BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

CREATING NOVEL DEEPFAKE SPEECH DATASET

TVORBA NOVÉ DEEPFAKE DATOVÉ SADY

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **MAROŠ SZTOLARIK**
AUTOR PRÁCE

SUPERVISOR **Ing. ANTON FIRC**
VEDOUCÍ PRÁCE

BRNO 2024

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

153353

Institut:	Department of Intelligent Systems (DITS)
Student:	**Sztolarik Maroš**
Programme:	Information Technology
Title:	**Creating Novel Deepfake Speech Dataset**
Category:	Security
Academic year:	2023/24

Assignment:

1. Learn about the technology of diffusion models and their use for creating deepfakes. Focus on speech synthesis.
2. Study available datasets containing speech deepfakes and describe their structure, content and usability.
3. Design the structure and content of a new dataset containing deepfake recordings created by diffusion models. Utilize at least two speech synthesis tools.
4. Create the proposed dataset.
5. Use at least two state of the art deepfake speech detection solutions to investigate how diffusion model technology impacts the quality and accuracy of detection.
6. Discuss the possible implications of using diffusion models for creating speech deepfakes on the development of new speech deepfakes detection tools and the applicability of the created dataset for further research in the area of speech deepfakes and their detection.

Literature:

- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M. (2022).Diffusion Models: A Comprehensive Survey of Methods and Applications. *ArXiv* ./abs/2209.00796
- Jeong, M., Kim, H., Cheon, S. J., Choi, B. J., & Kim, N. S. (2021). Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. *ArXiv*. /abs/2104.01409
- Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2022). DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 11020-11028. https://doi.org/10.1609/aaai.v36i10.21350
- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:8599-8608, 2021.

Requirements for the semestral defence:
1 - 3

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

Supervisor:	**Firc Anton, Ing.**
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2023
Submission deadline:	9.5.2024
Approval date:	6.11.2023

Abstract

In the recent years, deepfake technology has advanced to a point where it can convincingly mimic human speech, posing significant challenges in distinguishing between real and synthetic voices.

In this thesis, we introduce a novel dataset comprising speech deepfakes generated using diffusion models. This dataset, created with two sophisticated text-to-speech tools, DiffSpeech and ProDiff, aims to provide insight into the threat that these new tools pose. Two more datasets are created with more mature tools, Glow-TTS and Tacotron2, to provide a point of comparison. Then all the generated samples are analyzed through two deepfake detectors in order to provide a direct comparison into how much of a threat each tool is to these detectors.

The results show that even though the tools utilizing the diffusion models are threatening, the use of diffusion models did not provide these tools any meaningful advantage in evading the detection.

Abstrakt

V posledných rokoch deepfake technológia postúpila do bodu kedy je schopná uveriteľne napodobniť ľudský hlas, čím predstavuje významné výzvy v rozslišovaní medzi skutočnými a syntetickými hlasmi.

V tejto práci predstavujeme novú dátovú sadu obsahujúcu deepfake reč generovanú pomocou difúznych modelov. Táto dátová sada, vytvorená s pomocou dvoch sofistikovaných nástrojov pre prevod textu na reč, DiffSpeech a ProDiff, mieri poskytnúť náhľad do hrozby tieto nové nástroje predstavujú. Dve ďalšie dátové sady sú vytvorené s viac vyspelými nástrojmi pre poskytnutie bodu porovnania. Potom sú všetky vygenerované vzorky analyzované dvomi deepfake detektormi pre priame porovnanie akú veľkú hrozbu každý nástroj predstavuje.

Výsledky ukazujú, že aj keď nástroje ktoré využívajú difúzne modely predstavujú hrozbu, použitie difúznych modelov neposkytlo týmto nástrojom nijakú významnú výhodu vo vyhýbaní sa detekcii.

Keywords

Diffusion Models, Deepfake, Machine learning, Text-to-Speech, LJSpeech

Klíčová slova

Difúzne modely, Deepfake, strojové učenie, prevod textu na reč, LJSpeech

Reference

SZTOLARIK, Maroš. *Creating Novel Deepfake Speech Dataset.* Brno, 2024. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Anton Firc

Rozšířený abstrakt

V posledných rokoch sa umelá inteligencia dostala do skoro všetkých odvetví. Tieto sofistikované nástroje vytvorené pomocou strojového učenia nám poskytujú veľa spôsobov ako si uľahčiť život, prácu alebo sa zabaviť. Schopnosť generovať video, obrázky, a zvuk je nielen nesmierne užitočná ale v nesprávnych rukách aj nebezpečná, napríklad vo forme deepfakes.

Deepfaky sú syntetické média vytvorené umelou inteligenciou a to konkrétne hlbokým strojovým učením. Využívajú hlas alebo vzhľad reálnych ľudí na zobrazenie niečoho čo v skutočnosti neexistuje. Deepfaky v dnešnej dobe dosahujú kvalitu kedy ľudia nedokážu s istotou určiť, čo je deepfake a čo nie. To predstavuje hrozbu rozširovania dezinformácií. Sociálne siete nie sú pripravené na hrozbu takejto veľkosti. Preto je dôležité vyvíjať nástroje schopné odhaliť nie len obrazové, ale aj hlasové deepfakes s čo najvyššou presnosťou.

Táto práca je zameraná na novú technológiu - difúzne modely. Difúzne modely sú momentálne využívané hlavne pri generovaní obrázkov, ale používajú sa aj pri generovaní zvuku. Konkrétne sa snažíme zistiť aký vplyv majú difúzne modely na detekciu hlasových deepfakes.

Cieľom tejto práce je vytvoriť novú dátovu sadu s použitím dvoch nástrojov, ktoré využívajú difúzne modely. Konkrétne text-to-speech nástroje, ktoré konvertujú text na hovorený hlas, DiffSpeech a ProDiff. Táto dátová sada obsahuje 13 820 nahrávok vygenerovaných každým nástrojom. Teda dokopy 27 640 vygenerovaných náhrávok, ktoré majú dokopy približne 48 hodín. Pre preskúmanie efektu generovania viet ktoré boli použité pri trénovaní nástroja a generovania nových viet sa generuje 13 100 nahrávok s použitím trénovaných viet ako vstupný text a 720 nahrávok s použitím nových viet ako vstupný text.

Pre porovnanie toho aký efekt majú difúzne modely na detekciu deepfakes je potreba vytvoriť ďalšie dátové sady za pomoci dvoch uznávaných nástrojov z minulosti, Glow-TTS a Tacotron2. Pre priame porovnanie medzi týmito 4 nástrojmi, generujú všetky nástroje nahrávky s rovnakým vstupným textom.

Pre zhodnotenie ako náročné je odhaliť či sú tieto náhravky sfalšované alebo skutočné sú využité dva nástroje na detekciu hlasových deepfakes, SSL Anti-spoof a AASIST. Tieto nástroje udelia každej nahrávke skóre ktoré určuje v akej miere si detektor myslí, že je nahrávka falošná. S týmto skóre sa dajú následne dajú zistiť 3 metriky, False Match Rate, False Non-Match Rate a Equal Error Rate.

False Match Rate (FMR) je metrika používaná v biometrických systémoch. Táto metrika určuje pravdepodobnosť, že systém nesprávne interpretuje sfalšovanú nahrávku ako pravú.

False Non-Match Rate (FNMR) je ďalšia metrika používaná v biometrických systémoch, ktorá určuje pravdepodobnosť, že systém nesprávne interpretuje pravú nahrávku za sfalšovanú.

Equal Error Rate (EER) je metrika, ktorá určuje bod kedy sa FMR a FNMR rovnajú. Je to prah na ktorom je šanca interpretovať sfalšovanú nahrávku ako pravú a šanca interpretovať pravú nahrávku ako falošnú rovnaká. EER je bežne používana pre porovnávanie presnosti a spoľahlivosti biometrických systémov, s nižším EER indikujúcim presnejší systém.

Cieľom experimentu bolo teda vytvoriť 13 820 deepfakes s každým nástrojom a následne zanalyzovať všetky deepfakes pomocou deepfake detektorov. Takýmto spôsobom sa dajú ekvivalentne porovnať všetky štyri nástroje a dá sa porovnať aj rozdiel medzi trénovanými vetami a novými vetami.

Výsledky z detektoru SSL Anti-spoofing sa nachádzajú v tabuľke 1. Prekvapivo, detektor najlepšie oklamal Tacotron2, ktorý je z vybraných nástrojov najstarší, až z roku

2017. Predpokladane, generovanie viet na ktorých bol nástroj trénovaný bolo výrazne efektívnejšie vo vyhýbaní sa detekcii ako nové vety. Pri tomto detektori dopadli nástroje ktoré nevyužívajú difúzne modely o trochu lepšie ako nástroje ktoré ich využívajú.

Table 1: Výsledky z SSL Anti-spoofing

	DiffSpeech	ProDiff	Glow-TTS	Tacotron2
EER-Trénované vety [%]	27.88	14.94	15.81	29.81
ERR-Nové vety [%]	11.94	4.02	10.7	19.85

Výsledky z AASIST, zobrazené v tabuľke 2 ukázali, že nie vždy sú nové vety menej efektívne ako trénované vety kedže v 2 prípadoch pri nových vetách stúpol EER. Proti DiffSpeechu je AASIST kompletne bezmocný keďže EER dosiahlo 50 percent, čo je rovnaká úspešnosť ako náhodné určovanie či je nahrávka autentická alebo sfalšovaná. Avšak ani proti ostatným nástrojom nebol AASIST veľmi efektívny, najnižšie EER bolo 26.28 percent, čo je v praktickom použití jednoducho príliš vysoké a teda neakceptovateľné. Ale ani pri tomto detektore nebolo vidieť vyznámný rozdiel medzi nástrojmi ktoré využívajú difúzne modely a ktoré nie.

Table 2: Výsledky z AASIST

	DiffSpeech	ProDiff	Glow-TTS	Tacotron2
EER-Trénované vety [%]	53.99	32.07	38.66	30.86
ERR-Nové vety [%]	50.41	26.28	47.08	36.38

Takže difúzne modely nepredstavujú vo sfére hlasových deepfakes väčšiu hrozbu ako ostatné spôsoby vytvárania deepfakov. Zistili sme, že vo večšine prípadov má generovanie nových viet značný negatívny vplyv na detekciu, avšak prekvapivo nie vo všetkých prípadoch. Dozvedeli sme sa aj, že deepfake detektory nie sú dostatočne robustné, keďže AASIST bez špecifického tréningu proti deepfakom vytvoreným v tejto práci bol v niektorých prípadoch úplne nepoužiteľný.

Creating Novel Deepfake Speech Dataset

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Ing. Anton Firc.

I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

.....
Maroš Sztolarik
May 6, 2024

Acknowledgements

I would like to thank my supervisor for his help and patience. Consultations with you helped me immensely and always guided me in the right direction.

I would also like to thank my girlfriend who has motivated me to finish this work and helped me through it.

I would also like to thank my parents who have never doubted me and support me in my studies.

Contents

1 Introduction **4**

2 Deepfakes **6**
2.1 Deepfakes and their creation 6
2.2 Speech Deepfakes 9
2.3 TTS models 11

3 Diffusion Models **12**
3.1 Definition 12
3.2 Principles 12
3.3 Training 12
3.4 Architectures of Diffusion Models 13
3.5 Applications of Diffusion Models 13
3.6 Comparison with other generative models 14
3.7 TTS Models incorporating Diffusion Models 15

4 Datasets **17**
4.1 What are datasets 17
4.2 Datasets suitable for training TTS 18
4.3 Datasets with synthesized speech 19

5 Deepfake detection **23**
5.1 Need for detection software 23
5.2 Audio spoofing detection tools 25
5.3 Performance evaluation metrics 26

6 Experiment **27**
6.1 Experiment design 27
6.2 Experiment overview 28
6.3 Synthetizing datasets 29
6.4 Working with Audio Spoof Detectors 31
6.5 SSL Anti-Spoofing results 32
6.6 AASIST results 36
6.7 Results evaluation 40

7 Conclusion **42**

Bibliography **43**

A Contents of the included storage media **46**

List of Figures

2.1 Generative Adversarial Network diagram [15] 8
2.2 Linear and polynomial regression on the same data points [26] 9
2.3 Simlified TTS visualization [10] 10
2.4 Voice Conversion flow chart. Pink box represents training of the Mapping function. [34] 11

3.1 Forward diffusion and reverse denoising process[9] 13

4.1 Samples from the FakeAVCeleb dataset [22] 21

5.1 Screenshot from the livestream scam featuring Elon Musk [29] 24
5.2 Screenshot from the TikTok advertisement featuring deepfake of MrBeast [30] 24

6.1 DiffSpeech trained senteces results from SSL Anti-Spoofing 33
6.2 DiffSpeech novel sentences results from SSL Anti-Spoofing 33
6.3 ProDiff trained sentences results from SSL Anti-Spoofing 34
6.4 ProDiff novel sentences results from SSL Anti-Spoofing 34
6.5 Glow-TTS trained sentences results from SSL Anti-Spoofing 34
6.6 Glow-TTS novel sentences results from SSL Anti-Spoofing 35
6.7 Tacotron2 trained sentences results from SSL Anti-Spoofing 35
6.8 Tacotron2 novel sentences results from SSL Anti-Spoofing 36
6.9 DiffSpeech trained sentences results from AASIST 37
6.10 DiffSpeech novel sentences results from AASIST 37
6.11 ProDiff trained sentences results from AASIST 37
6.12 ProDiff novel sentences results from AASIST 38
6.13 Glow-TTS trained sentences results from AASIST 38
6.14 Glow-TTS novel sentences results from AASIST 39
6.15 Tacotron2 trained sentences results from AASIST 39
6.16 Tacotron2 novel sentences results from AASIST 40

Chapter 1

Introduction

Technological advancement is progressing at an unprecedented pace, and one of its most striking manifestations is the ability of computers to mimic human speech with remarkable resemblance. This capability, holds immense potential for innovation, but also poses significant security threats.

The number of a deepfake identity fraud cases has increased by 1 740 percent in North America and 780 percent in Europe from 2022 to 2023 according to research by Sumsub[35]. That is an alarmingly high rate of growth. The risk of people using deepfakes to commit identity fraud, spreading misinformation or possibly bypassing biometric systems is substantial. As the creation of deepfakes becomes more accessible, it will likely worsen. This warrants the need for development of deepfake detectors to fight back against this threat.

The aim of this thesis is to develop a novel dataset composed of speech deepfakes generated by a text-to-speech tools which utilize diffusion models. This dataset is instrumental in assessing the extent of the potential threat that the diffusion models pose. The dataset is analyzed through two modern deepfake detectors in order to see how the deepfake detectors fare against the speech deepfakes generated with the use of diffusion models. This highlights the limitations and capabilities of current detection methods.

The 2. chapter delves into what the deepfakes are. Exploring their definition and how are they generated with special attention into how words are transformed into speech. Artificial intelligence has a wide array of applications, some of which are mentioned in this chapter.

In the 3. chapter shows a relatively novel approach in the deepfake domain - Diffusion Models. These models have shown remarkable ability in generating lifelike images and voices. We discuss how they operate and compare them with more established approaches in the field. Apart from image and voice generation, the applications of Diffusion Models are vast and we examine these potential uses. We explore the currently available text-to-speech tools using diffusion models and tools using different approaches.

For the generation of deepfakes, a varied dataset is crucial. The 4. chapter examines existing datasets, including those voiced by real people and those synthesized by other means. A dataset is proposed, incorporating both authentic and generated audio, to create a dataset that can be used to evaluate the threat to the detection of deepfakes.

The 5. chapter focuses on the deepfake detection tools that will be used on the newly created dataset along with datasets generated with tools that use different approaches in order to compare the results and draw conclusions whether the diffusion models pose a significant security risk.

The 6. chapter describes the experiment. The experiment is made up of two parts. The creation of the datasets and the application of deepfake detectors on these datasets. This chapter explores the essential data preparation required for generating and detecting deepfakes. It takes a closer look into making each tool work properly and adjusting it to work with different datasets. Then the results of these deepfake detectors are graphed to show how each dataset performed.

The experiment provided valuable insight into diffusion models and effect on detection of deepfakes created with them. At the end, results of each text-to-speech tool is evaluated and compared with the results of others.

Chapter 2

Deepfakes

Deepfakes have rapidly emerged as a significant concern. Deepfakes began gaining notoriety around year 2017, as advancements in artificial intelligence and machine learning made it feasible to generate convincing fake videos and audio recordings.

2.1 Deepfakes and their creation

Deepfakes are artificially created or manipulated media, typically video or audio, generated using artificial intelligence techniques, specifically deep learning algorithms. These deep learning models are trained to replace one person's likeness or voice with another, resulting in realistic, yet fabricated, media content. The name is a combination of deep learning and fake. This technology has gained attention for its potential use in creating misleading or false representation in various contexts, from entertainment to politics. Deepfakes pose significant challenges in discerning real media from manipulated ones, raising concerns about their implications for misinformation and privacy. [38]

2.1.1 Machine learning

Machine learning is a branch of artificial intelligence that focuses on building systems capable of learning from and making decisions based on data. Presence of massive amount of data necessitates automated methods for data analysis, and machine learning is at the forefront of this movement. Machine learning is characterized by its ability to autonomously identify patterns within large data sets. These identified patterns are then leveraged for predicting future trends, making decisions and handling uncertainties. The power of machine learning lies in its capacity to adapt and improve over time, making it an essential tool for navigating and interpreting the vast and complex landscape of modern data. [26]

2.1.2 Deep learning

Deep learning is a subset of machine learning in artificial intelligence that mimics the working of the human brain in processing data and creating patterns for use in decision making. It is a field of learning based on artificial neural networks, which are algorithms inspired by the structure and function of the brain. Deep learning network can learn to perform tasks by considering examples, generally without task-specific programming. [13]

At the core of deep learning is the neural network architecture. The „deep“ in deep learning refers to the number of layers through which the data is transformed. More layers

allow for more complex representations of data. Unlike shallow machine learning algorithms, deep networks can actually discover the features to be used for classification or regression. [13]

Deep learning models are used in industries from automated driving to healthcare. Deep learning is also used in the creation of deepfakes. [13]

2.1.3 Types of machine learning

There are three types of machine learning.

Supervised machine learning, otherwise known as predictive learning approach, involves training a model on a labeled dataset. In this approach, the model learns from input data that is explicitly tagged with the correct output, allowing the algorithm to understand the relationship between the input and the output. The goal is to enable the model to make accurate predictions or decisions when presented with new, unseen data. This type of learning is widely used in application where historical data predicts future events, such as image recognition and financial forecasting. [26]

Unsupervised machine learning, also known as descriptive learning, is a type of machine learning where the models are trained on that are not labeled. This means the model must discern patterns, structures or features within the data without guidance on the desired outcome. This is called knowledge discovery. Common applications include clustering, where model groups similar data points together, and dimensionality reduction, where the model simplifies data without losing critical information. Unlike supervised learning, unsupervised learning does not work towards predicting a specific output but rather focuses on exploring the underlying structure of the data. [26]

Reinforcement learning is less common than the other two, where models learn from consequences of it s actions. It receives rewards or penalties based on its actions, guiding it to learn the best strategies over time. It is similar to how humans learn from trial and error. [26]

2.1.4 Generative Adversarial Networks introduction

Generative Adversarial Networks are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks competing against each other in a game. These two networks, known as the generator and the discriminator, engage in a continuous dynamic. The generator creates new data instances, synthesizing them from random noise inputs, aiming to mimic the real data present in the training dataset. Simultaneously, the discriminator evaluates both the real data from the training set and the synthetic data produced by the generator, learning to discern between the two. Diagram displaying how GAN works can be seen in figure 2.1. [14]

2.1.5 Variational Autoencoders introduction

Variational Autoencoders (VAEs) are a sophisticated type of autoencoder used for generating complex models. Unlike traditional autoencoders that simply compress and reconstruct data, VAEs introduce a probabilistic twist. They convert input data into a distribution of possible values in a latent space, typically a Gaussian distribution. This allows for more diverse and generalized data creation. VAEs use a technique known as the reparameterization trick to ensure that the model remains differentiable and thus trainable via backpropagation. The training process for VAEs not only focuses on reconstruction accuracy but also on

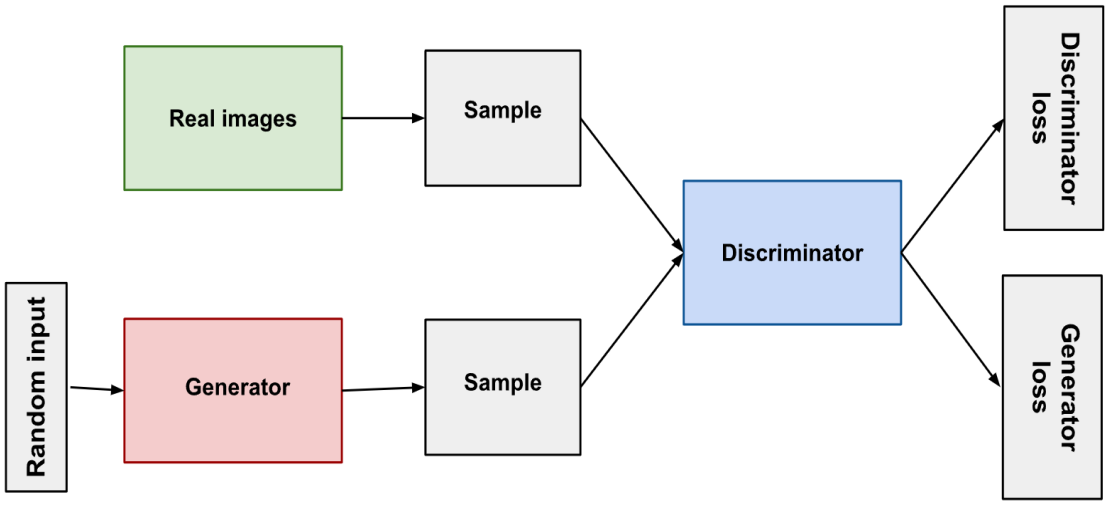


Figure 2.1: Generative Adversarial Network diagram [15]

smoothing and regularizing the latent space to enable the generation of new, coherent samples. This dual objective is achieved by adding a regularization term to the loss function, which encourages the latent space to mimic a predefined distribution, helping to prevent overfitting and ensuring a continuous, connected latent space. [8]

2.1.6 Classification

Classification is a core task in machine learning that involves predicting the category or class of a given input data point. It falls under the category of supervised learning, where the model is trained on a labeled dataset containing input-output pairs. The goal of classification is to accurately predict the output class for each input data point by learning from the training data. [26]

2.1.7 Regression

Regression is another fundamental machine learning method used for predicting continuous outcome variable based on one or more predictor variables. Regression also falls under the category of supervised learning with model being trained on a labeled dataset. The aim of regression analysis is to find the relationship between the input variables (independent variables) and the output variable(dependent variable) which allows predictions for new data points. [26]

The simplest and most widely form used of regression is linear regression, where the relationship between the input variables and the output variable is assumed to be linear. Polynomial regression is a form of regression analysis in which the relationship between the independent variable and the dependent variable is modeled as an nth degree polynomial. Polynomial regression is useful when the relations between independent and dependent variable is not linear. By adjusting the degree of the polynomial, you can find a curve that fits the data better. In the Figure 2.2 on the left is linear regression on 1d data and on the left there is polynomial regression on the same data. However, there are many types of

regression analysis, each with its own specific model to handle different types of data and relationships. [26]

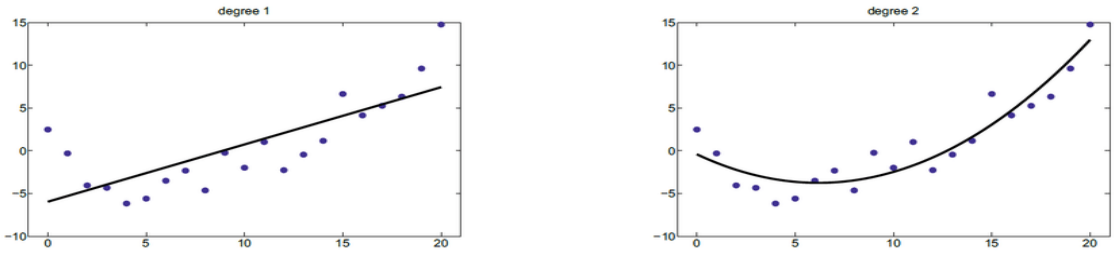


Figure 2.2: Linear and polynomial regression on the same data points [26]

2.1.8 Clustering

Clustering is a type of unsupervised learning method that involves grouping a set of objects in such a way that objects in the same group are more similar to each other that to those in other groups. It's a method of identifying similar instances and grouping them together without prior knowledge of the group labels. The goal of clustering is to discover underlying patterns in the data. Most commonly used clustering alghorithm is K-means. [26]

2.2 Speech Deepfakes

A speech deepfake is a type of synthetic media where a person's voice is artificially generated or manipulated using artificial intelligence and machine learning techniques. This technology enables the creation of audio recordings that sound like a particular individual, even though that person never said the words in the recording.

Text-to-Speech (TTS) is a subclass of speech synthesis in artificial intelligence and computer linguistics, focusing on the conversion of written text into spoken words.

2.2.1 Text-to-Speech

Deep learning-based speech synthesis systems typically involve two main components:

- **TTS Model:** The TTS model's primary function is to convert input text into an intermediate acoustic representation, often a mel-spectrogram. A mel-spectrogram is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time and is widely used because it closely approximates human auditory system responses. [20]
- **Vocoder:** The Vocoder takes the acoustic representation from the TTS model and synthesizes the actual audible speech waveform. It essentially reconstructs the audio from the mel-spectrogram, trying to produce natural-sounding speech. Advanced vocoders aim to generate speech that closely mimic human voice quality, including natural variations and subtleties. [20]

This process can be seen in the figure 2.3.

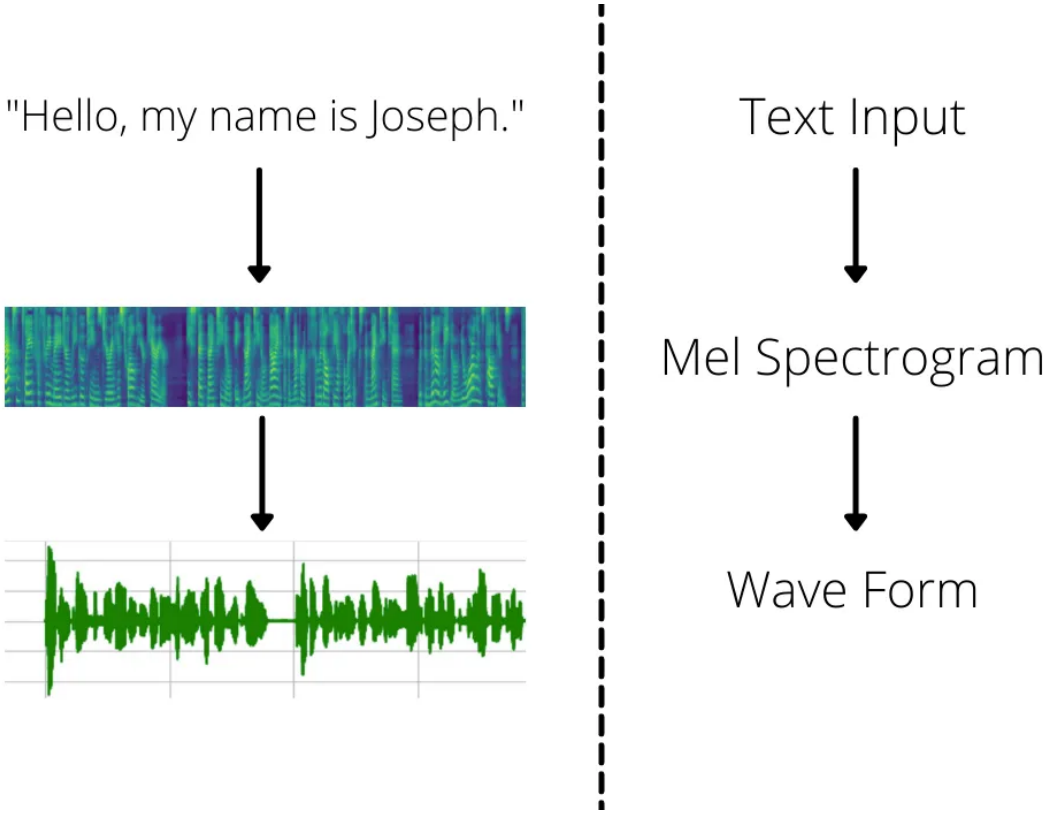


Figure 2.3: Simlified TTS visualization [10]

The ultimate goal is to produce a TTS system that not only accurately pronounces words and phrases but also captures the emotional and expressive qualities of natural speech.

Modern TTS models are generally divided into two categories based on their formulation:

- **Autoregressive (AR) Models:** AR models, are known for gererating high-quality speech samples. They achieve this by decomposing the output distribution into a sequence of conditional distributions. One of the main strengths of AR models is their ability to produce natural-sounding speech. However, they have notable limitation: the time it takes to infer increases linearly with the length of the mel-spectrograms. Additionally, AR models something struggle with robustness, exhibiting issues like word skipping or repetition due to cumulative prediction errors. [20]
- **Non-Autoregressive (Non-AR) Models:** Non-AR models offer more stable speech synthesis and significantly faster inference compared to AR models. However, these models also have their drawbacks. One of the key limitations of non-AR models is their lack of diversity in synthetic speech. This limitation arises because these feed-forward models are optimized using a simple regression objective function and do not incorporate probabilistic modeling, which restricts their ability to produce varied speech outputs. [20]

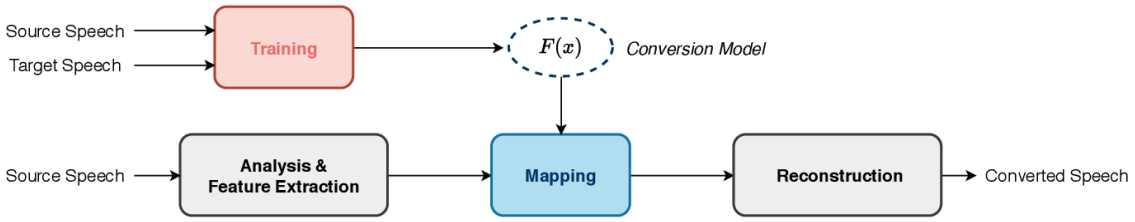


Figure 2.4: Voice Conversion flow chart. Pink box represents training of the Mapping function. [34]

2.2.2 Voice conversion

Voice conversion, is an important aspect of artificial intelligence. Voice conversion is a technology that involves altering a source speaker's voice to sound like that of a target speaker. This process typically involves capturing the unique vocal attributes of the target speaker, such as pitch, tone and speaking style, then applying them to the source speaker's voice. The result is a synthesized voice that maintains the content and language of the source speaker's speech but adopts the acoustic characteristics of the target speaker. Typical Voice Conversion flow can be seen in figure 2.4. [34]

2.3 TTS models

This section describes TTS models from recent years that utilize various generative mechanisms to achieve high quality samples.

2.3.1 Glow-TTS

Glow-TTS utilizes flow-based architecture to model the conditional distribution of mel-spectrograms. This setup allows it to perform efficient, parallel synthesis of speech, bypassing the sequential generation limitations of traditional autoregressive models. This models is particularly notable for not requiring any external aligner to learn its alignments, which distinguishes it from many other TTS models that rely on pre-trained autoregressive models for alignment guidance. Thanks to it parallel generation capabilities, Glow-TTS achieves significant speed improvements over traditional TTS systems. [23]

2.3.2 Tacotron2

The frontend of Tacotron2 uses a sequence-to-sequence framework with attention. This part of the model takes a sequence of characters as input and outputs a mel-spectrogram. The backend is a WaveNet-based vocoder that synthesized time-domain waveforms from the mel-spectrograms produced by the sequence-to-sequence model. [33]

Chapter 3

Diffusion Models

3.1 Definition

Diffusion models are a family of generative models that have emerged in the past few years in the field of deep learning. They are used to generate similar data to those which they are trained on. These models operate by gradually transforming data from a simple form like Gaussian noise into complex data like high-resolution image or sophisticated waveform.

Gaussian noise, also referred to as white noise, is a type of random noise where every value(e.g., every pixel in an image, every point in a sound wave) follows a normal distribution also known as Gaussian distribution.

This process is inspired by the physical process of diffusion, which discribes how particles move from areas of higher concentration to lower concentration over time. [16]

3.2 Principles

As illustrated in the Figure 3.1, the working principle of a diffusion model involves two key phases, the forward process and the reverse process, both of which are a Markov chain.

Markov chain is a stochastic process, meaning it is a process of some values changing with random probability. Difference between normal stochastic process and Markov chain is that the probability of transitioning to the next state depends only on the current state and not on the previous events. This property is called a Markov property. [2]

In the forward process, illustrated from left to right, the model incrementally adds noise to the data until only random noise remains, effectively destroying the structure of the original data. The reverse process, depicted from right to left in the figure, then aims to reconstruct the original data from the noise. This reverse process is what is learned by the model through training. By doing so, the model learns the patterns and structures of the target data distrubution. [16]

3.3 Training

Training diffusion models involves optimizing the neural network to effectively perform the reverse process of the Markov chain. This is typically done using a variant of the variational autoencoder framework. The training aims to minimize the difference between the original data and the data generated by the reverse process, ensuring that the model can accurately reconstruct or generate realistic data. [16]

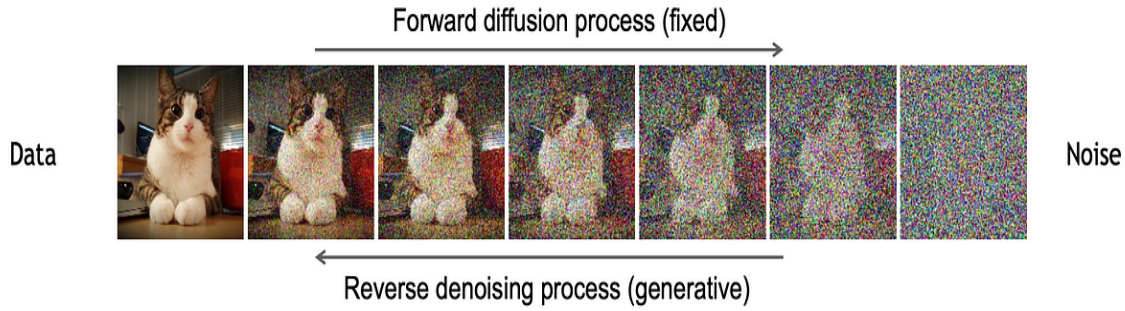


Figure 3.1: Forward diffusion and reverse denoising process[9]

3.4 Architectures of Diffusion Models

Diffusion Models can be implemented using various architectures. This section explores two important architectures used in these models.

3.4.1 Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPMs) operate with two Markov chains: a forward chain that progressively adds noise to data, and a reverse chain that transforms this noise back into data. The forward chain is custom-designed to convert any data distribution into a simpler form, typically a standard Gaussian distribution. The reverse chain, on the other hand, undoes this process using transition kernels learned by deep neural networks. Generation of new data points involves initially drawing a random vector from this simpler distribution, and then reconstructing the original data through the reverse chain via ancestral sampling. [41]

3.4.2 Score-Based Generative Models (SGMs)

Score-based generative models are centered around the Stein score concept, utilizing a technique where data is altered by progressively intensifying Gaussian noise. These models estimate score functions across all levels of noisy data distribution via a deep neural network trained on various noise levels, knows as noise-conditional score network. The generation of samples is achieved by sequentially applying these score functions, reducing noise levels through methods like Langevin Monte Carlo and various differential equations. Notably, the training and sampling processes in these models are independent, allowing for diverse sampling techniques post score function estimation. [41]

3.5 Applications of Diffusion Models

Diffusion models can be successfully utilized on multiple challenging real-world tasks thanks to their flexibility. Here are some of their exciting applications.

- **Image Super Resolution, Inpainting, Restoration**: Generative models have been used to tackle a variety of image restoration tasks including super-resolution and inpainting. Image super-resolution aims to restore high-resolution images from low-resolution inputs, while image inpainting revolves around reconstructing missing or damaged regions in an image. [41]

- **Text to image generation**: an integral part of vision-language models that has gained significant traction. This task involves creating visual images from textual desciptions. It showcases the remarkable ability of these models to interpret and visualize textual data. [41]
- **Anomaly detection:** Anomaly detection is a vital and intricate challenge. Generative models play a pivotal role in identifying anomalies. [41]
- **Text to audio generation**: Text to audio generation involves converting written language into spoken voice. ProDiff [18] is designed to enhance this process by directly predicting clear data. This helps in maintaining high-quality audio outputs even when the sampling process is accelerated, thereby preventing any significant degradation in the quality of the generated. [41]

3.6 Comparison with other generative models

Diffusion models have emerged as a significant breakthrough, offering unique advantages and characteristic that set them apart from other models. The next sections delve into the core differences from other well-established models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

The generator's objective is to produce data so convincing that the discriminator cannot reliably distinguish from real data. Conversely, the discriminator aims to enhance its ability to differentiate genuine data from the counterfeit ones. This process leads to the generator improving its data fabrication capabilities until the discriminator is often unable to tell the real and fake apart. [14]

3.6.1 Diffusion Models compared to Generative Adversarial Networks

GANs may encounter convergence difficulties during training. This challenge arises from the possibility of either generator or discriminator network becoming entrapped in local minimum, preventing global optimization. In contrast, Diffusion Models adopt a maximum likelihood estimation approach for training, which is generally more stable and converges reliably. [14]

Hyperparameter optimization in GANs is intricate due to the need to balance multiple parameters, such as learning rates and the equilibrium between generator and discriminator losses. This complexity can make the tuning process laborious and time-consuming. Diffustion Models are characterized by a simpler hyperparameter landscape, facilitating a more streamline and efficient tuning process. [14]

In terms of quality, both models are adept at generating compelling samples, Diffusion Models are inclined to producing images with sharper and more precise features. GANs are susceptible to mode collapse, a condition where the generator produces limited diversity, hindering the model's ability to span the full distribution of the training data. [14]

3.6.2 Diffusion Models compared to Variational Autoencoders

In comparing Diffusion Models with Variational Autoencoders, it must be said that VAEs have an advantage in terms of the speed of sample generation. This contrasts with the slower, multi-step sample generation process of Diffusion Models. [8]

In VAEs, the encoder generates a latent code distribution for each input. Occasionally, these distributions might overlap, causing different inputs to share similar latent codes. When this happens, the decoder's attempt to reconstruct the original inputs can result in an averaged output, often leading to blurred or less distinct samples. Diffusion Models generally offer higher quality samples than VAEs. [12]

VAEs are notable for generating a broad spectrum of samples, circumventing the mode collapse issue common in GANs. Similarly, Diffusion Models also excel in producing a diverse range of outputs and are not prone to mode collapse. Both models ensure a varied representation of data from their respective training sets. [8]

3.6.3 Comparison summary

Each model has different advantages and drawbacks, making each of them particularly well-suited for distinct use cases. Advantages and drawbacks of each model are summarised in the table 3.1.

High fidelity samples means that the model generates high quality and detailed samples.

Fast sampling refers to the time it takes to generate a sample. Fast sampling enhances the practical usability of a model in production environments.

Mode coverage assesses model's capability to capture and reproduce the diversity of a dataset. It ensures that the model can generate samples across all classes in the data distribution, rather than focusing on a limited subset.

Generative model	**High fidelity samples**	**Fast Sampling**	**Mode coverage**
Diffusion Models	yes	no	yes
GANs	yes	yes	no
VAEs	no	yes	yes

Table 3.1: Comparison summary

3.7 TTS Models incorporating Diffusion Models

Despite diffsuion models being a relatively recent advancement in the field of artificial intelligence, there is already a number of tools leveraging their capabilities while also attempting to mitigate their limitations.

3.7.1 Grad-TTS

Grad-TTS employs a score-based decoder to produce high-quality mel-spectrograms. This is achieved through a gradual transformation of noise, which is predicted by the encoder and aligned with the text input via Monotonic Alignment Search (MAS). The model's decoder converts Gaussian noise, parameterized by the encoder outputs, into a mel-spectrogram. Grad-TTS uses a generalized version of forward and reverse diffusion to reconstruct data from Gaussian noise effectively. A distinctive feature of Grad-TTS is its ability to control the balance between the quality of the output mel-spectrogram and the inference speed. Notably, it can generate high-quality mel-spectrograms with as few as ten reverse diffusion iterations, highlighting its efficiency. The Grad-TTS model as it can be trained as an end-to-end TTS system. This is accomplished by modifying the model's output from mel-spectrograms to raw audio waveforms, integrating both vocoder and feature generator

functionalities into a single, cohesive model. This approach simplifies the TTS pipeline while maintaining the model's effectiveness in generating high-quality speech output. [32]

3.7.2 FastDiff

FastDiff is a fast conditional diffusion model for high-quality speech synthesis. It addresses the challenges of tradiitional denoising diffusion probabilistic models, which typically require a large number of iterations, by introducing innovations for efficiency and quality improvement. FastDiff utilizes time-aware location-variable convolutions to effectively model long-term dependencies in audio data and employs a noise schedule predictor to reduce the number of reverse iterations required, thus speeding up the process. This approach allows FastDiff to synthesize high-quality speech at significantly faster rate compared to conventional methods, making it practical for real-world applications. The model also includes an end-to-end TTS synthesizer, FastDiff-TTS, which further simplifies the speech synthesis process. [17]

3.7.3 ProDiff

introduces an innovative approach to TTS synthesis. Addressing the limitations of AR models, which require many iterations and computational power, and Non-AR models, which generate samples with limited sample diversity and sample quality, ProDiff employs a progressive fast diffusion method. To avoid considerable drop in quality when reducing the number of iterations, this model directly predicts clean data, bypassing the need to estimate gradient for score matching, a common challenge in AR models. Additionally, ProDiff incorporates knowledge distillation techniques to enhance model convergence with fewer diffusion iterations. It utilizes a generated mel-spectrogram from an N-step denoising diffusion implicit model as a training target for a new model with half the steps, leading to sharper predictions and significantly faster sampling speeds. This approach effectively combines the benefits of both AR and Non-AR models, offering a balanced solution for high-quality, efficient TTS synthesis. [18]

3.7.4 DiffSpeech

DiffSpeech is an extension of the DiffSinger model, focusing on TTS tasks. DiffSinger, originally designed for singing voice synthesis, employs a diffusion probabilistic model to generate mel-spectrograms from music scores. It iteratively converts noise into mel-spectrograms, improving voice quality and inference speed with a shallow diffusion mechanism and boundary prediction methods. DiffSpeech adapts these techniques for TTS, proving its generalization and effectiveness in producing realistic speech outputs. It demonstrated superior performance compared to state-of-the-art TTS models and highlights the versatility of the diffusion probabilistic model in voice synthesis tasks. [24]

Chapter 4

Datasets

In machine learning and artificial intelligence, datasets play a pivotal role that fuel the development and refinement of models, including those designed for creating and detecting deepfakes. This chapter delves into the essence of deepfakes while focusing mainly on voice datasets.

4.1 What are datasets

Dataset is a collection of data that is used to train, validate and test machine learning models. It typically consists of a large number of examples, each of which includes one or more features and, in supervised learning, a label. In the case of deepfakes, these datasets usually consist of a substantial number of images, videos or audio files, along with associated data that the algorithms use to learn how to generate deepfakes. [13]

For training speech synthesis alorithms, datasets can include large and diverse collection of spoken words, sentences and longer speech segments. The quality and size of the dataset can significantly impact the performance of the resulting deepfake model. For instance, a more extensive and varied dataset can lead to a more robust and accurate deepfake detection model. A large dataset with diverse speech samples or facial expressions equips the algorithm with a better understanding of human speech and facial dynamics, enhancing its ability to generate convincing deepfakes. [43]

It is important to recognize that, even with seemingly large dataset, the actual number of relevant data points for specific cases of interest can be quite limited. This phenomenon is observed across various domains and is characterized by a distribution known as the long tail. In this distribution, a small number of elements are extremely frequent, while the majority are relatively rare. This pattern suggests that while common cases are well-represented, rare or unique instances may not be adequately covered in the dataset. [26]

There are two primary categories of datasets utilized in deepfake technology: real datasets and synthesized datasets. Typically, real datasets are employed in the creation of deepfakes, providing authentic examples for the algorithms to learn from. On the other hand, synthesized or fake datasets are predominantly used in the detection of deepfakes, as they contain examples of manipulated media that help train algorithms for deepfake detection. [43]

4.2 Datasets suitable for training TTS

These datasets contain audio clips narrated by real people and can be used to train TTS models. The quality and diversity of these datasets are fundamental in developing TTS technologies capable of generating natural, human-like speech. There is a variety in datasets, each with unique characteristics in terms of language, accent and emotional range. Understanding the nuances of each dataset helps in selecting the right on for specific TTS training objectives, ensuring the development of robust and efficient speech synthesis models. Summary of these datasets is in the table 4.2.

4.2.1 LJ Speech

This dataset is in the public domain and consists of 13 100 short audio clips from a single female speaker who read passages from seven non-fiction book. All clips are recorded in English. Each clip ranges from 1 to 10 seconds and totals around 24 hours. Each clip is labeled. The source texts, which are also in the public domain, were published from 1884 to 1964. The recordings were made in the period of 2016-2017. Each audio file is a single-channel 16-bit PCM WAW with sample rate of 22050 Hz. The total amount unique words in this dataset is 13 821. The clips are in good quality without significant background noise. This dataset can be used for TTS or automatic speech recognition. [19]

4.2.2 LibriTTS

LibriTTS is a speech dataset specifically designed for TTS applications. It is an extension of the LibriSpeech[31] dataset, which itself is derived from audiobooks read by volunteers for the LibriVox project. LibriTTS offers cleaner audio recording and includes both the original audiobook text and the spoken audio. This dataset is particularly useful for training and evaluating TTS system due to its diverse range of speakers, accents and speaking styles, making it a valuable resource in the field of speech synthesis. It contains almost 586 hours of read English speech from 2456 speakers structured to achieve a balance in gender representation and the duration of recordings per speaker. Data in LibriTTS is divided into 7 subsets shown in Table 4.1. [42]

LibriTTS has addressed some problems of the LibriSpeech. In LibriSpeech the audio files are at 16 kHz sampling rate, as opposed to 24 kHz in LibriTTS, which is too low to achieve high quality TTS. Modern high quality TTS systems use sampling rate between 24 to 48 kHz. Instead of splitting the data at silence intervals longer than 0.3 seconds, LibriTTS is splitting the data into sentences. LibriSpeech contained audio files which had significant background noise, which were removed in LibriTTS. [42]

4.2.3 Mozilla Common Voice

Mozilla Common Voice is an open-source, multi-language dataset that focuses on collecting voice data from volunteers globally. Its goal is to help create and train voice-enabled technologies in a wide variety of languages and accents. The project emphasizes diversity and inclusivity, aiming to represent a broad spectrum of speech patterns and dialects. Common Voice is unique for its community-driven approach, where anyone can contribute their voice, validate and verify the speech of others, helping to build one of the most diverse and accessible voice datasets available. As of writing this Common Voice has clips is 120 languages, spanning almost 20 thousand hours of labeled speech. Often voice clips also

Subset	Hours	Female Speakers	Male Speakers	Total Speakers
dev-other	6.43	16	17	33
dev-clean	8.97	20	20	40
test-other	6.69	17	16	33
test-clean	8.56	19	20	39
train-clean-100	53.78	123	124	247
train-clean-360	191.29	430	474	904
train-other-500	310.08	560	600	1160
Total	585.80	1185	1271	2456

Table 4.1: LibriSpeech Dataset Summary [42]

include demographic data like age, sex and accent that can further improve the usability of this dataset. [4]

4.2.4 Voice Conversion Toolkit

VCTK Corpus is a speech dataset featuring 110 English speakers with various accents, each reading approximately 400 sentences from selected texts, including the Herald Glasgow newspaper, the Rainbow Passage, and an elicitation paragraph from the speech accent archive. Being recorded in a controlled environment with high-quality microphones, it offers a rich resource for studying regional accents and speech synthesis advancements. [39]

4.2.5 VoxCeleb

The VoxCeleb audio dataset is a large-scale speaker identification dataset that includes 153 516 utterances from 1 251 celebrities extracted from YouTube videos. It is curated to maintain a balance between male and female speakers, with 45 percent of speakers being female. Speakers feature a vast array of ethnicities, accents, professional backgrounds and ages with the information about the speaker's nationality and gender included. [28]

Dataset	Utterances	Female Speakers	Male Speakers	Release Year
LJSpeech[19]	13 100	1	0	2017
LibriTTS[42]	200 000+	1185	1271	2019
VCTK[39]	88 328	63	47	2016
VoxCeleb[28]	153 516	563	688	2017

Table 4.2: Real datasets summary table

4.3 Datasets with synthesized speech

These datasets contain synthesized speech and are usually used for deepfake detection. Datasets with synthesized speech are crucial in the realm of deepfake detection. They provide a rich source of data that helps in training and improving algorithms designed to identify and differentiate between real and synthetic speech. By incorporating a wide range of synthesized voices and speech patterns, these datasets enable more accurate and robust detection systems. As deepfake technology becomes more sophisticated, having access to

diverse synthesized speech datasets becomes essential in keeping up with the advancements in deepfake creation, ensuring the continued effectiveness of detection methods. Summary of these synthesized datasets is in the table 4.3.

4.3.1 ASVspoof

The ASVspoof dataset is designed to foster research in anti-spoofing for automatic speaker verification systems (ASV) and provide platforms for the assessment and comparison of spoofing countermeasures. ASV is the most intuitive and user-friendly method for biometric person recognition, however it is susceptible to spoofing attacks. It addresses three types of spoofing attacks: TTS synthesis, voice conversion and replay attacks. Replay attacks involve recording legitimate access attempt, usually done secretly. This recorded voice is then played back to the ASV system. Goal of such attack is to deceive the system into believing that the recorded voice is authentic input from the original speaker, therefore granting unauthorized access. This dataset was created using utterances from 107 speakers, consisting of 46 male and 61 female speakers, from the Voice Cloning Toolkit corpus. [37]

This dataset focuses on two specific scenarios. Logical Access (LA) and Physical Access (PA), each with its own distinct characteristics and attack types. The LA scenario simulates situations where a remote attacker attempts to gain unauthorized access to a system protected by ASV using synthetic or converted speech, as in the case of remote banking services. This scenario assumes the microphone is chosen by the user, not controlled by the system. The database includes attacks generated by different TTS and VC systems., It consists of a training set, a development set, and an evaluation set, each containing bona fide and spoofed utterances generated using serveral TTS and VC algorithms. [37]

The PA scenario reflects attacks where an unauthorized person attempts to gain access to a physical space or a device protected by ASV by replaying recorded genuine speech. This scenario assumes control over the microphone by the system, such as in a secured facility entry. The replay attacks are simulated with controlled variability, such as room size, speaker-to-microphone distance, and reverberation time to study their impact on ASV systems. To generate a large and diverse set of replay recording, simulatations are used instead of real replay recordings. This approach allows for precise control over environmental factors and replay device characteristics. Replay attacks are categorized by attacker-to-talker distance and replay device quality, ranging from perfect to low quality devices. [37]

4.3.2 DEEP-VOICE

This dataset was created for the purpose of detecting synthesized speech, specifically focusing on deepfake voice conversion. It includes real human speech from eight well-known figures and their speech converted using Retrieval-based Voice Conversion. This dataset is used for binary classification to determine if the speech is real or synthesized. It facilitates the training of machine learning models to detect synthesized speech, boasting a high classification accuracy and the capability for real-time classification. The dataset is publicly available for further research in synthesized speech detection. [7]

4.3.3 FakeAVCeleb

The FakeAVCeleb dataset is a comprehensive collection encompassing four types of audio-visual media. It includes authentic videos with their original audio and three variants of audio-video deepfakes. These deepfakes are differentiated by their composition: one

combines real video with synthesized audio, another pair combines fake video with authentic audio and the the third features both fake video and audio. The synthesized audio is created using a transfer learning-based real-time voice cloning tool, which transforms real audio and text into the targeted individual's synthetic voice. For video creating, three tools are employed: FaceSwap and FSGAN for face-swapping and Wav2Lip for lip syncing audio to video. This diversity makes the dataset valuable tool for developing and testing deepfake detection technologies. [22]

Figure 4.1: Samples from the FakeAVCeleb dataset [22]

4.3.4 WaveFake

The WaveFake dataset is a specialized collection designed for the detection of audio deepfakes. It consists of a substantial amount of generated audio clips, approximately 196 hours in total. The dataset primarily utilizes the LJSpeech dataset as its foundation. WaveFake includes multiple samples from different state-of-the-art network architectures, offering a diverse range of audio for research and development in deepfake detection. This allows comparing same clip generated by different network architectures. Its extensive size and variety position WaveFake as a significant resource in the ongoing effort to imrpove and refine deepfake detection methodogologies. [11]

4.3.5 In-The-Wild Audio Deepfake Dataset

In-The-Wild Audio Deepfake Dataset comprised of audio deepfake and corresponding genuine, unmanipulated audio recording from 58 celebrities and politicians, gathered from public sources. It includes 20.8 hours of genuine and 17.2 hours of spoofed audio, aiming to aid in evaluating deepfake detection of anti-spoof machine learning models. This dataset is particularly well designed to assess a model's ability to generalize to realistic, in-the-wild, audio samples. [27]

Dataset	Utterances	Female Speakers	Male Speakers	Release Year
ASVspoof2019[37]	122 299	61	46	2019
DEEP-VOICE[7]	64	2	6	2023
FakeAVCeleb[22]	500	250	250	2021
WaveFake[11]	117 985	NA	NA	2021
In-The-Wild[27]	31 778	6	48	2022

Table 4.3: synthesized datasets summary table

Chapter 5

Deepfake detection

The rapid progress of technology in synthetic image and audio generation and manipulation has reached a point where distinguishing between genuine and fabricated content is increasingly challenging. These deepfakes can be used to strategically spread misinformation or fake news, potentially causing significant social, political, and economic repercussions. Examples include the creation of deceptive images, such as a political figure being depicted as arrested without that ever actually happening, or audio audio clips where a high ranking executive appears to make false statements. Moreover, video and audio deepfakes can be combined to convincingly mimic celebrities or public figures to perpetrate scams or manipulate public opinion.

5.1 Need for detection software

In recent years, there has been an uptick in deepfake scams online. For instance, in 2022, hackers gained access to popular YouTube channels to livestream deepfake videos of Elon Musk promoting a cryptocurrency scam, deceiving viewers into sending money to a fradulent digital wallet with promise of doubling their currency. This scheme netted 243 000 dollars in just over a week, exploiting the platform's late response in removing the malicious livestreams, staying up for hours before taking an action and removing these harmful livestreams. Screenshot from this livestream can can be seen in figure 5.1. [29]

In another case in 2023, a deepfake video advertisement on TikTok featuring popular popular YouTube creator MrBeast, known for philanthropy and giving away money, falsely promoting an iPhone giveaway, which led users to a malicious link. This harmful advertisement again exploited the late response of the platform, staying up for hours before being removed. The full extent of the damage caused by this scheme is unknown. Screenshot from this video can be seen in figure 5.2. [30]

These incidents highlight the urgent need for robust deepfake detection mechanisms. Social media platforms currently struggle to adequately detect and mitigate the spread of such synthetic content promptly. This gap necessitates the development of advanced detection software that can either remove these harmful videos or adequately label them to inform its viewvers of their inauthenticity. Effective deepfake detection is crucial not only for maintaining the integrity of information but also for safeguarding individual and public security against the malicious use of artificial intelligence.

The implications of unchecked deepfake content are vast, affecting not only individual reputations but also national security, democratic processes, and public trust. The abil-

ity of deepfakes to create convincing false realities can lead to misinformation campaigns designed to incite public fear, influence elections, or even provoke international conflicts. There is a critical need for sophisticated detection technologies that can effectively differentiate between real and synthetic content. These technologies muse evolve rapidly to keep pace with the increasing sophistication of deepfake generation methods, which continually improve in creating more convincing deepfakes.

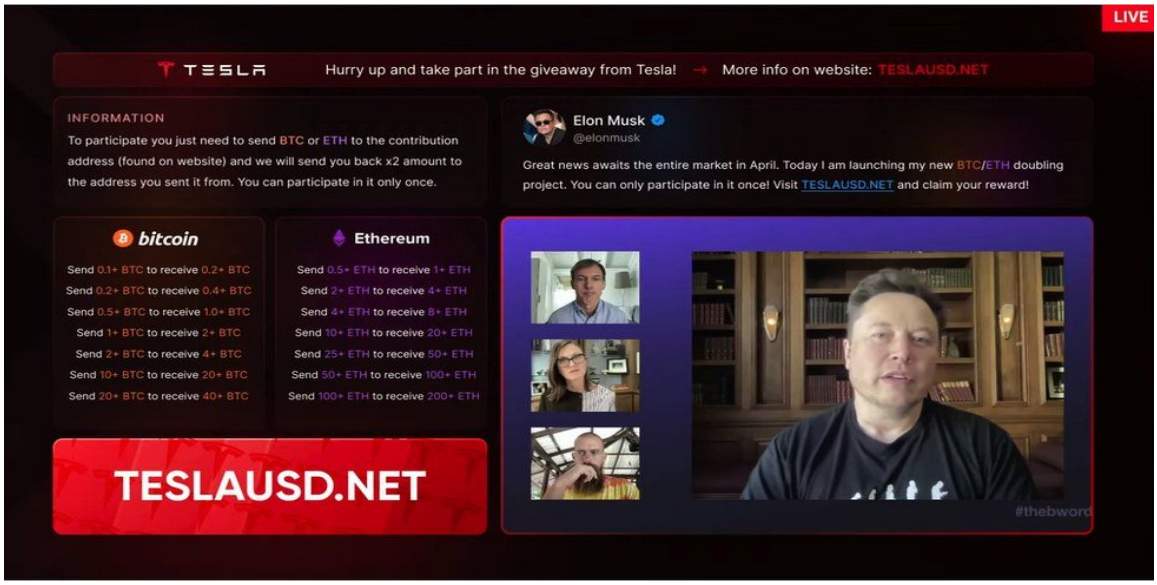


Figure 5.1: Screenshot from the livestream scam featuring Elon Musk [29]

Figure 5.2: Screenshot from the TikTok advertisement featuring deepfake of MrBeast [30]

5.2 Audio spoofing detection tools

Human senses are adept at recognizing familiar patterns, such as nuances in a voice or the subtleties of a face. However, as technology advances, so does the capability of fooling these human senses. This is where audio spoofing detection become crucial. These tools are designed to inspects artifacts and patterns that are imperceptible to the human ear.

Spoofing detection tools employ sophisticated algorithms trained on extensive datasets that include a wide array of both genuine and fake audio samples. By learning from these examples, they develop an acute sensitivity to the spectral and temporal discrepancies that characterize manipulated audio.

5.2.1 AASIST

AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) was released in 2021. AASIST addresses the challenge of distinguishing between genuine and spoofed audio by employing a novel graph attention network strcuture that models both spectral and temporal domains efficiently. The system avoids the need for computationally expensive ensemble systems by adopting a single, cohesive model capable of detecting a broad range of spoofing attacks. [21]

AASIST improves upon traditional spoofing detection by introducing a Heterogeneous Stacking Graph Attention Layer(HS-GAL), which incorporates a modified attention mechanism tailored to handle heterogeneity in the data. This layer effectively processes temporal and spectral information. Furthermore HS-GAL includes a stack node, that accumulated and integrated information across different nodes, effectively bridging the gap between the data domains of spectral and temporal graphs. This node enhances the layer's ability to synthesize information from all parts of the graph, ensuring a comprehensive analysis of potential spoofing artifacts. [21]

The system inculdes an operation termed max graph operation, which uses a mechanism to prioritize significant features during the detection process. This operation allows AASIST to focus on the most relevant artefacts withing the audio for more accurate spoof detection. [21]

Recongizing the need for practical deployment in resource-constrained environments, AASIST also offers a lightweight variant with significantly reduces computational requirements, while still maintaining competetive performance, named AASIST-L. [21]

5.2.2 SSL Anti-spoofing

This detection system uses wav2vec 2.0 model, which is pre-trained in a self-supervised way using only genuine data, meaning no spoofed data is used during the initial training phase. This model is chosen for its ability to learn robust generalizable features from a massive corpus of unlabeled audio data. [36]

After pre-training, the model is fine-tuned on a mix of genuine and spoofed data that helps adapt the wav2vec 2.0 model to the specific task of spoofing detection. The system includes data augmentation techniques that introduce variability in the training data, which helps in improving the model's robustness and generalization capabilities. These techniques include modifying audio samples through methods like adding noise, varying the pitch, and other signal transformations. [36]

Important component of the system is the attention-based aggregation layer, which effectively captures the most relevant features from the input data. This layer weights

the importance of different features, focusing more on those that are indicative of spoofing attacks. This system is designed to be robust against various forms of audio spoofing attacks, making it highly effective in practical scenarios where the nature of attacks can be unpredictable and varied. [36]

5.3 Performance evaluation metrics

Measuring the effectivenes of spoofing detection tools is important. The success of these tools hinges on their ability to distinguish between genuine and manipulated audio with precision. For this reason, a variety of performance evaluation metrics are employed, each offering unique insights into the tool's.

False Match Rate (FMR) or also False Acceptance Rate, are terms used to describe the rate at which the security system incorrectly accepts an access attempt by an unauthorized user, mistaking it for a legitimate one or system erroneously recognizing a non-authentic sample as authentic. [6]

False Non-Match Rate (FNMR), also known as False Reject Rate, is metric that measures the rate at which a security system incorrectly rejects an access attempt by an authorized user, or incorrectly classifies an authentic sample as non-authentic. Essentially, it represents a probability that the system fails to recognize a legitimate input. [6]

Equal Error Rate (ERR) is a common metric used to evaluate the performance of biometric systems, including voice recognition, fingerprint scanners, and facial recognition systems. ERR represents the point at which match rate equal the non-match rate. If the EER of a system is 5 percent, this indicates that at the threshold level where FMR equals FNMR, both of these rates are at 5 percent. [6]

Chapter 6

Experiment

This chapter describes a experiment designed to assess the effectiveness of audio spoofing detection tools in identifying deepfakes generated by diffusion models. The primary objective of this experiment is to explore whether the diffusion models impact the performance of advanced spoofing detection tools and measure how significantly.

6.1 Experiment design

This section talks mainly about the design of the dataset, motivation behind creating this dataset and the tools used to create it.

6.1.1 Motivation

Diffusion Models represent a significant advancement in deepfake technology, but it is still a very recent compared to other network architectures. While visual deepfakes have gained considerable attention, the domain of speech deepfakes has not received such a prevalent spotlight. This lack of focus has resulted in scarcity of publicly available speech datasets. Currently there are not any publicly available speech datasets containing synthesized speech generated by diffusion models.

The objective of this work is to develop a dataset that effectively aids in the detection of deepfake speech. By focusing on this goal, the dataset aims to become a vital tool in identifying and analyzing generated speech, particularly that created through diffusion models. The creation of such dataset is critical in the current digital landscape, where the spread of deepfake technology poses significant challenges to autenticity and security in digital communication.

Given that Diffusion Models are a new and different method for creating deepfake speech, there is a significant knowledge gap regarding the effectiveness of existing detection methods against this new form of synthetic speech. This work, therefore, seeks to provide an insight into how the current detection tools can handle this challenge. It is important to stay ahead in the race against malicious use of deepfakes.

6.1.2 Dataset description

The dataset is designed to include pairs of audio clips, each pair consists of an original voice recording and two corresponding synthesized counterparts, each synthesized by a different model. Additionally, there are only synthesized pairs, meaning there are only two

synthesized audio clips generated by a different model, with the same input text without the original audio recording to provide insight on the difference when the model generated audio clip based on a sentence that it was trained on and when it generates audio clip on uncovered sentence. The synthesized voice is produced using a model specifically trained on the respective original voice. This approach ensures a direct comparison between authentic and generated speech.

Original recordings are sourced from the LJSpeech[19] dataset, featuring single female speaker narrating audiobooks in English. These original voice recordings will be complemented by synthesized versions, created by using two distinct models: ProDiff[18] and DiffSpeech[24]. Both of these models are trained specifically on the LJSpeech dataset.

The synthesized-only audio pairs are created by ProDiff[18] and DiffSpeech[24] models, both trained on the LJSpeech dataset. This inclusion of sentences not covered in the training set tests the model's generalization capabilities, a critical factor in real-world applications where the ability to produce coherent speech from novel inputs is very important factor. Additionally the generated audio pairs utilize Harvard Sentences to ensure consistent and standardized testing conditions.

The dataset consists of a total 27 640 audio clips, evenly divided between the ProDiff[18] and DiffSpeech[24] models. Each models contributes 13 820 clips, with 13 100 being direct synthesized of existing LJSpeech[19] recordings. The remaining 720 clips from each model are synthesized from Harvard sentences that were not a part of the training dataset. All recordings in the dataset use the same sampling rate as the original LJSpeech recordings, which is 22 050 Hz.

6.1.3 Harvard Sentences

Harvard sentences are a set of phonetically balanced phrases that are widely used in the testing of audio equipment and speech processing algorithms. Developed by the Institute of Electrical and Electronics Engineers (IEEE), these sentences are grouped into lists, each containing ten phrases constructed to have a similar phonemic distribution. This phonetic balance ensures that each sentence places a similar load on the speech processing systems being tested, allowing for consistent and meaningful comparisons across different tests. [1]

The design of Harvard sentences reflects everyday conversational English, incorporating a range of common phonemes and intonations. As such, they are considered an effective tool for objectively evaluating the performance of speech related technologies and this design ensures that they are broadly applicable and allow for repeatable, comparable results in testing environments. This makes the Harvard sentences particulalry valuable in this study, where the goal is to compare and analyze the synthetic speech output of models trained on the LJSpeech dataset. [1]

6.2 Experiment overview

The primary objective of this experiment is to evaluate the performance of two audio spoofing detection systems when exposed to datasets synthesized by four different TTS models. This setup is designed to explore the specific impact of diffusion-based synthesis on the effectiveness of spoofing detection tools.

The initial part of the experiment is synthetizing the datasets, each TTS model has its own dataset, consisting of 13 100 trained sentences and 720 novel sentences. To provide

a direct comparison, each TTS model synthesizes the same set of sentences. This ensures consistency across datasets, allowing for an accurate assessment of each model's capabilities.

The second part of the experiment involves running each dataset through the two audio deepfake detection systems. Output of these detection systems are scores for every single recording that can be transformed into graphs. These graphs provide valuable insights, allowing for detailed analysis.

6.2.1 Selection of TTS models

ProDiff and DiffSpeech are recent TTS models utilizing diffusion models to generate high quality audio outputs. These TTS models were chosen because they are among the best performing TTS models when talking about audio quality.

Glow-TTS is a recent TTS model not incorporating diffusion models. This model is chosen to compare the performance of spoofing detection system on both TTS model incorporating diffusion models and TTS model not incorporating diffusion models.

Tacotron2 is a TTS model from 2017, chosen to provide a baseline comparison with newer models, illustrating how advances in TTS technologies might impact the effectiveness of spoofing detection systems. This comparison helps to assess the progress in TTS technologies over time and their implications for audio security measures.

These TTS models were chosen for their leading performance.

6.2.2 Selection of audio spoofing detection systems

AASIST, spoofing detection system designed to deliver robust performance in identifying spoofed audio through spectral analysis. AASIST has achieved an exceptional EER of 0.83 percent in the ASVspoof2019 LA dataset and has managed to outperform other state-of-the-art spoof detection systems. [21]

SSL Anti-Spoofing, state-of-the-art spoofing detection system designed with focus on versatility and robustness, this system employs self-supervised learning strategies to enhance its detection capabilities across a wide spectrum of spoofing scenarios. SSL Anti-Spoofing has managed to score the lowest reported EER for the ASVspoof2021 Deepfake Database and ASVspoof2021 LA Database, proving that the SSL Anti-Spoofing is a highly capable spoofing detection system. [36]

6.3 Synthetizing datasets

This section details the process to synthesize the datasets using the four distinct TTS models.

6.3.1 Preparing metadata

LJSpeech is a labeled dataset, all the recordings are labeled in the file named metadata.csv, where each line represents a single audio recording. On each line there are three items divided by the pipe character.

Each line contains identifier for each recording and corresponds directly to the name of the audio file, for example LJ001-0001 refers to file named LJ001-0001.wav.

Next item is transcription, which are the exact words spoken in the audio recording, providing a direct script of the spoken content.

Third item is normalized transcription, which is a version of the transcription with expanded numbers, ordinals, and monetary units into full words. For example „5th“ is expanded into „fifth“.

Dataset for each model is synthesized from its own unique metadata file. Only change in the metadata between models is in the identifier so that it reflects the model synthetizing the dataset. For example, the original identifier LJ001-0001 is changed to ProDiff-LJ001-0001 for ProDiff, for DiffSpeech it is changed to DiffSpeech-LJ001-0001, and for Glow-TTS it is changed to Glow-TTS-LJ001-0001. Transcription and normalized transcription remain identical across all metadata files.

In the original metadata, the first number in the name of the file represents different groupings, like a section or a chapter. Originally there is 50 of these groupings. In order to analyze the difference between synthetizing sentences already present in the training data and entirely new sentences, there is one more grouping added consisting of 720 new entries, all of which are sourced from the Harvard Sentences, a standardized set of phrases used for testing audio equipment and speech processing algorithms. This addition allows for assessment of how well each TTS model can handle familiar and novel material.

6.3.2 Synthetizing with DiffSpeech

To synthesize the dataset using DiffSpeech, some minor modifications were necessary to the source code to enable row-by-row synthesis from the metadata file. While it is possible to train the model on the LJSpeech dataset, the author provided pre-trained model already trained on the LJSpeech, which was utilized instead. Generating all 13 820 audio samples took 4 hours, 53 minutes, and 55 seconds on Nvidia RTX 3070.

6.3.3 Synthesizing with ProDiff

Synthetizing the dataset with ProDiff required minor changes to the source code to enable synthesizing from the metadata file as well. Although it is possible to train ProDiff on the LJSpeech dataset, a pre-trained model provided by the author already trained on LJSpeech was utilized instead. It took 3 hours, 54 minutes, and 49 seconds to generate all 13 820 audio samples, which is approximately 20 percent faster than DiffSpeech.

6.3.4 Synthetizing with Glow-TTS

Sythetizing with Glow-TTS model was done via Python package called 'TTS'[3] which is a comprehensive library commonly used for TTS or Voice Conversion. This library allows TTS synthetization with multiple models, pre-trained on one or more datasets, including Glow-TTS pre-trained on LJSpeech. Generating the samples took only 59 minutes and 52 seconds, which is substantially faster than both DiffSpeech and ProDiff.

6.3.5 Synthetizing with Tacotron2

Synthetizing with the Tacotron2 followed the same procedure as with Glow-TTS, but utilized pre-trained Tacotron2 model instead of pre-trained Glow-TTS model. Synthetizing took 1 hour, 48 minutes, and 3 seconds.

6.4 Working with Audio Spoof Detectors

This section provides insight into the preparation of data and setup when working with audio spoof detectors.

6.4.1 Preparing the protocol for SSL Anti-Spoofing

SSL Anti-Spoofing is originally designed for the ASVspoof2021[40] dataset, concentrating mainly on the Logical Access portion, which involves TTS and voice cloning attacks. To use this detector with alternative datasets, it is necessary to modify the protocols to accommodate the changes from dataset to dataset.

SSL Anti-Spoofing operates using three protocols: training file list, development trials, and evaluation trials. In this specific scenario, only the evaluation trials protocol is relevant since the training and development protocols are not utilized. Original evaluation protocol named `ASVspoof2021.LA.cm.eval.trl.txt` contains the names of the audio files without the file extension to be evaluated by the detector, listed one per line.

For this project, the protocol needs adjusting to include the outputs from three different TTS models. Therefore there are three distinct versions of the evaluation trials protocol required. Each modified version of the protocol includes 13 100 lines corresponding to the names of the original LJSpeech audio files, which serves as the control or baseline in the evaluation, and 13 820 lines for synthesized audio files specific to each TTS model.

6.4.2 Using SSL Anti-Spoofing

SSL Anti-Spoofing is designed to work with the ASVspoof2021[40] dataset, which uses the FLAC audio format, therefore the SSL Anti-Spoofing is looking for files with the .flac file extension, this needs to be changed to .wav in the source code, because all three datasets used in this project synthesize audio files in the wav audio format.

Before running the detector, pre-trained wav2vec 2.0 XLS-R model is required. XLS-R are a series of extensive models that utilize self-supervised learning to develop cross-lingual speech representations, building upon the wav2vec 2.0 framework. The models were pre-trained on an expansive dataset consisting of approximately 436 000 hours of speech from 128 different languages. Through fine-tuning, XLS-R models have reached state-of-the-art performance in various speech-related tasks. [5]

Additionally, SSL Anti-Spoofing requires either a training phase or use of pre-trained model for operation. Author provides a pre-trained model trained on the ASVspoof dataset which was utilized in this project.

6.4.3 Preparing protocols for AASIST

AASIST is designed to work with the ASVspoof2019[37] dataset, focusing on the Logical Access part of the dataset. AASIST requires three protocols to work: training protocol, development trials protocol, and evaluation trials protocol. Since this work focuses only on the evaluation, the training and development trials do not require any changes, but they do have to be present otherwise the detector will not execute. The evaluation protocol is named `ASVspoof2019.LA.cm.eval.trl.txt`.

The structure of the file is vastly different from the ASVspoof2021 version of the file. Each line in the evaluation protocol contains 5 columns divided by space. First column is speaker id, since LJSpeech is a single speaker dataset, this column is always the same.

Second column is name of the audio file without file extension. Third column is id of the speech spoofing system, left blank in case of genuine audio file. Fourth column is also blank because it is not used for logical access part of dataset. Fifth column is marked either „bonafide“ for genuine speech, or „spoof“ for spoofed speech.

There are three version of the protocol required, one for each TTS model. Each contains 13 100 rows from LJSpeech dataset marked as „bonafide“ and 13 820 rows from the respective TTS model marked „spoof“.

6.4.4 Using AASIST

As AASIST is designed to work with ASVspoof2019[37] dataset, it is looking for files with .flac file extension, this needs to be changed in the code to .wav for it to work with these datasets.

AASIST provides two pre-trained models, AASIST and AASIST-L. AASIST-L is lighter version of the model with sigificantly less parameters. Nevertheless for this project, the AASIST model was utilized in order to provide the best results.

6.5 SSL Anti-Spoofing results

This section presents the outcomes of the SSL Anti-Spoofing system when tasked with detecting synthetic speech produced by different TTS models.

6.5.1 PyEER

PyEER Python library was employed for generating the graphs presented. PyEER is a Python package designed primarily for assessing the performance of biometric systems, however it can also be used to evaluate binary classification systems. [25]

Generated graphs are Score Distribution graph, which shows the distribution of genuine and impostor scores, and FMR and FNMR curve graph which shows the two curves and the point where they cross is EER.

6.5.2 DiffSpeech trained sentences

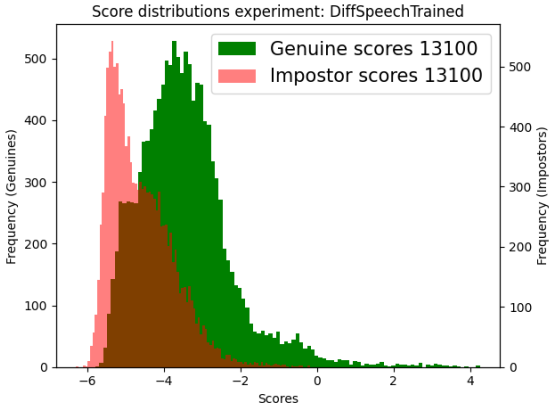
This test consists of 13 100 genuine audio clips and 13 100 spoofed audio clips. The spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.1a and FMR and FNMR 6.1b. The Equal Error Rate (EER) is 27.88 percent.

6.5.3 DiffSpeech novel sentences

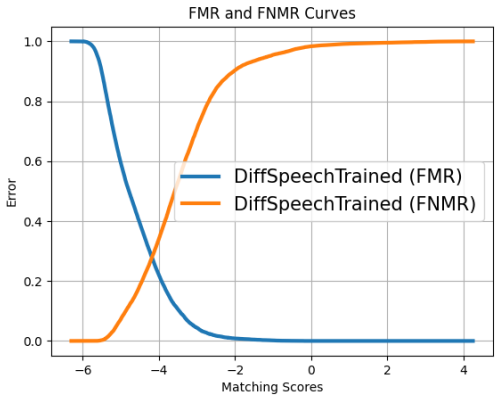
This test consists of 13 100 genuine audio clips and 720 spoofed audio clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test are displayed on the score distribution graph 6.2a and FMR and FNMR graph 6.2b. The EER is 11.94 percent.

6.5.4 ProDiff trained sentences

This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results

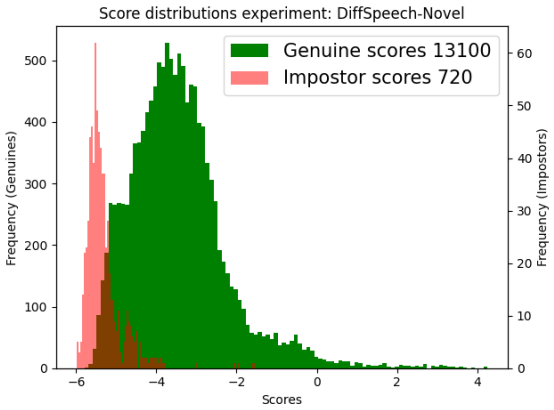


(a) Score distribution

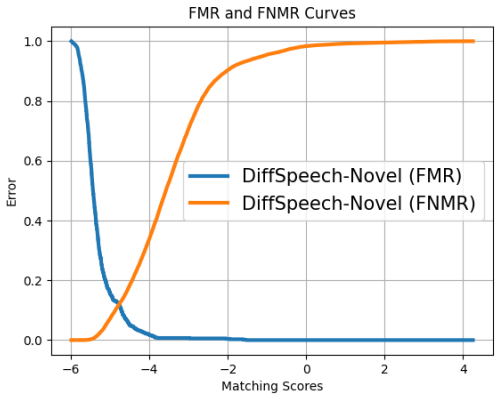


(b) FMR and FNMR curves

Figure 6.1: DiffSpeech trained senteces results from SSL Anti-Spoofing



(a) Score distribution



(b) FMR and FNMR Curves

Figure 6.2: DiffSpeech novel sentences results from SSL Anti-Spoofing

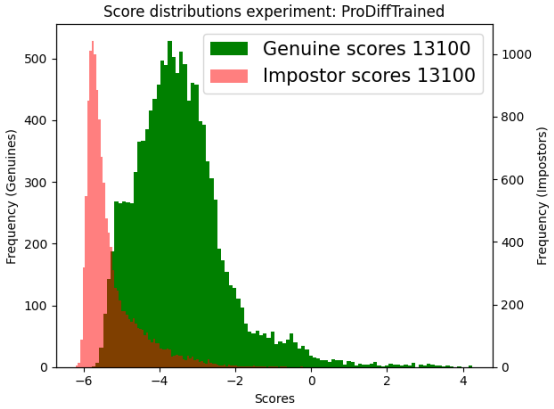
of this test are displayed on the score distribution graph 6.3a and FMR and FNMR graph 6.3b. The EER is 14.94 percent.

6.5.5 ProDiff novel sentences

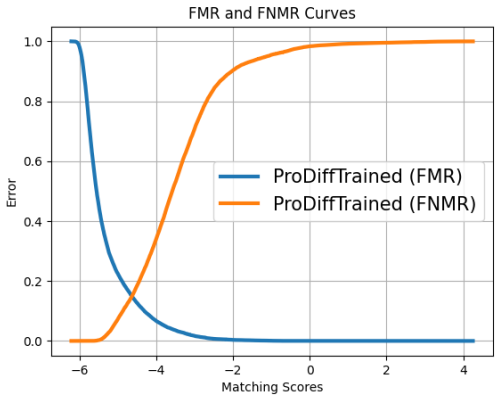
This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test are displayed on the score distribution graph 6.4a and FMR and FNMR graph 6.4b. The EER is 4.02 percent.

6.5.6 Glow-TTS trained sentences

This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.5a and FMR and FNMR graph 6.5b. The EER is 15.81 percent.

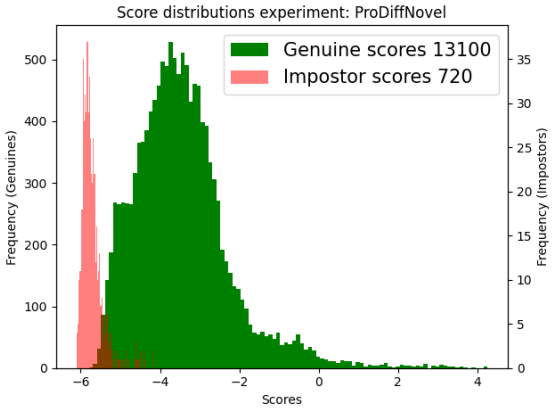


(a) Score distribution

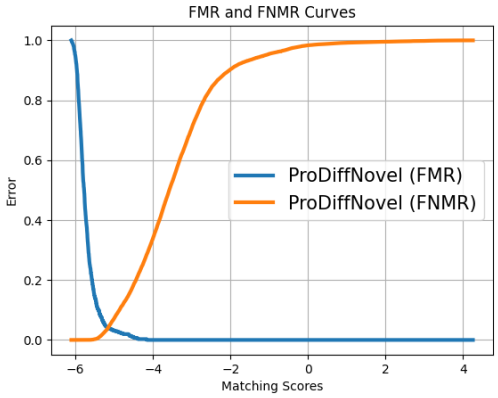


(b) FMR and FNMR Curves

Figure 6.3: ProDiff trained sentences results from SSL Anti-Spoofing

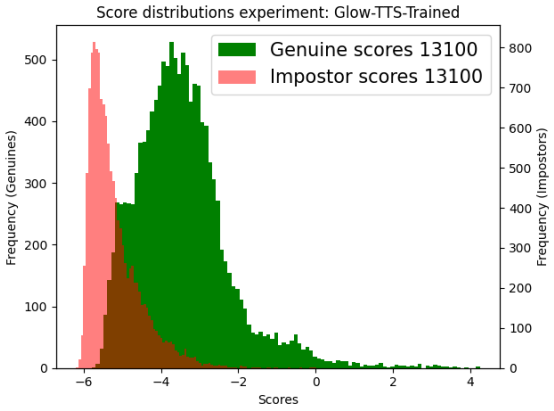


(a) Score distribution

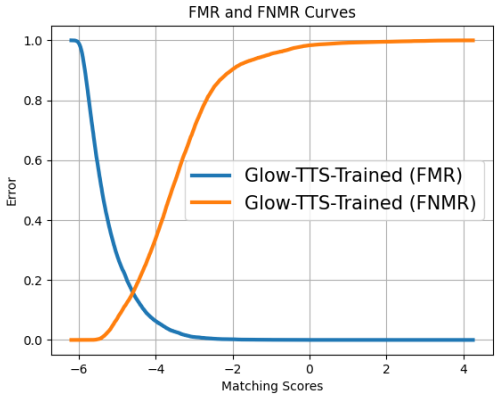


(b) FMR and FNMR Curves

Figure 6.4: ProDiff novel sentences results from SSL Anti-Spoofing



(a) Score distribution

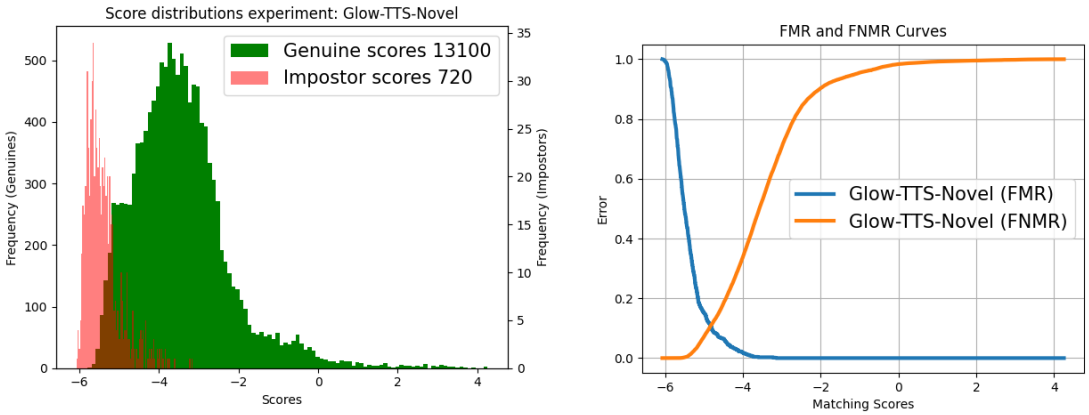


(b) FMR and FNMR Curves

Figure 6.5: Glow-TTS trained sentences results from SSL Anti-Spoofing

6.5.7 Glow-TTS novel sentences

This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test can are displayed on the score distribution graph 6.6a and FMR and FNMR graph 6.6b The EER is 10.7 percent.

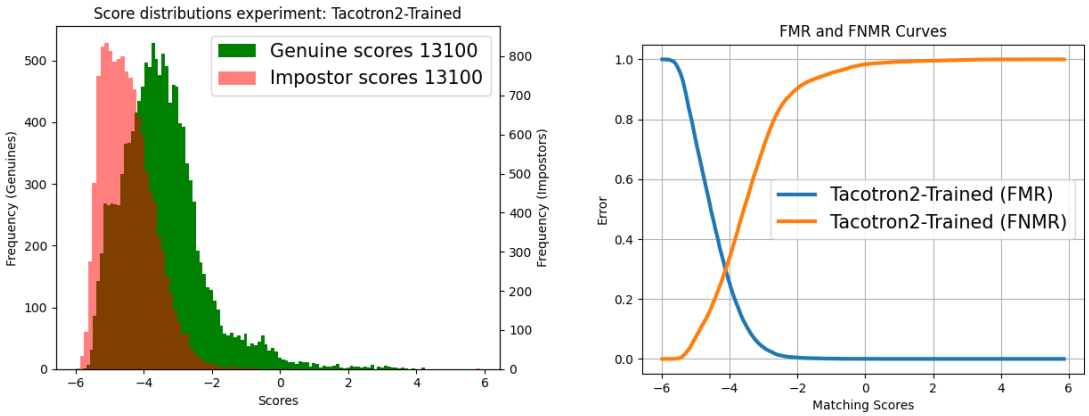


(a) Score distribution (b) FMR and FNMR Curves

Figure 6.6: Glow-TTS novel sentences results from SSL Anti-Spoofing

6.5.8 Tacotron2 trained sentences

This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.7a and FMR and FNMR graph 6.7b. The EER is 29.81 percent.

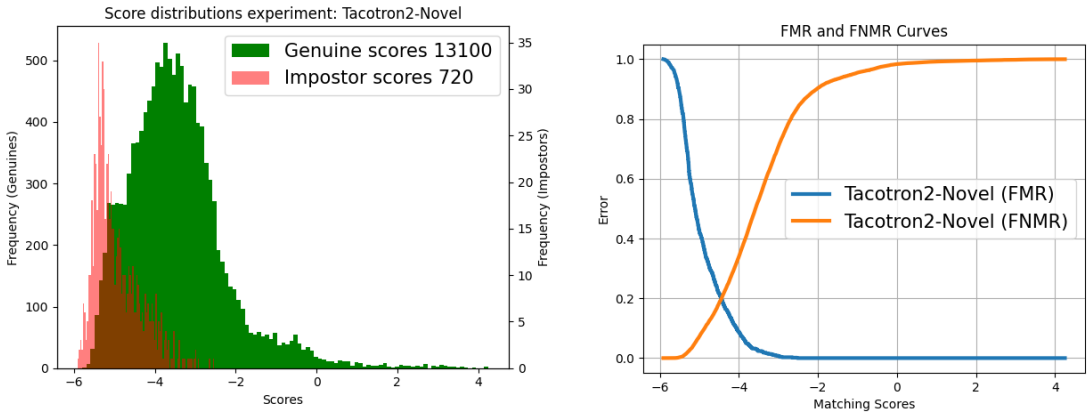


(a) Score distribution (b) FMR and FNMR Curves

Figure 6.7: Tacotron2 trained sentences results from SSL Anti-Spoofing

6.5.9 Tacotron2 novel sentences

This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test can are displayed on the score distribution graph 6.8a and FMR and FNMR graph 6.8b. The EER is 19.85 percent.



(a) Score distribution

(b) FMR and FNMR Curves

Figure 6.8: Tacotron2 novel sentences results from SSL Anti-Spoofing

6.6 AASIST results

This section shows the results of the AASIST audio spoofing detection system in identifying synthetic speech for each of the four TTS models.

6.6.1 DiffSpeech trained sentences

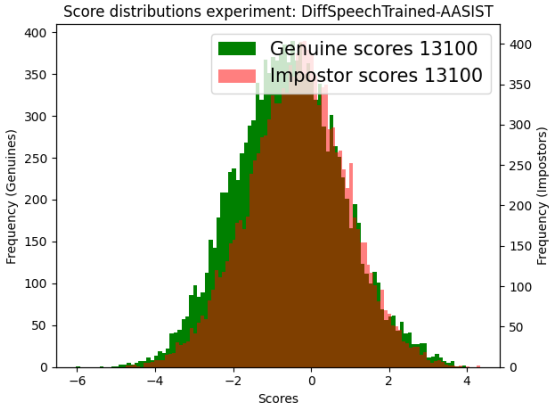
This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.9a and FMR and FNMR graph 6.9b. The EER is 53.99 percent.

6.6.2 DiffSpeech novel sentences

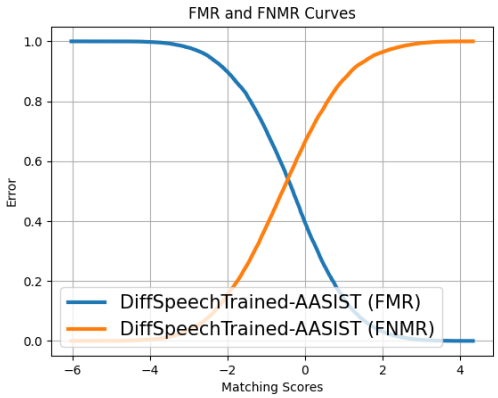
This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test can are displayed on the score distribution graph 6.10a and FMR and FNMR graph 6.10b. The EER is 50.41 percent.

6.6.3 ProDiff trained sentences

This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.11a and FMR and FNMR 6.11b. The EER is 32.07 percent.

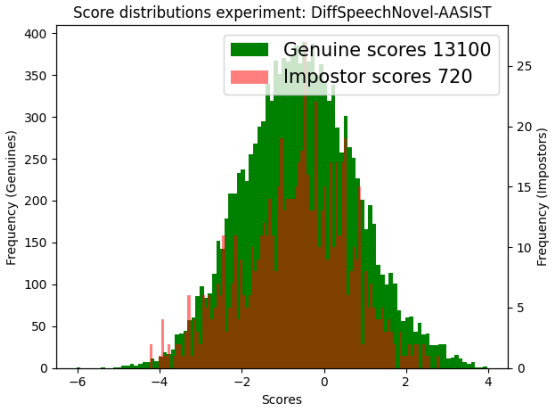


(a) Score distribution

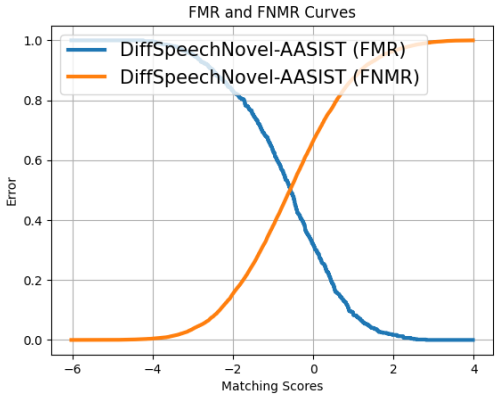


(b) FMR and FNMR curves

Figure 6.9: DiffSpeech trained sentences results from AASIST

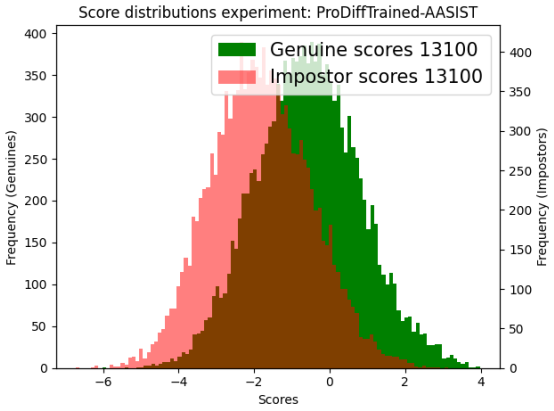


(a) Score distribution

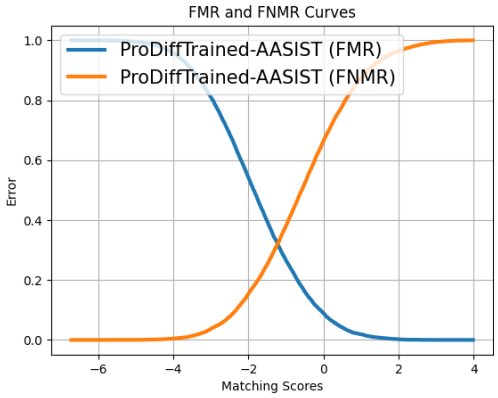


(b) FMR and FNMR Curves

Figure 6.10: DiffSpeech novel sentences results from AASIST



(a) Score distribution

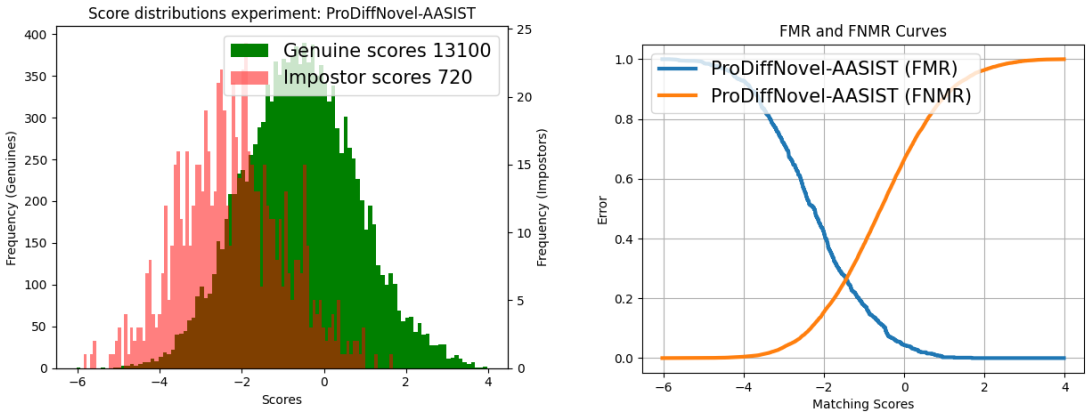


(b) FMR and FNMR Curves

Figure 6.11: ProDiff trained sentences results from AASIST

6.6.4 ProDiff novel sentences

This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test can are displayed on the score distribution graph 6.12a and FMR and FNMR 6.12b. The EER is 26.28 percent.

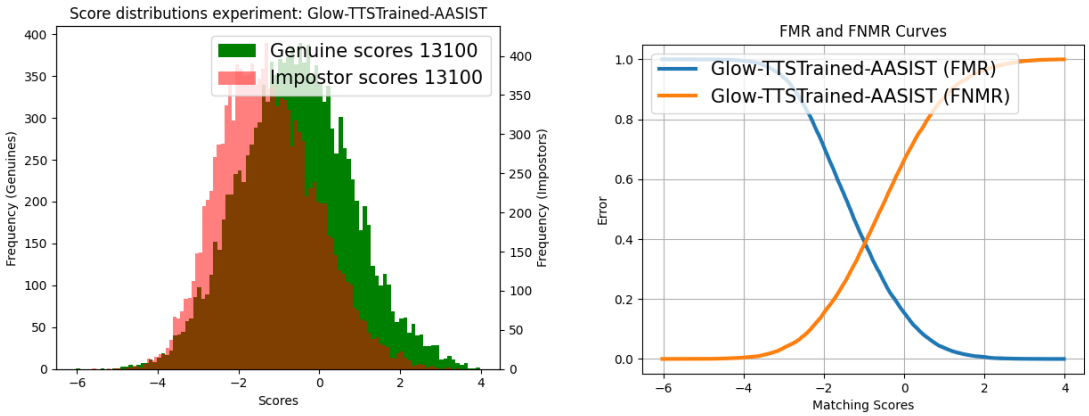


(a) Score distribution (b) FMR and FNMR Curves

Figure 6.12: ProDiff novel sentences results from AASIST

6.6.5 Glow-TTS trained sentences

This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.13a and FMR and FNMR graph 6.13b. The EER is 38.66 percent.

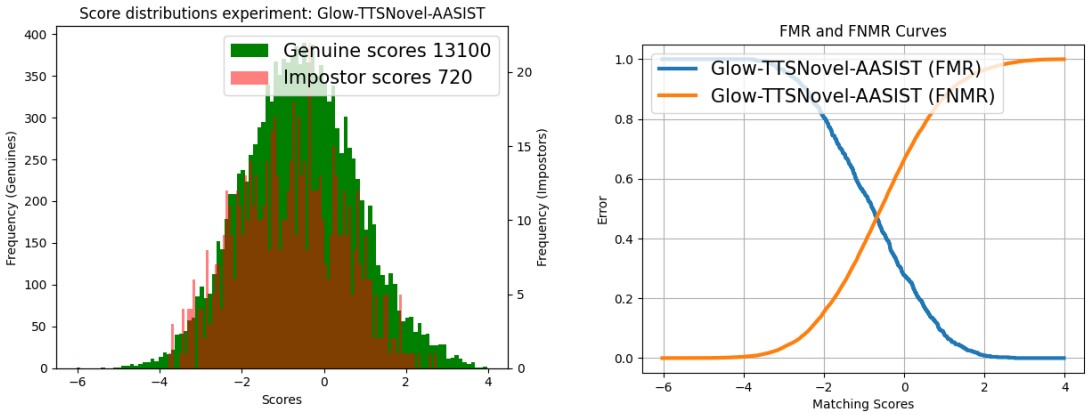


(a) Score distribution (b) FMR and FNMR Curves

Figure 6.13: Glow-TTS trained sentences results from AASIST

6.6.6 Glow-TTS novel sentences

This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test can are displayed on the score distribution graph 6.14a and FMR and FNMR graph 6.14b. The EER is 47.08 percent.

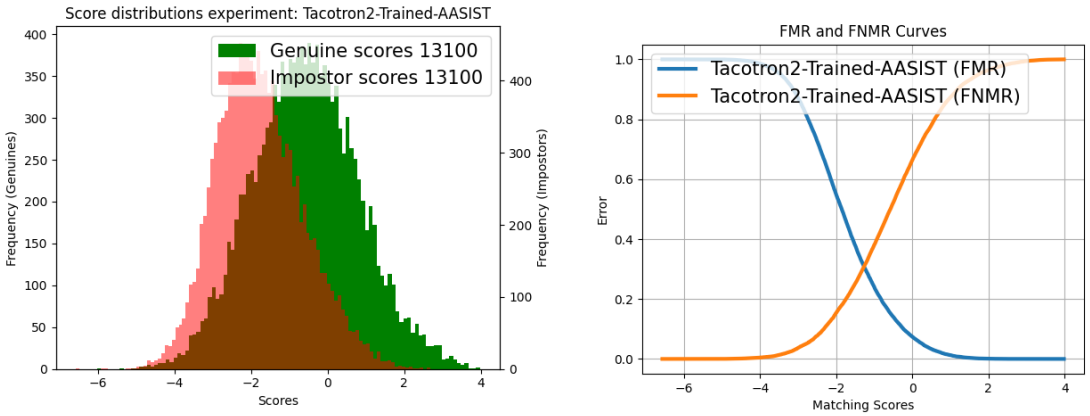


(a) Score distribution (b) FMR and FNMR Curves

Figure 6.14: Glow-TTS novel sentences results from AASIST

6.6.7 Tacotron2 trained sentences

This test consists of 13 100 genuine clips and 13 100 spoofed clips. All spoofed audio clips have been generated using the same transcription as the genuine audio clips. The results of this test can are displayed on the score distribution graph 6.15a and FMR and FNMR graph 6.15b. The EER is 30.86 percent.

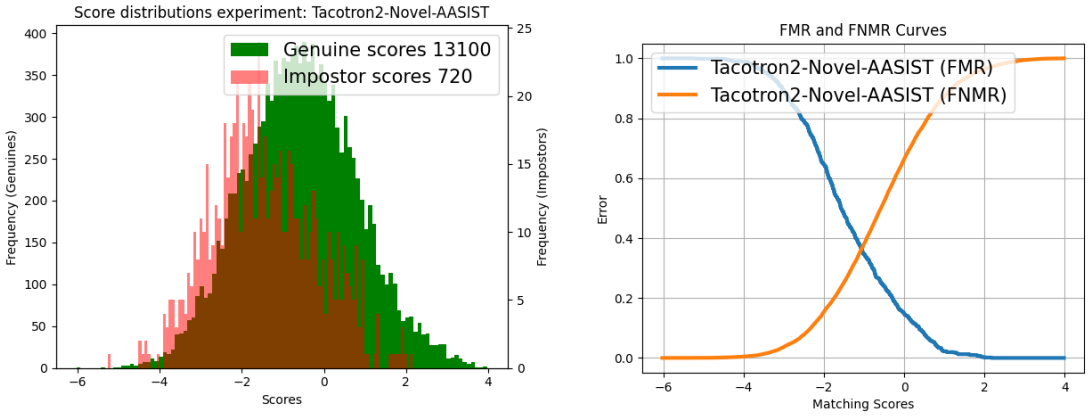


(a) Score distribution (b) FMR and FNMR Curves

Figure 6.15: Tacotron2 trained sentences results from AASIST

6.6.8 Tacotron2 novel sentences

This test consists of 13 100 genuine clips and 720 spoofed clips. All spoofed audio clips are novel sentences that the model was not trained on. The results of this test can are displayed on the score distribution graph 6.16a and FMR and FNMR graph 6.16b. The EER is 36.38 percent.



(a) Score distribution

(b) FMR and FNMR Curves

Figure 6.16: Tacotron2 novel sentences results from AASIST

6.7 Results evaluation

DiffSpeech performed very well against SSL-Antispoof, when synthetizing trained sentences, but its effectiveness dropped sharply with novel sentences, although 11.94 percent EER is still high. DiffSpeech has managed to completely fool AASIST with both trained and novel sentences, as EER around 50 percent is as good as guessing.

ProDiff showed weaker results against SSL-Antispoof with trained sentences and dropped even sharper with novel sentences to only 4.02 percent EER, becoming the lowest rated TTS model against SSL-AntiSpoof. ProDiff demonstrated strong performance against AASIST but not as strong as other TTS models, again becoming lowest rated TTS models against AASIST.

Glow-TTS performed very similarly as ProDiff against SSL-AntiSpoof with trained sentences but did not drop as hard as ProDiff, staying at 10.7 percent EER. Glow-TTS showed strong performance against AASIST with trained sentences and suprisingly even better performance with Novel sentences, managing to render the detector almost completely useless with 47.08 percent EER.

Tacotron2 managed to perform the best against SSL-Antispoof, with both trained and novel sentences, with 29.81 and 19.85 percent EER respectively. Tacotron2 delivered strong results against AASIST as well, with novel sentences again outperforming the trained sentences.

TTS models incorporating Diffusion Models showed very similar performance in spoof detection systems as TTS models not incorporating Diffusion Models, therefore the use of Diffusion Models in speech generation does not have significant impact on the audio spoof detection systems. Additionally, synthesizing trained sentences does provide less convincing results that synthetizing novel sentences in majority of tests.

Even though there is no proof that using Diffusion Models provides an advantage against detection systems, TTS models both with and without Diffusion Models have managed to score high EER in majority of tests, while rendering AASIST compeletely irrelevant in some. This proves the need for more robust audio spoof detectors. The new dataset generated in this study could be used in the development of these enhances detection systems.

Chapter 7

Conclusion

Human voice plays an integral role in our daily lives and with the technology of mimicing human speech advancing to unrecognizable levels there is a need for reliable detection methods for speech deepfakes. The goal of this work was to create a novel dataset, using recent technology - diffusion models. The aim is to evaluate the ability of this novel dataset to evade the detection by deepfake detection systems against the datasets created by previous methods.

This thesis delved into the domain of deepfakes, exploring their various forms and the underlying technologies, particularly focusing on diffusion models and innovative tools that employ these models. This technology has made large strides in the image generation category. We investigated existing datasets consisting of recordings from real people, as well as datasets consisting of generated recordings. We proposed a novel dataset, with the purpose of investigating whether the diffusion models are such a large technological leap that will render the deepfake detectors useless.

In the experiment we have managed to generate large number of deepfake recordings with 4 different TTS tools. Two of which utilized diffusion models and two of which were older more established models using different techniques. These deepfakes generated by the TTS tools were then evaluated by two deepfake detector systems.

The results of this evaluation revealed that diffusion models, while advanced in their capabilities to generate realistic speech deepfakes, did not provide a significant advantage in evading detection compared to traditional TTS technologies. They also revealed that in most cases, novel sentences are more likely to be marked as faked than trained sentences. While that is not surprising, it is surprising that in some cases it had an opposite effect, meaning the novel sentences were less likely to be spotted as fake in these cases.

Bibliography

[1] *IEEE Recommended Practice for Speech Quality Measurements*. 1969. DOI: 10.1109/TAU.1969.1162058.

[2] *Markov Chains* [online]. Brilliant, 2023 [cit. 2024-01-10]. Available at: https://brilliant.org/wiki/markov-chains/?quiz=transition-matrices.

[3] AI, C. *TTS: Text-to-Speech library by Coqui*. GitHub, 2023. Available at: https://github.com/coqui-ai/TTS.

[4] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M. et al. *Common Voice: A Massively-Multilingual Speech Corpus*. 2020.

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

[6] Benzaoui, A., Khaldi, Y., Bouaouina, R., Amrouni, N., Alshazly, H. et al. A Comprehensive survey on ear recognition: Databases, approaches, comparative analysis, and open challenges. *Neurocomputing*. 2023, vol. 537, p. 236–270. DOI: https://doi.org/10.1016/j.neucom.2023.03.040. ISSN 0925-2312. Available at: https://www.sciencedirect.com/science/article/pii/S0925231223002825.

[7] Bird, J. J. and Lotfi, A. *Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion*. 2023.

[8] Cheigh, J. *Generating Images Using VAEs, GANs, and Diffusion Models* [online]. Towards Data Science, may 2023 [cit. 2024-01-07]. Available at: https://towardsdatascience.com/generating-images-using-vaes-gans-and-diffusion-models-48963ddeb2b2.

[9] Chen, T. *Denoising Diffusion-based Generative Modeling: Foundations and Applications* [online]. Medium, june 2023 [cit. 2024-01-05]. Available at: https://medium.com/@timmy90617055/denoising-diffusion-based-generative-modeling-foundations-and-applications-3cab4a4fe374.

[10] Cottingham, J. *Text-to-Speech with Deep Learning: Introduction* [online]. Medium, november 2020 [cit. 2024-04-28]. Available at: https://josephcottingham.medium.com/text-to-speech-with-deep-learning-introduction-9d59b5b700cf.

[11] Frank, J. and Schönherr, L. *WaveFake: A Data Set to Facilitate Audio Deepfake Detection*. 2021.

[12] Gainetdinov, A. *Diffusion Models vs. GANs vs. VAEs: Comparison of Deep Generative Models* [online]. Towards AI, may 2023 [cit. 2024-01-07]. Available at: https://towardsai.net/p/machine-learning/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models.

[13] Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning*. MIT Press, 2016. ISBN 9780262035613. Available at: http://www.deeplearningbook.org.

[14] Goodfellow, I. J., Pouget Abadie, J., Mirza, M., Xu, B., Warde Farley, D. et al. *Generative Adversarial Networks*. 2014.

[15] Google. *Understanding GAN Structure* [online]. Google, 2023 [cit. 2024-02-28]. Available at: https://developers.google.com/machine-learning/gan/gan_structure.

[16] Ho, J., Jain, A. and Abbeel, P. *Denoising Diffusion Probabilistic Models*. 2020.

[17] Huang, R., Lam, M. W. Y., Wang, J., Su, D., Yu, D. et al. *FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis*. 2022.

[18] Huang, R., Zhao, Z., Liu, H., Liu, J., Cui, C. et al. *ProDiff: Progressive Fast Diffusion Model For High-Quality Text-to-Speech*. 2022.

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

[20] Jeong, M., Kim, H., Cheon, S. J., Choi, B. J. and Kim, N. S. *Diff-TTS: A Denoising Diffusion Model for Text-to-Speech*. 2021.

[21] Jung, J. weon, Heo, H.-S., Tak, H., Shim, H. jin, Chung, J. S. et al. *AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks*. 2021.

[22] Khalid, H., Tariq, S., Kim, M. and Woo, S. S. *FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset*. 2022.

[23] Kim, J., Kim, S., Kong, J. and Yoon, S. *Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search*. 2020.

[24] Liu, J., Li, C., Ren, Y., Chen, F. and Zhao, Z. *DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism*. 2022.

[25] Martínez, M. A. *Pyeer: A Python package for biometric performance evaluation*. GitHub, 2021 [cit. 2024-04-25]. Available at: https://github.com/manuelaguadomtz/pyeer.

[26] Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020.

[27] Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A. and Böttinger, K. *Does Audio Deepfake Detection Generalize?* 2022.

[28] Nagrani, A., Chung, J. S. and Zisserman, A. *VoxCeleb: A Large-Scale Speaker Identification Dataset*. ISCA, august 2017. DOI: 10.21437/interspeech.2017-950. Available at: http://dx.doi.org/10.21437/Interspeech.2017-950.

[29] News, B. *YouTube accused of not tackling Musk Bitcoin scam streams* [online]. BBC, june 2022 [cit. 2024-04-12]. Available at: https://www.bbc.com/news/technology-61749120.

[30] News, N. *MrBeast look-alike used in AI-generated TikTok ad to trick fans* [online]. NBC, march 2023 [cit. 2024-04-12]. Available at: https://www.nbcnews.com/tech/mrbeast-ai-tiktok-ad-deepfake-rcna118596.

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

[32] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T. and Kudinov, M. *Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech*. 2021.

[33] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N. et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018.

[34] Sisman, B., Yamagishi, J., King, S. and Li, H. *An Overview of Voice Conversion and its Challenges: From Statistical Modeling to Deep Learning*. 2020.

[35] Sumsub. *Sumsub Research: Global Deepfake Incidents Surge Tenfold from 2022 to 2023* [online]. Sumsub, april 2023 [cit. 2024-04-28]. Available at: https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/.

[36] Tak, H., Todisco, M., Wang, X., Jung, J. weon, Yamagishi, J. et al. *Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation*. 2022.

[37] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A. et al. *ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech*. 2020.

[38] Westerlund, M. The emergence of deepfake technology: A review. *Technology innovation management review*. 2019, vol. 9, no. 11.

[39] Yamagishi. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit* [online]. University of Edinburgh, november 2019 [cit. 2024-03-09]. Available at: https://datashare.ed.ac.uk/handle/10283/3443.

[40] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J. et al. *ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection*. 2021.

[41] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R. et al. *Diffusion Models: A Comprehensive Survey of Methods and Applications*. 2023.

[42] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J. et al. *LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech*. 2019.

[43] Zobaed, S., Rabby, M. F., Hossain, M. I., Hossain, E., Hasan, S. et al. *DeepFakes: Detecting Forged and Synthetic Media Content Using Machine Learning*. 2021.

Appendix A

Contents of the included storage media

- Technical Report - contains source code of the technical report and its compiled version in PDF
- DiffSpeech dataset - all synthetic media generated by DiffSpeech along with authentic media from LJSpeech.
- DiffSpeech metadata.csv - metadata for the DiffSpeech dataset
- ds_ljspeech.py - adjusted source file (original file is called ds.py)
- ProDiff dataset - all synthetic media generated by ProDiff along with authentic media from LJSpeech
- ProDiff metadata.csv - metadata for the ProDiff dataset
- ProDiff_Teacher_LJSpeech.py - adjusted source file (original file is called ProDiff_Teacher.py)
- DiffSpeech SSL Anti-Spoofing - DiffSpeech score files and graphs
- DiffSpeech AASIST - DiffSpeech score files, graphs, and protocol
- ProDiff SSL Anti-Spoofing - ProDiff score files, graphs, and protocol
- ProDiff AASIST - ProDiff score files, graphs, and protocol
- Glow-TTS SSL Anti-Spoofing - Glow-TTS score files, graphs and protocol
- Glow-TTS AASIST - Glow-TTS score files, graphs and protocol
- Tacotron2 SSL Anti-Spoofing - Tacotron2 score files, graphs, and protocol
- Tacotron2 AASIST - Tacotron2 score files, graphs, and protocol