and Evaluation*, 2023, vol. 57, no. 4, p. 1645–1670. Available at: https://doi.org/10.1007/s10579-023-09650-7.

[74] VoxForge. *VoxForge Speech Dataset*. 2025. Available at: https://www.voxforge.org/home. Accessed: 2025-01-04.

[75] Wakefield, J. Deepfake presidents used in Russia-Ukraine war. *BBC News*, 2022. Available at: https://www.bbc.com/news/technology-60780142. Accessed: 2025-05-08.

[76] Walczyna, T. and Piotrowski, Z. Overview of Voice Conversion Methods Based on Deep Learning. *Applied Sciences*, 2023, vol. 13, no. 5. ISSN 2076-3417. Available at: https://www.mdpi.com/2076-3417/13/5/3100.

[77] Wang, C.; Riviere, M.; Lee, A.; Wu, A.; Talnikar, C. et al. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In: Zong, C.; Xia, F.; Li, W. and Navigli, R., ed. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, August 2021, p. 993–1003. Available at: https://aclanthology.org/2021.acl-long.80/.

[78] Wang, X.; Delgado, H.; Tak, H.; Jung, J. weon; Shim, H. jin et al. *ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale*. 2024. Available at: https://arxiv.org/abs/2408.08739.

[79] Wang, X.; Delgado, H.; Tak, H.; Jung, J. weon; Shim, H. jin et al. *ASVspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech*. 2025. Available at: https://arxiv.org/abs/2502.08857.

[80] Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A. et al. *ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech*. 2020. Available at: https://arxiv.org/abs/1911.01601.

[81] Wang, Y.; Skerry Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J. et al. *Tacotron: Towards End-to-End Speech Synthesis*. 2017. Available at: https://arxiv.org/abs/1703.10135.

[82] Wester, M. *The EMIME Bilingual Database*. EDI-INF-RR-1388. The University of Edinburgh, 2010. Available at: https://www.emime.org/participate/emime-bilingual-database.html.

[83] Wu, Z.; Das, R. K.; Yang, J. and Li, H. *Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks*. 2020. Available at: https://arxiv.org/abs/2009.09637.

[84] Yamagishi, J.; Todisco, M.; Sahidullah, M.; Delgado, H.; Wang, X. et al. *ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database*. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019. Available at: https://doi.org/10.7488/ds/2555. [sound].

[85] Yamagishi, J.; Veaux, C. and MacDonald, K. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)* Sound. 2019. Available at: https://doi.org/10.7488/ds/2645.

[86] Yi, J.; Fu, R.; Tao, J.; Nie, S.; Ma, H. et al. ADD 2022: the first Audio Deep Synthesis Detection Challenge. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, p. 9216–9220.

[87] Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C. Y. et al. *Audio Deepfake Detection: A Survey.* 2023. Available at: https://arxiv.org/abs/2308.14970.

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

[89] Zhang, Q.; Wen, S. and Hu, T. Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier. In: *Proceedings of the 32nd ACM International Conference on Multimedia.* New York, NY, USA: Association for Computing Machinery, 2024, p. 6765–6773. MM '24. ISBN 9798400706868. Available at: https://doi.org/10.1145/3664647.3681345.

[90] Zhang, Y.; Li, Z.; Lu, J.; Hua, H.; Wang, W. et al. The Impact of Silence on Speech Anti-Spoofing. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* IEEE Press, august 2023, vol. 31, p. 3374–3389. ISSN 2329-9290. Available at: https://doi.org/10.1109/TASLP.2023.3306711.

[91] Zhang, Z.; Gu, Y.; Yi, X. and Zhao, X. *SynSpeechDDB: A new synthetic speech detection database.* IEEE Dataport, 2020. Available at: https://dx.doi.org/10.21227/ta8z-mx73.

[92] Zhao, G.; Sonsaat, S.; Silpachai, A.; Lucic, I.; Chukharev-Hudilainen, E. et al. L2-ARCTIC: A Non-native English Speech Corpus. In: *Proc. Interspeech.* 2018, p. 2783–2787. Available at: http://dx.doi.org/10.21437/Interspeech.2018-1110.

Appendix A

Contents of the submitted directory

```
/
├── text/ ........................ Directory with LaTeX source for this report
├── dataset/ .................................................. Dataset content
│   ├── waveforms/ .......................................... Audio utterances
│   │   └── [cd|de|en|es|fr]/ ..................... Subdirectories per language
│   │       └── [male|female]/ ........................ Subdirectories per gender
│   ├── metadata/ ................................... CSV files with metadata
│   └── utils/ .................... Utility scripts for processing the dataset
├── README.md ...................... Readme file with additional information
└── xsrnak00.pdf ......................................... This thesis in pdf
```