BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

DIFFERENTIAL-BASED DEEPFAKE SPEECH DETECTION

DIFERENČNÍ DETEKCE DEEPFAKE ŘEČI

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. VOJTĚCH STANĚK**
AUTOR PRÁCE

SUPERVISOR **Ing. ANTON FIRC**
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

152826

Institut:	Department of Intelligent Systems (DITS)
Student:	**Staněk Vojtěch, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Cybersecurity
Title:	**Differential-based deepfake speech detection**
Category:	Security
Academic year:	2023/24

Assignment:

1. Get familiar with deepfakes and methods for their detection. Focus on deepfake speech.
2. Study differential detection based approaches for detecting morphing attacks and analyze their technical implementation and quality.
3. Design a custom solution for deepfake speech detection based on differential detection.
4. Implement the proposed solution.
5. Test the accuracy and robustness of your implementation and compare it with other state-of-the-art deepfake speech detectors (at least three).
6. Discuss the feasibility of this approach to deepfake speech detection and its advantages and disadvantages.

Literature:

- FIRC Anton, MALINKA Kamil a HANÁČEK Petr. Creation and detection of malicious synthetic media - a preliminary survey on deepfakes. In: *Sborník příspevků z 54. konference EurOpen.CZ, 28.5.-1.6.2022*. Radešín, 2022, s. 125-145. ISBN 978-80-86583-34-1.
- Handbook of Digital Face Manipulation and Detection. (2022). In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), Advances in Computer Vision and Pattern Recognition. Springer International Publishing. https://doi.org/10.1007/978-3-030-87664-7
- SCHERHAG, Ulrich, et al. Deep face representations for differential morphing attack detection. *IEEE transactions on information forensics and security*, 2020, 15: 3625-3639.

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:	17.5.2024
Approval date:	6.11.2023

Abstract

Deepfake speech technology, which can create highly realistic fake audio, poses significant challenges, from enabling multi-million dollar scams to complicating legal evidence's reliability. This work introduces a novel method for detecting such deepfakes by leveraging bonafide speech samples. Unlike previous strategies, the approach uses trusted ground truth speech samples to identify spoofs, providing critical information that common methods lack. By comparing the bonafide samples with potentially manipulated ones, the aim is to effectively and reliably determine the authenticity of the speech. Results suggest that this innovative approach could be a valuable tool in identifying deepfake speech, especially recordings created using Voice Conversion techniques, offering a new line of defence against this emerging threat.

Abstrakt

Technologie deepfake řeči umožňuje vytvářet velmi realistické sytnetické nahrávky. Tato možnost představuje významné riziko, neboť hrozí její zneužití v mnoha oblastech, od milionových podvodů po rozporování pravosti důkazních materiálů. Tato diplomová práce představuje inovativní metodu pro detekci takových deepfake nahrávek, a to s využitím reálných nahrávek řečníka. Na rozdíl od ostatních přísupů využívá pravé nahrávky k získání důležité dodatečné informace o mluvčím. Porovnáním opravdových nahrávek s potenciálně upravenými nebo vygenerovanými lze efektivně a spolehlivě určit pravost řeči na nahrávce. Dosavadní výsledky ukazují, že tento inovativní přístup může být hodnotným nástrojem v rozpoznávání deepfake řeči, zejména nahrávek vytvořených s využitím technologie konverze hlasu (*voice conversion*), čímž nabízí zcela nový způsob obrany proti hrozbě deepfake zločinů.

Keywords

Differential detection, Deepfake, Synthetic speech, Deepfake Detection, AI

Klíčová slova

Diferenční detekce, Deepfake, Syntetická řeč, Detekce deepfake, AI

Reference

STANĚK, Vojtěch. *Differential-based deepfake speech detection.* Brno, 2023. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Anton Firc

Rozšířený abstrakt

Příchod *deepfakes* přináší řadu dosud bezprecedentních výzev. Tato převratná technologie, známá a medializovaná kvůli svému potenciálnímu zneužití, je dvojsečnou zbraní v digitálním prostoru. Na jednu stranu je možné neškodné využití – pro zábavu s přáteli, ve veřejném audiovizuálním vysílání, ke zprostředkování nevídaných představení a zážitků. Nástup umělé inteligence (AI) obecně může sloužit i k prospěšným cílům, jako je například prostředí filmu, ale také v oblasti vzdělávání, kde může strojové učení zrychlit výukový proces a zároveň šetřit čas a peníze učitelů. Na druhou stranu, případy zneužití deepfake technologie ke škodlivým účelům nejsou bohužel ojedinělé. Známy jsou různě závažné případy, počínaje finančními podvody (a to jak velkých společností s vysokým obratem, tak hanebné zneužití lidské laskavosti a dobrosrdečnosti zranitelných osob) a konče rozsáhlými dezinformačními kampaněmi s cílem ovlivnit geopolitickou situaci ve světě.

Obavy z dopadu deepfake technologie se týkají i státních orgánů a jednotek. Případy podvrhů podkopávají důvěru v digitální materiály. Zejména palčivý problém nastává, když je během důkazního řízení napadena pravost nahrávky, fotografie nebo videa s tvrzením, že médium bylo upraveno nebo kompletně vygenerováno za pomocí deepfake technologie. V takových případech padá důkazní břemeno na forenzní experty, aby dokázali, zda se jedná o médium pravé či nepravé.

Tato práce se zaměřuje na řečové deepfakes. Tradičním přístupem k odhalení pravosti řečové nahrávky je použití jednoho z moderních detektorů. Ačkoliv je tato metoda efektivní a používaná, tyto nástroje běžně pracují pouze s jedním testovaným vzorkem bez výhody porovnání s pravou nahrávkou toho stejného řečníka. Toto omezení implikuje otázku: mohla by dodatečná informace ve formě reálné, pravé nahrávky řečníka zlepšit přesnost a spolehlivost deepfake detektorů?

Inspirováno diferenční detekcí morphingových útoků v doméně obličejů, tato práce představuje první diferenční detektor deepfake řeči, který zahrnuje zpracování pravé nahrávky zároveň s testovanou nahrávkou k odhalení podvrhů. Pravou nahrávku lze v mnoha případech jednoduše získat – biometrická kontrola na letišti, výslech u policie, předkládání důkazů u soudu apod. Zpracováním obou nahrávek zároveň je možné získat důležité informace o rozdílech mezi nahrávkami, které běžné metody získat nemohou. Nejen, že tento přístup koresponduje s praktickými potřebami orgánů činných v trestním řízení, ale otevírá bránu k novému paradigmatu v boji proti řečovým deepfakes.

Navrhnuté systémy pracují zhruba následovně: nejprve jsou z nahrávek vyextrahovány příznaky, tzv. *speaker embeddings* pomocí systému XLSR založeného na Wav2Vec 2.0 a poté sdruženy pomocí MHFA (*Multi-head Factorized Attentive pooling*). Následně jsou získané vektory příznaků zkombinovány do jednoho vektoru představující rozdíly mezi nahrávkami. Rozdílový vektor je následně zpracován jednoduchou dopřednou neuronovou sítí, která testovaný vzorek klasifikuje jako pravý nebo podvrh.

Prozkoumány jsou dvě hlavní strategie pro kombinaci příznakových vektorů. *Diferenční* přístup používá metriku, která určuje rozdíl mezi vektory. Testovanými metrikami je rozdíl $A - B$, absolutní rozdíl $|A - B|$ a kvadratický rozdíl $(A - B)^2$. Při kombinaci *řetězením* (konkatenací) jsou příznakové vektory zřetězeny za sebe v různých fázích zpracování, jmenovitě tedy před extrakcí příznaků, po extrakci příznaků ale před sdružením MHFA a po sdružení MHFA. Experimentováno je také s buňkami LSTM, které po rozšíření a opravení původního návrhu dosahují nejmenší chybovosti.

Vyhodnocení odhaluje nové zajímavé poznatky. Nejprve je dokázáno, že použití detektoru s využitím pravé nahrávky je validní přístup, aplikovatelný v mnoha reálných situacích. Přidaná informace v referenční nahrávce pozitivně ovlivňuje výkonnost systému, zejména

při detekci nahrávek upravených nebo vytvořených pomocí technologie klonování hlasu (Voice Conversion – VC). Jedná se o zásadní objev, neboť moderní detektory nerozlišují původ a způsob vzniku syntetických nahrávek. Ukazuje se však, že by mohlo být výhodné použít některé detekční přístupy jen pro určitý typ deepfake řeči.

Dalším důležitým zjištěním je zvýšená schopnost generalizace párových detektorů v porovnání s konvenčními. Zatímco systémy pracující pouze s jednou testovanou nahrávkou dosahují lepších výsledků na datech se známým rozdělením, párové systémy vykazují menší chybovost na neznámých datech. Tato skutečnost otevírá slibnou možnost navazujícího výzkumu, konkrétně v oblasti přizpůsobení a odolnosti diferenčních a řetězících modelů.

Hledání strategie kombinování informací ze dvou vstupních nahrávek je další zajímavou oblastí, jak vylepšit párové systémy – nové poznatky mohou být potenciálně využity i v jiných odvětvích, např. při vytváření fúzí klasifikátorů a systémů. Dále je možné experimentovat s novými technikami využití *attention* mechanismu, ne nutně jen v LSTM buňkách, které by mohly nadále rozšířit detekční schopnosti představených systémů. Stejně tak by bylo vhodné zkoumat možnosti sloučení diferenčních a řetězících modelů tak, aby vzájemně pokryly slabé stránky druhého přístupu a společně dosáhly lepších výsledků.

Differential-based deepfake speech detection

Declaration

I hereby declare that this Master'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. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

.....
Vojtěch Staněk
May 9, 2024

Acknowledgements

I would like to thank everyone who supported me during the darker times of my life and during the writing of this thesis – most importantly my family, but also life-long friends who never let me down. I would like to especially thank my supervisor Ing. Anton Firc for his wise guidance, scientific insight, practical advices, but also patience and effort in individual approach, for which I am very grateful.

Contents

List of Figures

List of Tables

List of abbreviations

API Application Programming Interface. 18

CNN Convolutional Neural Network. 11, 17, 25–28

CQCC Constant-Q Cepstral Coefficients. 25

CRNN Convolutional Recurrent Neural Network. 26

DF DeepFake. 27, 42, 43

DNN Deep Neural Network. 25, 26, 30

EER Equal Error Rate. 38, 40, 42, 43, 46, 50

GAN Generative Adversarial Net. 12, 13, 15, 17, 19, 28

GAT Graph Attention Networks. 27, 43

GMM Gaussian Mixture Model. 13, 25–27

GPT Generative Pre-trained Transformer, a text-synthesis model. 20, 26

HMM Hidden Markov Models. 11, 13

LA Logical Access. 27, 36, 43

LFCC Linear Frequency Cepstral Coefficients. 24, 25

LSTM Long Short-Term Memory, a type of neural cell used e.g. in Recurrent Neural Networks. 4, 5, 26, 32, 34, 46, 47, 53

MFCC Mel Frequency Cepstral Coefficients. 24, 25

MHFA Multi-head Factorized Attentive pooling. 35, 47

ML Machine Learning. 24, 25, 34

MLP Multi Layer Perceptron, equivalent to feedforward neural network. 26

ReLU Rectified Linear Unit, an activation function used in classification neural networks. 36

RNN Recurrent Neural Network. 14, 17, 26

SSL Self-Supervised Learning. 25, 34, 43, 46

SVM Support Vector Machines. 25, 26, 29

TDNN Time Delay Neural Network. 26

TTS Text-to-Speech (system). 9, 10, 24, 25, 27, 30, 47, 50, 51

UI User Interface. 9

USD United States Dollar. 21

VC Voice Conversion. 3, 8, 9, 12, 13, 24, 27, 30, 40, 47, 50, 51

VCC Voice Conversion Challenge. 27

XLSR Cross-language Speech Representation, a Wav2Vec2 based model. 34, 36

Chapter 1

Introduction

In recent times, AI-powered systems showcased their powerful capabilities in producing high-quality deepfakes in many domains of digital media – non-existent faces, familiar voices speaking words they never uttered [27], videos of events that never happened [11], an essay that was never written [59]. All of these are real possibilities of deepfakes created with remarkable fidelity. The synthesized media can be indistinguishable by human perception, and even competing AI systems have trouble reliably exposing them [64, 106].

Digitally created or modified media can be used in a plethora of situations. On one hand, the technology can be used for harmless entertainment purposes. One such example might be modifying an individual's voice in the digital space. The process is straightforward; leveraging one of the available tools generally suffices for quick personal amusement [30]. Another example might be exploiting the possibilities of this technology in broadcasts and shows to provide unprecedented experiences for the audience, seemingly making the impossible a reality [85].

On the other hand, deepfakes can also be used with wicked intentions, dishonoring a specific person or a group of people, deceiving individuals, and luring money or benefits. For example, a deceptive advertisement imitating a famous individual circulated on social media platforms. While the impact of the fraudulent deepfake campaign is unknown, the incident caught the attention of mass media and instigated a wide-ranging social debate about the truthfulness of digital media [35]. There are also worries about broadcasting fake news or spreading disinformation, which we can already observe today. In addition, trust in online content is inevitably declining [75].

The side-effect of the medialized misuse of deepfake technology is undermining legal evidence. The integrity of the media can be contested, claiming it is a deepfake or digitally modified. In such scenarios, confidence in audio and visual media, the foundation of evidentiary material, falls apart, and the responsibility may be delegated to the country's legal apparatus to prove or refute the authenticity of the medium.

The traditional approach to combat the mentioned problems may involve deploying state-of-the-art deepfake speech detectors. While effective, these tools typically operate by analyzing isolated input samples without the benefit of direct comparison to verified sources. This limitation poses a question: could the inclusion of reference or ground-truth speech samples as a basis for comparison enhance the accuracy and reliability of deepfake detection?

This thesis presents the first differential-based deepfake speech detector, which incorporates trusted ground-truth speech samples to identify spoofs, providing critical information that standard methods lack. A trusted sample can be easily obtained – such cases might

include biometric checks at an airport or police questioning [78]. By employing a differential detection approach, the designed systems aim to leverage examined samples and their ground-truth complements to enhance the detection process. This method not only aligns with the practical needs of law enforcement agencies but also heralds a novel paradigm in the fight against digital speech manipulation.

Firstly, the feasibility of typical feature difference schemes for combining the feature vectors is assessed, as typically used in the facial domain, specifically differential morphing attack detection [42]. Secondly, various concatenation methodologies were used to combine the tested and ground-truth recordings.

Through a comprehensive proof-of-concept study, the feasibility of this new approach is validated. The foundational premise of utilizing differential analysis in deepfake speech detection showcases the potential for significant advancements. The differential and pair-based techniques are transferred from the facial to the speech domain. Finally, one of the significant discovered benefits of differential deepfake speech detection is its superior ability to detect Voice Conversion (VC) samples when compared to their single-input counterparts.

Chapter 2

Speech, facial and other deepfakes

The term deepfake has no agreed-upon definition, it is however an amalgamation of two terms that describe the meaning. The first part comes from *deep learning*, alluding to models of deep neural networks. The second part is *fake*, clearly describing the untruthfulness of the media [3]. The term deepfakes can be used in various circumstances, referring to almost any media – visual, auditory, textual, etc. [4].

This chapter is dedicated to exploring the types of deepfakes and their fabrication. Firstly, speech deepfakes are discussed in Section 2.1. Overviewed are the approaches of text-to-speech, voice conversion, and speech morphing, as well as some popular system architectures used for the creation.

Secondly, facial deepfakes are presented in Section 2.2. Various categories of visual manipulations are explored: face synthesis, face manipulation, face swapping, and face morphing. The section does not delve into immense detail, as the purpose is to give a sufficient overview as an analogy to audio deepfakes.

Finally, deepfake videos are briefly mentioned, merging the audio and visual domain. Section 2.3 further presents the possibility of 3D deepfakes and synthetic text (including the notorious ChatGPT). Ultimately, the combination of textual and audiovisual deepfakes is addressed, highlighting fake news as a real and relevant problem today.

2.1 Speech deepfakes

The first presented deepfake modality is synthetic audio recordings, which harness the power of machine learning to fabricate or manipulate speech recordings beyond the recognition of an average person. The recent advancements in technology enable the creation of realistic vocal imitations that challenge our perception of authenticity while becoming progressively more accessible. Many of the available tools have intuitive user interfaces (UI), which can be leveraged by anyone, even non-technically inclined individuals [30].

There are two prevalent methods for producing artificial auditory recordings. As the name suggests, **Text-to-speech** (TTS) systems generate new speech (acoustic waveforms) from text. TTS systems are usually used to produce new sentences automatically. Therefore, such systems should be able to process potentially any given text and transform it to equivalent phonetic representation [77].

On the other hand, **Voice Conversion** (VC) is a mechanism of creating artificial speech to render the words uttered by a source person sound as if the exact words were spoken by a different voice, i.e., the target person. In comparison with TTS systems, VC could also

be labeled as a *Speech-to-speech translation* system [58]. Both of the mentioned methods are further discussed in this section.

In addition, **Speech Morphing** is relatively uncharted territory in audio deepfake creation. While similar to and often interchanged with Voice Conversion, the principal difference between the two is the intended result: Voice Conversion aims to modify an utterance of a source speaker to imitate the sound of a target speaker, whereas Speech Morphing combines two voices to create an intermediate one [32].

For completeness's sake, another approach for faking speech worth mentioning is **Replay Attack**. While not directly connected to deepfakes, this technique consists of playing a recording of a target speaker to a speaker verification system – for example, using a recording to unlock a smartphone secured by a voice access control system. This does not directly relate to this thesis's main topic and aim. It is a topic for different research; therefore, replay attacks will not be further discussed [110].

2.1.1 Text-to-speech

Parts of this subsection were taken over from the article *An Overview of Text-To-Speech Synthesis Techniques* [77] from 2010 by Rashad et al.

The earliest evidence of TTS systems dates back to the 18th century (i.e., long before the invention of electronic signal processing) and vastly differs from current-age systems.[1] The older sibling of the two predominant synthetic speech fabrication methods aims to create a sound not only intelligible but also unrecognizable from genuine human speech [57]. There are many appropriate applications, such as navigation systems or smart assistants. Unfortunately, a good-quality product could also be used for malicious purposes, which is the primary motivation for developing deepfake detectors like the ones discussed in this thesis.

There are multiple techniques and approaches for converting text to speech. **Formant synthesis**, as the name suggests, utilizes formants, i.e., resonance frequencies of the vocal cords and phonation system, to model the pole frequencies[2] of the resulting signal. Since the formants constitute the prominent frequencies that give a specific sound its traits, speech is synthesized using these estimated frequencies.

During **Articulatory synthesis**, speech is generated by directly modeling the human articulatory characteristics. Using this approach usually does not yield competitive results compared to the other techniques because, in practice, the method is one of the most difficult to implement. There are many articulatory control parameters, including many variables for the formation and position of lips and tongue.

The primary challenge for both formant and articulatory synthesis is not generating the speech itself but instead finding the appropriate parameters. To overcome the limitation, the most popular technique nowadays called **Concatenative synthesis** follows a different approach. Small, prerecorded speech units are concatenated into a resulting utterance, as seen in Figure 2.1. The length of the unit directly affects the quality of the outcome – the longer the units, the fewer parts ought to be concatenated, which results in a more natural-sounding audio. The downside is the need for more memory and units, which quickly become numerous.

[1] The system consisted of a mechanical model of a vocal tract that could produce vowel sounds. More information can be found, for example, on https://speechify.com/blog/history-of-text-to-speech/

[2] In simple terms, poles are associated with the resonant frequencies that shape the spectrum of the synthesized sound. The pole frequencies determine the characteristics of the formants, which are essential for creating realistic and intelligible synthetic speech.

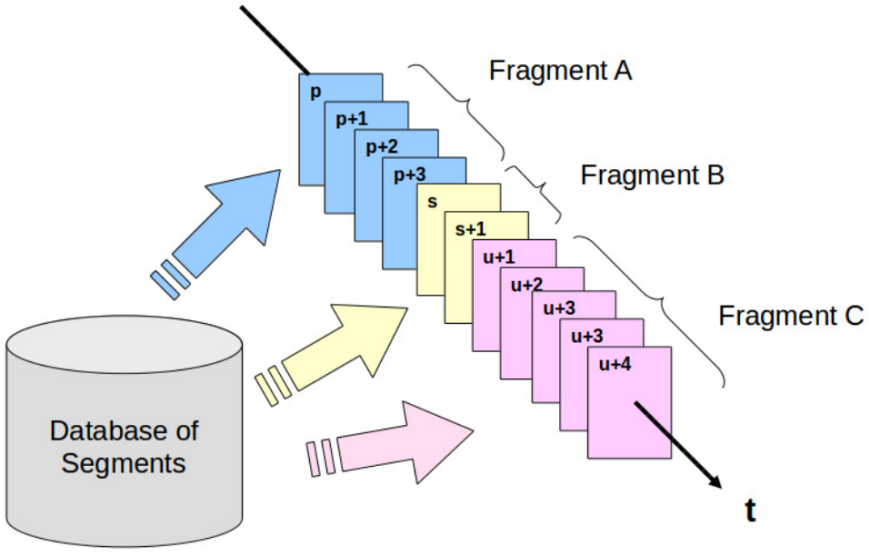


Figure 2.1: Concatenative synthesis illustration. Speech units extracted from a segment database are combined to create a sequence of frames to synthesize the resulting utterance. Image taken from *Two-Dimensional Expressive Speech Animation* [21] by Costa.

Additionally, multiple instances of each unit with varying expressions can be stored and used based on the context of a synthesized sentence. This approach is referred to as **Corpus-based (concatenative) synthesis**, or also **Unit selection synthesis**. This modification makes the process more complex, as an algorithm for selecting the most fitting unit is needed. However, the slight added burden is outweighed by significantly enhancing the resulting naturalness of the synthesized voice and overall quality.

The alternative to unit selection synthesis is to use parameterized and statistical techniques. Therefore, **Hidden Markov Models** (HMM) were predominantly used for a notable period. This stochastic finite-state machine model was used for various speech applications [57] (for example, voice conversion as well as for TTS). The main advantage against previously mentioned approaches is a significant reduction in memory requirements for storing the model parameters.

However, being a linear model, HMMs were recently superseded by **neural networks**. Those models usually consist of two stages, as presented in Figure 2.2 – text-to-spectrogram (also called *acoustic model* or *encoder*) and spectrogram-to-speech (also called *vocoder*), where the vocoder takes the spectrogram generated by the acoustic model to produce the resulting waveform [91]. There are several types of spectrograms representing the frequency spectrum of a signal, and the prevalent ones are *Mel-spectrograms*, which use a logarithmic-like mel-scale to simulate the human perception of sound better [57]. Probably the most famous example of these modern systems is from 2016 – WaveNet [67], which reportedly outperformed other contemporary systems [43]. WaveNet is an autoregressive CNN, meaning when an output sample is predicted, it is fed back to the input, influencing the generation of the following samples.

While autoregressive models are potent and influential, the forefront of recently published architectures consists of non-autoregressive models. Few-shot multi-speaker systems are another modern approach which seems to dominate the TTS field – only a short record-

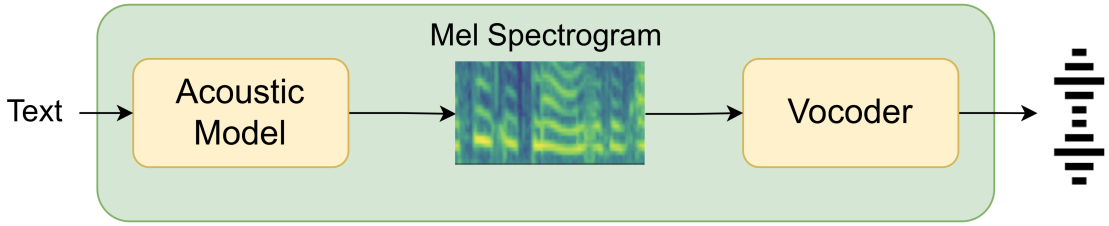


Figure 2.2: Two stage TTS system diagram. An acoustic model converts the input text to a spectrogram, which is, in turn, processed by a vocoder to generate waveforms. Image taken from *ParrotTTS* [86] by Shah et al.

ing of a target speaker is needed to synthesize speech in its voice. Similarly to facial deepfakes, approaches employing adversarial training schemes and GANs are also being used [32].

The latest trend in TTS is **end-to-end systems**, which enable the replication of voices with remarkable fidelity. These deep neural networks do not use explicit intermediate representations such as spectrograms and are trained to transform text to audio directly. The advantages of joint training, as opposed to separate training, are apparent – the systems are less prone to errors from each component as only one single component needs to be fine-tuned. On top of that, the data preparation phase is less demanding, and the training process is significantly simplified [57].

2.1.2 Voice Conversion

As mentioned in the section introduction, VC is a method employed to alter the characteristics of a provided speech from a source speaker to align with the vocal attributes of a target speaker. Different techniques can be used to achieve this feat; for example, pitch shifting, contraction or dilation in the time domain, frequency wrapping, or applying various sound filters [1]. These modifications are generally referred to as *Voice Transformation* [58]. Some cutting-edge VC frameworks have the ability to independently process and modify various speech components, such as pitch, rhythm, or language content [76, 32].

Voice conversion is well known to the general public through entertainment means, most notably the film industry – one such example is *Alvin and the Chipmunks*, where the voice actors recordings are modulated (and formant-corrected) to create the notorious chipmunk voice [58]. Additionally, VC usage can be found in online games, voice parodies, remixed songs, etc. Conventional available VC tools are able to alter various characteristics of the source voice (e.g., age, gender) to conceal the true identity of an individual [1].

VC system should account for both *timbre* and *prosody* of source as well as target speaker. Timbre generally refers to a sound's quality, *color* or *texture*, while prosody describes the intonation, stress, and rhythm patterns in a sound [28]. In the context of VC, timbrel features typically pertain to the dynamic spectral envelope of the voice, whereas prosody is attributed by pitch, rhythm, and overall energy of the signal [58].

The problem of voice conversion is usually approached using two primary methods, which differ mainly in the training process. **Parallel VC** relies on a *parallel dataset* to train, i.e., both the source and target recordings contain the same sentence. Training is usually conducted for all combinations of source and target speakers [32, 58]. Firstly, recordings are time-aligned (using, for example, *Dynamic Time Warping*); acoustic features are therefore synchronized and finally concurrently mapped between recordings [58]. The

training and the conversion process are depicted in Figure 2.3. Over the years, methods for parallel VC have evolved significantly, beginning with statistical techniques, expanding into Gaussian Mixture Model (GMM) or HMM models, ultimately reaching the present-day apex of neural networks or classification and regression trees [32].

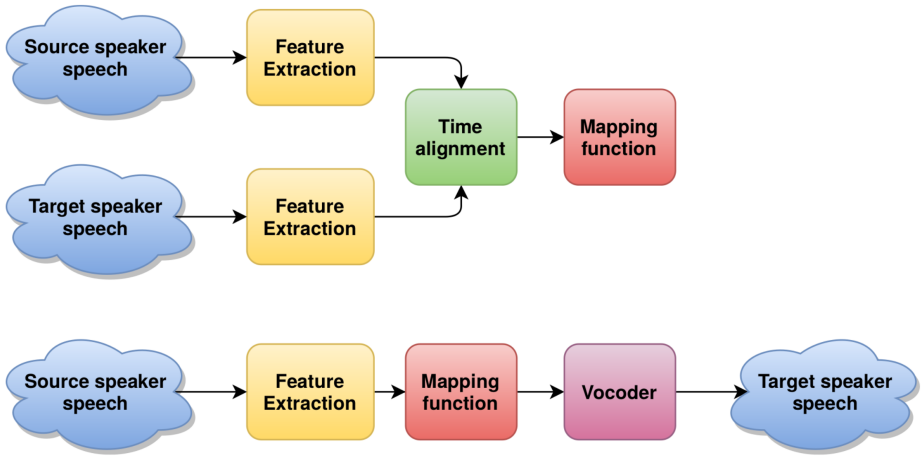


Figure 2.3: Schema of a basic parallel VC system. The top describes the training workflow, and the bottom depicts the conversion process. Image taken from *Non-parallel voice conversion* master's thesis [12] by Brukner.

On the other hand, the more recent **Non-parallel VC** systems do not require identical sentences to be recorded, enabling the utilization of more available datasets without particular necessities. These systems need to learn the mapping of one speaker to another while not being able to rely on aligning similar frames with the same content. For such a complex task, neural network architectures are almost exclusively used, primarily GANs and Autoencoders. In addition, one-shot VC was recently developed; one-shot refers to the robustness property of the system, which needs as little as one short embedding utterance from a speaker to perform the transformation. In effect, these systems are both independent of the dataset and do not require to be trained for a specific speaker, thus allowing voice conversion from a source to a target speaker even if the training set did not contain recordings from any of them [12, 32].

Lastly, an intriguing aspect of VC is related to the *phonetic content* of different languages in general. Employing recordings within the same language or between different languages for both training and actual conversion can significantly impact the result. In this context, it is worth mentioning that within a single language, the sets of phonemes[3] in the source and target recordings are almost identical. However, some phonemes may not correspond to each other in different languages, hence cross-lingual VC brings many new challenges to the field of voice conversion [62].

2.1.3 Speech morphing

As the review of relevant literature shows, speech morphing is often interchanged with voice conversion [1, 68]. While acknowledging the similarity, the difference lies in the expected outcome. VC systems try to make the resulting audio as similar as possible to the target speaker. The primary objective of speech morphing is to achieve a seamless transition

[3]The smallest unit of sound in a language that can distinguish words.

from one sound to another. This involves blending two sounds to create a new signal with an intermediate timbre [13]. The resulting synthetic speech exhibits speaker-dependent characteristics relating to two different speakers whose recordings were used to generate the morph [71].

Speech morphing is usually achieved by interpolating the parameters of the source recordings. Those parameters are obtained primarily from analysis or synthesis techniques such as Linear Predictive Coding, Sinusoidal Model Synthesis, or Short-time Fourier Transform [40]. Modern neural networks are used in one of the few recent publications on speech morphing. Pani et al. [71] propose a technique called Voice Identity Morphing, which uses DeepTalk [20] encoder to extract speaker embeddings from two source identities. Afterward, feature-level fusion is performed, forming a morphed embedding, which is then fed into a *Tacotron 2*-based synthesizer and *WaveRNN* vocoder, both utilizing an RNN architecture to create the resulting morph.

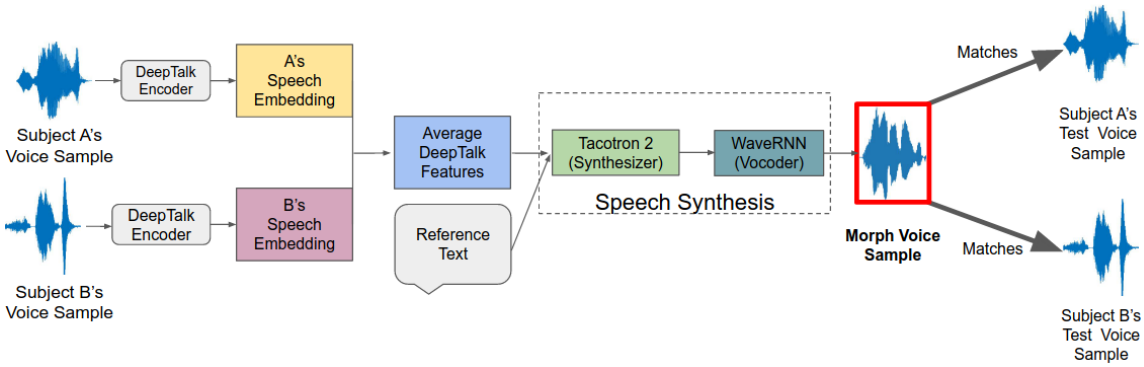


Figure 2.4: Voice Identity Morphing system by Pani et al.; image taken from the article *Voice Morphing: Two Identities in One Voice* [71]. The architecture consists of a DeepTalk encoder, a Tacotron 2-based synthesizer, and a WaveRNN vocoder.

The research in the area of speech morphing fell almost silent over a decade ago [32]. However, there are some emerging concerns regarding the robustness and vulnerability of biometric speaker recognition systems to morphing attacks, motivating the reestablishment of research, particularly in detecting morphed speech samples [71]. This thesis aims to join in, as the resulting differential-based deepfake detection system could also be used to reveal morphed attacks that are otherwise difficult to detect due to their significant similarity to authentic recordings.

2.2 Face deepfakes

The second type of synthetic media discussed in this thesis is deepfakes in the form of images. The primary domain of this visual content is most often the human face, as appearance is crucial for human identity and recognition [63], allowing for easy consumer confusion, both for malicious and entertainment purposes [10].

In addition to tampering with faces, methods for modifying or simulating complete body movement exist. Movie and game studios use, for example, *motion capture and transfer* to create realistic animations of high quality in their products. Using an existing video as a *driver* to transfer the motion to another subject is also possible [14].

Various manipulation techniques for creating graphical deepfakes, commonly split into four categories, are discussed in the following subsections. The European Parliamentary Research Service describes these categories [41] as seen in Table 2.1.

Table 2.1: Overview of techniques used for creating facial deepfakes as described by The European Parliamentary Research Service [41]

Technique	**Description**
Face synthesis Face generation	Creating a partial or entirely new image of people that do not exist. In the vast majority of cases, Generative adversarial networks (GANs) are used to perform such tasks [19].
Face manipulation	Techniques used for modifying specific parts of the target's face. Unlike the other techniques, the identity of the target is persevered. The aim is to, for example, transfer an actor's expression to the target (also called facial reenactment) or adjust the movement of the target's lips using an audio or text input (visual dubbing).
Face swapping	As the name suggests, two different people swap their faces. The face of a target person is replaced by the face of the source person and vice versa.
Face morphing	In general, morphing refers to transforming one image into another. This process can be stopped midway, creating a *morph* – an image resembling both the input images simultaneously.

2.2.1 Face synthesis

For generating parts or even entirely non-existent faces, usually GANs or other convolutional neural networks are used [45]. GANs are a unique type of deep net, where a generative model is confronted against an *adversary*, i.e., a discriminative model that learns to decide whether an image comes from the modeled distribution (newly generated faces) or the training data. These two models compete, where the generative part aims to produce an image as similar as possible to the ones from the training dataset. In contrast, the discriminative model tries to detect fake, generated images reliably. This competition improves both models until the counterfeit images are indistinguishable from the genuine, original ones [36].

GANs can be used in many other fields of interest but remain primarily visible in places where synthesis of images takes place [19]. There are many tools available for generating a new face. Take, for example, the website `https://thispersondoesnotexist.com/`, which produces a new, non-existent, never-before-seen face of a person each time it is loaded in a browser. Figure 2.5 presents two examples of such faces.

2.2.2 Face manipulation

Many titles refer to the same general process, which can be further separated into two categories. While terms such as *face editing*, *face retouching* and *face manipulation* refer to techniques used to modify some attributes of a face, such as color of hair and skin,

Figure 2.5: Two examples of non-existent faces generated by a GAN model from the website https://thispersondoesnotexist.com/

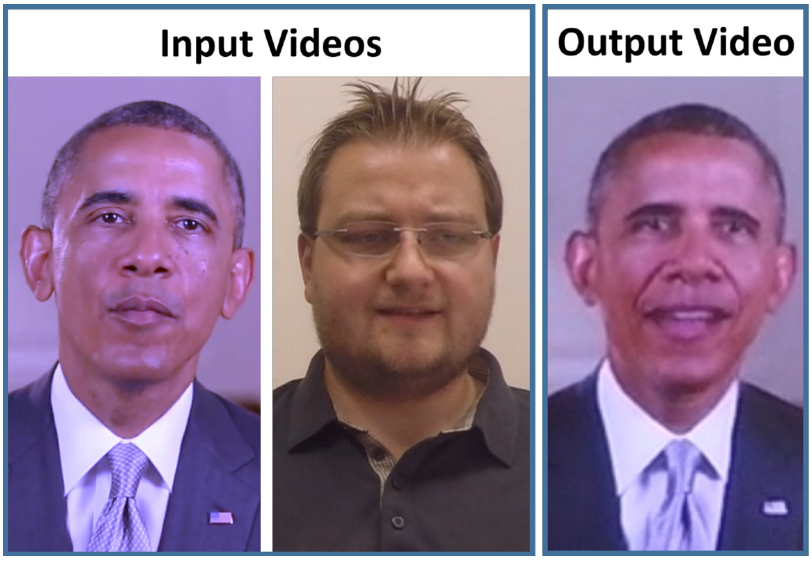


Figure 2.6: Real-time video face reenactment. Image taken from *Demo of Face2Face: Real-Time Face Capture and Reenactment of RGB Videos* [98] by Thies et al.

age, gender, adding glasses etc., *face reenactment* and *expression manipulation*, visible in Figure 2.6, usually describe techniques for altering the facial expression of a subject [98].

An outlying case is *visual dubbing*, also known as *lip-sync*, which denotes the process of adjusting the movement of a target's lips to match an arbitrary input, including written text [79]. Before the rise of deep neural networks and other semi-autonomous and semi-automatic tools and techniques, manual activity was the leading way to carry out several of the mentioned techniques. This process has been branded with the famous name *photoshopping* based on the renowned program Adobe Photoshop, which was (and still is) a well-suited commercial product for such modifications.

Nowadays, a more automatic and reliable approach triumphs over manual labor. Similarly to other discussed categories, the capabilities of GANs are harnessed for the modifications and manipulations, but other technologies are also used. For example, NeuralTextures model utilizing *Deffered neural rendering* for manipulating expressions [97] or recurrent neural networks (RNN) based on Long Short-term Memory (LSTM) cells can be employed in various methods, e.g., to synthesize high-quality lip-sync [79].

2.2.3 Face swapping

Transferring a face from a source image or video onto a target image or video can be challenging. Many factors come into play when attempting to create a realistic and visually appealing result – uneven lighting, various rotations in multiple axes, and complex facial geometry all need to be accounted for.

A wide variety of approaches for face swapping have been discovered. Usually, CNNs are involved in some way to extract facial features and geometry, render the target face, or even create a 3D model for fitting the target face to a different shape. Many of the processes are computationally demanding, however recently, some new applications allowing real-time utilization have been presented [105]. Similarly to other approaches, GANs are also utilized for face swapping [66].

Popular apps are very accessible through conventional means, i.e., through a web browser (for example, Face Swapper[4]), on mobile phones (e.g., Reface[5]) etc. An example of a face-swapped image is presented in Figure 2.7.

2.2.4 Face morphing

The objective of face morphing is to create facial images that resemble multiple identities simultaneously [31]. The resulting image contains features and characteristics of all the input images – even though only two images are commonly combined to produce the morph, as seen in Figure 2.8. Creating this morph can be achieved in a variety of ways, both manually (for example, using GIMP[6] or Adobe Photoshop[7]) or by automatic means (e.g., using triangulation or averaging). Of course, deep learning, specifically GANs and convolutional neural networks (CNN), can also be used for such purposes [24].

Face morphing has been popular for years with a wide range of applications, mainly for the film industry and, more recently, for entertainment purposes on mobile phones. Many available morphing tools allow almost anyone to create decent-quality morphs. Those

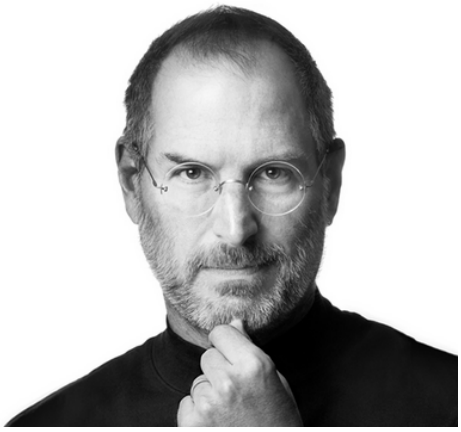
[4] https://faceswapper.ai

[5] https://play.google.com/store/apps/details?id=video.reface.app

[6] GNU Image Manipulation Program, https://www.gimp.org/

[7] In the newer versions, there is a feature called *Generative fill* which leverages deep learning for generating new content. https://www.adobe.com/products/photoshop.html

(a) Source face (Rowan Atkinson)



(b) Target image (Steve Jobs)

(c) Resulting face swap deepfake

Figure 2.7: Swapping the face of Rowan Atkinson onto Steve Jobs, created by using Stable Diffusion API. Images taken from https://blog.segmind.com/faceswap-api-guide/.

(a) Subject 1 (b) Morph (c) Subject 2

Figure 2.8: An example of face morphing. Images of Subject 1 (a) and Subject 2 (c) are combined to create a morphed image (b). Image taken from *Detection of Face Morphing Attacks Based on PRNU Analysis* [82] by Scherlag et al.

applications are accessible both on smartphones (e.g., Face Morph[8]) and from a web browser (e.g., FaceMorph.me[9]) [82].

2.3 Other deepfake modalities

In the previous sections, both visual and audio deepfakes were discussed separately. It is, however, important to remember that both modalities are being combined for creating complete **deepfake videos**. Some examples have already been presented in Section 2.2 – lip-sync or visual dubbing [116], full body puppetry [14] as seen in Figure 2.9, etc. Practically anyone is able to use the available tools to produce an authentic-looking deepfake video, mainly utilizing the power of GANs and Autoencoders. This raises several concerns, as even recently, audiovisual material used to be considered strong evidence of captured information. Many are starting to be skeptical in this regard, and the authenticity of videos should be questioned [50].

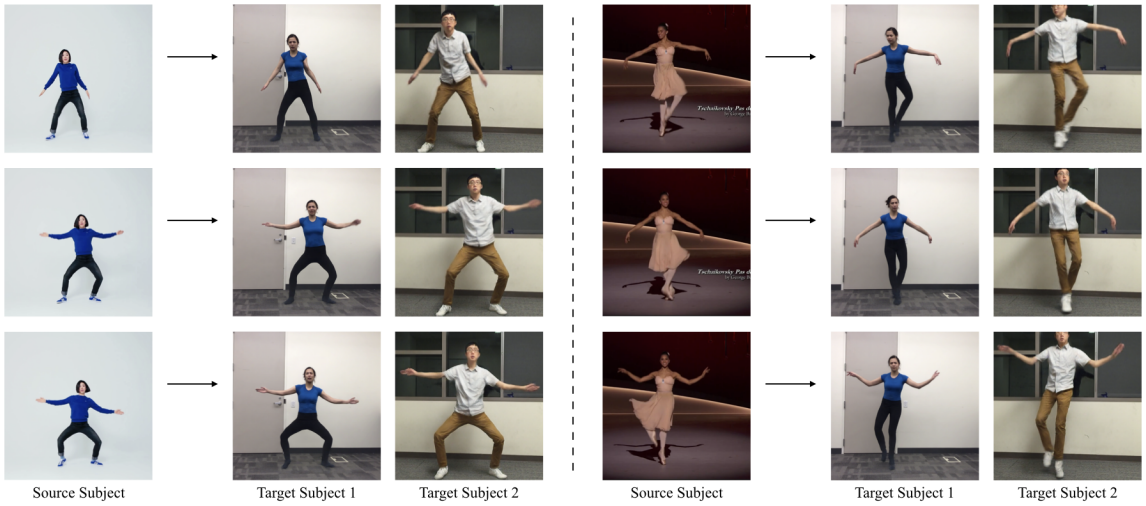


Figure 2.9: Everybody Dance Now by Chan et al. [14] showcasing full-body puppetry, image taken from the project's presentation website[10].

In addition, it is possible to leverage 3D animation technology together with capturing the desired movement to create virtual **3D deepfakes**. Technological giants like Facebook are investing in development in this field, which has already yielded results. For example, Facebook Codec Avatar allows users to control a 3D realistic but virtual representation of themselves in virtual or augmented reality [41].

Another vital area of deepfake landscape is language models and generated **synthetic text** with high linguistic quality. There are many applications, be it for entertainment purposes, text summarization, or more malicious generation of misinformation and fake online reviews to gain immoral, unfair business advantage [112]. Text synthesis technology can also be used to imitate a person's specific writing or speaking style. Language models lean heavily on natural language processing to 'understand' not only words or sentences, but also the emotions and intentions expressed by the text [41].

[8] https://play.google.com/store/apps/details?id=com.hamsoft.face.morph
[9] https://facemorph.me/
[10] https://carolineec.github.io/everybody_dance_now/

The most famous models for text generation are versions of Generative Pre-trained Transformer (GPT), including its freely available online chatbot variant ChatGPT created by OpenAI [75]. The building stone of this architecture is a deep learning algorithm called the *Transformer*, which essentially transforms input text into a new text. The model learns how words and sentences relate to each other within different contexts [41]. In practice, the decision on what words to generate is based on the highest appearance probability, determined by the learned parameters and input text.

Both deepfake videos and synthesized text alone, but even more so when combined, can be used to create convincing **fake news** articles. This can significantly simplify conducting large-scale disinformation campaigns, which can even be observed nowadays. Furthermore, further decline in trusting online content is inevitable [75]. On a slightly brighter note, fake news detectors already have promising performance [34, 114, 117]. Social media platforms have also issued several guidelines to prevent the distribution of fake news and, in addition, implemented automatic means to disrupt the spread of disinformation even further [60].

2.4 Deepfakes – toy, tool, threat?

Deepfakes are a relevant problem for today's society, as the resulting fake media are being used for legitimate but also malicious purposes; nowadays, it is primarily fake news and spreading misinformation [30]. In the context of this thesis, the field of speech deepfakes is essential. Still, other modalities are also relevant – the practices are similar in many respects, so one approach could influence the others.

Deepfake technology can be used as a solution for real-world problems, especially in the film industry. For example, Hollywood studios capitalize on the ability of de-aging older star actors and actresses[11]. Additionally, modeling a complete digital image based on previous appearances[12] is also plausible. [39]

Secondly, commercial application of deepfakes enables a portrayal of a generally liked, potentially famous person participating in activities that would otherwise require a lot of training and preparation. This burden can be instead shifted to an underlying actor with a specific skill set, who actually performs the stunt or acts out a specific situation. The deepfake of the celebrity can then be transferred to the underlying actor, saving hours of time otherwise needed for training or performing on set. [51]

Additionally, this allows influential people with a significant follower base on social media to provide more personalized content with little effort. For example, recording a targeted message for thousands of individuals without the actual need to record each message separately. In a similar fashion, deepfakes can break down linguistic barriers, enabling a wider reach of a media campaign, appealing to a broader audience all around the globe. One such example can be a deepfake of David Beckham and his *Malaria No More* campaign, which was able to deliver the message in nine different languages, significantly enhancing the reach and influence on various media platforms. [51]

Similarly, deepfake technology was used in the auditory domain to create a voice clone of the pop artist Andy Warhol, who left behind over 20,000 diary pages. These became the source material for a Netflix documentary *The Andy Warhol diaries*, where the artist's thoughts are represented by the created voice clone, significantly enhancing the authenticity of the documentary. What's even more astonishing, *Resemble AI*, the company behind the

[11] For example, the actors Patrick Stewart and Ian McKellen in *X-Men: The Last Stand* [39]

[12] This was utilized in the case of Carrie Fisher passing away during the long filming process of *Rogue One: A Star Wars Story* [39].

deepfake, only had about 3 minutes and 12 seconds of original audio data of the artist. The company successfully extrapolated missing phonemes from the few present in the recordings to create a well-resembling voice, together with the support of a voice actor to fill inconsistencies and help the system bridge the gap to learn the proper delivery. This astounding feat led the film to receive an Emmy nomination for Best Television Documentary. [49, 109]

Apart from legitimate utilization, malicious actors can perpetrate a wide range of fraudulent activities using deepfakes, and unfortunately, these do not occur sparsely. The threat of deepfake attacks is quickly escalating due to the constant advancements of deepfake creation tools, and we are rapidly approaching a situation where almost anyone can create credible deepfakes. On top of that, with the attempts to migrate the deepfake creation suites to end devices, such as smartphones, the availability of deepfake creation would allow a broader spectrum of attackers to exploit the power of digital manipulation at the grasp of their hands. Even more, usable audio and visual media for deepfake creation are easily obtainable from social networks, streaming platforms or even locally, i.e., capturing a person on the street. With the growth of social networks, collecting a sufficient amount of personal images, videos, or recordings is a simple task. Therefore, combining the availability of deepfake creation instruments and the accessibility of personal audiovisual material imposes almost no limits on who or where can create deepfakes, but also makes almost anyone a suitable target for a deepfake attack. [32]

To give a better perspective of the possibilities, several examples of deepfake attacks are presented. The first example of an unethical misuse concerns a high-ranking manager of a multinational corporation who received a phone call from someone impersonating the company's director using a voice clone. The voice ordered authorization for a substantial bank transfer. The manager perceived the request as legitimate because he recognized the voice and the information matched with previous (also deceptive) email correspondence. In the end, the company lost about 35 million USD. [49]

Criminals do not hesitate to exploit a person's kindness and vulnerability. The second example relates to a scam of an elderly lady who believed her grandson was in dire need of help. The criminals asked for an increasing amount of money based on various excuses (car accident, following legal demands). Finally, she checked with her grandson using his real phone number to discover that the presented situations were fictional, unfortunately only after losing a significant financial amount. [27]

The third example differs from the previous ones, as the aim was not a money scam, but rather to influence a geopolitical situation in the world. This pertains to a video deepfake of Ukraine's president Zelenskyy, which was used in a disinformation campaign designed to conceal and justify the actions of Russia. In the video, Zelenskyy is depicted admitting defeat and urging surrender. Hackers published the video on Ukrainian websites and social media, and it also appeared on TV broadcasts. Due to its low audio and visual quality, the video was quickly debunked and removed. This case raises significant concerns regarding a new digital front of battle in future conflicts. [11]

Another interesting application regarding deepfakes is the sophisticated natural language processing systems and large language models, specifically the chatbot model ChatGPT, which vastly exceeded expectations in many areas while keeping track of a broad context of the discussion. The possibility of both misuse and its positive impact has led to discussions about its influence, for example, in education. ChatGPT was able to pass university courses required for a university degree, which might threaten the integrity of exams, submissions, and the quality of the educational process in general. On the other hand, it can accelerate the learning process due to its excellent abilities – this includes text

comprehension and summarization, as well as a personalized approach based on the needs of an individual. Anyone can have a private assistant to discuss problems encountered, boosting performance while saving resources and time for a teacher. [59]

Additionally, it is expected that more and more attacks misusing the combinations of deepfake speech and video will be experienced by the ordinary people, not only targeted on celebrities or high-ranking individuals. One such potential use-case might be intercepting a corporate video meeting (for example a Teams call), where the attacker joins looking and sounding like one of the employees, which could lead to fraud, confidential information leakage, etc. The increasingly better quality of the deepfake and the ability to respond almost instantly allows the attacker to not only listen, but interactively join the conversation and subtly influence the flow to topics focused by the attacker. [32]

The presented examples and expectations only scratch the surface of potential deepfake usage. The technological advancements commence a race between the deepfake creation and detection. While it can be suspected that the development of creating higher-quality media will not stop, it is necessary to continue enhancing the detection to timely uncover as many deepfake attacks as possible, in addition to mitigating the damages caused by such attacks.

As a side note, apart from developing modern detectors, one way anybody can help in the battle against fraudulent deepfake usage is raising awareness. Focusing on the possibly vulnerable groups (for example children and elderly, as well as employees with bank access) might be the first, but also potentially robust defense line against deepfake attacks. Of course, in the long term, a more complex solution might be needed (such as legislative regulation), however, that will probably not come in the nearest future, so at the moment, the burden lies on individuals to take initiative.

Chapter 3

Audio deepfake detection and differential-based approach

While the previous chapter dealt with deepfake creation, it is time to shift focus to, figuratively speaking, the other side of the barricade on the deepfakes' battlefield. In the evolving landscape of digital forensics, the advent of deepfake speech technology poses unprecedented challenges [32]. This technology, renowned for its potential misuse, represents a double-edged sword. On the one hand, it offers innovative advancements in entertainment, education, and customer service through realistic voice synthesis. On the other hand, its malicious applications have drawn significant concern, particularly from Law Enforcement Agencies (LEAs).

These concerns are not unfounded, as deepfake speech technology can undermine the integrity of media evidence, a cornerstone of criminal investigations. A particularly vexing issue arises when defendants contest the authenticity of media evidence, claiming it has been tampered with or entirely fabricated using deepfake technology. In such scenarios, the onus falls on LEAs to conclusively demonstrate the genuineness of the evidence presented.

The traditional approach to tackling this dilemma may involve deploying state-of-the-art deepfake speech detectors. This chapter addresses the endeavor of creating such a detector that needs to be reliable and robust to all the various types, modifications, and qualities audio deepfakes come in.

Firstly, the challenge of detecting deepfake speech is researched in Section 3.1. A general schema of a detection system is presented, and individual stages are further described. Discussed are voice characteristics that can be used as features for discriminating between real and fake recordings. In addition, state-of-the-art system architectures are concisely presented. Covered are both shallow machine learning models and systems utilizing deep neural networks. Unconventional methods are also briefly discussed, including custom features and approaches to classification.

Secondly, differential-based detection is explored in Section 3.3. This area is unexplored in the auditory domain; however, inspiration could be drawn from facial deepfake detection methods. Those are, therefore, discussed in a concise yet comprehensive manner. Mentioned are both approaches of inverting the deepfake creation process and the idea of computing a feature difference vector for classification. Finally, Section 3.4 summarises the essential research takeaways and outlines possible approaches to adopting differential-based detection in the auditory domain.

3.1 Detecting deepfake speech

The availability of deepfake creation tools has recently drawn considerable attention to research regarding synthetic audio detection, even across different languages [4]. This section presents modern and state-of-the-art efforts in detecting manipulated and artificially generated voices. Both conventional machine learning (ML) methods and deep learning approaches utilizing various neural network architectures are introduced.

In the relevant literature, methods for detecting synthetic audio appear not to be exclusively designed for detecting a specific type of deepfake speech, such as TTS, VC, or speech morphing – in fact, quite the opposite. The proposed methods address the detection problem in general, not limited to a single technology, and include all forms of deepfake speech creation [32].

A high-level detector system schema can be seen in Figure 3.1. Generally, it consists of three stages: *feature extraction*, *model training/fitting* and finally *classification*. The complexity of individual stages differs from system to system based on the technologies and approaches in general.

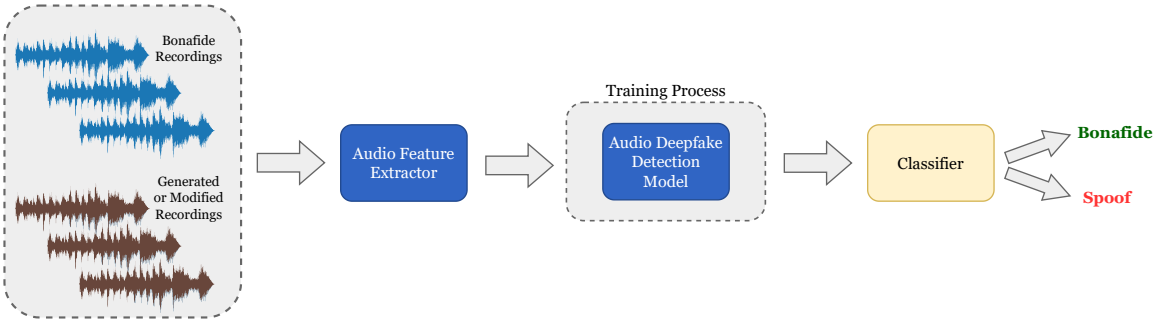


Figure 3.1: A high-level schema of a deepfake speech detector, visualizing the discussed process of feature extraction, model training, and classification.

For the sake of clarity, features, shallow and deep classification models are presented separately, however additional information about the complete mentioned system can be found in the respective references. The discussed features are conveniently shown in Table 3.1. This section is not exhaustive, and many systems have not been described because of the sheer amount; the aim is to picture the diverse landscape of audio deepfake detection and state-of-the-art methods.

Features

One of the popular characteristics to describe speech is **cepstrum**, or more specifically **cepstral coefficients** (which collectively make up the cepstrum), which are a representation of the short-term power spectrum of a signal. The first commonly used variant is Mel Frequency Cepstral Coefficients (MFCC) utilizing a mel-scale, which better corresponds to human perception of sound [57]. It is widely used in many systems [2, 5, 54] and is considered a de facto standard when comparing systems using different features [4, 89]. Similarly, Linear Frequency Cepstral Coefficients (LFCC) are present in today's methods as well [44, 53], exhibiting better performance with higher frequencies in speech recordings due to its linear filter bank (as opposed to logarithmic-like mel-scale of MFCC) [118].

[0] https://en.wikipedia.org/wiki/Audio_deepfake

Table 3.1: Overview of discussed popular audio features used for deepfake classification

Type	**Features**	**Description**
Cepstral coefficients	LFCC, MFCC, CQCC	Representation of a short-term power spectrum of a sound
Spectrograms	Linear- and Mel-spectrogram	Visualization of the spectrum of frequencies of a sound
Embeddings	i-vectors, x-vectors, speaker embeddings	A vector of values representing voice characteristics, extracted by DNNs

Analogously, albeit with notable differences, Constant Q Cepstral Coefficients (CQCC) can also reliably describe speech characteristics in the context of automatic speaker recognition. In contrast to LFCC and MFCC, CQCC is based on the constant Q transform, which employs geometrically spaced frequency bins resulting in higher frequency resolution at lower frequencies while offering a higher temporal resolution at higher frequencies [99]. CQCC are common features in the diverse detection tools ecosystem [5, 16].

As discussed in Section 2.1.1 in the context of TTS, **spectrograms** (or similarly speech histograms [8]) are visualizations that can represent speech. Therefore, it is no surprise that they are also being used for deepfake detection. Similarly to cepstral coefficients, depending on the frequency scale used in the visualization and signal processing, there are two main types of spectrograms: linear and Mel-spectrograms [57]. They are instrumental when combined with a classifier handling structured data, such as the prevalent CNNs, because of their 2D nature [9, 46, 70].

Modern, state-of-the-art and arguably the most efficient way to represent voice characteristics of a speaker are features called **i-vectors** [47], its successor **x-vectors** [89] or more generally called **speaker embeddings**. These are high-level speech properties, usually without semantic meaning sensible for a human[1]. Embeddings are generally extracted by a neural network of various architectures, for example, ResNet (discussed below) [15]. Additionally, cutting-edge approaches, such as Self-Supervised Learning (SSL), are being utilized [61]; similarly, large pre-trained models can be fine-tuned to extract better-suited embeddings, resulting in better spoofing detection performance [96].

ML (shallow) classifiers

Classical machine-learning approaches still have their place in synthetic speech detection. Following is a short list of the most prevalent models with references to exemplar systems:

- **Gaussian Mixture Model** (GMM) [2]
- **K-Nearest Neighbors** [46]
- **Logistic Regression** [80]
- **Support Vector Machines** (SVM) [54] and its extension **Quadratic SVM** [87]

[1]The individual values of the embedding cannot be assigned to a specific real-world feature.

In addition, a combination of multiple classifiers to form a more robust one is also possible. For example, Chettri et al. [17] propose *Ensemble Models*, fusing GMM and SVM classifiers with convolutional and recurrent neural networks. Some models are used multiple times with different input features to upgrade the system further. The reported results indicate a significant boost in performance when compared to using any of the classifiers separately.

DNN architectures

As apparent, most solutions consist of **Convolutional Neural Networks** or employ some convolutional layers in their architecture. There are many specialized custom-tailored CNN architectures [8, 9, 52, 54]. Popular architectures also include a combination of CNN and RNN called **Convolutional Recurrent Neural Network** (CRNN) [18, 115].

Many systems incorporate proven popular models, such as **ResNet** of various sizes [15, 16, 44, 70]. ResNet, a.k.a. Residual Network, consists of blocks of convolutional layers. The original *input* of the block is added to the *output* of the block, creating a shortcut connection, as seen in Figure 3.2. This allows for learning only the difference (called *residual*) between the input and output of the block instead of direct mapping [38]. These 'skip' connections are also used, for example, in LSTM cells or transformer models mentioned in the following paragraphs.

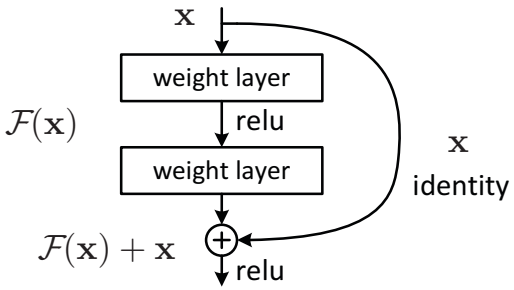


Figure 3.2: Building block for residual learning used by ResNet. Image taken from *Deep Residual Learning for Image Recognition* [38] by He et al.

Simpler **Feedforward Network**, also called **Multi Layer Perceptron** (MLP) can sparsely be found in some specific cases [55]; more powerful architectures supersede them. A more significant share is covered by **Time Delay Neural Networks** (TDNN), which can comprehend the context and surroundings of input cells – this proved to be convenient when working with sound [22].

Present are also **LSTM** and its variant Bidirectional LSTM in RNNs. The main advantage of this approach is the ability of LSTM network to 'remember' information from distant layers, leading to better context comprehension and processing [52].

The ability to remember and comprehend context is shown to be important when processing audio files. This feature is generally called *attention mechanism* and, as mentioned, can be found in RNNs and LSTMs (which calculate the attention weights sequentially). In 2017, Vaswani et al. [102] presented a novel innovative approach called the **Transformer** architecture, which is able to calculate attention weights effectively in parallel. Apart from their use in notorious GPT models in text synthesis (see Section 2.3), state-of-the-art deepfake detection approaches also utilize attention mechanism [101]. Similarly, **Conformers**,

a combination of Transformers and CNN, begin to emerge in the field, and its potential is being researched [92].

Another interesting approach is the use of **Graph Attention Network** (GAT), which also leverages the attention mechanism. GATs are able to attend over their neighborhoods' features and operate on graph-structured data [103]. It is also applicable in spoofing detection, as proposed by Tak et al. [94], to model the spectral and temporal relationship between neighboring sub-bands or segments in a recording. The graph can be constructed from speaker embeddings, where the nodes represent both temporal and frequency segments.

Other approaches

Apart from the orthodox approaches discussed above, many interesting ideas and novel techniques reporting promising results can be found in the literature. Additionally, experiments with extracting custom speech features are also being conducted. For example, Ahmed et al. [2] built a detection system based on extracting the spectral power of high and low-frequency bands in a recording. These features are then used with a GMM classifier.

Among the less traditional approaches lies the DeepSonar framework. Proposed by Wang et al. [106], the aim is to catch layer-wise neuron activations (with high values) in different network layers in a neural-network-based system. Depending on which neurons activate for bonafide and fake speech, a relatively simple classifier suffices to detect spoofing. Another interesting method is proposed by Lei et al. [53]. The architecture includes two siamese (i.e., identical) parallel CNNs concatenated by a fully connected layer.

Finally, as mentioned, some combinations of features and classifiers fit together well, such as the presented spectrograms and CNNs. However, the trend is to remove as many intermediate stages and create **end-to-end** systems [33, 93], which are able to transform raw audio waveforms into the desired output directly. As an example, RawNet is a specifically designed neural network that can handle inputs of raw waveforms (4 seconds of 16kHz audio, i.e., 64.000 samples) that can also be used for anti-spoofing measures [95].

3.2 Deepfake speech datasets

An essential prerequisite for deepfake speech detection is a dataset of deepfake speech samples. These datasets are used to train and validate implemented deepfake speech detection methods. The paramount deepfake speech dataset may be considered the Logical Access (LA) subset of the ASVspoof 2019 challenge dataset [108]. The dataset is divided into training, development, and evaluation subsets, consisting of samples synthesized using 19 different tools, including TTS and VC.

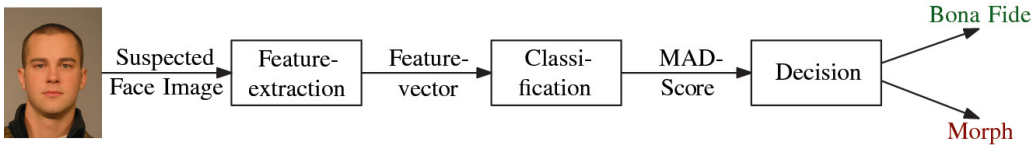
The successor, ASVspoof 2021 DF (DeepFake) eval database [111], was intended as an evaluation subset only for deepfake speech detectors. It consists of samples from the ASVspoof 2019 LA evaluation subset and Voice Conversion Challenge (VCC) 2018 [56] and 2020 [113] data.

In contrast to datasets prepared by experts in the field, the 'In-the-Wild' Audio Deepfake dataset [64] collects samples from publicly available sources. This dataset is currently the most challenging validation dataset for deepfake speech detectors as it contains very different samples (novel tools, post-processing, compression) from the ones in other datasets. This dataset thus best reflects the real-world deployment of deepfake speech detectors.

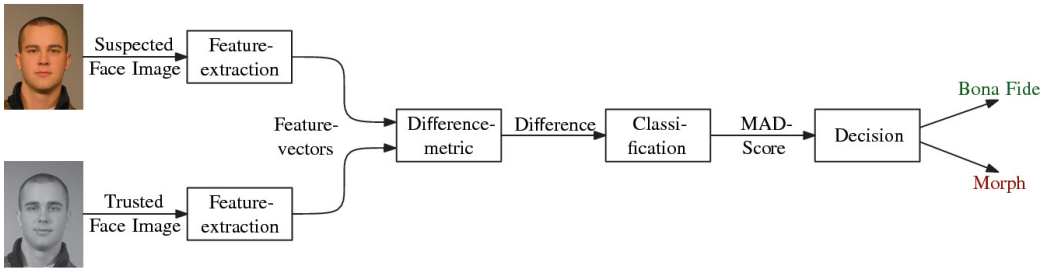
3.3 Differential-based detection

So far, the described detectors and systems focus on a single-input classification, meaning the decision, whether the recording is real or fake, is made based on the properties of the input itself. However, in many cases, a trusted recording is available (or can be easily obtained) and might provide essential additional information. This becomes even more apparent in the context of morphing, both speech and facial (discussed in Section 2.1.3 and Section 2.2.4, respectively.).

In the environment of audio deepfake detection and automatic speaker verification, pair-based or differential approaches are uncharted territory. However, some methods exist in the area of face deepfakes, mainly for detecting morphing and retouching. A comparison of the two approaches – single input and differential – is presented in Figure 3.3. The differential techniques can be broadly divided into two major categories: *Demorphing* and *Feature difference* based detection [79]. A few facial deepfake detection systems from both categories are further discussed.



(a) no-reference morphing detection scheme



(b) differential morphing detection scheme

Figure 3.3: Comparison of single input (a) and differential (b) morphing attack detection (MAD). Image taken from *Handbook of Digital Face Manipulation and Detection* [79] by Rathgeb et al.

Demorphing

The face demorphing techniques try to invert the morphing procedure, revealing the image components used for the morph generation [104]. The first demorphing system by Ferrara et al. [29] introduced the demorphing approach by subtracting the trusted live image from the tested one and comparing the facial landmarks[2].

More recent works utilize deep CNNs, for example the work by Peng et al. [73] is based on a symmetric dual network architecture, a custom GAN. The goal is to restore the hidden face in the morphed image. Alternatively, Ortega et al. [69] use an autoencoder to extract deep face embeddings, concatenate them, and use the decoder part of the used autoencoder to reconstruct the hidden face.

[2]Unique key points on a human face, typically anatomical features such as nose, eyes, mouth, etc.

Demorphing techniques are robust when the conditions and image quality is good, however, the performance is sensitive to outside factors such as facial poses, lighting, etc. These limitations render this approach challenging to use in real-life conditions. Additionally, prior knowledge of a *blending factor*, i.e., the coefficient of morphing intensity, is usually required [104].

Feature difference based detection

As the name suggests, detectors in this category depend on calculating the difference between two feature vectors. The fundamental idea of feature difference-based detection could already be seen in Figure 3.3 and can be described in three steps:

1. Extract feature vectors (both trusted and tested images)
2. Compute the difference between extracted features
3. Perform classification based on the difference vector

Evidently, this method allows for many approaches in any of the three presented steps. Scherlag et al. [81] and Damer et al. [23] propose a detection algorithm based on landmark positions, angles, and shifts. Facial landmarks are further used in combination with an SVM classifier [83].

Apart from facial landmarks, other features are also used. The system presented by Singh et al. [88] decomposes the input images into a diffuse reconstructed image (which contains information, for example, about the reflectance of material) and a corresponding normal map, which represents the shape of the face. Rathgeb et al. [78] extract multiple features, such as texture descriptors and facial landmarks, but also deep face representations, i.e., face embeddings. In addition, the input faces' normalization (by scaling, rotating, and cropping) is implemented for more effective feature extraction. SVM classifier is then trained separately for each of the features, and a weighted score-level sum obtains the resulting detection score.

Deep face embeddings are probably the most popular feature, as they can be found in many presented systems [84, 90]. Additionally, Ibsen et al. [42] propose three fusion schemes for obtaining the feature differences: for deep face embeddings A and B, presented variants are *subtraction* $A - B$, *squared subtraction*, i.e., $(A - B)^2$ and *absolute subtraction* $|A - B|$. Based on the results, there is not much difference between the fusions, with ordinary subtraction performing only insignificantly better than the alternative schemes.

3.4 From images to speech

While the previous section discusses differential-based detection in the visual domain, this thesis aims to adopt this approach for detecting speech deepfakes. Inspiration for the foundation could be taken from the state-of-the-art systems already employed in spoofed audio detection. Those systems may be modified or extended to leverage the additional information from the trusted ground-truth sample to recognize bonafide and synthetic recordings efficiently.

Similarly, inspired by morphing attack detection [42], the suggested approach seems to be feature difference-based. The feature-difference vector might be obtained by subtracting the feature vectors extracted from the probe and tested recordings.

Instead of subtracting the feature vectors, another approach of differentiating could be explored as an alternative method. Perhaps concatenation would be a good fit, as it would allow for the discovery of more complex relations between the tested and ground truth recordings using a robust model that is able to take full advantage of the additional information.

With the ambition to design a reliable and well-rounded detector, various types of recordings must be allowed to be processed. In the general case, the only precedent that can be relied on is having pairs of recordings – one to be tested and a second one with a trusted probe, the ground-truth voice of the tested speaker. Therefore, no restrictions on the recording properties should be imposed. That means: be able to handle different lengths, regardless of spoken language, independent of deepfake type (i.e., do not limit the system for speech morphing detection, but also enable uncovering recordings synthesized by TTS, modified by VC) etc.

Research takeaways

As apparent, in most cases, modern systems for deepfake speech detection are based on DNN architectures. That includes both the stage of feature extraction and classification. Similarly, due to its modular paradigm, end-to-end models also utilize neural networks.

Self-supervised learning models are gaining popularity in the deepfake speech detection field and reportedly outperform other models [61], so it would suggest using one of the state-of-the-art speaker embedding extractors. This should provide the highest quality features to date and, combined together with a proper attention-respecting pooling technique, it should theoretically supersede alternative approaches.

Without proper investigation, it is unclear what the resulting distribution of extracted and pooled features would look like. Therefore, the classification part of the system should be able to reliably separate the bonafide and spoofed classes using a non-linear decision boundary. Therefore, both neural architectures (with appropriate activation function [26]) and robust shallow classifiers (e.g., SVM utilizing a non-linear kernel function [72]) are suitable for the task.

The attention mechanism seems to play a relevant role in many systems. It is utilized in many places, e.g., Wav2Vec systems use it for feature extraction, but also in the classification part of the systems, for example, in graph attention networks [96]. However, it is not absolutely necessary to implement attention, as there are also quite a few examples of systems with competitive performance without the need for attention mechanism [54].

Chapter 4

Design of a differential-based deepfake speech detector

This chapter delves into the intricacies of designing a competitive differential-based deepfake speech detector. Building upon prior research in Chapter 2 and Chapter 3 and motivated by the cases described in Section 2.4, the focus shifts towards the technical overview of crafting a solution tailored to utilize the additional information in the form of trusted ground-truth sample. Harnessing the power of advanced machine learning models, multiple models capable of unveiling differences between real and fake audio are engineered.

The very first approach originates from the feature-difference-based morphing attack detection systems employed in the domain of facial deepfakes [79]. As presented in Figure 4.1, features are first extracted from both ground-truth and tested recordings, followed by computing a difference metric. A classifier is trained on the obtained difference feature vector to recognize whether the tested recording is real or fake.

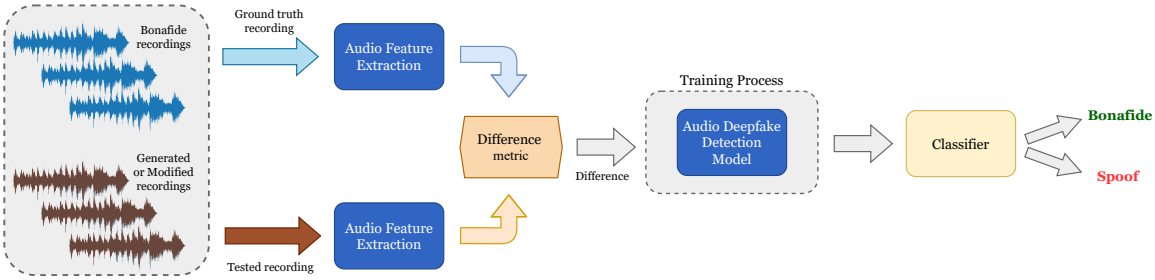


Figure 4.1: A high-level schema of the proposed differential-based deepfake speech detector.

The second approach explores the possibility of fusing the recordings with alternative techniques based on concatenation. A variety of combinations are explored, as in the high-level schema, there are multiple places where the concatenation could occur.

4.1 Models overview

As mentioned, there are two main approaches proposed in this thesis. Differential-based systems assess the influence of different subtraction schemes: *subtraction*, *squared subtraction*, and *absolute subtraction*. Concatenation-based systems are a bit more versatile,

allowing for various concatenation schemes, as well as utilizing advanced machine learning techniques, such as the attention mechanism in LSTM.

Additionally, a baseline single-input system is implemented to compare with the pair-input models. This model works the same way as *classical* deepfake speech detectors. It takes only one input – an examined sample – and decides on its authenticity. This baseline model is used to assess whether the additional information in the form of a ground truth sample increases detection performance. Figure 4.2 presents an overview of all implemented systems.

4.1.1 Differential-based models

Feature-difference-based models very closely resemble the ones employed in the domain of deepfake faces. As proposed by Ibsen et al. [42], the three models utilize the following fusion schemes to produce the resulting feature-difference vectors:

- *FFDiff* model uses a simple subtraction of extracted and pooled features.
- *FFQuadratic* model employs a squared subtraction, i.e., taking the second power of the subtracted features.
- *FFAbs* model takes the absolute value of the subtraction of extracted and pooled features.

As apparent, the models are very similar to each other. One important fact to mention is that the models always subtract the tested features from the ground-truth ones. This makes the models more stable, especially in the case of *FFDiff*, as *FFAbs* and *FFQuadratic* are using operations invariant to the order of subtraction.

4.1.2 Concatenation-based models

The first obvious alternative way of combining the two input recordings is concatenation. Multiple places in the processing pipeline are suitable for the concatenation to happen. Therefore, the designed models experiment with the following options:

- *FFConcat1* model concatenates the raw recordings before passing them to the feature extractor module.
- *FFConcat2* model first extracts the features for both recordings separately and concatenates the features before pooling.
- *FFConcat3* model concatenates the extracted and pooled features before passing it to the classifier.
- *FFLSTM* model follows an alternative pooling scheme to concatenate the features before passing them to two LSTM cells with the aim of finding attention. The final output of the last cell is used for classification.

Concatenation-based models rely on robust machine-learning techniques to unveil the difference between the ground truth and tested recordings. Unlike feature-difference-based models, a specific strategy to identify discrepancies between the two inputs is not enforced in the models, where the difference metric is well-defined as subtraction (or its variants)

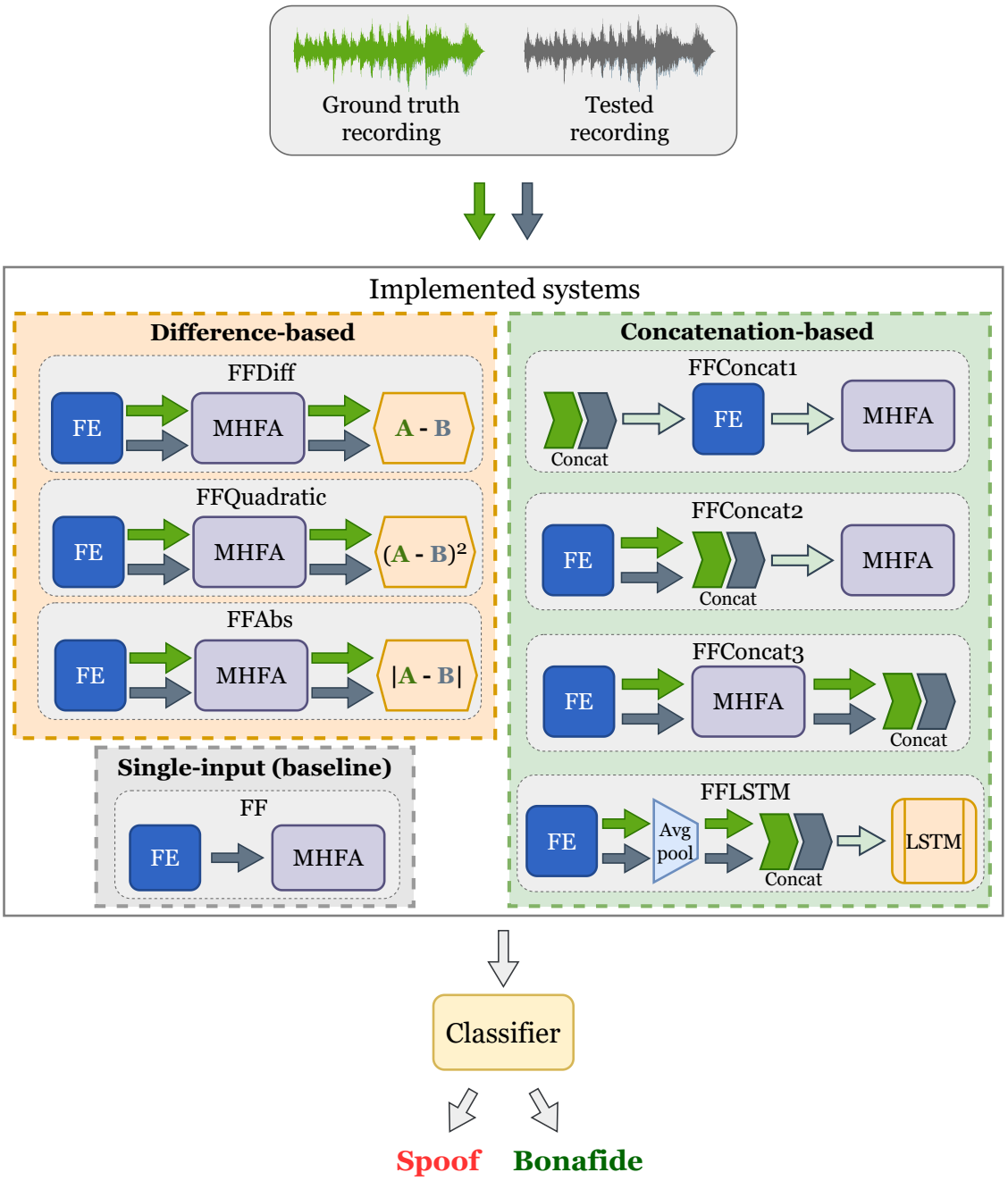


Figure 4.2: Overview of the proposed differential detector architectures. The dashed-border boxes represent different approaches, and the dotted-border boxes represent different combination schemes of ground truth and tested recordings.
FE – Feature Extraction, *MHFA* – Multi-head Factorized Attentive pooling [74].

of feature vectors. Therefore, it is fundamental that the system can capture the vital information through other means. This can be achieved by a robust classifier, employing a powerful ML model (such as LSTM), or utilizing potent pooling techniques, such as Multi-head Factorized Attention pooling [74].

4.2 Processing pipeline

The proposed system architecture follows standard processes in deepfake detection, already presented in Figure 3.1. Apart from obtaining the training and evaluation data, the steps can be summarized as:

1. Extract features from the input data
2. Optionally pool or compress the extracted features to a suitable shape or format
3. Classify, determine if samples are real or fake

Of course, there are many nuances in all of the steps as well as between them, as they can vastly differ based on the intended task and approach of the model. The three steps mentioned are briefly described in more detail in the following paragraphs.

Feature extraction

As Self-Supervised Learning (SSL) models are gaining popularity in the deepfake speech detection area and reportedly outperform other models, it was decided to use them to build our differential-based detector. For the extraction, a pre-trained *Wav2vec 2.0* system was selected as a state-of-the-art embedding extractor, used in many current detectors [61, 96].

The framework consists of a multi-layer convolutional neural network, which creates latent speech representations (embeddings), which are consequently fed into a transformer network that builds contextualized representations of the embeddings [7].

The chosen model is the Cross-lingual Speech Representation (XLSR) adaptation [6] due to its broad language support. The minor variant with 300M parameters proved sufficient for the purpose of this thesis. However, larger versions (with 1B and 2B parameters) are also available but require extensive hardware resources and computing power to fully utilize their power. The feature extraction module XLSR-300M operates on 50ms time frames and extracts a feature vector of 1024 values for each frame. The features can be retrieved from each of the 24 transformer layers to extract the maximum possible information from the recording.

Pooling

The obtained features needed to be pooled due to their complex shape and structure to enable the processing of varied-length recordings. The feature vector is multidimensional, as there is a dimension for the 24 transformer layers, a temporal dimension of the 50ms time frames, and a feature dimension containing the features from the specific transformer layer for the particular time frame.

Initially, it was experimented with simple pooling techniques, such as average or max pooling, which led to a significant loss of information stored in the feature vector. These inferior approaches were superseded by a more robust technique – Multi-head Factorized

Attentive Pooling (MHFA) as described by Peng et al. [74], allowing for context-aware pooling along both transformer and temporal dimensions.

It was decided to use the default parameters as proposed by the paper, with the slightest alteration of enlarging the final projection layer. The reason behind this is to have the same size output as the feature extractor for experimenting with alternative pooling techniques. Therefore, the parameters of the MHFA were selected as follows: 32 heads, a compression layer size of 128 neurons, and a projection layer of 1024 neurons.

The only pooling exception was the *FFLSTM* model, where keeping the temporal dimension (frames) to pass through the LSTM cells was necessary. Therefore, instead of MHFA, only average pooling was conducted along the transformer dimension.

Classification

The resulting and pooled feature vector was fed to a downstream feed-forward neural network classifier, the schema of which is visible in Figure 4.3. It was decided to use a relatively simple architecture but include an activation function to allow the classifier to learn a non-linear boundary between the bonafide and spoofed classes. This resulted in three linear

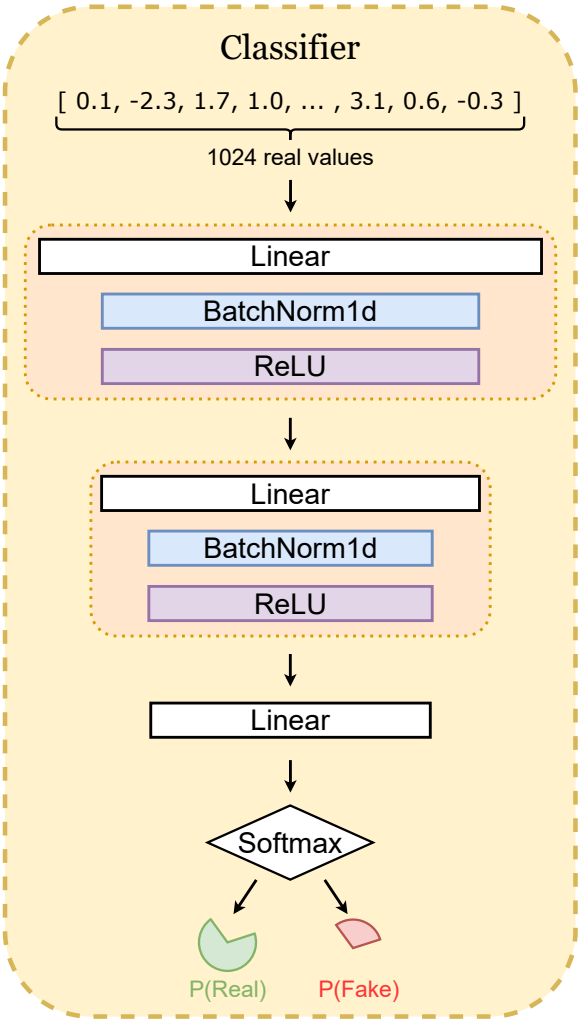


Figure 4.3: Schema of the Feed-forward neural network classifier.

(fully connected) layers with the Rectified Linear Unit (ReLU) [37] non-linearity between them. Batch normalization is also performed after each of the hidden layers. The classifier configuration is shown in Table 4.1. The softmax function is finally applied to the output of the final layer to obtain the classification probabilities for bonafide and spoofed classes.

Table 4.1: Architecture of the Feed-forward neural network classifier. For shapes in format (X/Y), X denotes shape for traditional combination, and Y for the *FFConcat3* model, which concatenates the features before passing them to the classifier, resulting in double the input size.

	Layer (Input shape)	**Output shape**
Feed-forward Neural Network Classifier	Linear (1024/2048)	(512/1024)
	BN & ReLU	(512/1024)
	Linear (512/1024)	(256/512)
	BN & ReLU	(256/512)
	Linear (256/512)	2

4.3 Implementation remarks

The designed systems are implemented in the PyTorch framework (v2.2.0) with Python ver. 3.10. The default Adam optimizer [48] was used for training with a simple Cross-entropy loss function. Implementation and more details can be found in the PyTorch documentation[1].

It was decided to keep the pretrained XLSR feature extractor frozen for this work so as not to fine-tune the model parameters. The main reason is the necessity of extensive hardware resources and computing power required to do so, seconded by an intention of not interfering with the objectivity of results. For similar reasons, no data augmentation procedures took place. Both techniques bring notable benefits and should be considered for further experiments.

The presented systems were trained on the training subset of the ASVspoof2019 LA dataset [108]. However, the classes in the dataset are heavily unbalanced; there are about 90% of spoofed recordings and only about 10% of bonafide samples. For this reason, a technique for ensuring a meaningful training process was necessary. Supplying weights (priors) to the cross-entropy loss function was considered, but in the end, a more versatile solution was decided upon. Weighted Random Sampler[2] from PyTorch, which can under-sample the prevalent class and at the same time over-sample the under-represented class, is able to provide very similar number of samples from both classes and was therefore employed to balance the classes during training.

The dataloader has to provide pairs of samples from the same speaker, where the first sample is always bonafide. To load samples obeying this condition, each time a random sample from the set is drawn, a pool of samples from the same speaker is gathered, and the second, bonafide sample is randomly selected. This procedure is applied during both

[1] https://pytorch.org/docs/
[2] https://pytorch.org/docs/stable/data.html#torch.utils.data.WeightedRandomSampler

the training and evaluation process. All samples are drawn solely from the corresponding subsets.

After training, the models were evaluated on all three mentioned datasets:

1. ASVspoof2019 LA eval [100]
2. ASVspoof2021 DF eval [111]
3. In-the-Wild [64]

This diverse evaluation setup ensures a comprehensive assessment of model performance across various spoofing scenarios and real-world environments and a comparison to modern detection systems.

Chapter 5

Results

With the systems implemented and trained, it is finally time to evaluate the novel approach of differential-based and concatenation-based deepfake speech detection. The pair-input systems and the single-input baseline *FF* system are compared to one another. Additionally, the detectors should be evaluated on their own, but also showcase how well they compete against state-of-the-art competition.

As the primary evaluation metric, Equal Error Rate (EER) is used as a standard for evaluating deepfake speech detection. Moreover, it is threshold-independent and provides reliable information about how well the detector can separate bonafide and spoof classes. As an additional metric, accuracy is disclosed but not used to evaluate or draw conclusions. The accuracy depends on a selected threshold to make a hard decision between the two classes. The threshold was set as $t = 0.5$, meaning the more-likely class with higher probability was assigned to the tested sample. Measured metrics and evaluations of the implemented systems are presented in Table 5.1, and the results are shortly discussed in the following section.

5.1 Implemented systems evaluation

Firstly, the differential-based approaches, performing various forms of subtraction of pooled feature vectors: *FFDiff*, *FFQuadratic* and *FFAbs*, achieve a decent performance below 0.4% EER on the ASVspoof2019 dataset. The absolute difference ($|A-B|$) performs the best with an EER of 0.28%. However, the *FF* baseline performs significantly better, with a relative percentual improvement of approximately 60%. The ASVSpoof2021 DF eval dataset results reveal the *FFDiff* and *FFQuadratic* models as the best performers. However, during the In-the-Wild evaluation, the differential-based models perform worse than the single-input baseline again.

Secondly, the concatenation-based models *FFConcat1*, *FFConcat2*, *FFConcat3* and *FFLSTM* report varying performance on the ASVSpoof2019 dataset. The best-performing *FFConcat3* achieves an EER of 0.22%, while the worst-performing *FFLSTM* achieves an EER of 0.85%. However, even the best-performing model still lags behind the baseline by approximately 22% in relative percentual improvement. The ASVspoof2021 evaluation reveals that the *FFConcat1* and *FFConcat3* models outperform the *FF* baseline by a relative percentual improvement of 26% and 14% respectively. The *FFConcat3* is the best-performing overall with an EER of 5.5%

Table 5.1: Accuracy (percentage of correctly classified samples) and Equal Error Rate (EER) of developed systems. The best-performing system (based on EER) is highlighted in **bold**. *Italics* emphasize the best-performing pair-input system.

Dataset	Model	EER	Accuracy
ASVspoof2019 LA	FFDiff	0.2975%	96.4344%
	FFQuadratic	0.3913%	99.5283%
	FFAbs	0.2844%	99.5031%
	FFConcat1	0.5166%	98.5078%
	FFConcat2	0.6258%	99.1283%
	FFConcat3	*0.2176%*	99.8680%
	FFLSTM	0.8548%	89.6753%
	FF	**0.1776%**	99.5438%
ASVspoof2021 DF	FFDiff	5.9318%	87.8231%
	FFQuadratic	5.9722%	90.0041%
	FFAbs	7.3641%	88.9380%
	FFConcat1	***5.5068%***	98.7208%
	FFConcat2	7.5123%	98.3973%
	FFConcat3	6.0662%	98.7163%
	FFLSTM	14.2128%	97.2152%
	FF	6.9276%	98.7172%
In-the-Wild	FFDiff	15.3502%	85.7768%
	FFQuadratic	23.3994%	77.2649%
	FFAbs	18.7545%	83.2059%
	FFConcat1	19.8875%	42.3613%
	FFConcat2	15.2998%	62.0945%
	FFConcat3	***12.3637%***	43.8875%
	FFLSTM	21.9729%	63.0196%
	FF	13.0496%	53.4567%

Finally, the single-input *FF* system stands out as the top performer on the ASVspoof 2019 dataset, achieving an impressively low EER of 0.18%. However, its effectiveness diminishes with previously unseen data from the ASVSpoof2021 DF and In-the-Wild datasets, where systems that use paired inputs demonstrate superior performance. This decline in the *FF* model's performance can likely be attributed to its overfitting on the ASVspoof2019 dataset. In contrast, the pair-input systems maintain robust performance, even with new data. The trade-off here is between achieving high performance on familiar data at the expense of adaptability and robustness with new datasets.

Ultimately, the results of the ASVspoof2021 database were further explored. This databse integrates samples from previous challenges: ASVspoof2019 in addition to Voice Conversion Challenges VCC2018 and VCC2020. A significant portion of this database consists of Voice Conversion (VC) samples, leading to an investigation into whether pair-based systems perform better with VC samples.

As illustrated in Table 5.2, pair-input systems better detect VC samples than single-input systems. In contrast, the single-input system performs the best on TTS samples. This improvement can be attributed to the additional context by comparing a ground-truth sample with a VC sample. The suspected reason for this behavior is caused by VC samples containing leaked speaker information from source and target utterances used as inputs [25].

This issue should become even more pronounced in zero-shot VC scenarios. Therefore, analyzing ground truth and Voice Conversion samples together uncovers inconsistencies that positively influence their detection. However, this method does not yield the same results as Text-to-Speech samples, likely because the synthesis process does not involve mimicking a specific speaker on the fly.

Table 5.2: Comparison of pair-input and single-input systems regarding speech synthesis methods. *VC* denotes the ASVspoof2021 DF eval data originating from the Voice Conversion systems, *TTS* denotes the data from Text to Speech systems from the same dataset. The best results are highlighted in **bold**.

Model	*EER VC*	*EER TTS*
FFDiff	6.15%	4.03%
FFQuadratic	6.35%	1.38%
FFAbs	7.86%	2.51%
FFConcat1	**4.98%**	3.72%
FFConcat2	7.01%	4.85%
FFConcat3	5.97%	1.29%
FFLSTM	14.08%	5.07%
FF	6.99%	**1.22%**

5.2 System fusion

Upon examining the score distributions, it became evident that systems that employ the differential-based techniques performed better in accurately identifying bonafide samples. Conversely, concatenation-based systems demonstrated a heightened ability to distinguish spoofed samples effectively. An example of the score distributions of differential-based system *FFDiff* and concatenation-based system *FFConcat1* is visually presented in Figure 5.1. This observation prompted a follow-up exploration of potential system fusion as a strategic approach to further enhance overall performance.

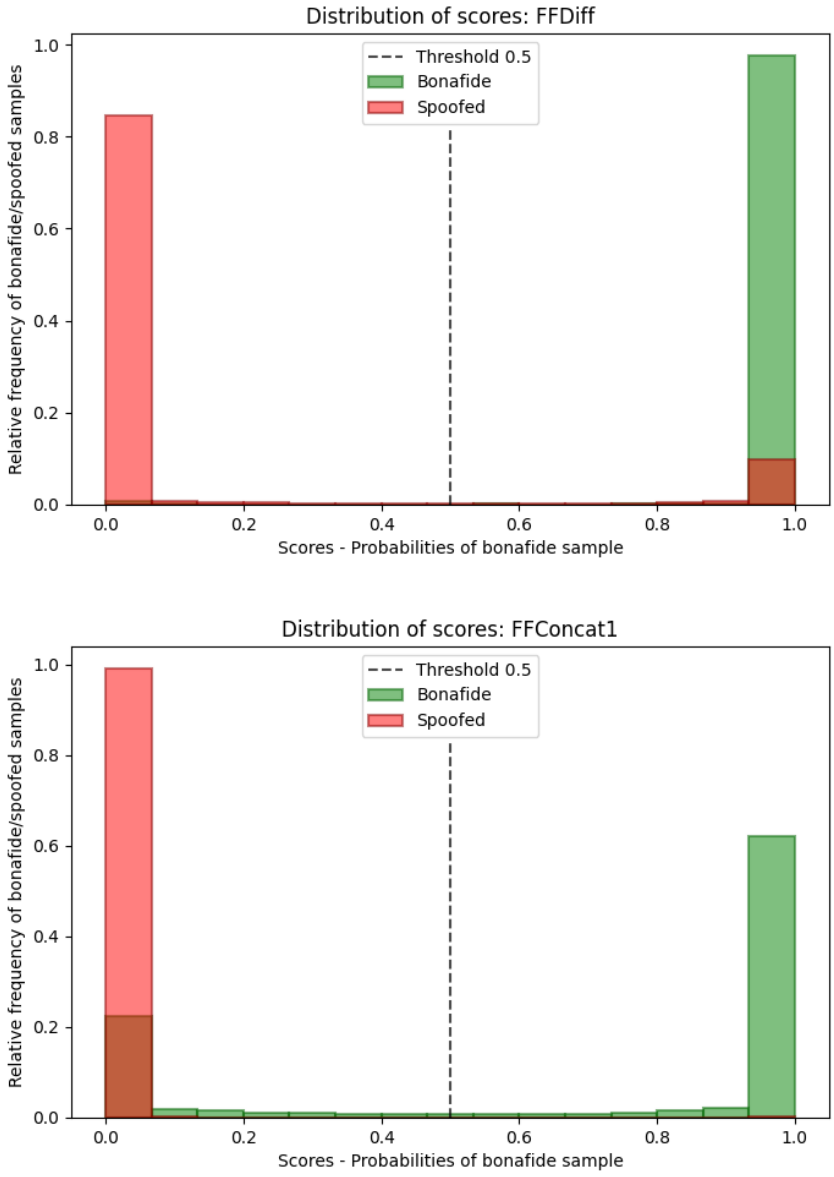


Figure 5.1: Histograms of score distributions of the *FFDiff* and *FFConcat1* models on the ASVspoof2021 DF dataset. Scores are the models' outputs, i.e., the predicted probability of a sample being a bonafide one.

To this end, a series of experiments were executed, focusing on various methodologies of combining the resulting scores to boost the performance further. Explored are the following schemes:

- *Mean fusion* – taking the score average of the fused systems
- *Maximum fusion* – taking the highest score
- *Minimum fusion* – taking the lowest score
- *Root fusion* – taking the root of the product of scores, i.e., $\sqrt[n]{a_1 a_2 \dots a_n}$ also known as geometric mean

Each technique was tested to assess its efficacy in synergistically combining the strengths of the differential-based and concatenation-based systems. The aim was to explore the outcomes of various fusion strategies to determine their feasibility for future application, even on unseen or unlabeled data. Firstly, only combinations of two systems – one differential-based and one concatenation-based – were investigated to prove the initial idea. Secondly, a fusion of multiple systems, inspired by state-of-the-art detectors [17], was examined to determine the best-performing combination. Table 5.3 presents the best performing fusions.

Table 5.3: Best performing fusion schemes for the ASVspoof2021 DF and In-the-Wild. Presented are the best performing differential-concatenation pairs as well as the best multi-system ensembles for both datasets. *D-C* stands for Differential-concatenation pair, while *MSE* stands for Multi-system ensemble, and the fusion scheme is listed in parentheses.

Dataset	*ASVspoof2021 DF*		*In-the-Wild*	
Fusion	D-C (root)	MSE (mean)	D-C (root)	MSE (root)
Models	FFConcat1 FFQuadratic	FFConcat1 FFConcat2 FFConcat3 FFDiff FFQuadratic	FFConcat3 FFDiff	FFConcat1 FFConcat2 FFConcat3 FFAbs FFDiff FFQuadratic
EER	3.73%	3.10%	11.31%	9.59%

For ASVspoof2021 DF, the differential-concatenation pair yielding the best results was the combination of *FFConcat1* and *FFQuadratic* systems using the root fusion scheme with observed EER of 3.73%. Furthermore, an ensemble of *FFConcat1*, *FFConcat2*, *FFConcat3*, *FFDiff* and *FFQuadratic* was the best performer using the mean fusion on ASVspoof2021 DF dataset with measured EER of 3.10%.

The best differential-concatenation pair results for In-the-Wild were achieved using the root fusion of *FFConcat3* and *FFDiff* systems, with an observed EER of 11.31%. It was further experimented with combining multiple systems, achieving an even lower EER of 9.59% by root fusion of *all* systems except *FFLSTM*.

For the ensemble systems that showed the best improvement, a notable enhancement in performance was observed, with the Equal Error Rate decreasing by more than 2.5%

in absolute values. The results varied depending on the chosen fusion method and the combined systems, with EERs generally ranging between 3% and 6% on the ASVspoof2021 DF dataset and between 9.5% and 20% on the In-the-Wild dataset. This indicates that combining concatenation-based and differential-based systems is a promising avenue for future research.

5.3 Comparison to other systems

To put the achieved results into the context of the existing work, the five top-performing systems from the ASVspoof2019 LA, ASVspoof2021 DF, and In-the-Wild benchmarks were selected. The complete comparison of all three evaluations is presented in Table 5.4 for ASVspoof2019, Table 5.5 for ASVspoof2021, and finally, Table 5.6 for the In-the-Wild dataset.

Table 5.4: Comparison of systems evaluated on ASVspoof2019 LA [100]. The top five performing systems from the respective evaluation papers and implemented systems are presented side-by-side. *Ensemble* refers to systems using an ensemble of features and classifiers, and *Single* refers to single systems, such as the ones submitted in this thesis. The best result is highlighted in **bold**.

ASVspoof2019 LA					
Ensemble		**Single**		**Implemented**	
ID	*EER*	*ID*	*EER*	*Model*	*EER*
T05	0.22%	T04	5.74%	**FF**	**0.18%**
T45	1.86%	T08	6.38%	FFConcat3	0.22%
T60	2.64%	T38	7.51%	FFAbs	0.28%
T24	3.45%	T54	7.71%	FFDiff	0.30%
T50	3.56%	B02	8.09%	FFQuadratic	0.39%

The results of ASVspoof2019 LA most obviously unveil the evolution over the past few years. The ensemble systems all consist of multiple classifiers combined with various features. They also leverage neural architectures in some way, contrary to single systems, where all but one (T54) are based on conventional (*shallow*) machine learning models and features [65]. The discrepancy in performance highlights the power of deep learning and its rapid development and adoption. The single-input *FF* baseline is the best performer against all compared models. The performance of *FFConcat3* model is the same as the T05 system, which is an ensemble system. This documents how powerful Self-supervised learning models are for deepfake speech detection.

The contrast is also visible in the ASVspoof2021 DF evaluation in Table 5.5. The best-enrolled system, T23, scored an equal error rate of 15.64%, but only about a year later, it was brought down to under 3% EER using cutting-edge techniques such as the SSL model Wav2Vec2 and attention mechanism in GAT [96]. However, it is essential to note that such a system could not be submitted to the challenge itself, as the Wav2Vec2 frontend was pre-

Table 5.5: Comparison of systems evaluated on ASVspoof2021 DF [111]. The top five performing systems from the respective evaluation papers and implemented systems are presented side-by-side. ASVspoof2021 DF *Post-challenge* features selected systems evaluated after the challenge itself took place. Additionally, the ensemble systems from Section 5.2 are presented for comparison. The best results are highlighted in **bold**.

ASVspoof2021 DF			
Challenge		**Implemented**	
ID	*EER*	*ID*	*EER*
T23	15.64%	**FFConcat1**	**5.51%**
T20	16.05%	FFDiff	5.93%
T08	18.30%	FFQuadratic	5.97%
-	18.80%	FFConcat3	6.06%
T06	19.01%	FF	6.93%
Post-challenge			
Description			*EER*
Wav2Vec2 + GAT [96]			**2.85%**
Wav2Vec2 + Temporal pooling [61]			4.98%
XLS-R + LCNN + LSTM [107]			6.18%
Implemented ensemble systems			
Description			*EER*
FFConcat1 + FFQuadratic			3.71%
FFConcat1 + FFConcat2 + FFConcat3 + FFQuadratic			3.10%

Table 5.6: Comparison of systems evaluated on In-the-Wild [64] dataset. The top five performing systems from the evaluation paper and implemented systems are presented side-by-side. Additionally, the ensemble systems from Section 5.2 are presented for comparison. The best results are highlighted in **bold**.

In-the-Wild			
Evaluated		**Implemented**	
ID	*EER*	*ID*	*EER*
RawNet2	33.94%	**FFConcat3**	**12.36%**
RawGAT-ST	37.15%	FFConcat2	15.30%
MesoInception	37.41%	FF	13.05%
CRNNSpoof	41.71%	FFDiff	15.35%
Transformer	43.78%	FFAbs	18.75%
Implemented ensemble systems			
Description			*EER*
FFConcat3 + FFDiff			11.31%
FFDiff + FFAbs + FFQuadratic + FFConcat1 + FFConcat2 + FFConcat3			**9.76%**

trained on multiple datasets, which violates the challenge rules, just as the implemented systems presented in this thesis.

The post-challenge models outperform the implemented models significantly in terms of EER; however, this direct comparison is a bit misleading, as the post-challenge models used data augmentations for training, which demonstrably increases the performance [55]. Secondly, the weights of the post-challenge SSL models were fine-tuned, significantly boosting the system's performance.

Finally, the most extensive trial of the three, the In-the-Wild dataset (Table 5.6), presents a significant challenge to real-world examples of deepfake data. The models evaluated in the paper were not fine-tuned, nor did they use data augmentation. This also corresponds to the approach taken with the proposed models. All but one (MesoInception) of the top five performing systems utilize the attention mechanism or residual blocks and use either raw waveform or Constant-Q spectrogram as input [64]. The proposed models significantly outperform the models evaluated during the In-the-Wild benchmark, with over 20% absolute difference in final EER between the first places and even more when considering the ensemble systems.

5.4 Expanding pair-based systems

In the previous sections, several implemented deepfake speech detectors were evaluated. The results indicate their competitiveness with relatively simple architectures and without the need for additional struggle with techniques like data augmentation when compared to other state-of-the-art systems. While the implemented and evaluated systems have showcased commendable performance, as with any endeavor, there is still room for enhancement and refinement.

The results presented in the previous sections were achieved in advance with a considerable time reserve before the submission date. Recognizing the potential for improvements with feasible effort, this section briefly delves into a pair of extensions that build on the foundation of the initial design and implementation. Explored is the possibility of elevating the efficacy and robustness of current systems with the aim of pushing the detection performance even further.

5.4.1 Revisiting attention and LSTM

The most obvious deficiency in the evaluation is the *FFLSTM* model. While one of the more complicated models from the implemented ones, its underwhelming results drew considerable attention to engage in a closer inspection. When comparing the resulting EERs, it turns out it is the worst-performing model of the designed ones. Further investigation of the scores revealed that the *FFLSTM* model almost always predicted one of the classes with great confidence while neglecting the second class.

This could be a sign of overfitting; however, even with fewer training epochs, the results were very similar, and of course, during the early stages of the training, the predictions were almost random guesses, i.e., the model was underfitting. This led to the conclusion that the model design was insufficient for the desired task. The main reason is probably the fact that after passing the two whole recordings through the model, the LSTM cells *forget* the relevant information by the time it would be helpful during the processing of the following segments. This means that for the LSTM cells, only a concise context was available, almost certainly resulting in the inability to compare the two recordings at all.

On top of that, taking only the final output of the last cell harshly cuts the information that was passed through the model and is likely important for classification, seconded by the used average pooling, which proved inferior to other techniques such as MHFA used in the other implemented models.

Redesign of the model – FFLSTM2

These observations culminated in redesigning and creating a new model *FFLSTM2*, the schema of which is presented in Figure 5.2. Similarly to the original model, first and foremost, features are extracted (FE) using the XLSR-300M model from both ground-truth and tested recordings. Secondly, the obtained feature vectors are concatenated along the temporal dimension (time frames).

To avoid losing information, the 24 transformer layers are flattened, i.e., stacked after one another, along the time frame dimension. This results in a feature vector with two dimensions: time and transformer layer, and feature dimension, i.e., the 1024 values representing the characteristics at that specific timeframe within the particular transformer layer.

This flattened vector is passed through two LSTM cells. Contrary to the initial approach, all outputs of the last cell are captured, resulting in an output of the same shape as the input vector. The output is unflattened, i.e., reshaped to the original shape with three dimensions: transformer layers, time frames, and feature dimension.

During the last step of the pipeline, the feature vector is pooled using MHFA, just as in the other models. Finally, the result is processed by the feed-forward neural network classifier described in Section 4.2, which outputs the scores (probabilities) of the bonafide or spoofed samples.

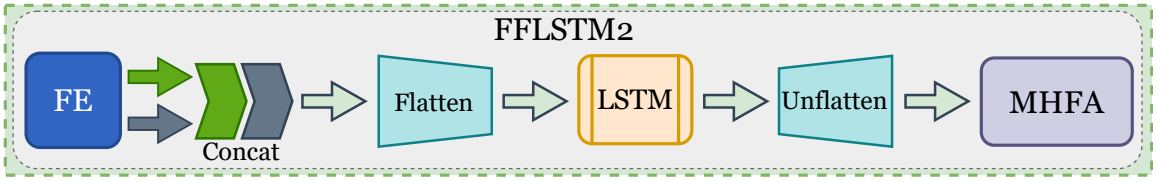


Figure 5.2: Schema of a redesigned model *FFLSTM2*.

Evaluation and results

The new model was evaluated using the same process described in earlier in this chapter. The results for all the benchmark datasets are presented in Table 5.7 for ASVspoof2019 LA, Table 5.8 for ASVspoof2021 DF and Table 5.9 for In-the-Wild dataset. Additionally, a comparison of achieved results on TTS and VC is also presented in Table 5.10. For convenience, the best-performing systems from the respective categories are presented together to put the results into perspective.

It seems the *FFLSTM2* model amplifies the findings discussed in Section 5.5. The results are not stellar on the seen data from ASVspoof2019 – it is the least performant model from the implemented ones (apart from the original *FFLSTM*). However, it shows a heightened generalization ability on the unseen samples from the ASVspoof2021 and In-the-Wild dataset, where out of the implemented ones, *FFLSTM2* performs the best.

Similarly, it became a crucial part of the ensemble systems, and its inclusion further boosted the performance. Most notably, the mean fusion of the three systems *FFConcat1*,

Table 5.7: Comparison of *FFLSTM2* performance on ASVspoof2019 LA. The most performant system from each category is selected. The best result is highlighted in **bold**.

ASVspoof2019 LA		
	ID	*EER*
Ensemble	T05	0.22%
Single	T04	5.74%
Implemented	**FF**	**0.18%**
FFLSTM2	FFLSTM2	0.74%

Table 5.8: Comparison of *FFLSTM2* performance on ASVspoof2021 DF. The most performant system from each category is selected. The best results are highlighted in **bold**.

ASVspoof2021 DF		
	ID	*EER*
Challenge	T23	15.64%
Post-challenge	**Wav2Vec2 + GAT**	**2.85%**
Implemented	FFConcat1	5.51%
FFLSTM2	FFLSTM2	5.31%
Ensemble without FFLSTM2		
Description		*EER*
FFConcat1 + FFQuadratic		3.71%
FFConcat1 + FFConcat2 + FFConcat3 + FFQuadratic		3.1%
Ensemble with FFLSTM2		
Description		*EER*
FFQuadratic + FFLSTM2		3.48%
FFConcat1 + FFQuadratic + FFLSTM2		**2.92%**

Table 5.9: Comparison of *FFLSTM2* performance on In-the-Wild dataset. The most performant system from each category is selected. The best results are highlighted in **bold**.

In-the-Wild		
	ID	*EER*
Evaluated	RawNet2	33.94%
Implemented	FFConcat3	12.36%
FFLSTM2	**FFLSTM2**	**12.28%**
Ensemble without FFLSTM2		
Description		*EER*
FFConcat3 + FFDiff		11.31%
FFConcat1 + FFConcat2 + FFConcat3 + FFDiff + FFAbs + FFQuadratic		9.59%
Ensemble with FFLSTM2		
Description		*EER*
FFDiff + FFLSTM2		10.26%
FFConcat1 + FFConcat2 + FFConcat3 FFDiff + FFAbs + FFLSTM2		**9.02%**

Table 5.10: Comparison of single-input, *FFLSTM2* and best pair-input systems regarding speech synthesis methods. *VC* denotes the ASVspoof2021 DF eval data originating from the Voice Conversion systems, *TTS* denotes the data from Text to Speech systems from the same dataset. The best results are highlighted in **bold**.

Model	*EER VC*	*EER TTS*
FFConcat1	4.98%	3.72%
FF	6.99%	**1.22%**
FFLSTM2	**4.79%**	5.50%

FFQuadratic and *FFLSTM2* achieves a remarkable EER of 2.92% on the ASVspoof2021 DF dataset, almost beating the very best post-challenge competitor. Even more impressively, this was accomplished without using the techniques of data augmentation and fine-tuning of the pre-trained feature extractor, which the post-challenge model used. In a similar fashion, the mean fusion system, which included *FFLSTM2*, almost broke the barrier of 9% EER on the In-the-Wild dataset.

Regarding the TTS versus VC performance, as visible in Table 5.10, the mentioned improvement in detecting VC samples is further magnified as *FFLSTM2* overshadowed the former champion *FFConcat1* and scored the lowest EER out of all the implemented systems. Properly using the attention mechanism can support unveiling the leaked information in VC samples.

5.4.2 Alternative fusion scheme – weighted score-level sum

With the success of ensemble systems in state-of-the-art detectors [17], as well as in the presented results, a question arises: is there an even better fusion scheme than the ones already evaluated? As this goes beyond the main aim of this thesis, it is not necessary to conduct an in-depth exploration to develop the most optimal fusion strategy. But, as it is a relevant and apparently important part of creating competitive systems, some alternative approaches can be explored. Therefore, for simplicity, only score-level fusion will be considered.

The tested alternative approach is a weighted sum of the scores produced by the single models. To obtain the best-suited weights, it was decided to employ a single linear layer that takes the scores as input and produces the resulting probabilities (again using the softmax function) of the bonafide/spoofed classes. It was trained with the objective of minimizing the Cross-entropy loss function.

It is important to note that this approach is biased towards better results, as the weights are trained and evaluated on the same data – the outputs from the single models for the specific database. Therefore, the fusion parameters that best fit one of the datasets might not be optimal for the other. The results should be taken cautiously but provide an exciting insight into the potential pinnacle of performance.

As the training process unfolded, the phenomenon of differential-based techniques performing better in identifying bonafide samples and concatenation-based models performing better in detecting spoofed samples could be observed. Weights belonging to differential-based models were getting high for obtaining the score prediction for the sample being bonafide and low for the spoofed sample, while weights attributed to concatenation-based models were evolving in the opposite direction.

Scores from all models except *FFLSTM* were taken as the input and the weighted score-level sum was performed. The resulting fusion scores were used for the final evaluation and classification. For ASVspoof2021 DF, the lowest achieved EER was 2.99%. For the In-the-Wild dataset, the results break the 9% barrier with EER of **8.96%**.

5.5 Discussion

The results clearly indicate that pair-based deepfake speech detection is a relevant and promising field. Both differential-based and concatenation-based approaches are feasible and achieve competitive performance, overshadowing many state-of-the-art systems.

The major improvement seems to be in detecting Voice Conversion samples. The presented results unveil that the performance is better than that of single-input systems. Speaker information from both speakers must be leaked into the final deepfake utterance, which the pair-input deepfake speech detector may pick up [25]. This is an interesting observation, as up to date, the detection of TTS and VC-generated samples was considered to be equal. However, it is shown that some detection methods might be better suited for specific types of deepfakes. This might be an important aspect of designing robust deepfake speech detectors – employing diverse techniques that capture different parts of available information in the recording.

The original use case for differential-based detection may further support this finding: morphing attack detection. As morphs include a portion of both identities, the difference between ground truth and morphed sample reveals discrepancies. The same seems to apply to VC samples, where speaker information may get somewhat mixed. In contrast, the TTS samples should contain only the synthesized identity and mimic the target more closely; thus, the differential analysis seems less suitable. However, this behavior may substantially change with zero-shot usage of TTS models, which may be more susceptible to leaking speaker information. This behavior should be analyzed in greater detail in future research.

It is expected that differential- and concatenation-based approaches will be extremely efficient in detecting morphed speech. Such recordings contain speaker information of both contributing speakers. Thus, the effect of detecting Voice Conversion samples might be amplified. Unfortunately, no publicly available dataset contains morphed speech at the moment. There is only a minimal amount of speech morphs available on the internet, of unclear quality, with no metrics to back it up, and from random sources. These samples are probably created as an individual's hobby project. The implemented systems were tested on a few morphing samples, coming from an untrusted source of a public GitHub repository[1], so the results do not provide any meaningful insight. All of the differential-based and concatenation-based systems, as well as the single-input baseline *FF* model, correctly classified all of the samples from the repository.

Finally, the pair-based models seem more robust to overfitting, as demonstrated by the comparison with the single input *FF* model. In the case of the pair-based models, there is a slight tradeoff between the performance over data from known distributions (ASVspoof2019 on which the systems were trained) and robustness, i.e., performance over data from different distributions (ASVspoof2021, In-the-Wild). This tradeoff makes the differential-based and concatenation-based models a bit less efficient over the ASVspoof2019 data, but generalize better over ASVspoof2021 and In-the-Wild data.

[1] https://github.com/itsuki8914/Voice-morphing-RelGAN/tree/master/result_examples

Chapter 6

Conclusion

The massive advancements in computing power, data availability, and machine learning techniques blur the boundaries between digital space and reality. This advent of deepfake technology poses unprecedented challenges in today's society. As with almost any emerging technology, it can be used to uplift and advance humankind in a broad spectrum of areas; however, *with great power comes great responsibility*. The potential of malicious applications and wicked influence has alerted individuals across various social classes [35].

Therefore, to aid the battle over the truthfulness and integrity of digital media, this thesis aims to elevate the current methods of deepfake speech detection by proposing a novel approach of pair-input detectors. In auditory domain and deepfake speech detection, this approach is an uncharted territory, nevertheless, some experiments were conducted in the domain of faces. The methodology is inspired by differential face morphing attack detection [78], which focuses specifically on detecting morphs and digital retouching of a human face.

The systems introduced in this thesis leverage a tested sample in addition to a probe, ground-truth recording, which can often be obtained without a significant hindrance, e.g., during a biometric check at airports, police questioning, submission of legal evidence, etc. Two main strategies are presented: *differential-based*, utilizing a difference metric between the features of two recordings, and *concatenation-based*, which combine the trusted and examined recordings using concatenation in various manners.

After conducting extensive research in the field of deepfakes and their detection, multiple models are designed and implemented. During the assessment, several intriguing findings have come to light. Firstly, employing differential-based and concatenation-based approaches proves to be a viable strategy for detecting deepfake speech that is applicable in many real-world scenarios. The additional information in the form of supplied ground-truth recording indicates a boost in performance, especially in detecting Voice Conversion samples. This is an interesting observation, as, to date, the detection of deepfake speech has been considered equal regardless of the method of its creation. However, the evaluation reveals that some approaches might be better suited for detecting only a specific type of deepfake speech.

On top of that, it is suspected that the pair-based approach would be even more efficient in detecting morphed speech. Theoretically, the effect of leaking speaker information into the resulting recording in Voice Conversion samples should be amplified in morphed speech, as the audio contains the information about both contributing speakers. This was unfortunately not possible to verify, as up to date, no public dataset of morphed speech is available.

Finally, from a more technical perspective, the proposed models seem more robust to overfitting. While the single-input baseline outperforms all the other systems on data from known distributions, it is overshadowed by the generalization ability of the pair-input models on unseen data from unknown distributions. This dynamic suggests a promising avenue for future research, specifically focusing on enhancing the adaptability and robustness of pair-input models.

In conclusion, this thesis explores the uncharted waters of differential-based deepfake speech detection and opens many possibilities for follow-up research. Finding a proper way of combining the information from the two input recordings unveils a promising domain for conducting additional experiments. Similarly, an even more robust and powerful application of the attention mechanism, not necessarily only in LSTM, can further enhance the detection performance. Exploring the possibilities of merging differential-based and concatenation-based approaches is also an exciting field for more profound elaboration, and the emerging schemes may even be applicable in alternative areas of deepfake detection.

Bibliography

[1] Ahmed, I.; Sadiq, A.; Atif, M.; Naseer, M. and Adnan, M. Voice morphing: An illusion or reality. In: IEEE. *2018 International Conference on Advancements in Computational Sciences (ICACS)*. IEEE, 2018. ISBN 978-1-5386-2172-1.

[2] Ahmed, M. E.; Kwak, I.-Y.; Huh, J. H.; Kim, I.; Oh, T. et al. Void: A fast and light voice liveness detection system. In: USENIX Association. *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, August 2020, p. 2685–2702. ISBN 978-1-939133-17-5. Available at: https://www.usenix.org/conference/usenixsecurity20/presentation/ahmed-muhammad.

[3] Akhtar, Z. Deepfakes Generation and Detection: A Short Survey. *Journal of Imaging*. 1st ed., 2023, vol. 9, no. 1. ISSN 2313-433X. Available at: https://www.mdpi.com/2313-433X/9/1/18.

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

[5] Alzantot, M.; Wang, Z. and Srivastava, M. B. *Deep Residual Neural Networks for Audio Spoofing Detection*. 2019.

[6] Arun Babu, T.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q. et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In: Ko, H. and Hansen, J. H. L., ed. *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. September 2022, p. 2278–2282.

[7] Baevski, A.; Zhou, H.; Mohamed, A. and Auli, M. Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. and H.Lin, ed. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020. NIPS '20. ISBN 9781713829546.

[8] Ballesteros, D. M.; Rodriguez Ortega, Y.; Renza, D. and Arce, G. Deep4SNet: deep learning for fake speech classification. *Expert Systems with Applications*. 1st ed. Pergamon Press, Inc., 2021, vol. 184, C, p. 115465. ISSN 0957-4174. Available at: https://www.sciencedirect.com/science/article/pii/S0957417421008770.

[9] Bartusiak, E. R. and Delp, E. J. *Frequency Domain-Based Detection of Generated Audio*. 2022.

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

[11] Boháček, M. and Farid, H. *Protecting President Zelenskyy against Deep Fakes*. June 2022. Available at: https://doi.org/10.48550/arXiv.2206.12043.

[12] Brukner, J. *Non-Parallel Voice Conversion*. Brno, CZ, 2020. Master's thesis. Brno University of Technology, Faculty of Information Technology. Available at: https://www.fit.vut.cz/study/thesis/19207/.

[13] Cano, P.; Loscos, A.; Bonada, J.; Boer, M. and Serra, X. *Voice Morphing System for Impersonating in Karaoke Applications*. August 2002.

[14] Chan, C.; Ginosar, S.; Zhou, T. and Efros, A. *Everybody Dance Now*. 2019.

[15] Chen, T.; Khoury, E.; Phatak, K. and Sivaraman, G. *Pindrop Labs' Submission to the ASVspoof 2021 Challenge*. 2021.

[16] Chen, Z.; Xie, Z.; Zhang, W. and Xu, X. ResNet and Model Fusion for Automatic Spoofing Detection. In: Association for Computing Machinery. *Proc. Interspeech 2017*. 2017, p. 102–106. ISBN 9781450395540.

[17] Chettri, B.; Stoller, D.; Morfi, V.; Ramírez, M. A. M.; Benetos, E. et al. *Ensemble Models for Spoofing Detection in Automatic Speaker Verification*. 2019.

[18] Chintha, A.; Thai, B.; Sohrawardi, S. J.; Bhatt, K.; Hickerson, A. et al. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE Journal of Selected Topics in Signal Processing*. 1st ed., 2020, vol. 14, no. 5, p. 1024–1037.

[19] Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S. et al. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. 2018.

[20] Chowdhury, A.; Ross, A. and David, P. DEEPTALK: Vocal Style Encoding for Speaker Recognition and Speech Synthesis. In: Conference Management Services, I., ed. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, p. 6189–6193. ISBN 978-1-7281-7605-5.

[21] Costa, P. *Two-Dimensional Expressive Speech Animation*. 2015. Dissertation. University of Campinas, School of Electrical and Computer Engineering.

[22] Cáceres, J.; Font, R.; Grau, T. and Molina, J. *The Biometric Vox System for the ASVspoof 2021 Challenge*. 2021.

[23] Damer, N.; Boller, V.; Wainakh, Y.; Boutros, F.; Terhörst, P. et al. Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts. In: Brox, T.; Bruhn, A. and Fritz, M., ed. *Pattern Recognition*. Cham: Springer International Publishing, 2019, p. 518–534. ISBN 978-3-030-12939-2.

[24] Damer, N.; Saladié, A. M.; Braun, A. and Kuijper, A. MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In: IEEE. *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 2018, p. 1–10. ISBN 978-1-5386-7180-1.

[25] Deng, J.; Chen, Y.; Zhong, Y.; Miao, Q.; Gong, X. et al. Catch You and I Can: Revealing Source Voiceprint Against Voice Conversion. In: USENIX. *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, August 2023, p. 5163–5180. ISBN 978-1-939133-37-3. Available at: https://www.usenix.org/conference/usenixsecurity23/presentation/deng-jiangyi-voiceprint.

[26] Dubey, S. R.; Singh, S. K. and Chaudhuri, B. B. *Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark*. 2022.

[27] Dudha, A. *Scammers can easily use voice-cloning AI to con family members: Expert | CBC News*. CBC/Radio Canada, Jun 2023. Available at: https://www.cbc.ca/news/canada/saskatoon/fraudsters-likely-using-ai-to-scam-seniors-1.6879807.

[28] Edelson, L. and Diehl, J. Prosody. In: Volkmar, F. R., ed. *Encyclopedia of Autism Spectrum Disorders*. Springer New York, 2013, p. 2413–2417. ISBN 978-1-4419-1698-3. Available at: https://doi.org/10.1007/978-1-4419-1698-3_366.

[29] Ferrara, M.; Franco, A. and Maltoni, D. Face Demorphing. *IEEE Transactions on Information Forensics and Security*. 1st ed., 2017, vol. 13, no. 4, p. 1008–1017.

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

[31] Firc, A.; Malinka, K. and Hanáček, P. Creation and detection of malicious synthetic media - a preliminary survey on deepfakes. In: Šustr, Z., ed. *Sborník příspevků z 54. konference EurOpen.CZ, 28.5.-1.6.2022*. Czech Open Systems User's Group, 2022, p. 125–145. ISBN 978-80-86583-34-1. Available at: https://www.fit.vut.cz/research/publication/12741.

[32] Firc, A.; Malinka, K. and Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*. 1st ed., 2023, vol. 9, no. 4, p. e15090. ISSN 2405-8440. Available at: https://www.sciencedirect.com/science/article/pii/S2405844023022971.

[33] Ge, W.; Patino, J.; Todisco, M. and Evans, N. *Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection*. 2021.

[34] Gehrmann, S.; Strobelt, H. and Rush, A. GLTR: Statistical Detection and Visualization of Generated Text. In: Costa jussà, M. R. and Alfonseca, E.,

ed. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, July 2019, p. 111–116. ISBN 978-1-950737-49-9. Available at: https://aclanthology.org/P19-3019.

[35] Gerken, T. *MrBeast and BBC stars used in Deepfake scam videos*. BBC, Oct 2023. Available at: https://www.bbc.com/news/technology-66993651.

[36] Goodfellow, I. J.; Pouget Abadie, J.; Mirza, M.; Xu, B.; Warde Farley, D. et al. Generative Adversarial Networks. In: Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. and Weinberger, K., ed. *Advances in neural information processing systems*. Curran Associates, Inc., 2014, vol. 27. ISBN 9781510800410. Available at: https://arxiv.org/abs/1406.2661.

[37] Hara, K.; Saito, D. and Shouno, H. Analysis of function of rectified linear unit used in deep learning. In: INNS, IEEE-CIS. *2015 International Joint Conference on Neural Networks (IJCNN)*. 2015, p. 1–8. ISBN 9781479919611.

[38] He, K.; Zhang, X.; Ren, S. and Sun, J. *Deep Residual Learning for Image Recognition*. 2015.

[39] Holliday, C. Retroframing the Future: Digital De-Aging Technologies in Contemporary Hollywood Cinema. *Journal of Cinema and Media Studies*. 1st ed. University of Texas Press, july 2022, vol. 61, no. 5, p. 210–237. ISSN 2578-4919.

[40] Huang, D.-Y.; Rahardja, S. and Ong, E. P. *High level emotional speech morphing using straight*. 2010.

[41] Huijstee, M.; Boheemen, P.; Das, D.; Nierling, L.; Jahnel, J. et al. *Tackling deepfakes in European policy*. 1st ed. European Parliament, 2021. ISBN 978-92-846-8400-7.

[42] Ibsen, M.; Gonzalez Soler, L. J.; Rathgeb, C.; Drozdowski, P.; Gomez Barrero, M. et al. Differential Anomaly Detection for Facial Images. In: IEEE. *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2021, p. 1–6. ISBN 978-1-6654-1717-4.

[43] Kahn, J. *Google's DeepMind Achieves Speech-Generation Breakthrough* online. 9. september 2016. Available at: https://www.bloomberg.com/news/articles/2016-09-09/google-s-ai-brainiacs-achieve-speech-generation-breakthrough. [Accessed 17-12-2023].

[44] Kang, W. H.; Alam, J. and Fathan, A. *CRIM's System Description for the ASVSpoof2021 Challenge*. 2021.

[45] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J. et al. Analyzing and Improving the Image Quality of StyleGAN. *CoRR*. 1st ed. IEEE Computer Society, 2019, abs/1912.04958, no. 2. Available at: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00813.

[46] Khochare, J.; Joshi, C.; Yenarkar, B.; Suratkar, S. and Kazi, F. A deep learning framework for audio deepfake detection. *Arabian Journal for Science and Engineering*. 1st ed. Springer, 2021, no. 3, p. 1–12.

[47] Khoury, E.; Kinnunen, T.; Sizov, A.; Wu, Z. and Marcel, S. Introducing I-vectors for joint anti-spoofing and speaker verification. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 1st ed., january 2014, no. 1, p. 61–65.

[48] Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*. 1st ed., december 2014, no. 9.

[49] Kinsella, B. and Herndon, A. *Deepfake and Voice Clone Consumer Sentiment Report (October 2023)*. Pindrop, VoiceBot.AI, October 2023.

[50] Korshunov, P. and Marcel, S. Subjective and Objective Evaluation of Deepfake Videos. In: IEEE. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, p. 2510–2514. ISBN 978-1-7281-7605-5.

[51] Lalla, V.; Mitrani, A. and Harned, Z. *Artificial Intelligence: Deepfakes in the entertainment industry*. Jun 2022. Available at: https://www.wipo.int/wipo_magazine/en/2022/02/article_0003.html.

[52] Lataifeh, M.; Elnagar, A.; Shahin, I. and Nassif, A. B. Arabic audio clips: Identification and discrimination of authentic Cantillations from imitations. *Neurocomputing*. 1st ed., 2020, vol. 418, no. 1, p. 162–177. ISSN 0925-2312. Available at: https://www.sciencedirect.com/science/article/pii/S0925231220312881.

[53] Lei, Z.; Yang, Y.; Liu, C. and Ye, J. Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection. In: ISCA. *INTERSPEECH*. 2020, p. 1116–1120. ISBN 9781713820697.

[54] Liu, T.; Yan, D.; Wang, R.; Yan, N. and Chen, G. Identification of Fake Stereo Audio Using SVM and CNN. *Information*. 1st ed., 2021, vol. 12, no. 7. ISSN 2078-2489. Available at: https://www.mdpi.com/2078-2489/12/7/263.

[55] Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H. et al. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 1st ed., 2023, vol. 31, no. 1, p. 2507–2522.

[56] Lorenzo Trueba, J.; Yamagishi, J.; Toda, T.; Saito, D.; Villavicencio, F. et al. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods . In: Odyssey. *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*. 2018, p. 195–202.

[57] Luner, M. *Text-to-Speech Personalization*. Brno, CZ, 2022. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology.

[58] Machado, A. F. and Marcelo, Q. *Voice Conversion: A Critical Survey*. Zenodo, July 2010. ISBN 2518-3672. Available at: https://doi.org/10.5281/zenodo.849853.

[59] Malinka, K.; Perešíni, M.; Firc, A.; Hujnák, O. and Janus, F. On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree? In: ITiCSE. *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. New York, NY, USA: Association for Computing Machinery, 2023, p. 47–53. ITiCSE 2023. ISBN 9798400701382. Available at: https://doi.org/10.1145/3587102.3588827.

[60] Mannan, I. and Nova, S. N. An Empirical Study on Theories of Sentiment Analysis in Relation to Fake News Detection. In: Bangladesh University of Professionals. *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. 2023, p. 79–83. ISBN 979-8-3503-5866-7.

[61] Martín Doñas, J. M. and Álvarez, A. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In: ICASSP. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, p. 9241–9245. ISBN 978-1-6654-0540-9.

[62] Mohammadi, S. H. and Kain, A. *An overview of voice conversion systems*. 2017. ISSN 0167-6393. Available at: https://www.sciencedirect.com/science/article/pii/S0167639315300698.

[63] Moore, K.; Dalley, A. and Agur, A. *Clinically Oriented Anatomy*. 6th ed. Lippincott Williams & Wilkins, 2009. ISBN 978-1605476520. Available at: https://books.google.cz/books?id=J3VVPgAACAAJ.

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

[65] Nautsch, A.; Wang, X.; Evans, N.; Kinnunen, T. H.; Vestman, V. et al. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*. 1st ed. Institute of Electrical and Electronics Engineers (IEEE), april 2021, vol. 3, no. 2, p. 252–265. ISSN 2637-6407. Available at: http://dx.doi.org/10.1109/TBIOM.2021.3059479.

[66] Nirkin, Y.; Keller, Y. and Hassner, T. *FSGAN: Subject Agnostic Face Swapping and Reenactment*. 2019.

[67] Oord, A. van den; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O. et al. *WaveNet: A Generative Model for Raw Audio*. 2016.

[68] Orphanidou, C.; Moroz, I. and Roberts, S. *Wavelet-based voice morphing*. 2004.

[69] Ortega Delcampo, D.; Conde, C.; Palacios Alonso, D. and Cabello, E. Border Control Morphing Attack Detection With a Convolutional Neural Network De-Morphing Approach. *IEEE Access*. 1st ed., 2020, vol. 8, no. 1, p. 92301–92313.

[70] P, R. T.; Aravind, P. R.; C, R.; Nechiyil, U. and Paramparambath, N. *Audio Spoofing Verification using Deep Convolutional Neural Networks by Transfer Learning*. 2020.

[71] Pani, S.; Chowdhury, A.; Sandler, M. and Ross, A. *Voice Morphing: Two Identities in One Voice*. September 2023.

[72] Patle, A. and Chouhan, D. S. SVM kernel functions for classification. In: ICATE. *2013 International Conference on Advances in Technology and Engineering (ICATE)*. 2013, p. 1–9. ISBN 9781467356183.

[73] Peng, F.; Zhang, L. bing and Long, M. *FD-GAN: Face-demorphing generative adversarial network for restoring accomplice's facial image*. 2018.

[74] Peng, J.; Plchot, O.; Stafylakis, T.; Mošner, L.; Burget, L. et al. An Attention-Based Backend Allowing Efficient Fine-Tuning of Transformer Models for Speaker Verification. In: IEEE. *2022 IEEE Spoken Language Technology Workshop (SLT)*. 2023, p. 555–562. ISBN 9798350396911.

[75] Pu, J.; Sarwar, Z.; Abdullah, S. M.; Rehman, A.; Kim, Y. et al. Deepfake Text Detection: Limitations and Opportunities. In: IEEE. *2023 IEEE Symposium on Security and Privacy (SP)*. 2023, p. 1613–1630. ISBN 978-1-6654-9336-9.

[76] Qian, K.; Zhang, Y.; Chang, S.; Cox, D. and Hasegawa Johnson, M. Unsupervised Speech Decomposition via Triple Information Bottleneck. In: Daumé, H. and Singh, A., ed. *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020, vol. 119. ICML'20. ISBN 9781713821120.

[77] Rashad, M.; El Bakry, H.; Isma, R. and Mastorakis, N. An overview of text-to-speech synthesis techniques. *International Conference on Communications and Information Technology - Proceedings*. 1st ed., july 2010, no. 1.

[78] Rathgeb, C.; Satnoianu, C.-I.; Haryanto, N. E.; Bernardo, K. and Busch, C. Differential Detection of Facial Retouching: A Multi-Biometric Approach. *IEEE Access*. 1st ed., 2020, vol. 8, no. 1, p. 106373–106385.

[79] Rathgeb, C.; Tolosana, R.; Vera Rodriguez, R. and Busch, C. *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. 1st ed. Springer International Publishing, 2022. Advances in Computer Vision and Pattern Recognition. ISBN 9783030876647.

[80] Rodríguez Ortega, Y.; Ballesteros, D. M. and Renza, D. A machine learning model to detect fake voice. In: Springer. *International Conference on Applied Informatics*. 2020, p. 3–13. ISBN 978-3-030-61702-8.

[81] Scherhag, U.; Budhrani, D.; Gomez Barrero, M. and Busch, C. Detecting Morphed Face Images Using Facial Landmarks. In: Mansouri, A.; El Moataz, A.; Nouboud, F. and Mammass, D., ed. *Image and Signal Processing*. Cham: Springer International Publishing, 2018, p. 444–452. ISBN 978-3-319-94211-7.

[82] Scherhag, U.; Debiasi, L.; Rathgeb, C.; Busch, C. and Uhl, A. Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*. 1st ed., 2019, vol. 1, no. 4, p. 302–317.

[83] Scherhag, U.; Rathgeb, C. and Busch, C. Towards Detection of Morphed Face Images in Electronic Travel Documents. In: IEEE. *2018 13th IAPR International*

Workshop on Document Analysis Systems (DAS). 2018, p. 187–192. ISBN 978-1-5386-3346-5.

[84] Scherhag, U.; Rathgeb, C.; Merkle, J. and Busch, C. Deep Face Representations for Differential Morphing Attack Detection. *IEEE Transactions on Information Forensics and Security*. 1st ed., 2020, vol. 15, no. 1, p. 3625–3639.

[85] Schwartz, E. H. *Watch a virtual Simon Cowell perform for "America's got talent"*. Feb 2023. Available at: https://voicebot.ai/2022/06/07/watch-a-virtual-simon-cowell-perform-for-americas-got-talent/.

[86] Shah, N.; Kosgi, S.; Tambrahalli, V.; Sahipjohn, N.; Pedanekar, N. et al. *ParrotTTS: Text-to-Speech synthesis by exploiting self-supervised representations*. 2023.

[87] Singh, A. K. and Singh, P. Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics. In: IEEE. *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2021, p. 412–417. ISBN 978-1-6654-1865-2.

[88] Singh, J. M.; Ramachandra, R.; Raja, K. B. and Busch, C. Robust Morph-Detection at Automated Border Control Gate Using Deep Decomposed 3D Shape & Diffuse Reflectance. In: IEEE. *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 2019, p. 106–112. ISBN 978-1-7281-5686-6.

[89] Snyder, D.; Garcia Romero, D.; Sell, G.; Povey, D. and Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, p. 5329–5333. ISBN 978-1-5386-4658-8.

[90] Soleymani, S.; Dabouei, A.; Taherkhani, F.; Dawson, J. and Nasrabadi, N. M. Mutual Information Maximization on Disentangled Representations for Differential Morph Detection. In: IEEE. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, 2021, p. 1730–1740. ISBN 978-1-6654-0477-8. Available at: https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00177.

[91] Song, K.; Cong, J.; Wang, X.; Zhang, Y.; Xie, L. et al. *Robust MelGAN: A robust universal neural vocoder for high-fidelity TTS*. 2022.

[92] Ta, B. T.; Nguyen, T. L.; Dang, D. S.; Le, D. L. and Do, V. H. A Multi-task Conformer for Spoofing Aware Speaker Verification. In: IEEE. *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*. 2022, p. 306–310. ISBN 978-1-6654-9745-9.

[93] Tak, H.; Jung, J. weon; Patino, J.; Kamble, M.; Todisco, M. et al. *End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection*. 2021.

[94] Tak, H.; Jung, J. weon; Patino, J.; Todisco, M. and Evans, N. *Graph Attention Networks for Anti-Spoofing*. 2021.

[95] Tak, H.; Patino, J.; Todisco, M.; Nautsch, A.; Evans, N. et al. *End-to-end anti-spoofing with RawNet2*. 2021.

[96] 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.

[97] Thies, J.; Zollhöfer, M. and Niessner, M. Deferred Neural Rendering: Image Synthesis using Neural Textures. *CoRR*. 1st ed., 2019, abs/1904.12356, no. 1. Available at: http://arxiv.org/abs/1904.12356.

[98] Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C. and Niessner, M. Demo of Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In: University of Erlangen-Nuremberg. *ACM SIGGRAPH 2016 Emerging Technologies*. New York, NY, USA: Association for Computing Machinery, 2016. SIGGRAPH '16. ISBN 9781450343725. Available at: https://doi.org/10.1145/2929464.2929475.

[99] Todisco, M.; Delgado, H. and Evans, N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*. 1st ed., 2017, vol. 45, no. 1, p. 516–535. ISSN 0885-2308. Available at: https://www.sciencedirect.com/science/article/pii/S0885230816303114.

[100] Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H. et al. *ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection*. 2019.

[101] Ulutas, G.; Tahaoglu, G. and Ustubioglu, B. Deepfake audio detection with vision transformer based method. In: IEEE. *2023 46th International Conference on Telecommunications and Signal Processing (TSP)*. 2023, p. 244–247. ISBN 979-8-3503-0396-4.

[102] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. et al. *Attention Is All You Need*. 2017.

[103] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P. et al. *Graph Attention Networks*. 2018.

[104] Venkatesh, S.; Ramachandra, R.; Raja, K. and Busch, C. *Face Morphing Attack Generation & Detection: A Comprehensive Survey*. 2020.

[105] Wang, H.; Xie, D. and Wei, L. Robust and Real-Time Face Swapping Based on Face Segmentation and CANDIDE-3. In: Geng, X. and Kang, B.-H., ed. *PRICAI 2018: Trends in Artificial Intelligence*. Cham: Springer International Publishing, 2018, p. 335–342. ISBN 978-3-319-97310-4.

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

[107] Wang, X. and Yamagishi, J. *Investigating self-supervised front ends for speech spoofing countermeasures*. 2022.

[108] Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A. et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*. 1st ed., 2020, vol. 64, no. 1,

p. 101114. ISSN 0885-2308. Available at: https://www.sciencedirect.com/science/article/pii/S0885230820300474.

[109] Watercutter, A. *Why "The Andy Warhol Diaries" recreated the artist's voice with AI*. Conde Nast, Mar 2022. Available at: https://www.wired.com/story/andy-warhol-diaries-artificial-intelligence-voice/.

[110] Woubie, A. and Bäckström, T. Voice Quality Features for Replay Attack Detection. In: IEEE. *2022 30th European Signal Processing Conference (EUSIPCO)*. 2022, p. 384–388. ISBN 978-90-827970-9-1.

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

[112] Yao, Y.; Viswanath, B.; Cryan, J.; Zheng, H. and Zhao, B. Y. Automated Crowdturfing Attacks and Defenses in Online Review Systems. In: ACM. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA: Association for Computing Machinery, 2017, p. 1143–1158. CCS '17. ISBN 9781450349468. Available at: https://doi.org/10.1145/3133956.3133990.

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

[114] Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A. et al. *Defending against neural fake news*. 2019.

[115] Zhang, C.; Yu, C. and Hansen, J. H. L. An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing. *IEEE Journal of Selected Topics in Signal Processing*. 1st ed., 2017, vol. 11, no. 4, p. 684–694.

[116] Zheng, R.; Song, B. and Ji, C. Learning Pose-Adaptive Lip Sync with Cascaded Temporal Convolutional Network. In: IEEE. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, p. 4255–4259. ISBN 978-1-7281-7605-5.

[117] Zhong, W.; Tang, D.; Xu, Z.; Wang, R.; Duan, N. et al. Neural Deepfake Detection with Factual Structure of Text. In: Webber, B.; Cohn, T.; He, Y. and Liu, Y., ed. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, November 2020, p. 2461–2470. ISBN 978-1-952148-60-6. Available at: https://aclanthology.org/2020.emnlp-main.193.

[118] Zhou, X.; Garcia Romero, D.; Duraiswami, R.; Espy Wilson, C. and Shamma, S. Linear versus mel frequency cepstral coefficients for speaker recognition. In: IEEE. *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. 2011, p. 559–564. ISBN 978-1-4673-0367-5.

Appendix A

Contents of the attached storage medium

```
latex_src                <- directory with LaTeX source for this report

src                      <- Python source code for the project
├── classifiers          <- contains the classes for models
│   ├── differential         <- pair-input models
│   └── single_input         <- single-input baseline
├── datasets             <- contains Dataset classes
├── extractors           <- contains various feature extractors
├── feature_processors   <- contains pooling implementation (avg pool, MHFA)
├── scripts              <- output directory for script_generator.py
├── trainers             <- contains classes for training models
├ Makefile               <- Makefile for cleaning cache, packing, script gen
├ README.md              <- repository readme
├ common.py              <- common code, enums, maps, dataloaders
├ config.py              <- hardcoded config, paths, batch size
├ eval.py                <- script for evaluating trained model
├ parse_arguments.py     <- argument parsing script
├ requirements.txt       <- requirements to install in conda environment
├ runner.sh              <- script for running jobs in parallel
├ scores_utils.py        <- functions for score analysis and evaluation
├ script_generator.py    <- generator of job scripts for cloud computation
└ train_and_eval.py      <- main script for training and evaluating models

xstane45.pdf             <- this report in .pdf format
```