BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

MULTILINGUAL VOICE DEEPFAKE DATASET

VÍCEJAZYČNÁ DATOVÁ SADA HLASOVÝCH DEEPFAKES

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **EVA TRNOVSKÁ**
AUTOR PRÁCE

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

BRNO 2024

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

154478

Institut:	Department of Intelligent Systems (DITS)
Student:	**Trnovská Eva**
Programme:	Information Technology
Title:	**Multilingual Voice Deepfake Dataset**
Category:	Security
Academic year:	2023/24

Assignment:

1. Study the area of voice deepfakes, their creation, and existing datasets, including research areas that use datasets.
2. Based on the analysis of existing datasets and the needs of current research in the area of security implications of deepfakes, define the size and parameters for a new dataset to enable new areas of research, e.g., detector behavior on different languages.
3. Create the proposed dataset (of a sufficient number of hours and other parameters needed to be used for training models, etc.).
4. Compare the features of the new dataset with the existing ones.
5. On a selected research use case, run an experiment using the new dataset.

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.

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

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

Supervisor:	**Malinka Kamil, Mgr., Ph.D.**
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2023
Submission deadline:	31.7.2024
Approval date:	6.11.2023

Abstract

This thesis examines the area of voice deepfakes: their creation and detection. It describes the state of current research and the methods of creating fake recordings. Furthermore, it provides a comprehensive analysis of available voice deepfake datasets, based on which a new multilingual dataset is designed and compiled. The dataset aims to enable further research on the generalization of deepfake detection across languages and the differences in the accuracy of male and female voice detection. The results of the experiments show that for the models tested, it is possible to replace detectors trained to detect in a single language with detectors trained on a multilingual set, with an accuracy loss of a few percent. The tested models were generally more accurate in detecting recordings with female voices, but this property was not demonstrated for all tested detectors.

Abstrakt

Tato práce se zabývá oblastí hlasových deepfakes: jejich vytvářením a detekcí. Popisuje aktuální stav výzkumu v této oblasti a metody pro vytváření falešných nahrávek. Dále poskytuje širší analýzu dostupných datových sad obsahující hlasové deepfakes, na jejímž základě je navržena a vytvořena nová vícejazyčná datová sada. Tato sada má za cíl umožnit další výzkum v oblasti zobecňování detekce deepfakes napříč jazyky a rozdílech v přesnosti detekce mužského a ženského hlasu. Výsledky experimetů ukazují, že u testovaných modelů je možné nahrazení detektorů trénovaných pro detekci v jediném jazyce detektory, jež jsou natrénované na vícejazyčné sadě, a to se ztrátou přesnosti v jednotkách procent. Testované modely byly obecně přesnější při detekci nahrávek s ženskými hlasy, ovšem tato vlastnost se neprokázala u všech testovaných detektorů.

Keywords

voice deepfakes, deepfake detection, text-to-speech, voice conversion, multilingual dataset, dataset analysis

Klíčová slova

hlasové deepfakes, detekce deepfakes, převod textu na řeč, konverze hlasu, vícejazyčná datová sada, analýza datasetů

Reference

TRNOVSKÁ, Eva. *Multilingual voice deepfake dataset.* Brno, 2024. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.

Rozšířený abstrakt

Pojem *deepfakes* označuje počítačem generovaná média, která vyobrazují události nebo osoby, které nikdy neexistovaly. Tato práce se zaměřuje na hlasové deepfakes: výroky, které nikdy nebyly proneseny těmi, kteří vypadají, že je vyslovují. Ačkoli mají uměle vygenerované nahrávky i pozitivní využití např. v zábavním průmyslu, stále přibývá podvodů, které je využívají k vylákání peněz z nic netušících obětí či k ovlivnění veřejného mínění.

Některé deepfakes mohou lidé snadno odhalit, jiné je snazší detekovat pomocí deepfake detektorů založených na neuronových sítích. Syntéza lidské řeči se stále zlepšuje a je tedy potřeba detektory trénovat na aktuálních datech. Datasetů, které zfalšované nahrávky obsahují, vzniká každý rok několik, ty ovšem trpí několika nedostatky, jako je používání zastaralých technologií, nedostatečná velikost, zastoupení pohlaví či různých jazyků. Trénování detektorů nad těmito datasety tedy vyvolává pochyby o jejich přesnosti a nezaujatosti, tím spíše, pokud obsah datasetu není detailně anotován.

Tato práce se snaží poskytnout přehled o současném výzkumu hlasových deepfakes a identifikovat potenciální mezery. Stručně popisuje dostupné nástroje pro syntézu řeči a korpusy, na kterých jsou trénovány. Poté analyzuje existující datasety a shrnuje je v určitých klíčových aspektech, jako jsou velikost, použité nástroje, jazyky a další statistiky.

Analýza přinesla poznatky o silných a slabých stránkách těchto datasetů, které spolu se současnými mezerami ve výzkumu vytvářejí základ pro kladení nových výzkumných otázek. Je snazší identifikovat deepfake mužského, nebo ženského hlasu? Jak detektory reagují, setkají-li se s novým typem deepfake nebo s deepfakes v jiném jazyce? K jejich zodpovězení je navržen a sestaven nový dataset.

Zahrnuje pravé i vygenerované nahrávky v pěti jazycích: angličtině, němčině, francouzštině, španělštině a italštině. Jedná se o jednu z největších detekčních sad, které byly vytvořeny; jediným veřejně dostupným datasetem s větším počtem nahrávek je ASVspoof4. Z hlediska zahrnutých jazyků se jedná o druhý nejrozmanitější dataset, přičemž je také jako jediný dokonale genderově vyvážený, a navíc byly tři ze čtyř použitých syntetizačních nástrojů zveřejněny v uplynulém roce. Nad tímto datasetem byly natrénovány ve vícero běžích dva typy detektorů, LCNN a RawNet3, se kterými byly provedeny dva experimenty.

V prvním experimentu byla stanovena hypotéza, že detektory trénované nad vícejazyčnými datovými sadami při detekci jsou stejně přesné, jako detektory trénované nad sadami obsahující nahrávky v jediném jazyce. Tato hypotéza byla vyvrácena: vícejazyčné detektory jsou méně přesné, pokles úspěšných detekcí je ale pouze v řádu jednotek procent.

Ve druhém experimentu bylo ověřováno, zda je detekovat falešné nahrávky hlasu jednoho z pohlaví jednodušší, za předpokladu že detektor je natrénován na vyvážené sadě. Po vyhodnocení všech běhů dohromady se zdá, že detektory jsou přesnější při odhalování deepfakes imitující ženské hlasy. To však ale nelze tvrdit o všech testovaných kombinacích, např. výsledky modelu RawNet3 testovaného s anglickou sadou nevykázaly významný rozdíl v přesnosti.

Na závěr byl proveden test schopnosti odhalit doposud neviděné typy deepfakes, sesbírané ze sociálních sítí v rámci datasetu In-the-wild. Tento pokus ukázal, že navržený dataset v kombinaci s testovanými modely není schopen reálné deepfakes spolehlivě určit – téměř všechny falešné nahrávky byly vyhodnoceny jako pravé. Z této práce tedy vyplývá, že schopnost detektorů zobecňovat, a celkově vliv pohlaví a jazyku na detekci by neměly být podceňovány. Další oblastí, kterou by se výzkum mohl zabývat, je zda zahrnutí pravých i falešných nahrávek znázorňujících téhož člověka má na detekci pozitivní (či negativní) vliv.

Multilingual voice deepfake dataset

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Mgr. Kamil Malinka, Ph.D. The supplementary information was provided by Ing. Anton Firc. I have listed all the literary sources, publications, and other sources, that were used during the preparation of this thesis.

.....
Eva Trnovská
July 31, 2024

Acknowledgements

I would like to thank my supervisor Mgr. Kamil Malinka, Ph.D. for his help, guidance, and providing computational resources, and Ing. Anton Firc for providing additional information. Moreover, I would like to thank my friends and family who supported me throughout the process of writing this thesis and the whole duration of my studies. Additional 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

List of Figures

Chapter 1

Introduction

The term *deepfakes* encompasses computer-generated media that feature events or people that never existed. This thesis focuses on speech deepfakes: statements never uttered by those who appear to be pronouncing them.

Synthetic speech improves user experience through text-to-speech services, funny filters, or a voice conservation service. The danger lurks in the credibility of the available speech synthesis and voice conversion tools. Do their products sound human-like? How well do they imitate the target speaker? Fooling voice biometric systems deployed in banks or built in smartphones may lead to money loss or identity theft.

Some deepfakes can be easily spotted by humans, some are easier to reveal using deepfake detectors based on neural networks. As technology evolves, artificial detectors must be trained on up-to-date and suitable datasets. The question arises: what is a suitable dataset? Can the detectors generalize? How do they react if met with a novel type of deepfake, or a deepfake in another language?

Researchers produce diverse deepfake datasets, though often with incomplete metadata. Training deepfake detectors on a limited number of datasets raises concerns about their fairness and accuracy, more so if the dataset's contents lack proper description. Only recently did the topic of multilingual deepfake detection and transferring deepfake detection capabilities across languages come up: experiments with an English-Chinese dataset have shown a steep increase in error rate when changing the language of the testing set [3].

The contents of this thesis strive to provide an overview of current audio-deepfake research and to identify potential gaps, such as gender bias or language influence in audio deepfake detection. It briefly describes available speech synthesis tools and corpora on which they are trained to further explore the landscape of audio deepfakes. After that, existing audio deepfake datasets are analyzed and summarized in certain key aspects such as size, used tools, languages, and other statistics.

The analysis gives insights into the strengths and weaknesses of these datasets, which together with current research gaps lay the ground for posing new research questions. To answer them, a suitable dataset is designed and compiled: a process that is documented as well. Finally, the questions are turned into experiments. To understand the concepts presented in this thesis, the reader is expected to have a basic understanding of how machine learning and neural networks work.

Chapter 2

Current deepfake research

Deepfakes have no precise definition. For some researchers, only video or face images are considered deepfakes. Sometimes, photoshopped images, edited videos, or human-imitated voice recordings are referred to as deepfakes as well. However, these low-quality imitations should be classified rather as *cheapfakes*, a term coined by Paris and Donovan [40]; the term *deepfake* is mostly reserved for media created using deep learning tools. According to a recent survey [1], Tencent[1] proposed using the term *deep synthesis* instead of calling them *deepfakes* to embrace both positive and negative aspects of their usage.

As mentioned earlier, synthetically generated media have many positive uses: from cost-effective dubbing and illustrations in the entertainment industry to accessibility features for the visually impaired. However, technology can be misused for bank fraud, identity theft, or changing public opinion, some examples of which can be found in Section 2.5.

This chapter gives a quick overview of existing deepfake types, continued in Chaper 3, and introduces some of the methods currently used in audio deepfake detection. To understand how their efficiency is measured, several metrics are defined for evaluating the detectors. The last two sections further discuss novel topics in the audio deepfake area: possible gender bias and the influence of languages on the detection process.

2.1 Deepfake types

Deepfake videos and images can be created using a wide range of techniques, from face swap and face morphing to lip syncing, in the case of videos. This thesis focuses on speech deepfakes and therefore face and other manipulations that do not change the voice or speech content will not be discussed. Audio deepfakes can include various types of sounds, from street sounds to speech to songs. A recent survey on audiovisual content manipulation defines 5 groups of audio modifications: voice conversion, text-to-speech synthesis, voice cloning, voice morphing, and replay attacks [31].

Voice conversion retains the speech contents and transforms the tone and other speaker characteristics to match a different speaker. Text-to-speech produces synthetic speech from a given text excerpt. Voice cloning advances TTS technology by generating speech with a particular speaker's characteristics. Voice morphing enables a smooth transformation of one voice to another, resulting in the creation of a new voice with mixed characteristics. Finally, replay attacks profit from prerecorded utterances that can be either replayed or cut and pasted together to form different statements.

[1] https://tech.sina.com.cn/roll/2020-07-14/doc-iivhvpwx5201226.shtml

In this thesis, voice cloning is considered as a part of text-to-speech. Text-to-speech and voice conversion are discussed in Chapter 3, while voice morphing and replay attacks will not be further addressed. Voice morphing has not been widely used to the present moment, and the tools allowing for the creation of morphed audio content are scarce. On the other hand, replay attacks are a real threat that produces altered – but not deepfake – audio. Before moving on to the topic of audio deepfake detection, the usage of some key terms must be clarified. The terms *spoofed*, *fake*, and *deepfake* will be used interchangeably to denote computer-generated audio, while *real*, *genuine* and *bonafide* will be reserved for unmodified utterances.

2.2 Audio deepfake detection

Deepfakes can be detected by both machines and humans. While testing the deepfakes on real human subjects can reveal the realistic effects the deepfakes can have on the public, automatized testing is a must in the case of examining larger amounts of recordings.

The development of deepfake detection tools is encouraged by two challenges: ASVspoof challenge[2], held biannually since 2015, and Audio Deep synthesis Detection challenges[3] (ADD) challenge, held in 2022 and 2023. The participants propose a classifier and submit the scores of the testing dataset. A survey on deepfake detection lists several commonly used deepfake detection approaches: machine learning-based methods such as decision trees or GANs, deep CNNs, RNNs, and methods based on statistical measurements (e.g. GMMs) or blockchains [43].

These architectures can be briefly described with the help of a speech processing handbook [7]. Generative adversarial networks (GANs) consist of two parts, a generator, that creates new samples, and a discriminator that evaluates these samples by trying to classify them either as genuine or as generated. These models are often used for text and image generation. Convolutional neural networks (CNNs) are widely used neural networks for image classification. They learn to recognize features of an image by progressively reducing it to an abstract representation using the convolutional layers. Recurrent neural networks (RNNs) are neural networks that use the output of the previous step as an input for the current time step, meaning that the model develops a certain type of long-term dependencies. The structure of RNNs and deep CNNs is illustrated in Figure 2.1.

Gaussian mixture models (GMMs) combine 2 approaches: modeling signals as Gaussian processes and using a mixture model, which supposes that the examined signal has multiple classes and each has its statistical model. The final distribution is the weighted sum of the class distribution, where weights are the frequencies of their occurrence in the signal. The models are called Gaussian because most of the classes typically represent normal – Gaussian – distributions.

Human evaluation is usually measured using MOS[5] when releasing a new TTS or VC tool. Deepfake datasets are mostly designed for automatized detection, and therefore the credibility of their contents for human ears is unknown. MOSNet[6] is a neural network developed to predict MOS on Voice Conversion Challenges' submissions to predict human ratings.

[2] https://www.asvspoof.org/
[3] http://addchallenge.cn/
[4] https://www.asimovinstitute.org/author/fjodorvanveen/
[5] MOS – mean opinion score
[6] https://github.com/lochenchou/MOSNet

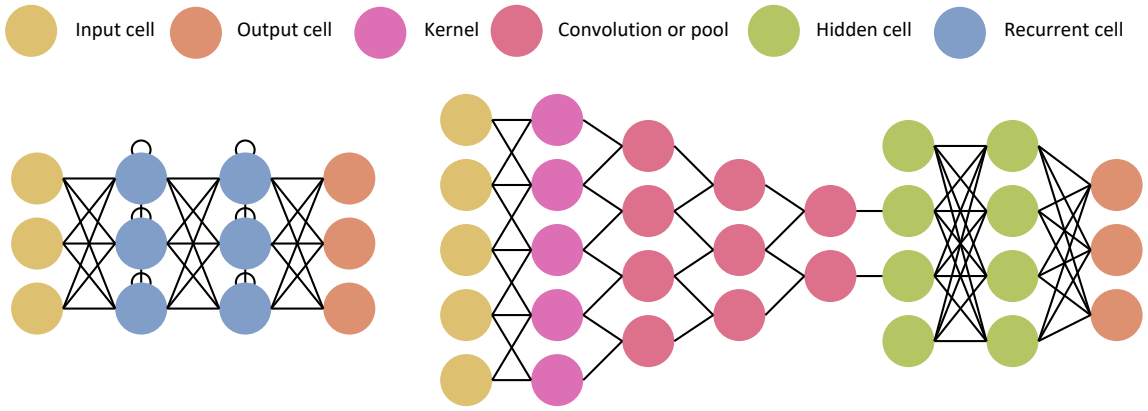

Figure 2.1: RNN (left) and deep CNN (right) structure, adapted from the Asimov Institute[4].

ASVspoof 2021 baselines

ASVspoof5[7] has not yet published its baselines, so the attention was shifted to the 4 CM[8] baselines of ASVspoof4: CQCC-GMM, LFCC-GMM, LFCC-LCNN, and RawNet2 [30]. CQCC-GMM and LFCC-GMM are Gaussian mixture models (GMM) using different features: CQCC refers to features extracted with the constant Q transform (CQT)[9], while LFCC is an abbreviation of linear frequency cepstral coefficients; LCNN stands for light convolutional neural network.

This LCNN implementation also uses LFCC as input features, while the RawNet extracts the embeddings directly from the waveform, relying on the hidden layers to learn to discriminate between the speakers. The original RawNet is a speaker embedding extractor for speaker verification based on a CNN-GRU, a convolutional neural network-gated recurrent unit architecture, an architecture similar to an RNN [21]. It was optimized twice, introducing the RawNet2 and RawNet3 models[10].

These models have been used as baselines for newer detection tools or correlated research, as seen in the work by Müller et al. [35], Li et al.[11], Wen et al.[12] or Kawa et al. [23]. The framework developed by the last mentioned was tested with a different dataset as a part of unpublished research and was found to be very accurate even in the case of modified recordings.

2.3 Evaluation metrics

Altuncu et al. list several deepfake-related performance metrics: confusion matrix, precision, recall, true and false positive rates, equal error rate, accuracy, F-score, ROC, AUC, Log loss, and PQA. In the context of audio deepfake detection, the most commonly used are accuracy and EER. The description of the metrics is adapted from the survey [1].

[7] https://www.asvspoof.org/

[8] counter-measures: in this context, audio deepfake detection

[9] https://doc.ml.tu-berlin.de/bbci/material/publications/Bla_constQ.pdf

[10] https://github.com/Jungjee/RawNet

[11] Li, Jing, et al. Advanced RawNet2 with Attention-based Channel Masking for Synthetic Speech Detection. In: *Proc. INTERSPEECH*. 2023. p. 2788-2792.

[12] Wen, Yan, et al. Multi-Path GMM-MobileNet Based on Attack Algorithms and Codecs for Synthetic Speech and Deepfake Detection. In: *INTERSPEECH*. 2022. p. 4795-4799.

Confusion matrix

The confusion matrix relates the actual class of the samples with the predicted one:

- TP, true positives, are the correctly predicted positive samples (genuine audio predicted to be genuine),
- TN, true negatives, are the correctly predicted negative samples (deepfake audio predicted to be deepfake),
- FP, false positives, are the samples incorrectly predicted to be positive (deepfake audio predicted to be genuine),
- FN, false negatives, are the samples incorrectly predicted to be negative (genuine audio predicted to be deepfake).

Precision and recall

Using the terms defined in the above section, other measures can be inferred: precision and recall. Precision specifies the fraction of *truly* positive samples among that were *predicted* to be positive. Inversely, the ratio of samples predicted positive to those that are truly positive is called recall, sensitivity, or true positive rate. Both metrics can be defined formally:

$$precision = \frac{TP}{TP + FP}$$

$$recall = TPR = \frac{TP}{TP + FN}$$

Accuracy

Accuracy can be described as the ratio of correctly classified samples to all classified samples. Using the terms from above, accuracy can be formally denoted as:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

However, accuracy is not the best metric when using a dataset with an unbalanced evaluation set. In the case of an unbalanced set, EER is preferred.

EER

Equal error rate is a metric calculated for a threshold t such that the false positive rate (FPR) and the false negative rate (FNR) are equal. FPR and FNR are defined as:

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

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

In other words, instead of setting the threshold to a specific value (usually 0.5 for a scale between 0 and 1), we select the threshold to reach equal ratios of incorrectly classified genuine and incorrectly classified spoofed samples.

This metric provides a better understanding of the model performance when testing it with unbalanced sets. However, EER is not the optimal metric in situations where increasing the number of correct predictions of one class even at the cost of increasing the error for the other class is beneficial, as it gives the same importance to FPR and FNR. Airport X-ray controls, tumor detection, and speaker verification in bank call centers can serve as an example where samples predicted as positives can be manually checked again for false positives, but to review them and find true positives, they must first be flagged by automated detection tools.

2.4 Gender bias in deepfake detection

Although the topic of fairness resonates in the area of face deepfakes, gender bias in audio deepfake detection remains an underestimated factor. If deepfake detectors or speaker verification systems are biased, the attackers could use that fact to increase their chances and use speakers of the gender that is less detectable for their deepfake recordings. Conversely, if there is a higher false positive rate for a specific gender, it could lead to discrimination in tools that automatically remove suspicious content.

Trinh and Liu state that many datasets are biased, leading to biased detectors. They performed experiments with 3 video deepfake detection tools and concluded that while the difference in results for speakers of different genders was small, the detectors were racially biased [53]. Nadimpalli and Rattani annotated 2 commonly used video deepfake datasets, FaceForensics++[13] and Celeb-DF[14] with gender labels. Their detectors performed worse on women, as the representation of men was much higher. Consequently, they propose a gender-balanced deepfake dataset, GBDF [38].

Bilika et al. conducted a set of experiments attacking voice assistants on Android and iOS mobile devices. They found that iOS devices were more than 3 times more susceptible to accept deepfake recordings using a male voice rather than a female voice (35.24% and 10.98% successful attacks, respectively) [4]. In contrast, another study reached 43.3 $\pm$ 16.1% and 61.8 $\pm$ 9.4% success rate for males and females respectively when attacking Resemblyzer[15], and 7.8 $\pm$ 13.9% and 47.3 $\pm$ 32.0% when attacking Microsoft Azure speaker recognition[16]. When it comes to human detection, the authors concluded that women and young people were more successful in identifying deepfakes [60]. Nevertheless, the speaker's gender is not the only factor that influences the detection ability.

2.5 Deepfakes and languages

The internet is full of deepfakes and articles recounting their malicious use, since deepfakes are already used in political campaigns and financial fraud. Recent examples include US Democratic voters receiving a wave of discouraging robocalls imitating the voice of the current US president, Joe Biden[17], or using the notoriety of the former UK PM, Rishi

[13] Rossler, Andreas, et al. Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[14] Li, Yuezun, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[15] https://github.com/resemble-ai/Resemblyzer

[16] https://azure.microsoft.com/en-us/products/ai-services/ai-speech/

[17] https://edition.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html

Sunak, for dozens of fake video advertisements on Facebook[18]. Getting a public figure to promote your business has never been easier: hop on Soundboard[19] and let Donald Trump voice your commercial. And while English is the lingua franca of the internet, deepfakes do not stop there.

A deepfake video of Serbian prime minister Miloš Vučević talking about non-existing government projects[20] was posted on Facebook, and audio messages with a forged voice of the leader of Slovak opposition, Michal Šimečka, were supposed to stir the election terrain with mentions of rising beer prices and rigging the election[21]. Current Czech president Petr Pavel and former Czech prime minister Andrej Babiš are among the public figures used in too-good-to-be-true investment advertisements[22], while Italian TV presenter Enrico Mentana and a businessman Giovanni Ferrero's imitations describing a fraudulent investment program appeared in a fake news release[23].

In April 2024, a South Korean woman revealed her love story[24] with a fake Elon Musk whom she befriended on social media. The impersonator sent her modified images and proclaimed his love to her over a video call, then persuaded her to transfer over $50,000. The fraudsters do not, however, only target the large public on social media, trying to influence elections or gain money from credulous citizens. Some specialize in personalized attacks on employees of large corporations, creating custom deepfakes of company executives ordering their subordinates to transfer large sums of money. In a recent case, a Hong Kong multinational firm lost $25 million this way[25].

Given the number of datasets containing deepfake speech in English, deepfake detectors for English speech can be trained to prevent some of these attacks. However, how reliable are the results for audio tracks in different languages? So far, the research in the area of deepfake detection has scarcely touched upon the topic of languages. Deepfake datasets are only available in a few languages (see Chapter 4) and therefore the influence of languages on deepfake detection remains mostly unknown, as acknowledged by a recent survey [14].

Ba et al. have examined the influence of introducing deepfakes in a different language to selected deepfakes detectors and have noted a substantial decrease in accuracy. Consequently, they proposed a solution for transferring deepfake detection capabilities across languages. However, their experiments are limited in several ways. They only used two languages that differ significantly: English and Chinese. Although they used the same methods for generating deepfakes in both languages, they used different types of reference speech sources, and the language sets were slightly imbalanced. The authors addressed the issue of lack of languages but decided to refrain from creating additional language sets for the time being [3].

[18] https://www.theguardian.com/technology/2024/jan/12/deepfake-video-adverts-sunak-facebook-alarm-ai-risk-election

[19] https://www.101soundboards.com/boards/696510-donald-trump-hq-tts-computer-ai-voice

[20] https://balkaninsight.com/2024/07/04/serbia-reacts-fast-over-ai-deepfake-video-of-pm-unlike-other-cases/

[21] https://edition.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html

[22] https://ct24.ceskatelevize.cz/clanek/domaci/pavel-varuje-pred-deepfake-nikdo-z-ustavnich-cinitelu-by-nedelal-reklamu-pochybne-komercni-350902

[23] https://www.youtube.com/watch?v=P1pULfU2g7M

[24] https://fortune.com/2024/04/29/ai-deepfakes-drake-elon-musk-baltimore-principal/

[25] https://www.scmp.com/news/hong-kong/law-and-crime/article/3250851/everyone-looked-real-multinational-firms-hong-kong-office-loses-hk200-million-after-scammers-stage

Chapter 3

Speech synthesis

Having introduced the topic of audio deepfakes and the underlying threats connected to this technology, this chapter dives deeper into how speech is generated by computers. The first two sections provide a brief overview of speech synthesis techniques, namely text-to-speech (TTS) and voice conversion (VC). In addition, it describes speech corpora used for training TTS and VC tools and creating deepfake datasets.

3.1 Text-to-speech

Text-to-speech represents a group of techniques that convert text to speech. Unless indicated otherwise, the information in this section was adapted from a survey on neural speech synthesis [51]. The first computer-based systems saw the light of day in the second half of the 20th century. Throughout time, different approaches were tried, starting with articulatory synthesis, a technique imitating the way humans speak: their vocal cords, lip movements, etc. Although in theory it could be the most effective type of synthesis, the results were far from optimal.

The next evolution stage is formant synthesis, specifying complex linguistic rules and producing intelligible, yet unnatural speech. Using a large database of speech units, concatenative synthesis generates speech by chaining the required speech units recorded by voice actors. More recently, statistical parametric speech synthesis (SPSS) and neural network-based speech synthesis have gained popularity. SPSS consists of 3 parts: a text analysis module, a parameter prediction module (acoustic model, usually based on Markov models), and a vocoder. In this thesis, only neural TTS will be further discussed, as it generates the most realistic speech.

For a few years, neural TTS has been evolving from 3-part systems composed of a text analysis module, an acoustic model and a vocoder, shown in Figure 3.1, to end-to-end models. The first step in the process of text-to-speech synthesis is to convert text into phonemes or linguistic features by the text analysis module. This module generally performs text normalization, word segmentation, part-of-speech tagging (annotating words based on their definition and context), prosody prediction, and grapheme-to-phoneme conversion.

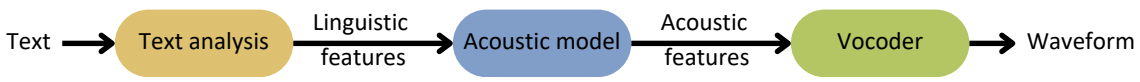

Figure 3.1: TTS pipeline, adapted from Tan et al. [51].

Afterward, acoustic features are generated by an acoustic model. Acoustic models in neural TTS are based on recurrent (RNNs) and convolutional neural networks (CNNs), both briefly described in Chapter 2.2, or on Transformer architecture. Transformer[1] is a model proposed by Google in 2017, using a mechanism called attention.

Finally, a waveform is generated by a vocoder. Most of the acoustic models in the survey convert phonemes or characters into Mel spectrograms, a visual representation of the signal in the Mel scale. The Mel scale maps frequencies in a way that human listeners consider them to be in equal distance from each other [7].

The vocoders in neural TTS can be divided into 6 groups (autoregressive, flow-based, GAN-based, VAE-based, diffusion-based and other models); however, the most popular ones, HiFiGAN[2], Parallel WaveGAN[3], and MelGAN[4] all fall into the GAN-based category. A few examples of different open-source architectures include Tacotron 2[5], a RNN converting characters into Mel spectrograms, SpeedySpeech[6], a CNN converting phonemes into Mel spectrograms, and more recently VITS [25], and ZMM-TTS [19].

VITS

VITS is a parallel end-to-end TTS proposed in 2019. The text is encoded into the International Phonetic Alphabet (IPA), the architecture itself is composed of two encoders, a decoder, and a module performing monotonic alignment search is used to identify the most likely alignment between speech and text [25]. VITS is a very popular TTS model, also implemented as a part of the Coqui.ai[7] TTS toolkit, where 38 of the 72 retrained models available are using the VITS architecture. Several other speech synthesis tools are based on it, including YourTTS[8] or FreeVC [29].

ZMM-TTS

ZMM-TTS, announced in late 2023, is a zero-shot multilingual multispeaker TTS framework. The paper proposed a 2-step architecture: a `txt2vec` module converting text to a discrete representation and a `vec2wav` module generating a waveform from this representation. The `txt2vec` module supports several ways of encoding the input text: either using the raw text as characters, using IPA phonemes, or delegating the phoneme representation on a pretrained large-scale multilingual language model. The speaker identity is captured by a pretrained speaker encoder model.

The `wav2vec` part either directly maps discrete representations to waveforms or divides the task into 2 submodules, `vec2mel` and vocoder modules. The `vec2mel` module generates a Mel spectrogram from the discrete representation which is later used as input for the HiFiGAN vocoder. The authors trained the model on a dataset composed of 6 languages (English, French, German, Portuguese, Spanish, and Swedish), with utterances assembled from different single and multispeaker datasets [19].

[1] Vaswani, Ashish, et al. Attention is all you need. *Advances in neural information processing systems*, 2017, 30.

[2] https://github.com/jik876/hifi-gan

[3] https://github.com/kan-bayashi/ParallelWaveGAN

[4] https://github.com/descriptinc/melgan-neurips

[5] https://github.com/NVIDIA/tacotron2

[6] https://github.com/janvainer/speedyspeech

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

[8] https://github.com/Edresson/YourTTS

3.2 Voice conversion

Voice conversion (VC) is the task of transforming an utterance pronounced by the source speaker in a way that the linguistic contents are persevered but the speech adopts the pitch and other characteristics typical for the target speaker. Unless marked otherwise, the information was adapted from an overview of voice conversion [47]. Voice conversion tools can be parallel or nonparallel, depending on the way they were trained: parallel voice conversion tools require databases with the same utterances spoken by both the source and the target speaker.

VC systems are typically made up of a speech analysis, a mapping, and a reconstruction module, analogously to the TTS pipeline in Figure 3.1. The first part of the pipeline divides the source utterance into features that represent supra-segmental (prosodic characteristics) and segmental information (spectrum). In speech analysis, the signal can be viewed as a set of time segments or segments divided into classes such as noise, harmonic, etc., or it can be described mathematically as a model with parameters changing over time. The module computes spectral (related to voice timbre) and prosodic features, influencing the fundamental frequency, intonation, and duration.

The mapping module then transforms the parameters to match those of the target speaker. This can be done with methods ranging from vector quantization (clustering) [7], dynamic time wrapping[9] to deep learning; using long-stort-term memory models (LSTMs, recurrent networks that discard irrelevant information [7]) or encoder-decoder architectures with attention. The reconstruction module – a vocoder – performs the resynthesis of the speech in the target's speaker voice. There are many state-of-the-art systems, but only the following 3 open-source tools will be briefly described: FreeVC [29], LVC-VC [22], and DDDM-VC [10].

FreeVC

FreeVC is the only VC system implemented in the Coqui.ai toolkit. It is a one-shot voice conversion system adopting features from the end-to-end TTS system VITS: it uses 3 encoders, of which one is a speaker encoder, a decoder, and a discriminator deciding if an utterance is genuine or generated. The *one-shot* keyword indicates that this model is capable of transforming a source utterance into the target's speaker voice given one reference recording [29].

LVC-VC

LVC-VC is a state-of-the-art zero-shot end-to-end voice conversion system, meaning that it can convert utterances to voices of speakers not seen during training. This architecture consists of a generator (a CNN), a speaker encoder, and a group of discriminators. It uses location-variable convolutions: using different convolution kernels for different input sequence intervals [22].

DDDM-VC

DDDM-VC is a voice conversion system based on decoupled denoising diffusion models proposed in 2023. The idea behind denoising diffusion models is to progressively corrupt the training input, transforming it into a sample from a Gaussian distribution. The deep

[9] https://rtavenar.github.io/blog/dtw.html

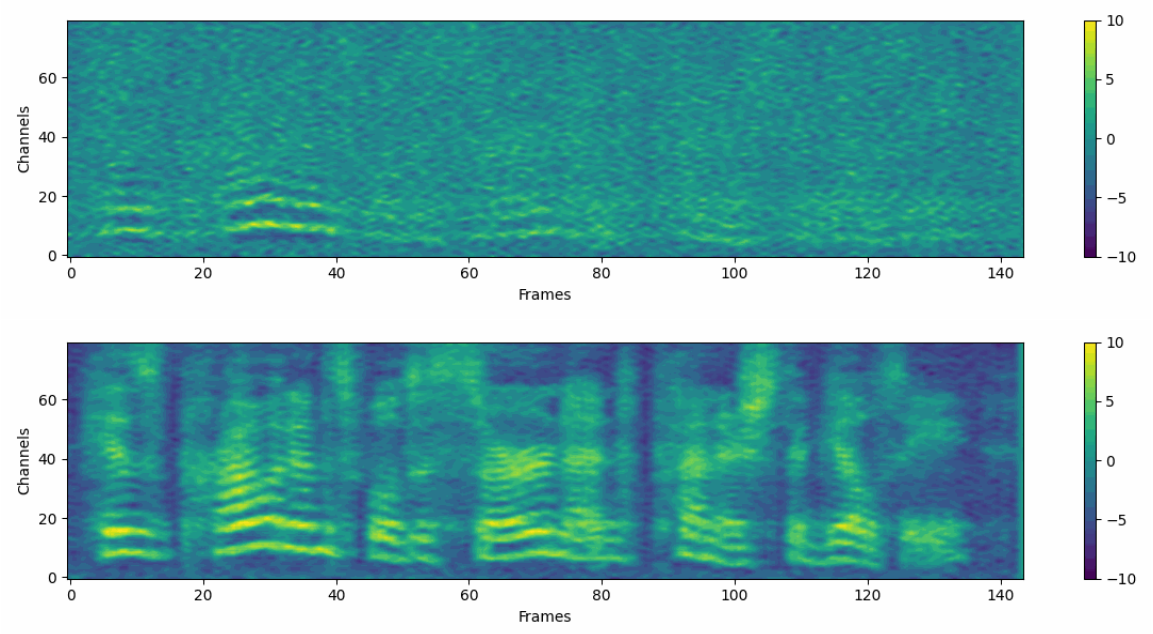

Figure 3.2: From a sample from a Gaussian distribution to a nearly reconstructed spectrogram. Denoising spectrograms with DDDM-VC, image adapted from the demo page[10].

neural network learns to reverse this process and thus generate new outputs from this distribution [5]. The denoising process is illustrated in Figure 3.2.

It decouples speech into 3 parts: content, pitch, and speaker. The content is represented as phonemes and to capture the pitch, the fundamental frequency is computed and then clustered using the F0 quantizer model from the Speech Resynthesis project[11]. Finally, capturing the speaker equals capturing his or her speaking style [10].

3.3 Speech corpora

To create a deepfake new dataset, a suitable speech source must be found. To avoid lengthy preprocessing, it is beneficial to use an existing speech corpus. Moreover, analyzing the corpora commonly used for training TTS and VC models helps to understand the key parameters for creating new ones. Some of the datasets listed have already been used to create deepfake datasets described in Chapter 4.

Monolingual speech datasets

The following list contains commonly used English, Chinese, Japanese, and German speech corpora for training TTS and VC tools, or for speech recognition. The key aspects of the datasets are summarized in Table 3.1, with the exception of the CMU Arctic, which was discarded due to incomplete information.

- **Aidatatang_200zh**: a Mandarin Chinese corpus with transcribed speech recorded from 600 speakers by Datatang[12] with a total duration of 200 hours.

[10] https://hayeong0.github.io/DDDM-VC-demo/
[11] https://github.com/facebookresearch/speech-resynthesis
[12] https://www.datatang.ai

Table 3.1: Key features of monolingual corpora.

Corpus	Language	Hours	Speakers	Transcribed
Aidatatang_200zh	Chinese	200	600	✓
AISHELL 1-4	Chinese	1370	2600+	✓
VCTK	English	44	109	✓
DAPS	English	4,67	20	✓
JSUT	Japanese	10	1	✓
LibriTTS	English	585	2,000+	✓
LJ Speech	English	24	1	✓
THCHS-30	Chinese	50	35	✓
Thorsten-Voice	German	11	1	✓

- **AISHELL 1-4**: a series of Mandarin Chinese speech recognition corpora. AISHELL-1 contains 165 hours recorded by 400 participants, mostly young people from Northern China, on a high-fidelity microphone or a mobile phone. The sentences mostly cover the topics of finance, science, and technology [6].
  AISHELL-2 provides additional 1,000 hours of clean speech by 1,991 speakers under 40 years, including children. Most speakers were recorded in a studio, the rest in a living room. The topics are among other entertainment, sports, and IoT device control commands [12]. AISHELL-3 contains 85 hours of speech read by 218 native speakers in a neutral tone [46], and AISHELL-4 adds 120 hours of multispeaker recordings [17].
- **CMU Arctic**: a collection of single speaker databases of short prompts read by a native English speaker, recorded in a studio environment [27].
- **CSTR's VCTK Corpus**: a corpus containing mostly newspaper articles read by 109 native English speakers with different accents. The speakers read different articles, total length is 44 hours [55].
- **DAPS**: a collection containing 14 minutes per speaker of aligned speech – read public domain texts – by 20 speakers with equal gender representation [34].
- **JSUT**: a single-speaker Japanese corpus covering most of the daily-use Japanese characters. Divided into multiple sections, the total amount of recorded speech is 10 hours [49].
- **LibriTTS**: a corpus derived from LibriSpeech[13], with 585 hours of speech by over 2,000 speakers [69].
- **LJ Speech**: an English single-speaker dataset commonly used for benchmarking TTS models. The speech is extracted from LibriVox[14] audiobooks read by a female speaker. The total length is almost 24 hours, and the clips are up to 10 seconds long, transcribed, and aligned [20].
- **THCHS-30**: a standard Mandarin speech database for speech recognition. It contains 35 hours recorded by 50 speakers, mostly college students, in 2000-2001. The set of 1000 sentences used in the recordings was selected from news articles [56].

[13] Panayotov, Vassil, et al. Librispeech: an asr corpus based on public domain audiobooks. *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015.
[14] https://librivox.org/

- **Thorsten-Voice**: a German single-speaker dataset with over 12,000 sentences making up 11 hours of speech. Also available in an emotional variant or a dialect [37].

Multilingual speech datasets

Although not as commonly used, multilingual corpora are an important source of training material for speech synthesis and recognition in multiple languages. This list contains 10 datasets, ranging from small to large collections, with 9 datasets summarized in Table 3.2.

- **Babel speech corpora**: provides 10-hour and 80-hour long packs of transcribed audio for 10 low-resource languages, such as Haitian Creole, Vietnamese, or Zulu. The data is recorded in various acoustic conditions and real-life scenarios [18].
- **CML-TTS**: a dataset for text-to-speech training containing segments from the same LibriVox audiobooks as MLS [42], leaving out the English part. Therefore, 7 European languages are present. The audiobooks were downloaded using the LibriVox API and resampled to 24 kHz instead of 16 kHz used in MLS. They were then split into smaller segments and realigned with aeneas[15] to facilitate training text-to-speech models. The acronym stands for CML-Multi-Lingual-TTS [39].
- **Common Voice**: a collection of transcribed speech primarily intended for speech recognition. As of December 2023, the website[16] provides 120 speech corpora in the category of launched language. The contents are crowdsourced by volunteers who read provided sentences, the recordings are then verified by other contributors [2].
- **CSS10**: a dataset divided into 10 language parts, ranging from 4 to almost 24 hours, but every language subset only features a single speaker reading one or up to 4 books. The audiobooks were extracted from LibriVox [41].
- **M-AILABS**: a corpus of transcribed speech in 8 European languages for machine learning purposes. All recordings except the Ukrainian subset come from LibriVox [48].
- **Multilingual LibriSpeech (MLS)**: a collection of segmented and transcribed audiobooks from LibriVox in 8 European languages. The subsets contain multiple speakers reading anywhere between a chapter and multiple books. The subsets differ significantly in size, with the English subset taking up 2.4 TB and the Polish 6.2 GB [42].

[15] https://www.readbeyond.it/aeneas/docs/index.html
[16] https://commonvoice.mozilla.org/en/languages

Table 3.2: Key aspects of multilingual corpora.

Corpus	Languages	Hours	Transcribed
Babel	10	800+	✓
CML-TTS	7	3,233+	✓
Common Voice	120	20,000+	✓
CSS10	10	141	✓
M-AILABS	8	999	✓
MLS	8	50,000+	✓
TUNDRA	14	60	✓
VoxLingua107	107	6,000+	

- **Spoken Wikipedia Corpus Collection**: the first version of this corpus contains partially-aligned speech in English and German [26].
- **TUNDRA**: a dataset of preprocessed audiobooks in 14 languages assembled from LibriVox and Project Gutenberg[17]. For each language, one audiobook read by a single speaker was chosen. Transcripts were manually corrected by native speakers. The total amount of speech is 60 hours [50].
- **VoxLingua107**: a dataset for speech recognition in 107 languages, the source of the utterances are videos retrieved from YouTube. Individual subsets span from 2 hours to 155 hours of speech, totaling over 6,000 untranscribed hours [54].
- **Other**: open-source database VoxForge[18] and partly open-source Magic Data corpora[19].

As seen in Tables 3.1, 3.2, most of the datasets are transcribed, making them usable for training TTS systems. Nevertheless, they vary noticeably in the number of speakers, languages, and their total size. With this information in mind, another type of dataset, often based on these corpora, can be examined: the deepfake datasets used to train deepfake detection tools.

[17] https://www.gutenberg.org/
[18] https://www.voxforge.org/
[19] https://www.magicdatatech.com/datasets/asr

Chapter 4

Voice deepfake datasets

In Chapter 2, gender bias and the underexamined language variable in audio deepfake detection were discussed. To establish whether there is a gender bias, language dependency, or how they can be detected, an overview of existing datasets is needed to select a suitable one for the task. Moreover, the analysis can be useful for pointing out common features and weak points of said datasets and help design a new generation of audio deepfake datasets.

Searching for audio and speech deepfake datasets on Google Scholar[1], Arxiv[2] and scanning relevant surveys resulted in finding more than 30 databases that contain computer-generated speech. Some surveys also include H-Voice[3], an audio deepfake spectrogram dataset, Baidu Neural Voice cloning samples [4], or recordings from other tools' demo pages in their list of deepfake datasets. In this thesis, only datasets containing at least hundreds of clips and easily downloadable as a single or a small number of archives were considered.

The analysis includes 25 speech deepfake datasets and basic information about 2 singing and 6 video datasets including spoofed speech. The most commonly used datasets for benchmarking audio deepfake detection systems are ASVspoof challenges 2019 [59] and 2021 [30], Fake or Real [44], WaveFake [16] and In-the-wild [35]. The latest ASVspoof challenge has not released the dataset or its description publicly yet, and therefore it cannot be included in the analysis.

The Fake or Real dataset is composed of English recordings and the spoofed samples were generated using commercial tools. It comes in 4 variants: the original collection, normalized dataset, the normalized dataset with 2-second long recordings, and a rerecorded set [44]. WaveFake is a vocoder-based dataset featuring only spoofed recordings of 2 female speakers, one Japanese and one English-speaking [16]. The corresponding genuine audio can be supplied by downloading the LJSpeech [20] and JSUT [49] datasets. Finally, In-the-wild is an audio deepfake dataset containing speech uttered by known public figures, such as politicians, and corresponding deepfake audio collected from social media and other sources. It is however an unbalanced dataset; most speakers are male and two-thirds of the recordings are genuine [35]. All datasets are listed in Table 4.1.

[1] https://scholar.google.cz/

[2] https://arxiv.org/

[3] Ballesteros, Dora M., Yohanna Rodriguez, and Diego Renza. A dataset of histograms of original and fake voice recordings (H-Voice). *Data in brief 29* (2020).

[4] https://sforaidl.github.io/Neural-Voice-Cloning-With-Few-Samples

Table 4.1: Datasets containing deepfake audio, divided into 3 parts: speech deepfake datasets, song deepfake datasets, and video datasets containing deepfake speech. The collected data includes the year of publishment, the accessibility of the dataset, the number of systems generating deepfake speech, and the number of languages represented.

Dataset	Year	Accessibility	DF tools	Languages
VCC 2016 [52]	2016	public	18	1
VCC 2018 [32]	2018	public	23	1
ASVspoof 2019 LA [59]	2019	public	17	1
Fake or Real [44]	2019	public	7	1
MC-TTS [57]	2020	restricted	1	1
Sprocket-VC [57]	2020	restricted	1	1
SynSpeechDDB [71]	2020	restricted	16	2
VCC 2020 [67]	2020	public	3	1
ASVspoof 2021 DF [30]	2021	public	100+	1
ASVspoof 2021 LA [30]	2021	public	13	1
FMFCC-A [72]	2021	restricted	13	1
Half-Truth (HAD) [64]	2021	public	1	1
WaveFake [16]	2021	public	7	2
AD for SFR [62]	2022	restricted	5	1
ADD challenge 1 [65]	2022	restricted	N/A	1
CFAD [33]	2022	public	11	1
F&M [13]	2022	public	1	2
In-the-wild [35]	2022	public	N/A	1
ADD challenge 2 [66]	2023	restricted	N/A	1
DECRO [3]	2023	public	10	2
LibriTTS-DF [28]	2023	restricted	3	1
PartialSpoof v1.2 [70]	2023	public	9+	1
TIMIT-TTS [45]	2023	public	12	1
Voc.v2,v3,v4 [58]	2023	public	4	1
MLAAD [36]	2024	public	19	23
FSD [61]	2023	restricted	5	1
SingFake [68]	2023	restricted	N/A	5
DFDC [11]	2019	restricted	1	N/A
FakeAvCeleb [24]	2021	restricted	1	1
AV-Deepfake1M [8]	2023	restricted	2	N/A
DefakeAVMiT [63]	2023	restricted	2	1
LAV-DF [9]	2023	public	1	1
SWAN-DF [28]	2023	restricted	4	N/A

4.1 Dataset parameters

Available speech deepfake datasets can be divided into groups or summarized based on their availability, year of publication, size, number of utterances, included deepfake types and tools, languages, real speech source, codec, postprocessing, demographics, utterance pairing by speakers, metadata, and recordings' quality assessment.

Availability

Publicly available datasets play a key role in deepfake detection. They can be examined and used by a diverse group of researchers, developers, or enthusiasts and combined into bigger structures, as in Attack Agnostic Dataset[5]. However, only 17 speech deepfake datasets can be freely downloaded by anyone. Some datasets can be claimed per request, some by creating an account at a specific site or by registering for a challenge, for others the download link isn't working or they were never meant to be publicly available.

Year of publication

Datasets containing audio that can be classified as deepfake started appearing almost 10 years ago [52]. While it was not the aim of the Voice Conversion Challenge to create

[5]Kawa, Piotr, et al. Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection. In: *Proc. Interspeech 2022*.

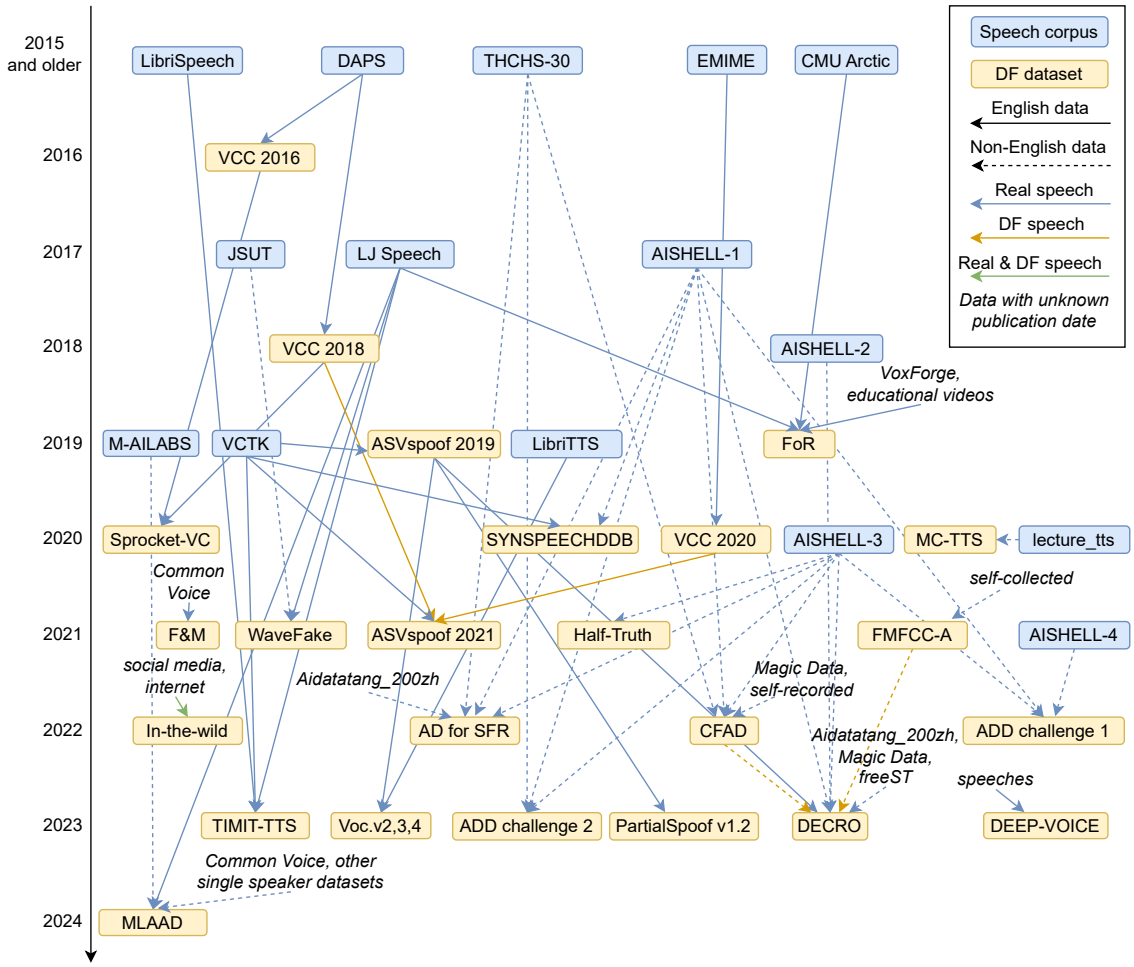

Figure 4.1: Audio datasets containing deepfake speech with the corpora they were made from. The time axis shows the year the dataset was published. The year of publication does not always align with the recording time: THCHS-30 [56] was recorded 15 years before it was made publicly available.

a deepfake dataset, the participants' submissions can be considered as such. FoR [44] and ASVspoof [59], the oldest datasets included in this study which were initially deepfake datasets, appeared in 2019. Speech deepfake datasets started rapidly appearing after 2021, see Figure 4.1.

The year of the dataset creation plays a crucial role in its practical usability, as older datasets do not include recordings generated by state-of-the-art tools. However, deepfake datasets commonly use synthesis systems that were available already for some time at the moment of their use or commercial tools that do not specify how recently their tools were updated. This means that if a dataset was published in a given year, the recordings can be partially generated by technology that was already considered old at that time.

Size and utterances

The size of audio datasets ranges from hundreds of MB to tens of GB. If the dataset is supposed to be used also for training deep learning models based on spectrograms, the training set should contain at least 70,000 recordings [15], so the whole dataset should have at least 80,000 recordings. The utterance counts per dataset are shown in Figure 4.2. Some datasets only include fake audio, but most also include genuine recordings. Balancing these two classes avoids the need for augmentation. In some cases, e.g. Wave-Fake [16], the claimed number of recordings does not match the number of utterances in the published archive which can make choosing the right dataset for a task challenging.

Deepfake types and tools

The examined datasets contain recordings generated by voice conversion and text-to-speech tools as well as partially fake recordings. 19 speech datasets were created using TTS tools

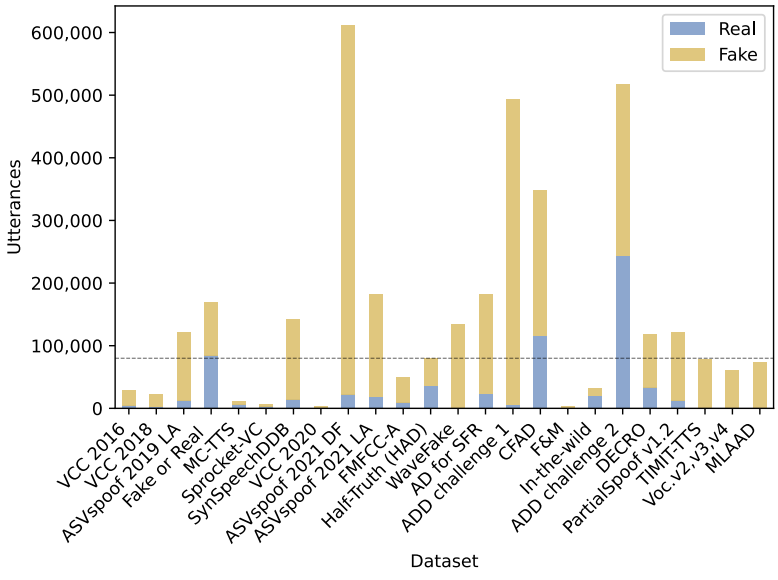

Figure 4.2: Number of utterances per speech deepfake dataset. The dashed line at 80,000 points out the recommended number of utterances for a deep-learning-based detector (small test set included).

to generate deepfake recordings, out of which 6 only used TTS [44, 62, 13, 45, 57, 36], 14 obtained utterances through voice conversion, out of which 4 datasets very purely VC-generated [52, 32, 67, 57], and 10 datasets combined these approaches [59, 71, 30, 72, 65, 66, 3, 28]. Two datasets are purely vocoder-based [16, 58] and four datasets also include partially fake recordings [33, 70, 65, 66].

The number of different tools used ranges from 1 to over 100 in the datasets that provided this information. Audio-visual datasets mostly use SV2TTS[6]. Commonly used tools for speech deepfake datasets include Tacotron[7], HiFiGAN[8], Parallel WaveGAN[9], MelGAN[10] and versions of VITS [25]. Some datasets also used commercial tools for generating deepfake recordings, such as BaiduTTS[11].

A wide range of tools used is important to address real-life detection scenarios. Not all dataset creators specify the tools utilized, which can negatively influence the credibility of cross-evaluation if training and evaluation datasets use recordings produced by the same or similar technologies. However, in the case of datasets that collect real-life deepfakes, such as In-the-wild [35] or SingFake [68], it is not possible.

Language

Until early 2024, deepfake datasets were only available in 6 languages (English, Chinese, Japanese, Spanish, Czech, and Persian), the most prominent being English and Chinese. Most datasets only include deepfakes in one language, yet four bilingual datasets can be found: WaveFake (English, Japanese) [16], SYNSPEECHDDB (English, Chinese) [71], F&M (English, Czech) [13], and DECRO (English, Chinese) [3]. Nevertheless, Chinese datasets are more often inaccessible. With the arrival of MLAAD, the language diversity was significantly improved. However, some of the sets presented in this dataset are of low quality or use machine translation to obtain input texts for TTS [36].

Real speech source

Most Chinese datasets use some of the AISHELL corpora [6, 12, 46, 17], THCHS-30 [56] or Aidatatang_200zh[12]. English datasets are more varied, including EMIME[13], DAPS [34], VCTK [55], LJSpeech [20], ARCTIC [27]. Sometimes, resources from social media or other parts of the internet are used. Self-recording is rather uncommon. For detailed information about speech-only datasets, see Figure 4.1. A large part of the datasets derive from common sources like the AISHELL corpora or LibriVox[14] which can cause issues in cross-validation and real-life deepfake detection scenarios.

[6] Jia, Ye, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[7] https://github.com/Kyubyong/tacotron

[8] https://github.com/jik876/hifi-gan

[9] https://github.com/kan-bayashi/ParallelWaveGAN

[10] https://github.com/descriptinc/melgan-neurips

[11] https://intl.cloud.baidu.com/product/speech.html

[12] https://www.datatang.ai

[13] Wester, Mirjam. *The EMIME bilingual database*. The University of Edinburgh, 2010.

[14] https://librivox.org/

Audio format

Most audio datasets contain WAV files, except for ASVspoof challenges that opted for FLAC. Both are lossless types. Some recordings in Fake or Real are in MP3. For videos, MP4 is sometimes used. Using the same format for recordings facilitates the development of detection tools and their cross-evaluation.

Postprocessing

Postprocessing is explicitly mentioned in several dataset descriptions and the processes include resampling [35, 44, 67, 62, 58], change of codec [35, 44, 30, 72, 62, 33, 45], normalizing volume [44, 64, 70], and noise addition [72, 33, 45]. These modifications reflect some of the uncertainties present in detecting real-life deepfake audio and others can be added, either directly by the researchers using the datasets or implemented as part of augmentations in deepfake detectors. However, it is important to mention the presence or lack of postprocessing in the dataset description to better understand the evaluation of deepfake detectors and the risks these modifications pose to it.

Demographics

Most datasets' descriptions do not pay closer attention to the age, gender, or accents of the speakers represented. Those who do usually only specify the number of speakers of each gender. Except for the Voice Conversion Challenges' datasets [52, 32, 67] and TIMIT-TTS [45], the genders aren't near balanced. Gender representation should be described in 2 or 3 ways: the number of distinct female and male speakers, the number of recordings uttered by female and male speakers, and ideally also the number of recordings per speaker.

Paired speaker recordings

Seven datasets mention representations of the same speakers in both utterance classes [35, 32, 59, 67, 30], rather than using generic TTS voices, and for WaveFake [16], the real samples could be supplied. Whether including genuine and deepfake samples by the same speaker in the training set positively influences the detectors' performance remains a question for future research.

Metadata

Nine datasets offer protocols, such as metadata about recordings or training splits. Ideally, this metadata should contain the speaker's ID, gender, synthesis tool used, postprocessing information, and if available, the speaker's accent, native language, age, or else.

Quality assessment

Human evaluation of naturalness and speaker similarity was only carried out on the Voice Conversion Challenges' datasets [52, 32, 67]. Therefore, it may be possible that the deepfakes that are used to train and test the detectors sound robotic and unnatural – and so, they are not enough to train the detectors to recognize more sophisticated attacks. Such deepfakes can be found in several datasets, for instance, DECRO [3] or Fake or Real [44]. As the datasets usually feature thousands of recordings, their human evaluation is not a

simple task. It would be however preferred to have a sample reviewed by human listeners or have the dataset evaluated using an automated tool.

4.2 Common weak points

Out of the 25 examined speech deepfake datasets, 14 are suitable for training deep-learning spectrogram-based detectors, and out of these 10 are publicly available. Most datasets lack in their descriptions: be it metadata, information about the speakers, or even the language used: that has to be inferred from the source of genuine recordings. Concerning the sources, the pitfall of the datasets as a whole is reusing the same corpora.

It is not easy to find different speech sources, so using tried-and-trusted corpora is understandable, yet the research would benefit from introducing new sources of genuine speech and training materials for the TTS and VC systems. The datasets' authors should also include samples generated by newer technologies to prevent their datasets from becoming obsolete soon after publishing.

Another big gap in the current dataset landscape is the lack of gender-balanced datasets and the lack of inclusion of languages other than English and Chinese, a necessity for the development of detection tools for other languages. The existing datasets make testing the language influence on detection accuracy and evaluating the gender bias on a gender-balanced dataset hard, as it would involve a lot of preprocessing and manual selection.

Chapter 5

Design of a new dataset

In the previous chapters, the current state of audio deepfake research, methods for creating voice deepfakes, and existing datasets were discussed. It was pointed out that the influence of the language spoken has been scarcely explored, and there are no suitable datasets for such experiments. The datasets are available in a small number of languages and many big languages are represented only by a few thousand recordings. As for possible gender bias, there are no gender-balanced datasets except the Voice Conversion Challenges' submissions.

5.1 Research directions

The objective of this thesis is the creation of a dataset that would facilitate addressing two gaps in audio deepfake detection: the ability of deepfake detectors to generalize across languages and the assessment of possible gender bias in deepfake detection. The dataset was therefore designed with the two following questions in mind.

Can audio deepfake detectors generalize across languages?

Current datasets use different tools to generate deepfakes and vary in several other parameters such as size or quality. Although it may be possible to cross-test audio deepfake detectors using existing datasets in different languages, the results could be influenced by other factors, such as the tools used, the real speech sources and the structure of the dataset.

On the other hand, using a novel dataset, the detectors could be trained on audio sets of similar quality, generated by the same tools, and using real speech from the same source. The results of testing Language A on a detector trained on Language B should be therefore less influenced by other factors than the chosen languages and their similarity.

Moreover, a detector could be trained on a multilingual set with the same structure and size as an individual language set. Comparison of its accuracy with a detector trained in a single language could indicate whether it would be preferable to train multilingual detectors and what decrease in accuracy could they possibly cause.

Is there a gender bias in audio deepfake detection?

Bias in deepfake detection has been examined primarily in face deepfake detection. Speech corpora are not annotated enough, but some include the information about speaker's gender, which could be determined manually in other datasets. Using a detector with a bal-

anced gender ratio, the error rates for male and female deepfake voices could be calculated separately.

5.2 Requirements

As recording and collecting speech with aligned transcripts are tedious and time-consuming activities, it is preferable to use one of the already existing speech corpora. There are several requirements for the dataset's contents:

- At least 3 languages. Bilingual speech datasets already exist (e.g. DECRO [3] or WaveFake [16]), even though their language subsets are not necessarily balanced. The more languages the new one has, the more combinations can be tested, and the more deepfake detection models can be trained.
- At least 10 speakers for each language, with at least 5 women and 5 men. This thesis aims to create a multilingual and gender-balanced dataset. To have speaker-disjoint training, development, and testing sets, the more speakers we can collect, the more varied the subsets can be.
- Multiple hours of speech per speaker, totaling at least 24 hours evenly distributed between the speakers. Most text-to-speech (TTS) tools use the LJSpeech dataset [20] for benchmarking, and therefore we assume that the number of hours it contains is enough to train a TTS tool. Voice conversion tools (VC) commonly make use of VCTK [55] with 44 hours of speech, however, to keep the training time realistic considering available computational resources, 24 hours of training material will be used.
- Speech in English. English is included as a reference language that we suppose most researchers in the field of audio deepfake detection understand and, therefore, can assess the features and quality of this new dataset themselves.
- Aligned transcripts. This is necessary to train text-to-speech models. Transcription tools could be used at the risk of language-dependent errors.

MLS [42] is the only speech corpus examined that meets all the mentioned requirements.

5.3 Dataset proportions

To train certain spectrogram-based audio deepfake detectors, at least 70,000 utterances are needed [15]. Therefore, to be able to use individual parts separately, each language subset should contain more than 70,000 utterances for training. The development and testing sets should be smaller to maintain a reasonable dataset size.

MLS [42] contains 8 language subsets. After a closer look, four to five of the subsets seem to have enough speakers with enough hours and therefore are apt for creating a deepfake dataset. Given that training synthesis tools require a noticeable amount of computational resources, we select only four of them: English (the reference language), German, French, and Spanish. However, the Italian recordings could be used to create a smaller set to complement the multilingual detector. The Dutch, Portuguese, and Polish subsets have parameters, specifically the amount and the distribution of recordings among speakers, that are incompatible with this project.

Table 5.1: Real speech and speakers selected for a language subset.

	Gender	Synthesis	DF train	DF dev	DF eval
Speakers	M	24	5	2	5
	F	24	5	2	5
Real speech [min./sp.]	–	30	180	70	70

The Italian subset is the smallest and most limited of the suitable datasets. However, there are more than 24 female and 24 male readers with more than 0.5 hours of read audiobooks. VCTK [55] contains approximately 24 minutes of speech per speaker, therefore to imitate this, approximately 25-30 minutes of speech per speaker to train the synthesis models will be used. To avoid possible bias, the recordings used for training the synthesis tools were excluded from appearing in the deepfake dataset as genuine recordings. Unlike in some other datasets, all parts (training, development, and testing) of a language subset contain genuine and deepfake recordings of the given speakers and are speaker-disjoint.

The material available for genuine recordings is enough to use 180 minutes of real recordings per speaker in the training and 70 minutes in the development and test set. As some of the speakers only have a little more than 30 minutes of recorded speech, the number of speakers in the dataset is reduced to 24; 12 of each gender. The use of the real recordings is summarized in Table 5.1.

5.4 Synthesis tools used

To ensure that the subsets are as similar as possible, training custom synthesis models is crucial. The Coqui.ai[1] TTS toolkit was selected for its unified interface and the many tools it provides. To provide a certain diversity in deepfake tracks and their quality, different neural architectures were chosen. The framework provides pretrained models for these tools, so the outputs can be subjectively assessed beforehand. However, only the VITS model's training worked correctly on a custom dataset, thus different systems had to be selected. The number of trained models is limited to 5 (2 end-to-end tools, 2 encoders, and 1 vocoder) for each language due to limited computational resources.

- TTS: VITS (end-to-end) [25], ZMM-TTS [19] with HiFiGAN[2]
- VC: LVC-VC (end-to-end) [22], DDDM-VC [10] with HiFiGAN

The tools were selected based on their year of publication, code availability (GitHub), compatible license (typically MIT, Apache...), and the estimated difficulty of getting them to work. All of the selected systems are multispeaker – they can generate speech with different speaker characteristics. Training a separate model for each speaker is not feasible with the available data and computational resources. All tools used are described in Chapter 3.

[1] https://coqui.ai/
[2] https://github.com/jik876/hifi-gan

Chapter 6

Dataset compilation

This section describes the technical side of creating the proposed dataset: selecting speakers and their respective speech segments, diving the speakers and segments into disjoint groups based on their future use, realigning the training segments for training of text-to-speech (TTS) systems, discarding problematic recordings and equalizing the language sets in terms of audio length. With these in hand, the TTS and voice conversion (VC) models can be trained. To generate deepfake recordings, input is needed. For TTS, the text was extracted from unused MLS dataset [42] recordings, for VC the remaining unused recordings were divided into groups and split into shorter segments. The whole process can be seen in Figure 6.1.

6.1 Speaker and segment selection

First, the uncompressed MLS subsets were downloaded for all 5 selected languages. The speakers used in the dataset were selected based on the length of audio included in MLS and based on their gender. For smaller language sets, this task was challenging: for each speaker, there must be around 30 minutes to train TTS and VC tools, enough reference speech to be used as genuine recordings in the new dataset, and finally enough spare recordings to generate converted speech, either by using the audio clips directly for voice conversion or by extracting the transcripts for TTS.

To ensure enough reference audio is left for each represented speaker, on top of the 24 speakers directly used in the deepfake dataset, additional 24 speakers are used to train the tools to reduce the audio needed per speaker by half. The reference audio length corresponds with the ratio of the training, development, and test set. For speakers in the train set, 3 hours are used, for speakers in the development and test set, it is 70 minutes. The speaker and segment selections are implemented in `speakers.py` and `segments.py` respectively.

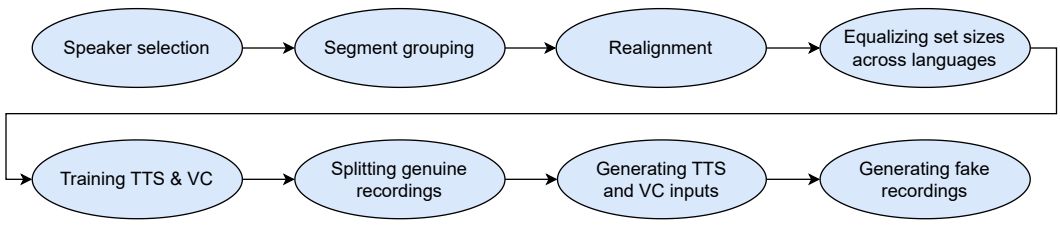

Figure 6.1: The process of compiling a new speech deepfake dataset.

Table 6.1: Total audio length and number of utterances in training sets.

Language	English	German	French	Spanish	Italian
Utterances	14,379	14,224	13,804	14,256	13,945

6.2 Realignment

After the segments were divided into groups, those that were assigned to be used for training synthesis tools needed to be split, as too-long recordings are not suitable for speech synthesis. These segments must also be realigned with a part of the original transcript, a requirement for TTS models. In the beginning, the same method for splitting the clips was considered as in the CML-TTS [39] dataset description was considered. However, manually testing aeneas[1] yielded largely inaccurate results on the tested Italian recordings.

The original paper claims to have split recordings longer than 15 seconds and discarded clips with less than 90% similarity. This would mean splitting approximately half of the recordings in the whole Italian subset and resulting in a 50% increase in clip count (in case of no errors) in comparison to MLS. However, the CML-TTS Italian subset has fewer recordings than the original MLS set, suggesting about 60% or higher error rate. Therefore, this method cannot be applied when the available audio length is already limited.

As force alignment tools, programs that match audio with its transcription, e.g. by adding timestamps to individual words, are language-dependent, a system supporting English, German, French, Spanish, and Italian is needed. Another tool with modules for all chosen languages is Montreal Forced Aligner[2] (MFA). MFA returned considerably more accurate results when manually tested, however aligning larger batches resulted in various errors in both alpha and stable versions, suggesting some transcriptions might be slightly inaccurate.

The third tested option was Whisper[3], a model providing both transcription and timestamping of audio recordings. The timestamping is not precise enough and sometimes results in clipping a word in half but the transcription worked better. Therefore, it was used to transcribe segments split on silence using pydub library[4]. To avoid a significant difference in the quality of the transcription between the language sets, the partial transcriptions were concatenated and compared with the original transcript. To compare them, their Levensthein (edit) distance with equal weights for substitution, deletion, and insertion was calculated. If the normalized Levensthein distance was smaller than 0.1, the split recording and its transcriptions were added into the training set, as seen in Figure 6.2.

After all sets were realigned, they were equalized based on the total audio length of the shortest training set: 22.82 hours of speech were retained for each language. The difference between the sets with the lowest and the highest number of utterances, the French and the English sets, is 575 utterances, which represents 4.17% of the French set's recordings, as shown in Table 6.1. Additionally, none of the recordings by one male Italian speaker and one male German speaker passed the 90% transcript similarity rule, meaning that these two speakers were not represented in the training set. The length distribution is visualized in Figure 6.3. The realignment process is implemented in `realign.py`, the set normalization in `equalize.py`.

[1] https://www.readbeyond.it/aeneas/docs/index.html
[2] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner
[3] https://github.com/openai/whisper
[4] https://github.com/jiaaro/pydub

6.3 Training synthesis tools

After the audio segments were split and realigned, the synthesis tools were ready to be trained. However, some of the originally considered tools turned out not to be usable for this dataset: SpeedySpeech[5] crashed after 60,000 training steps even with the reference LJSpeech dataset [20], OverflowTTS[6] could not be trained beyond a similar number of steps on the custom dataset and Glow-TTS[7] produced very noisy recordings. The tools not implemented in the Coqui.ai[8] toolkit, except VITS [25], happened to be better suited for this project.

Several modifications were required, mostly related to the environment, data loading, or logging. The changes to the original repository are listed in a separate file, as described in `README` of the accompanying storage media. The synthesis tools were trained until the recordings they produced reached an intelligible quality, a compromise between the number of steps used in the original paper and available computation resources. The number of steps, batch sizes, and training time are listed in Table 6.2.

For the ZMM-TTS text analysis module, the configuration using IPA transcription was selected and the linguistic features were then fed into the `vec2mel` module, leaving the synthesis to a self-trained HiFiGAN model, instead of synthesizing the text directly with `vec2wav`. Due to the lack of training scripts and incompatibility with already trained vocoders, the DDDM-VC vocoder was not trained. The models were trained on a single NVIDIA RTX A5000 GPU with 24GB memory. Additionally, training `f0_vq` requires significant CPU power.

6.4 Generating recordings

Before generating the recordings, protocols with texts for TTS and source recordings for VC systems were generated. Due to the limited resources for the less represented languages (Spanish, Italian), the texts for TTS are slightly augmented by using partially overlapping

[5] https://github.com/janvainer/speedyspeech
[6] https://github.com/shivammehta25/OverFlow
[7] https://github.com/jaywalnut310/glow-tts
[8] https://coqui.ai/

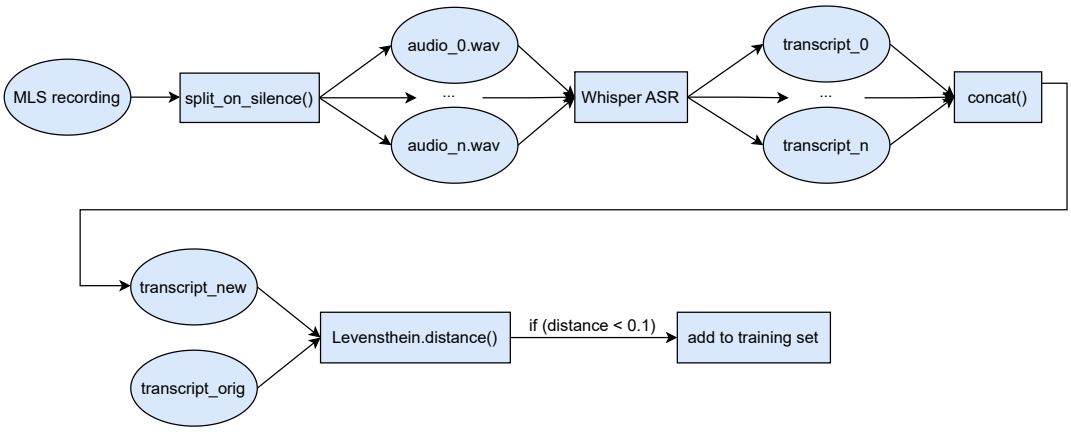

Figure 6.2: The process of splitting MLS recordings and adjusting their transcripts.

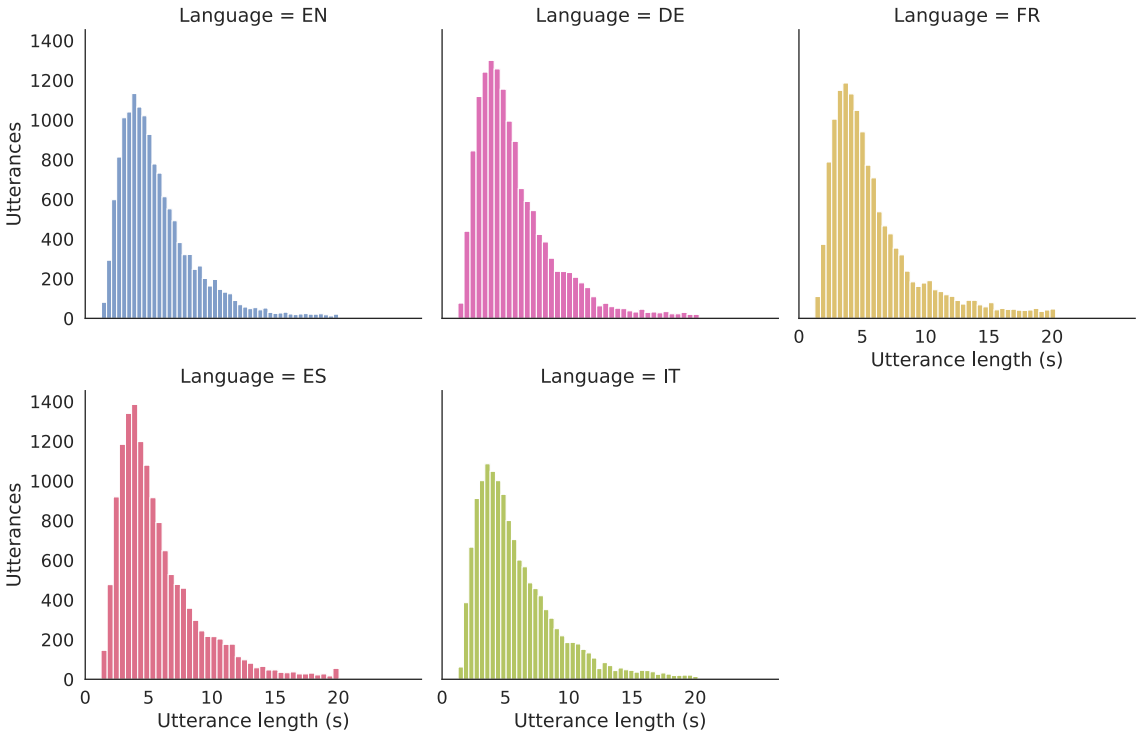

Figure 6.3: Duration distribution of clips per language in the training dataset for speech synthesis models.

text sequences from transcripts of unused recordings. For better extensibility, the texts are used for both TTS tools. However, the speakers are rotated for the second tool to avoid the same speaker pronouncing the sentence twice. To prevent the same phrase from appearing in two different sets (out of the training, development and test set), the speakers are rotated inside the specific set, as illustrated in Figure 6.4. Generating VITS utterances and protocols for generating ZMM-TTS utterances, including text augmentation and rotating speakers, are implemented in `tts.py`.

For VC tools, remaining unused sentences are split preferably on silence, if not possible then after under 6 seconds, and fed into the VC tool. To avoid using the same source speaker in different sets, the utterances are sorted by their speaker number, and each speaker is selected for one of the sets. The number of source recordings per subset is gender-balanced.

Table 6.2: Parameters of models trained for synthesis in each language.

Model	Steps/epochs	Batch size	Training time
VITS	140,000 steps	46	42 hours
LVC-VC	350 epochs	16	42 hours
ZMM-TTS txt2vec	200,000 steps	16	8 hours
ZMM-TTS vec2mel	50,000 steps	24	6 hours
HiFiGAN	100,000 steps	32	27 hours
f0_vq	350,000 steps	16	62 hours
DDDM-VC	170,000 steps	32	48 hours

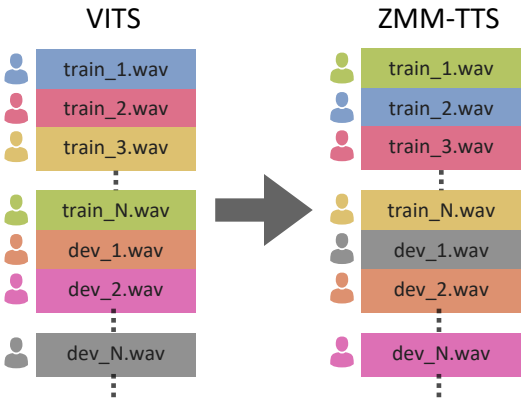

Figure 6.4: Rotating speakers when generating deepfake recordings.

Additionally, half of the generated recordings are produced by conversion of voices belonging to speakers of the same gender, the other half was created by converting a male voice to a female one or vice versa.

As some of the recordings were not split on silence but at an arbitrary timestamp to avoid too-long segments, there must be an auxiliary step to prevent silent segments from being added to the dataset. Thus, the last segment is only taken into consideration if it is longer than 2 seconds and the transcript does not match one of the suspicious transcriptions that occurred when the segment was silent during manual testing.

This task is divided into 2 scripts: `generate_train_lists.py` divides the unused utterances into speaker-disjoint gender-balanced sets and splits the recordings. As the process is not deterministic, using a random threshold under 6 seconds, the second script, `vc_selection.py` selects the adequate number of split utterances, and matches them with target speakers and their reference utterances. The bonafide utterances for all speakers were split the same way, using the `groundtruth.py` script. The dataset is finalized by generating the metadata files, including the genders and subsets of the speakers, with `metadata.py`.

The inference time for one language set of 13,500 generated recordings was 72.5 $\pm$ 6.58 minutes for VITS, 11.4 $\pm$ 0.5 minutes for the 3-step ZMM-TTS pipeline, 177 $\pm$ 4.21 minutes for LVC-VC and 132 $\pm$ 1.87 minutes for DDDM-VC with the pretrained HiFiGAN vocoder. For LVC-VC inference, the source speakers were unseen and the target speakers were set as seen. In total, 224,000 utterances were generated.

6.5 Comparison with existing datasets

The dataset design started in 2023 and the dataset was finalized in the summer of 2024. It contains 448,000 audio clips in 5 languages, as seen in Table 6.3, of which half are deepfakes recordings generated by 4 tools: 2 TTS and 2 VC systems. Three out of four synthesis tools were published in 2023 or later, the remaining tool, VITS, is the backbone of several other widely used TTS systems.

Some of the synthesis systems relied on a pretrained model as part of the synthesis process, but no commercial or completely pretrained system was used: all systems were trained on a custom training set of MLS [42] recordings. The genuine recordings were also adapted from MLS. Although MLS has not been used as a source for any other dataset,

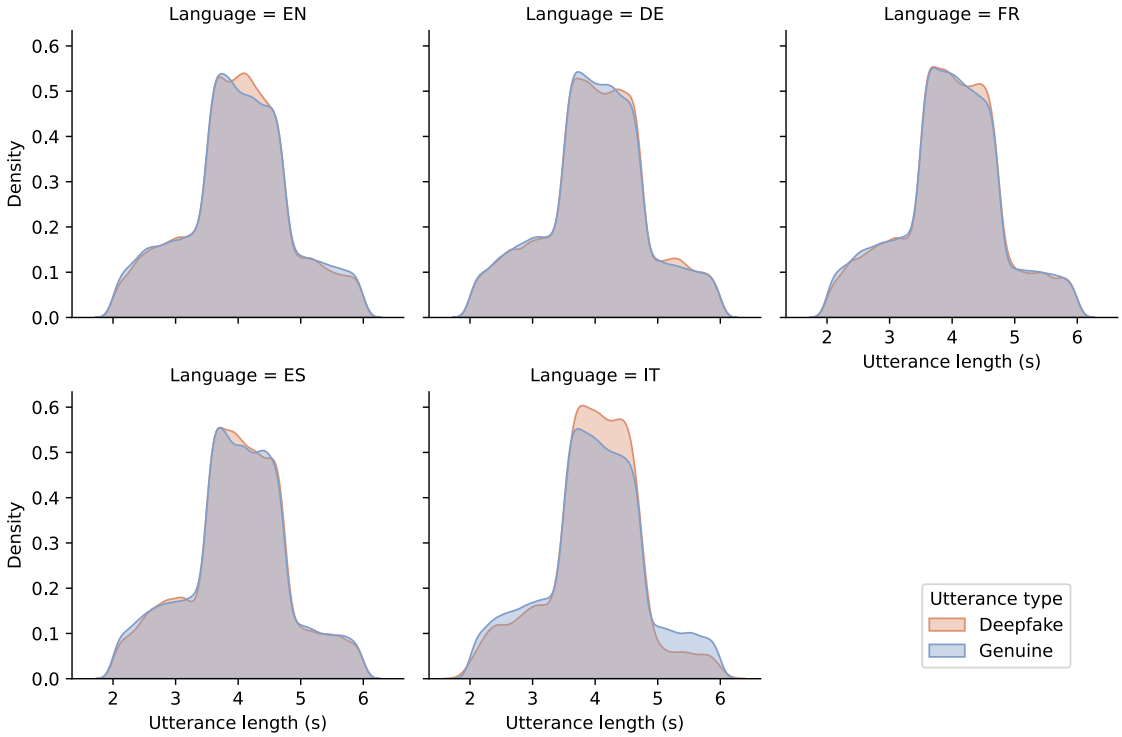

Figure 6.5: Length distribution of utterances in the proposed dataset.

the recordings come from LibriVox[9], a commonly used source for audio deepfake datasets. ASVspoof5[10] announced that they may use the English set of MLS for their dataset which is not publicly available yet.

The files were split into shorter segments and saved as WAV files instead of FLAC for compatibility reasons. Most recordings are 2-6 seconds long, and the length distributions across the language subsets and the two classes (genuine, deepfake) are very similar, as seen in Figure 6.5. No other postprocessing techniques were used. The dataset is gender-balanced in 3 ways: there is the same amount of speakers of both genders represented, the same number of recordings uttered by female and male speakers, and in the scope of the subsets (training, development, and test sets), the number of genuine and deepfakes

[9] https://librivox.org/
[10] https://www.asvspoof.org/

Table 6.3: Proportions of the new multilingual, gender-balanced speech deepfake dataset.

Language	Deepfake utt.	Genuine utt.	Male speakers	Fem. speakers	Tools
English	54,000	54,000	12	12	4
German	54,000	54,000	12	12	4
French	54,000	54,000	12	12	4
Spanish	54,000	54,000	12	12	4
Italian	8,000	8,000	5	5	4
Total	224,000	224,000	53	53	4

recordings per speaker generated by a particular tool are equal for male and female speakers. Additionally, the genuine and deepfake recordings are uttered by the same set of speakers.

Therefore, this dataset is one of the largest deepfake datasets created, the only publicly available dataset with more utterances is the ASVspoof2021 DF set [30]. In terms of languages included, it is the second most diverse speech deepfake dataset after MLAAD [36]. It is the only fully gender-balanced dataset available and, unlike the DECRO bilingual database [3], the 4 main language sets are perfectly balanced. The Italian part only contains a small training set. 3 out of 4 used tools were recently published and introduced for the first time as a part of a deepfake dataset. A possible improvement would be generating deepfake recordings with other TTS and VC systems which was not considered in the scope of this thesis due to limited computational resources.

Chapter 7

Experiments

With the dataset ready, the research questions from Chapter 5.1 can be further examined. As representative audio deepfake detection tools, 2 detector architectures based on the supervisor's recommendation will be used: LCNN with LFCC frontend and RawNet3. They were described in Chapter 2.2 and implemented as part of the Audio Deepfake Adversarial Attacks detection framework presented at Interspeech 2023 [23].

The models were trained on a single NVIDIA RTX A5000 GPU for 5 epochs, LCNN with batch size 256, and RawNet3 with batch size 64. The training took 20 minutes in the case of a LCNN model, and over 3 hours 15 minutes in the case of a Rawnet3 model. Each configuration was trained 3 times with a different seed to generalize the results.

7.1 Deepfake detection abilities across languages

This experiment is divided into 2 parts: monolingual and multilingual. By monolingual, a detector trained on utterances in only one language is meant, while a multilingual detector was trained on bonafide and spoofed utterances in multiple languages. In the first part, 4 detectors are trained, each on 80,000+ utterances in a given language, as seen in Table 7.1. Then, the accuracy on the corresponding test set is compared with the accuracy of tests in different languages. The goal is to determine if a detector trained on language A can be useful (and how accurate) to detect deepfakes in another (fairly similar) language B.

Table 7.1: Utterances used for experiments with monolingual and multilingual detectors. `G` refers to genuine recordings, `D` to the deepfakes.

Monolingual	Train (G)	Train (D)	Dev (G)	Dev (D)	Test (G)	Test (D)
Utt./speaker/tool	4,000	1,000	1,000	250	1,000	250
Speakers	10	10	4	4	10	10
Tools		4		4		4
Total	40,000	40,000	4,000	4,000	10,000	10,000
Multilingual	Train (G)	Train (D)	Dev (G)	Dev (D)	Test (G)	Test (D)
Utt./speaker/tool	4,000	1,000	1,000	250	1,000	250
Speakers/language	2	2	1	1	10	10
Tools		4		4		4
Languages	5	5	4	4	1	1
Total	40,000	40,000	4,000	4,000	10,000	10,000

In the second part of the experiment, a multilingual detector is trained on a language-balanced set of 80,000+ utterances in 4 languages. It is then tested with the individual language test sets and the results are compared to those of the monolingual detectors. The goal is to explore whether monolingual detectors could be substituted by multilingual detectors with a satisfying result, i.e. whether the detectors can be trained on multiple languages, with fewer utterances per language, and still efficiently distinguish deepfake and genuine recordings. The following hypotheses are taken into consideration:

$$H_0 : Acc(MULTI, x) = Acc(X, x)$$

$$H_1 : Acc(MULTI, x) \neq Acc(X, x)$$

Where $Acc(A, b)$ refers to the accuracy of a detector trained on set A and tested on set b. $MULTI$ stands for a multilingual detector trained in English, German, French, Spanish, and Italian, while X and x are the training and test sets, respectively, of a given language.

Results

As Table 7.2 suggests, the accuracy decreases when using a multilingual detector compared to a monolingual one, given that the training dataset is of the same size. This is confirmed by a χ^2 independence test on the ratio of correct and incorrect predictions, rejecting H_0, and showing that the distributions are statistically significantly different with a $p\text{-}value < 0.001$.

However, the accuracy decreases only by a few percent when using the multilingual detector, therefore it can be a viable solution for a multilingual environment with insufficient computational resources. In some cases, even better average accuracy can be reached with a monolingual detector trained in a different language, see the Spanish LCNN detector tested with German samples in Table 7.2. However, this is due to the large variance of the multilingual models, where two runs reached over 98%, and the last one 86.94% accuracy.

All monolingual LCNN models reached 95% or higher average accuracy for test sets in the four included languages, suggesting that this architecture can be used for detecting deepfakes in related languages if the same synthesis methods are used. The EERs for both architectures for all tests oscillate between 1% and 5%, providing sufficient distinguishability between deepfake and genuine samples.

7.2 Role of gender in audio deepfake detection

The monolingual detectors from Section 7.1 are tested with corresponding test sets and the accuracy for recordings by female and male voices is compared. The goal is to discover a potential gender bias on a gender-balanced dataset.

It is expected that tools trained on unbalanced datasets without over- or undersampling might struggle with identifying deepfakes or genuine recordings by the underrepresented gender. However, do balanced detectors discriminate? Many tools use spectrogram representation of utterances, while some datasets provide the clips in lossy formats. Speech uttered by male and female speakers has different frequency ranges, therefore some nuances could escape the detector's attention. The following hypotheses are tested:

$$H_0 : Acc(F) = Acc(M)$$

$$H_1 : Acc(F) \neq Acc(M)$$

Where $Acc(x)$ means the accuracy on class x, F stands for recordings of female speakers, M of male.

Table 7.2: EER & accuracy: experiments with languages, LCNN and RawNet3.

Accuracy (%): LCNN				
Train/test	English	German	French	Spanish
English	97.74 ± 0.29	96.10 ± 0.36	99.48 ± 0.01	97.91 ± 0.68
German	96.21 ± 1.25	97.76 ± 1.08	98.80 ± 0.35	96.27 ± 1.50
French	97.14 ± 0.54	95.85 ± 1.77	99.45 ± 0.29	96.83 ± 0.50
Spanish	98.62 ± 0.09	98.07 ± 0.52	99.45 ± 0.07	98.82 ± 0.41
Multilingual	96.52 ± 3.01	94.65 ± 5.45	98.91 ± 0.85	97.58 ± 1.42
Accuracy (%): RawNet3				
Train/test	English	German	French	Spanish
English	99.13 ± 0.46	88.49 ± 2.48	94.32 ± 0.15	94.64 ± 2.84
German	90.16 ± 0.73	94.54 ± 2.35	95.84 ± 0.45	96.03 ± 1.86
French	92.46 ± 0.58	89.12 ± 1.72	99.07 ± 0.20	94.16 ± 1.28
Spanish	94.99 ± 0.74	94.02 ± 0.78	93.96 ± 1.33	96.28 ± 0.75
Multilingual	96.13 ± 0.11	96.48 ± 0.70	98.24 ± 0.14	95.58 ± 0.62
EER: LCNN				
Train/test	English	German	French	Spanish
English	0.0159 ± 0.0021	0.0166 ± 0.0001	0.0042 ± 0.0003	0.0136 ± 0.0028
German	0.0183 ± 0.0020	0.0069 ± 0.0010	0.0050 ± 0.0007	0.0191 ± 0.0058
French	0.0186 ± 0.0028	0.0167 ± 0.0020	0.0035 ± 0.0013	0.0293 ± 0.0020
Spanish	0.0124 ± 0.0005	0.0132 ± 0.0012	0.0044 ± 0.0012	0.0103 ± 0.0017
Multilingual	0.0128 ± 0.0006	0.0179 ± 0.0063	0.0047 ± 0.0007	0.0161 ± 0.0041
EER: RawNet3				
Train/test	English	German	French	Spanish
English	0.0062 ± 0.0041	0.0255 ± 0.0021	0.0154 ± 0.0038	0.0311 ± 0.0101
German	0.0365 ± 0.0052	0.0177 ± 0.0038	0.0114 ± 0.0039	0.0215 ± 0.0059
French	0.0423 ± 0.0009	0.0397 ± 0.0019	0.0029 ± 0.0008	0.0487 ± 0.0073
Spanish	0.0382 ± 0.0051	0.0038 ± 0.0396	0.0039 ± 0.0023	0.0059 ± 0.0456
Multilingual	0.0307 ± 0.0004	0.0261 ± 0.0026	0.0120 ± 0.0006	0.0254 ± 0.0047

Results

The results of this experiment have shown that there is a statistically significant difference between the classification accuracy of male and female recordings on a gender-balanced dataset. Using the χ^2 test on a contingency table of (in)correctly predicted samples by gender, H_0 was rejected with a *p-value* < 0.001. In general, the female recordings were easier to classify. Practically, the difference was very small: out of 480,000 predictions made, 235,570 predictions of female recordings and 234,077 of male recordings were correct.

To further explore these findings, another χ^2 test was performed on the German set and RawNet3 models, since the accuracy for men and women differed noticeably, as seen in Table 7.3. The difference is again considered statistically significant, H_0 is rejected with a *p-value*< 0.001, contrary to the results of the English set, for which the distributions are not deemed to be significantly different – with a *p-value* of 0.83, H_0 is not rejected.

Therefore, even though the dataset was designed to be as balanced as possible, there is still room for error stemming from the quality of used source recordings, speaker selection, or other factors. Given the bigger accuracy variance of the German LCNN detector on the

Table 7.3: EER & accuracy: Experiments with gender, LCNN, and RawNet3.

LCNN				
Language	EER men	EER women	Acc. men (%)	Acc. women (%)
English	0.0151 ± 0.0024	0.0121 ± 0.0035	98.36 ± 0.29	97.19 ± 0.17
German	0.0093 ± 0.0024	0.0034 ± 0.0006	96.41 ± 2.00	99.10 ± 0.13
French	0.0049 ± 0.0018	0.0017 ± 0.0007	99.09 ± 0.46	99.79 ± 0.11
Spanish	0.0075 ± 0.0019	0.0103 ± 0.0015	98.70 ± 0.60	98.90 ± 0.18
RawNet3				
Language	EER men	EER women	Acc. men (%)	Acc. women (%)
English	0.0028 ± 0.0019	0.0077 ± 0.0041	99.14 ± 0.57	99.12 ± 0.36
German	0.0089 ± 0.0031	0.0173 ± 0.0077	92.68 ± 2.16	96.41 ± 2.56
French	0.0035 ± 0.0011	0.0016 ± 0.0003	98.59 ± 0.29	99.54 ± 0.25
Spanish	0.0156 ± 0.0048	0.0388 ± 0.0124	97.28 ± 0.94	95.28 ± 1.03

male recordings, it is also possible that the difference could be eliminated by running the training more than 3 times. The EER stayed below 5% for all languages and architectures tested, suggesting that the models are well capable of differentiating between deepfakes and genuine recordings for both genders.

7.3 Final note on the generalization ability

Deepfake detection is a complicated task: the detectors should not only be able to recognize deepfakes created by tools already seen, but they should also possess a certain ability to generalize and recognize deepfakes created by methods unknown to them, a feature many researchers are working on. The created dataset was generated using only 4 tools: how will it score facing other datasets generated by a more diverse set of synthesis systems?

Evaluation with In-the-wild

To show how the detectors would react when faced with a real-life deepfake, the English detectors were tested on the In-the-wild dataset [35]. The results have shown while identifying real samples did not pose a problem with a recall equal to $96.68 \pm 1.34\%$, the vast majority of deepfakes were also classified as genuine, with a precision of $62.19 \pm 0.30\%$, as seen in Figure 7.1.

Both datasets use deepfake and genuine samples with the voices of the same speakers. The usefulness of pairing the genuine and deepfake samples by speakers was not confirmed or disproved yet, leaving room for future research. The proposed dataset could be also enhanced by generating the deepfake samples with additional tools.

Evaluation with MLAAD

Testing the multilingual detectors with MLAAD [36] did not show a correlation between the ratio of VITS recordings to the total amount of recordings of a language set, nor higher accuracy for known languages. There were significant differences between the LCNN and RawNet3 models for the same test set: in the case of 1,000 Czech utterances generated by VITS, the accuracy was $0.07 \pm 0.09\%$, and $86.00 \pm 6.88\%$ for LCNN and RawNet3 respectively.

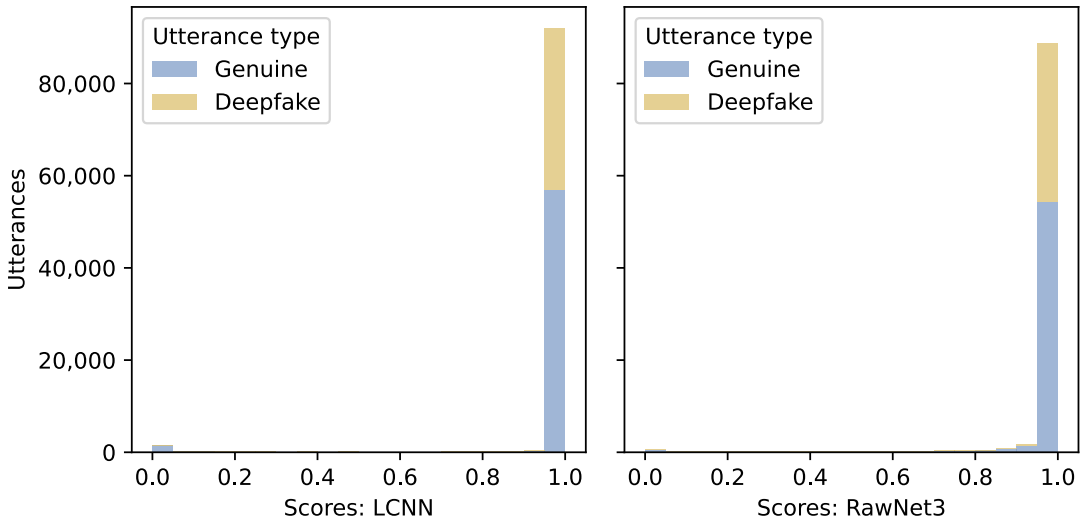

Figure 7.1: English detectors tested on In-the-wild, 3 runs. 0 for predicted deepfake recordings, and 1 for predicted genuine recordings.

Moreover, the two very similar languages, Czech and Slovak, both with test sets composed of 1,000 clips generated by VITS trained on the Common Voice Dataset [2], were classified differently: the classification accuracy of Slovak utterances was $59.17 \pm 16.96\%$, and $71.23 \pm 18.68\%$ for LCNN and RawNet3 respectively.

Chapter 8

Conclusion

This thesis presented the topic of speech deepfakes and the state of current research, including gender bias and language influence, followed by a brief description of ways how to generate them and available tools. The corpora needed to train these tools and create deepfake datasets were also introduced.

The analysis of existing audio deepfake datasets revealed several of their shortcomings: their unavailability, insufficient size, lack of equal gender representation, lack of language diversity, and others. It was concluded that to continue in the current research direction, a new dataset is needed. This dataset was designed and compiled with the revealed weaknesses and, most importantly, the following questions in mind: is there an inherent gender bias in audio deepfake detection? Does the language of the recording influence the detection ability of the detector? Can similar results be reached if instead of training a separate detector for each language, a multilingual detector is trained?

The dataset compilation process included extracting speech segments from an existing dataset, training open-source synthesis tools, inference, and creating protocols with the metadata. The dataset was then used to train and evaluate two state-of-the-art detection tools. The experiments revealed that the performance of a multilingual detector was statistically worse, although only by a few percent – a similar accuracy of monolingual LCNN detectors tested with languages different from the training one.

Concerning the gender bias, it seems that predicting deepfake of female voices is easier: this does however not work for all dataset subsets. Finally, a brief evaluation of the detector using another dataset was performed. Using a challenging dataset formed from collected real-life deepfakes showed the detector's inability to identify the deepfake utterances. Testing the detectors with a recent multilingual dataset showed that for some unknown languages, the accuracy was high, however, the results were not predictable and often differed among the two tested architectures. To conclude, this thesis points out that gender bias and language influence should not be an underestimated factor in audio deepfake detection, as well as the inability to generalize, and should be further investigated.

Bibliography

[1] Altuncu, E.; Franqueira, V. N. and Li, S. Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. *ArXiv preprint arXiv:2208.10913*, 2022.

[2] Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J. et al. Common Voice: A Massively-Multilingual Speech Corpus. In: Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C. et al., ed. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, p. 4218–4222. ISBN 979-10-95546-34-4. Available at: https://aclanthology.org/2020.lrec-1.520.

[3] Ba, Z.; Wen, Q.; Cheng, P.; Wang, Y.; Lin, F. et al. Transferring Audio Deepfake Detection Capability across Languages. In: *Proceedings of the ACM Web Conference 2023*. 2023, p. 2033–2044.

[4] Bilika, D.; Michopoulou, N.; Alepis, E. and Patsakis, C. Hello Me, Meet the Real Me: Audio Deepfake Attacks on Voice Assistants. *CoRR*, 2023, abs/2302.10328. Available at: https://doi.org/10.48550/arXiv.2302.10328.

[5] Bishop, C. M. and Bishop, H. Diffusion Models. In: *Deep Learning: Foundations and Concepts*. Cham: Springer International Publishing, 2024, p. 581–607. ISBN 978-3-031-45468-4. Available at: https://doi.org/10.1007/978-3-031-45468-4_20.

[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: IEEE. *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] Bäckström, T.; Räsänen, O.; Zewoudie, A.; Zarazaga, P. P.; Koivusalo, L. et al. *Introduction to Speech Processing*. 2nd ed. 2022. Available at: https://speechprocessingbook.aalto.fi.

[8] Cai, Z.; Ghosh, S.; Adatia, A. P.; Hayat, M.; Dhall, A. et al. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset. *ArXiv preprint arXiv:2311.15308*, 2023.

[9] Cai, Z.; Stefanov, K.; Dhall, A. and Hayat, M. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In: IEEE. *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2022, p. 1–10.

[10] 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. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024, vol. 38, no. 16, p. 17862–17870.

[11] Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R. et al. The deepfake detection challenge (dfdc) dataset. *ArXiv preprint arXiv:2006.07397*, 2020.

[12] Du, J.; Na, X.; Liu, X. and Bu, H. Aishell-2: Transforming mandarin asr research into industrial scale. *ArXiv preprint arXiv:1808.10583*, 2018.

[13] Firc, A. and Malinka, K. The Dawn of a Text-Dependent Society: Deepfakes as a Threat to Speech Verification Systems. In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. New York, NY, USA: Association for Computing Machinery, 2022, p. 1646–1655. SAC '22. ISBN 9781450387132. Available at: https://doi.org/10.1145/3477314.3507013.

[14] Firc, A.; Malinka, K. and Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*. Elsevier, 2023.

[15] Firc, A.; Malinka, K. and Hanáček, P. Deepfake Speech Detection: A Spectrogram Analysis. In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. New York, NY, USA: Association for Computing Machinery, 2024, p. 1312–1320. SAC '24. ISBN 9798400702433. Available at: https://doi.org/10.1145/3605098.3635911.

[16] Frank, J. and Schönherr, L. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2021.

[17] 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: *Interspeech*. 2021.

[18] Gales, M. J.; Knill, K. M.; Ragni, A. and Rath, S. P. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In: International Speech Communication Association (ISCA). *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*. 2014, p. 16–23.

[19] Gong, C.; Wang, X.; Cooper, E.; Wells, D.; Wang, L. et al. ZMM-TTS: Zero-shot Multilingual and Multispeaker Speech Synthesis Conditioned on Self-supervised Discrete Speech Representations. *ArXiv e-prints*, 2023, p. arXiv–2312.

[20] Ito, K. and Johnson, L. *The LJ Speech Dataset* online. 2017. Available at: https://keithito.com/LJ-Speech-Dataset/. [cit. 2023-12-25].

[21] Jung, J. weon; Heo, H.-S.; Kim, J. ho; Shim, H. jin and Yu, H.-J. RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification. In: *Proc. Interspeech 2019*. 2019, p. 1268–1272. ISSN 2958-1796.

[22] Kang, W.; Hasegawa Johnson, M. and Roy, D. End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions. In: *Proc. INTERSPEECH 2023*. 2023, p. 2303–2307. ISSN 2958-1796.

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

[24] Khalid, H.; Tariq, S.; Kim, M. and Woo, S. S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *ArXiv preprint arXiv:2108.05080*, 2021.

[25] Kim, J.; Kong, J. and Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: PMLR. *International Conference on Machine Learning*. 2021, p. 5530–5540.

[26] Köhn, A.; Stegen, F. and Baumann, T. Mining the Spoken Wikipedia for Speech Data and Beyond. In: Chair), N. C. C.; Choukri, K.; Declerck, T.; Grobelnik, M.; Maegaard, B. et al., ed. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), May 2016. ISBN 978-2-9517408-9-1.

[27] Kominek, J. and Black, A. W. The CMU Arctic speech databases. In: *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*. 2004, p. 223–224.

[28] Korshunov, P.; Chen, H.; Garner, P. N. and Marcel, S. Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes. In: *IEEE International Joint Conference on Biometrics (IJCB)*. September 2023.

[29] Li, J.; Tu, W. and Xiao, L. Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, p. 1–5.

[30] 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*. IEEE, 2023.

[31] Liz López, H.; Keita, M.; Taleb Ahmed, A.; Hadid, A.; Huertas Tato, J. et al. Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Information Fusion*, 2024, vol. 103, p. 102103.

[32] Lorenzo Trueba, J.; Yamagishi, J.; Toda, T.; Saito, D.; Villavicencio, F. et al. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. In: ISCA. *The Speaker and Language Recognition Workshop (Odyssey 2018)*. 2018, p. 195.

[33] Ma, H.; Yi, J.; Wang, C.; Yan, X.; Tao, J. et al. FAD: A Chinese dataset for fake audio detection. *ArXiv preprint arXiv:2207.12308*, 2022.

[34] 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*. IEEE, 2014, vol. 22, no. 8, p. 1006–1010.

[35] Müller, N.; Czempin, P.; Diekmann, F.; Froghyar, A. and Böttinger, K. Does Audio Deepfake Detection Generalize? In: *Proc. Interspeech 2022*. 2022, p. 2783–2787.

[36] Müller, N. M.; Kawa, P.; Choong, W. H.; Casanova, E.; Gölge, E. et al. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. *ArXiv preprint arXiv:2401.09512*, 2024.

[37] Müller, T. and Kreutz, D. *ThorstenVoice Dataset 2022.10*. Zenodo, october 2022. Available at: https://doi.org/10.5281/zenodo.7265581.

[38] Nadimpalli, A. V. and Rattani, A. GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection. In: *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges: Montreal, QC, Canada, August 21–25, 2022, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 320–337. ISBN 978-3-031-37741-9. Available at: https://doi.org/10.1007/978-3-031-37742-6_25.

[39] Oliveira, F. S.; Casanova, E.; Junior, A. C.; Soares, A. S. and Filho, A. R. Galvão. CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages. In: *Text, Speech, and Dialogue: 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 188–199. ISBN 978-3-031-40497-9. Available at: https://doi.org/10.1007/978-3-031-40498-6_17.

[40] Paris, B. and Donovan, J. *Deepfakes and cheap fakes* online. Data & Society Research Institute, 2019. Available at: https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf. [cit. 2024-06-23].

[41] Park, K. and Mulc, T. CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages. *Interspeech*, 2019.

[42] 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.

[43] Rana, M. S.; Nobi, M. N.; Murali, B. and Sung, A. H. Deepfake detection: A systematic literature review. *IEEE access*. IEEE, 2022, vol. 10, p. 25494–25513.

[44] Reimao, R. and Tzerpos, V. For: A dataset for synthetic speech detection. In: IEEE. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019, p. 1–10.

[45] 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.

[46] Shi, Y.; Bu, H.; Xu, X.; Zhang, S. and Li, M. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *ArXiv preprint arXiv:2010.11567*, 2020.

[47] Sisman, B.; Yamagishi, J.; King, S. and Li, H. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, p. 132–157.

[48] Solak, I. *The m-ailabs speech dataset* online. Munich Artificial Intelligence Laboratories GmbH, january 2019. Available at: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/. [cit. 2023-12-25].

[49] Sonobe, R.; Takamichi, S. and Saruwatari, H. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis. *ArXiv preprint arXiv:1711.00354*, 2017.

[50] Stan, A.; Watts, O.; Mamiya, Y.; Giurgiu, M.; Clark, R. A. et al. TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. In: *INTERSPEECH*. 2013, p. 2331–2335.

[51] Tan, X.; Qin, T.; Soong, F. and Liu, T.-Y. A survey on neural speech synthesis. *ArXiv preprint arXiv:2106.15561*, 2021.

[52] Toda, T.; Chen, L.-H.; Saito, D.; Villavicencio, F.; Wester, M. et al. The Voice Conversion Challenge 2016. In: International Speech Communication Association. *Interspeech 2016*. 2016, p. 1632–1636.

[53] Trinh, L. and Liu, Y. An Examination of Fairness of AI Models for Deepfake Detection. In: *International Joint Conference on Artificial Intelligence*. 2021. Available at: https://api.semanticscholar.org/CorpusID:233481637.

[54] Valk, J. and Alumäe, T. VoxLingua107: a dataset for spoken language recognition. In: IEEE. *2021 IEEE Spoken Language Technology Workshop (SLT)*. 2021, p. 652–658.

[55] Veaux, C.; Yamagishi, J.; MacDonald, K. et al. *CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit* online. University of Edinburgh. The Centre for Speech Technology Research (CSTR), november 2019. Available at: https://datashare.ed.ac.uk/handle/10283/3443. [cit. 2023-12-25].

[56] Wang, D. and Zhang, X. Thchs-30: A free chinese speech corpus. *ArXiv preprint arXiv:1512.01882*, 2015.

[57] Wang, R.; Juefei Xu, F.; Huang, Y.; Guo, Q.; Xie, X. et al. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices. In: *Proceedings of the 28th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1207–1216. MM '20. ISBN 9781450379885. Available at: https://doi.org/10.1145/3394171.3413716.

[58] Wang, X. and Yamagishi, J. Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders. In: IEEE. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, p. 1–5.

[59] 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. *Computer Speech & Language*. Elsevier, 2020, vol. 64, p. 101114.

[60] Wenger, E.; Bronckers, M.; Cianfarani, C.; Cryan, J.; Sha, A. et al. „ Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World.

In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, p. 235–251.

[61] Xie, Y.; Zhou, J.; Lu, X.; Jiang, Z.; Yang, Y. et al. FSD: An initial chinese dataset for fake song detection. *ArXiv preprint arXiv:2309.02232*, 2023.

[62] Yan, X.; Yi, J.; Tao, J.; Wang, C.; Ma, H. et al. System fingerprints detection for deepfake audio: An initial dataset and investigation. *ArXiv preprint arXiv:2208.10489*, 2022.

[63] Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z. et al. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security*. IEEE, 2023, vol. 18, p. 2015–2029.

[64] Yi, J.; Bai, Y.; Tao, J.; Ma, H.; Tian, Z. et al. Half-Truth: A Partially Fake Audio Detection Dataset. In: *Proc. Interspeech 2021*. 2021, p. 1654–1658.

[65] Yi, J.; Fu, R.; Tao, J.; Nie, S.; Ma, H. et al. Add 2022: the first audio deep synthesis detection challenge. In: IEEE. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, p. 9216–9220.

[66] Yi, J.; Tao, J.; Fu, R.; Yan, X.; Wang, C. et al. ADD 2023: the Second Audio Deepfake Detection Challenge. *ArXiv preprint arXiv:2305.13774*, 2023.

[67] Yi, Z.; Huang, W.-C.; Tian, X.; Yamagishi, J.; Das, R. K. et al. Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —. In: *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. 2020, p. 80–98.

[68] Zang, Y.; Zhang, Y.; Heydari, M. and Duan, Z. SingFake: Singing Voice Deepfake Detection. *ArXiv preprint arXiv:2309.07525*, 2023.

[69] Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J. et al. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In: *Proc. Interspeech 2019*. 2019, p. 1526–1530.

[70] Zhang, L.; Wang, X.; Cooper, E.; Evans, N. and Yamagishi, J. The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol. 31, p. 813–825.

[71] 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.

[72] Zhang, Z.; Gu, Y.; Yi, X. and Zhao, X. FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection. In: Springer Nature. *Digital Forensics and Watermarking: 20th International Workshop, IWDW 2021, Beijing, China, November 20–22, 2021, Revised Selected Papers*. 2022, vol. 13180, p. 117.