POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching

Tesi di Laurea Magistrale in
Music and Acoustic Engineering

Emma Coletta, 214391

Abstract: In recent years, the rise of AI-driven generative models has enabled the creation of highly realistic speech deepfakes—synthetic audio signals capable of mimicking a target speaker's voice with remarkable accuracy—raising critical security concerns. While these developments pave the way to new possibilities in various applications, they also raise serious security concerns, particularly in the domains of authentication, audio forensics, and misinformation. Existing methods for detecting speech deepfakes primarily rely on supervised learning, which suffers from two critical limitations: generalization to unseen synthesis techniques and interpretability regarding model decision-making processes. The need to overcome these limitations has prompted recent research toward novel strategies for detecting speech deepfakes. These strategies include one-class-based approaches to address generalization issues and anomaly localization techniques to enhance the interpretability of detection systems, enabling more robust and explainable detection of speech deepfakes. To address these issues, this thesis introduces a novel interpretable one-class detection framework, reframing speech deepfake detection as an anomaly detection problem. Our proposed approach leverages a model trained exclusively on real speech, learning to characterize its intrinsic distribution to distinguish genuine speech from out-of-distribution synthetic samples. A key feature of our framework is its ability to produce interpretable anomaly maps during inference, highlighting anomalous regions across both time and frequency domains. This is achieved through a Student-Teacher Feature Pyramid Matching (STFPM) system, enhanced with Discrepancy Scaling (DS) to improve generalization capabilities across diverse and unseen data distributions. Extensive evaluations indicate that our method outperforms the considered baselines, showcasing its effectiveness in both detecting deepfakes and localizing artifacts within an investigated speech sample. Our work highlights the potential of anomaly detection-based strategies in strengthening the reliability and interpretability of speech deepfake detection systems.

Advisor:
Prof. Paolo Bestagini

Co-advisors:
Davide Salvi
Viola Negroni

Academic year:
2023-2024

Key-words: Multimedia Forensics, Audio Forensics, Speech Deepfake, Explainability, Anomaly Detection

1. Introduction

In the age of rapid digital transformation we live in, the integration of Artificial Intelligence (AI) into everyday life has unlocked unprecedented creative and educational opportunities. However, it has also introduced significant challenges. The rapid advancement of AI-driven generative systems has made the creation of high-quality synthetic content easier than ever. While it does not go unnoticed how the democratization of such generative tools offers numerous benefits across many creative and scientific domains, our society's growing reliance on digital media content makes us deeply aware of the societal risks this proliferation entails. The ability to manipulate and generate hyperrealistic multimedia content has opened the door to new forms of digital deception, potentially enabling identity theft, financial fraud, and misinformation campaigns [5]. As sophisticated generative techniques become increasingly accessible, the need for robust countermeasures to detect and mitigate the risks associated with synthetic media has never been more pressing for the scientific community [16].
Among the various forms of synthetic media, content generated leveraging advanced DL-based architectures represents a particularly concerning threat. We refer to this specific class of manipulated content as deepfakes, i.e., synthetic data generated through deep learning techniques that can convincingly depict individuals performing actions or pronounce utterances that do not adhere to reality [26]. A particularly concerning aspect within this domain is the rise of speech deepfakes. Unlike visual manipulations, which often leave perceptible artifacts, synthetic speech can be generated with relatively low computational costs while maintaining high perceptual realism. This makes deepfake speech an effective and widely accessible tool for deception, with potential applications in identity theft, blackmail, and security breaches.

To counteract these threats, researchers within the multimedia forensics community addressed the challenge of speech deepfake detection proposing many and diverse strategies from the analysis of low-level acoustic features [11, 75], to that of higher-level semantic aspects [15]. Over time, synthetic speech detection methods have evolved to sophisticated machine learning and DL-based classification methods. Within this category, we can identify frameworks leveraging traditional machine learning techniques, [20], advanced deep learning architectures [49, 95], and pre-trained models [19, 68]. Convolutional Neural Networks (CNNs) and ResNet-based models proved particularly effective in distinguishing real from synthetic speech [22, 86, 99]. More recently, end-to-end architectures, which bypass traditional handcrafted features, have emerged as promising alternatives [77]. However, while demonstrating strong performance in detecting speech deepfakes, classification-based approaches face several limitations, including generalization and interpretability.

Generalization, in this context, refers to a model's ability to accurately detect deepfakes produced by previously unseen speech synthesis techniques, ensuring robustness across a broad range of generative methods. Many existing deepfake detection systems rely on supervised learning, where models are trained to perform classification on explicitly labeled datasets, thus learning the statistical distribution of both counterfeit and genuine speech. While effective in recognizing known deepfakes, a detector trained solely on a fixed set of deepfake examples, potentially produced employing the same manipulation strategy, may struggle to recognize artifacts introduced by novel or unseen generation techniques, leading to reduced detection performance in real-world applications. This lack of adaptability poses a significant challenge, as it requires constant retraining with new data to maintain effectiveness, a process often impractical and not always feasible given the rapid evolution of generative AI.

Interpretability, on the other hand, concerns the lack of transparency of deepfake detection models, particularly those based on deep learning. Many state-of-the-art (SOTA) detection methods rely on complex neural networks that function as "black box" systems, therefore providing predictions without offering clear explanations for their decision-making. This opacity hinders our understanding of why a particular speech sample is classified as real or fake, limiting the trustworthiness of these models in forensic and security applications. Furthermore, in scenarios such as legal proceedings or biometric authentication, the ability to provide human-understandable justifications is crucial [61]. Without interpretability, decision-makers may hesitate to rely on automated detection systems, thereby reducing their practical utility in combating the malicious deployment of synthetic media.

The imperative to overcome these limitations —poor generalization and lack of interpretability— has motivated recent research to explore novel strategies for detecting speech deepfakes. One promising direction to overcome generalization issues involves the use of One-Class (OC) systems. These models are trained exclusively on authentic data and characterize the distribution of normal speech by establishing a boundary around it. Any samples falling outside this boundary, thus deviating from the expected distribution, are flagged as synthetic. This approach helps address the challenge of detecting deepfakes generated by previously unseen methods.
Concurrently, efforts to improve interpretability have focused on understanding which specific features of the audio signal influence model predictions. For instance, recent work has sought to identify and highlight which temporal and/or spectral characteristics contribute the most to deepfake detection [65]. By uncovering the key factors that influence both the generation of synthetic speech and its classification, researchers aim to provide valuable insights into the vulnerabilities of generative models while enhancing the transparency and trustworthiness of the detection process.

In this thesis, we propose a novel method that integrates interpretable detection with OC learning principles to address these challenges. Specifically, we reframe the problem of speech deepfake detection as an Anomaly Detection (AD) task, wherein synthetic speech is conceptualized as real speech embedded with anomalies. Also known as outlier or novelty detection, anomaly detection involves identifying rare instances of data that significantly deviate from the majority and do not conform to a well-defined notion of normal behavior, i.e., anomalies [12]. In the context of speech deepfakes, these anomalies manifest as outliers or previously unseen patterns that diverge from the expected distribution of real speech.
Building on the work of Valjakka et al. [81], who successfully applied an OC method originally designed for visual anomaly detection to the audio domain, we present an adaptation of this approach for anomaly detection and localization in synthetic speech signals. Our proposed system employs the Student-Teacher Feature Pyramid Matching (STFPM) framework, enhanced with Discrepancy Scaling (DS), a normalization technique that improves the detection capabilities of the STFPM architecture. Following the student-teacher approach, we train a student network to replicate the feature representations of a teacher network pre-trained on speaker identification, allowing the system to learn the distribution of genuine speech. To optimize feature extraction, we use a modified ResNet-18 tailored for audio spectrogram analysis. During inference, the system computes discrepancies between student and teacher activations at multiple levels of abstraction. These discrepancies are then aggregated to generate an anomaly score, effectively capturing the presence of synthetic speech artifacts.

Extensive experimental evaluations validate the effectiveness of treating speech deepfake detection as an AD problem, demonstrating the superior performance of our framework compared to other SOTA methods. The combination of STFPM and DS proved instrumental in enabling accurate detection and localization of speech artifacts across both time and frequency domains. Beyond providing a robust solution to the detection challenge, our approach also enhances interpretability, providing deeper insights into the underlying factors influencing detection decisions. Additionally, by framing the problem as OC anomaly detection, our method eliminates the need for large labeled speech datasets, which are often scarce or unavailable in real-world scenarios.

The rest of this thesis is organized as follows:

- Section 2 provides background information in the audio forensics field.
- Section 3 formally defines the speech deepfake detection problem and outlines the proposed method we propose to address it.
- Section 4 describes the experimental setup, including datasets, evaluation metrics, and implementation details.
- Section 5 presents and discusses the results of our experimental campaign.
- Section 6 concludes the thesis, summarizing key findings and suggesting directions for future work.

2. State of the Art

This section provides an overview of the existing literature and recent research advancements relevant to the key topics of this thesis. We begin by examining the current state of audio deepfake generation techniques, focusing on the fundamental tools and methodologies used to create synthetic speech. Next, we explore the landscape of speech deepfake detection in Section 2.2, introducing the most common and state-of-the-art approaches for detecting manipulated speech content, commonly referred to as speech deepfakes. We continue by discussing the role of anomaly detection and localization in multimedia forensics (Section 2.3). This section first establishes the foundational concepts of Anomaly Detection (AD) and Anomaly Localization (AL) within the broader domain of multimedia forensic analysis, highlighting key methodologies adopted in the field. We then delve into their specific application to speech deepfake detection in Section 2.3.2. Finally, we briefly examine the significance of explainability in Audio Deepfake Detection (ADD), emphasizing the critical role AD and AL play in enhancing model transparency, interpretability, and trustworthiness.

Once completing this section, the reader should have gained a clear understanding of the current state-of-the-art (SOTA) in speech deepfake generation and detection, with particular focus on the roles AD and AL play in this domain.

2.1. Synthetic Speech Generation

Speech synthesis is a machine learning task focused on converting written text into human-like synthetic speech. In recent years, this field of study has witnessed remarkable progress, largely driven by the rapid development, widespread and deployment of powerful deep learning-based technologies. Deep learning architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have revolutionized the landscape of speech signal processing, achieving remarkable performance across various tasks, including speech synthesis [41]. These architectures enable the generation of highly realistic, human-like sounding audio, often indistinguishable from genuine speech recordings. Modern synthetic speech generation primarily relies on two distinct families of systems: Text-to-Speech (TTS) systems and Voice Conversion (VC) systems. While both aim to generate synthetic speech, they differ significantly in the input signal they receive and in their core functionalities. TTS systems follow a multistage process, leveraging the transformative power of advanced artificial neural networks to transform written text into spoken language. VC systems, on the other hand, take an existing speech signal as input and manipulate it to mimic the voice characteristics of a target speaker.

2.1.1 Text-To-Speech Generative Systems

Modern Text-to-Speech (TTS) systems leverage deep learning architectures, such as sequence-to-sequence models [6, 88] and neural vocoders, to synthesize natural-sounding speech from text. Recent advancements in TTS architectures, including Generative Adversarial Networks (GAN)-based [10, 32, 33, 90] and transformer-based systems [34, 60], have significantly enhanced the quality of synthetic speech enabling models to capture and replicate the subtle nuances of human expression, such as intonation and rhythm (prosody) with remarkable precision.

We can distinguish between two major families of TTS systems: **modular** (or cascated) and **end-to-end** models. Fully end-to-end models integrate the entire synthesis pipeline into a single deep learning framework, exploiting the powerful learning capabilities of deep neural networks which allow for the direct transformation from text to speech without defined intermediate steps [30, 71, 79, 88]. Modular models, on the other hand, follow a staged approach.

Early TTS systems relied on Statistical Parametric Speech Synthesis (SPSS), a method for synthetic speech generation consisting of three distinct stages [96]:

1. **Text-analysis module** – extracts linguistic features from text.
2. **Acoustic module** – chracterizes speech in terms of acoustic features.

3. **Vocoder** – generates the final waveform.

With the advent of deep learning, traditional SPSS-based TTS systems evolved in two major ways: some architectures retained the three-module structure but replaced each stage with deep neural networks, while others, such as WaveNet [50], Deep Voice [6] and HiFi-GAN [32], follow a two-stage pipeline consisting of:

1. **Acoustic Model: from Text to Spectrogram.** Input text is converted into a time-frequency representation of audio, namely the spectrogram. A spectrogram visually represents the frequency spectrum of an audio signal as it varies over time, making it an essential tool for capturing acoustic nuances of speech, such as rhythm and intonation. Widely used in modern speech synthesis is the mel-spectrogram, which transforms the linear spectrogram onto the Mel-scale to better mimic human auditory perception.
2. **Neural Vocoder: from Spectrogram to Waveform.** The neural vocoder is a deep learning-based generative technology that converts the spectrogram returned by the acoustic model into a waveform, producing high-fidelity, natural-sounding speech.

It should be noted that the two aforementioned modules are independent and are often used separately to accommodate different research and application needs (i.e. feature engineering, speech waveform synthesis).

The transition from SPSS-based synthesis to deep learning-driven architectures has led to significant improvements in naturalness, speaker adaptability, and overall synthesis quality of TTS generative systems. These enhancements paved the way for highly realistic speech generation, which proved versatile in a wide range of applications in speech-related tasks such as education, entertainment and accessibility [9, 28, 41].

2.1.2 Voice Conversion

Voice Conversion (VC) is a technique seeking to modify an input speech signal by manipulating the characteristics of a source speaker's voice to resemble that of a target speaker, while preserving linguistic content of the utterance [83]. Unlike TTS, which generates speech from text, VC operates on existing audio recordings.

Priori to the deep learning era, VC technologies relied on statistical methods and signal processing techniques such as Gaussian Mixture Modelss (GMMs) [74] and Vector Quantization (VQ) [1] to map and reproduce speaker-specific features.

VC techniques can be broadly categorized into **parallel** and **non-parallel** methods [43]:

1. **Parallel VC** [17] requires paired training data, where both the source and target speakers say the same utterances, ensuring the same linguistic content (i.e. parallel utterances).
2. **Non-parallel VC** [24, 55] removes the requirement for paired training data, allowing for VC even when speech data of source and target speakers is unpaired (i.e. unparallel utterances).

The advent of deep learning significantly enhanced the performance of VC systems. Deep VC systems exploit the powerful learning capabilities of deep neural networks to characterize speaker-specific attributes of both the source and target voices, to then perform transformation of acoustic features. This process leads to the generation of high-fidelity speech conversions that are both natural-sounding and virtually indistinguishable from the target speaker's voice. Deep VC techniques include Variational Autoencoderss (VAEs) [14], GANs [24, 55], and CycleGANs [25, 41, 94].

Research in VC is rapidly progressing towards systems capable of generating increasingly high-fidelity and natural-sounding speech, unlocking a wide range of beneficial applications, from enhancing entertainment with personalized voice acting to enabling accessibility tools for speech-impaired individuals [72]. However, the potential for misuse of VC, particularly in creating deepfakes for impersonation and misinformation, highlights the need for responsible development and deployment of these technologies.

2.2. Speech Deepfake Detection

Speech deepfakes can be broadly classified into two major categories: **fully fake speech signals** and **partially fake speech signals** [36]. Fully fake speech signals are entirely synthesized audio signals, generated from text or other non-speech inputs. Partial manipulations, on the other hand, involve altering existing speech recordings, such as VC or inserting synthetic artifacts into real audio, i.e. splicing artifacts [48]. Both categories pose significant challenges for deepfake detection systems, especially as generative models become increasingly sophisticated.

The challenge of speech deepfake detection has been tackled in various ways by the multimedia forensics community, including the analysis of spectral features [51, 63], as well as the use of Automatic Speaker Verification (ASV) [78], Automatic Speech Recognition (ASR) [68] and Speech Emotion Recognition (SER) [15] systems. These diverse approaches reflect the complexity of distinguishing between genuine and synthetic speech.

2.2.1 Supervised Learning Methods

Several existing detection methods fall into the broad category of **supervised learning methods**, which learn to distinguish real (i.e., bonafide) and fake (i.e., spoofed) speech signals using deep learning models [97] or hand-crafted features [11, 39, 91, 92]. Supervised learning approaches train classifiers on labeled datasets, learning to characterize and recognize patterns of both genuine and manipulated speech. While these methods demonstrated overall high performances on tasks such as ASV [35, 77] and ASR [76], their heavy reliance on fake samples in the training stage, along with the intrinsic assumption that real and synthetic speech have a similar distribution, severely limits their ability to generalize to novel deepfakes [98]. In essence, a model trained on a specific type of deepfake may fail to detect new, unseen manipulation techniques. These systems' lack of generalization capabilities across different manipulation techniques makes them extremely vulnerable in the rapidly evolving landscape of deepfake generation.

While some studies have proposed alternative solutions to improve generalization within the supervised learning domain [49, 67], others have explored unsupervised learning methods as an alternate approach to overcoming this limitation [4, 29, 98].

2.2.2 Unsupervised Learning Methods

Unlike supervised learning-based systems, unsupervised learning approaches learn to classify data by identifying underlying patterns within data itself, without relying on explicit labels. For instance, in the context of speech deepfake detection, unsupervised methods can learn the typical statistical properties of genuine speech and consequently flag any audio segments that exhibit significant pattern deviations as potential deepfakes.
Unsupervised methods for speech deepfake detection follow different approaches, including:

1. **Clustering-based algorithms**. Clustering algorithms are Machine Learning (ML) algorithms that seek to group similar data points based on specific criteria, thereby revealing underlying structures or patterns within a dataset. In unsupervised learning, these algorithms operate without prior knowledge of class labels, identifying patterns and structures within the data. Data instances are then grouped based on inherent similarities within the feature space [93]. Clustering strategies proved efficient in enhancing both robustness and explainability of ADD systems [13, 89].
2. **One-Class (OC) Classification.** OC classification-based methods learn to characterize the distribution of normal data (i.e., genuine speech in our discussion) in an unsupervised manner and establish a boundary around it. Any samples falling outside this boundary are flagged as potential anomalous cases, namely deepfakes [29]. This approach is particularly advantageous when fake speech samples are scarce or unavailable in the investigated dataset, as it requires training only on real speech. Furthermore, OC classification methods can detect novel deepfakes that deviate significantly from the learned normal distribution, offering improved generalization

capabilities. OC classification methods applied to speech deepfake detection obtained SOTA results in distinguishing between bonafide and spoofed samples [4, 18, 29, 98].

3. **Anomaly Detection.** Also referred to as outlier or novelty detection, anomaly detection is generally understood by the scientific community as the process of identifying rare instances of data (events or observations) that significantly deviate from the majority of the data and do not conform to a well-defined notion of normal behavior, namely anomalies [12]. Since labeled data is often non-existent or unavailable in standard AD settings, unsupervised learning is the predominant approach for this task. [62].

2.3. Anomaly Detection and Localization

We refer to **Anomaly Detection** as the process of identifying data patterns which suspiciously deviate from the expected or normal behavior of patterns within a dataset. A natural extension of the aforementioned approach is **Anomaly Localization**, a process which not only seeks to indicate whether anomalies are present, but also to pinpoint the specific location within the representation of data where anomalies occur. The performance of anomaly detection and localization is evaluated on **anomaly scores**, which serve as indicators of likelihood and/or severity of anomalies within the investigated collection of data. The higher the score, the greater the likelihood and/or severity of anomalies occurring.

The application of deep learning frameworks to the task of AD is commonly referred to as *deep anomaly detection* in the scientific community. Deep anomaly detection leverages deep neural networks to learn feature representations or directly compute anomaly scores for detecting outliers [53]. deep learning-based AD systems have demonstrated remarkable performance in addressing detection challenges across a variety of real-world applications including fraud detection in finance [2] and telecommunications [100], medical diagnosis [37] and industrial fault and damage detection [58, 80].

In summary, deep AD and AL offer a robust approach for identifying and localizing anomalies, paving the way for the development of advanced detection systems with improved explainability across various fields, including the critical domain of multimedia forensics.

2.3.1 Anomaly Detection and Localization in Image and Video

The challenge of deepfake detection within the realm of synthetic multimedia has often been tackled by the multimedia forensics community as a problem of AD. Recent research demonstrates how AD and AL methods play an instrumental role in identifying and localizing manipulations within synthetic image and video content representing real-world scenarios. For example, Schlegl et al. [70] managed to successfully identify and pinpoint anomalies within clinical imaging by leveraging the unsupervised learning capabilities of GANs. In [31], Kiran et al. provide an overview of deep AD and methods that have proven effective in detecting manipulations within video content, particularly in the field of video surveillance. Finally, in [84], Wang et al. apply AD and AL on a dataset of high-resolution images of industrial products, specifically created to benchmark AD systems. With an original implementation, the Student-Teacher Feature Pyramid Matching (STFPM) method, they achieve remarkable results in both image-level and pixel-level detection of anomalies within faulty images of the investigated dataset.

2.3.2 Anomaly Detection and Localization in Audio

Anomalous Sound Detection (ASD) is the process of automatically identifying audio events or segments that deviate from typical or expected acoustic patterns within an audio recording. The addressed challenge has gained significant attention from the multimedia forensics community in recent years due its wide range of real-life applications such as industrial monitoring [81], medical diagnosis and public-safety measures [42].
Building upon sound AD, sound AL aims to pinpoint the precise temporal or spectral location of the anomalous sound within an audio stream, providing detailed information about the onset, duration,

and frequency characteristics of the anomalous event (i.e. the synthetic artifact), which can be categorized into:

1. **Temporal localization:** identifying the specific time intervals where the anomaly occurs.
2. **Spectral localization:** identifying the specific frequency bands where the anomaly is concentrated.

In the context of speech deepfake detection, AD systems aim to identify speech audio segments that exhibit acoustic properties or patterns that significantly deviate from those characterizing genuine speech. When extended to AL, the objective is not only to detect suspicious patterns but also to localize the temporal and/or spectral regions within potentially manipulated speech samples.

Speech AD systems in the literature mostly leverage acoustic feature representations such as Mel Frequency Cepstral Coefficients (MFCCs) and mel-spectrograms [73] to characterize bona-fide speech and then reveal suspicious patterns (classified as synthetic artifacts) in the corresponding representation of manipulated audio.

2.4. Explainability for Audio

In general, the goal of explainability is to make some aspects of a system understandable by humans. Explainable Artificial Intelligence (XAI) is the field of study that seeks to develop methods and frameworks for elucidating the reasoning behind AI systems and their outputs. The increasing complexity and rapid advancement of AI architectures, particularly in ML and deep learning, has made XAI methods essential for enhancing the interpretability and transparency of AI-based systems [21, 38]. Recent studies [5, 27, 54] highlight the crucial role of explainability in ensuring accountability, trustworthiness and robustness of AI-driven systems.

The risks associated to the potential misuse of audio deepfakes for malicious purposes (impersonation through voice cloning) and broader ethical concerns arising from the fast development and unsupervised deployment of synthetic speech technologies, has drawn significant attention from the forensic scientific community towards XAI. In forensic investigations, every decision must be both reproducible and legally defensible, making the clarity provided by XAI indispensable [5].
In the context of speech deepfake detection, XAI methods play a fundamental role in attempting to understand the critical features of audio that drive and influence the predictions of a speech deepfake detection system [64].

In recent years, substantial progress has been made in applying XAI to ADD systems. Researchers have explored various methods to visualize and interpret the decision-making processes of these systems, including attention mechanisms in deep learning models, saliency maps, and feature importance analysis [3]. For instance, Becker et al.[7] investigated the explainability of deep models in the audio domain trained on raw audio and spectrograms across various classification tasks, while Salvi et al. [65] demonstrated that focusing on single spectral bands can improve artifacts detection accuracy. In essence, XAI method applied to the ADD field aim to highlight the specific audio characteristics that contribute to the detection of deepfakes, providing insights into the vulnerabilities of generative models. Reconstruction-based AD methods, such as autoencoders, can further enhance explainability by identifying segments of an audio signal with high reconstruction errors, often indicative of potential deepfake manipulations within the recording. When combined with AL, these methods can produce saliency maps (i.e., an image highlighting only a specific region) that pinpoint the specific regions in the audio signal responsible for the anomaly. These localized insights allow humans to, not only understand *where* the artifact occurs, but also *why* the system flagged those regions as anomalous, thereby increasing both transparency and trustworthiness of the detection process.

Building on this, in the following sections, we will examine whether our proposed ADD method, based on AD and AL, can similarly provide a clearer understanding of *where* speech generation artifacts occur in synthetic speech and, consequently, *why* a detection system identifies those regions as anomalous.

3. Proposed Method

This section formalizes the problem we address and presents our considered system.
In summary, we address the problem of speech deepfake detection and propose a novel Anomaly Detection (AD) framework to tackle this task. Our approach is based on the hypothesis that synthetic speech can be conceptualized as genuine, real, speech embedded with anomalies. These anomalies arise due to imperfections in the generative process, which introduce subtle artifacts that distinguish synthetic speech samples from real ones. Following this perspective, we train an anomaly detection model exclusively on authentic speech data, enabling it to learn the statistical distribution of genuine speech. During inference, any deviation from this learned distribution is considered an indicator of synthetic speech and is used to perform deepfake detection. Also, our approach enables anomaly localization in both time and frequency domains, providing additional forensic insights into the input speech signal and serving as a valuable Explainable Artificial Intelligence (XAI) tool.

3.1. Problem Formulation

The Audio Deepfake Detection (ADD) problem can be formally defined as follows. Let us consider a discrete-time input speech signal $\mathbf{x}$ associated with a class $y \in \{0, 1\}$, where $y = 0$ denotes that the signal is authentic and $y = 1$ indicates that it has been synthetically generated. The goal of this task is to develop a detector $\mathcal{D}$ that is able to estimate the class of the signal $\mathbf{x}$ as $\hat{y} \in [0, 1]$, where $\hat{y}$ is the likelihood of the signal $\mathbf{x}$ being fake.
Unlike traditional supervised learning approaches, we adopt a one-class AD strategy, wherein $\mathcal{D}$ is trained exclusively on authentic speech samples to learn their underlying distribution. During inference, the model evaluates how much a given speech signal deviates from this learned distribution. Signals that closely align with the distribution of genuine speech are assigned a low $\hat{y}$ value, indicating authenticity. Conversely, signals exhibiting significant deviations are assigned a higher $\hat{y}$ value, indicating a likely deepfake. Beyond classification, our framework incorporates anomaly localization, allowing the model to identify specific regions in both the time and frequency domains where synthetic artifacts are most concentrated. This enhances interpretability and provides deeper forensic insights into the specific regions of the signal that exhibit anomalous characteristics.

3.2. Considered System

In this section, we outline the core framework of our proposed approach to speech deepfake detection. We first introduce the Student-Teacher Feature Pyramid Matching (STFPM) method, an anomaly detection technique originally developed for images, and describe how we adapt it to the audio domain. We then present Discrepancy Scaling (DS), a normalization technique developed to enhance the detection and localization capabilities of the STFPM framework. Finally, we detail the architecture of our *feature extractor*, which serves as the backbone for capturing relevant speech-specific representations from acoustic features of raw audio samples.

3.2.1 Student-Teacher Feature Pyramid Matching

The STFPM is a method developed by Wang et al. [84] to address the challenges of image anomaly detection and localization. This method relies on two Convolutional Neural Network (CNN)-based networks with identical architectures, arranged in a teacher-student configuration [85]. The core idea behind STFPM is knowledge distillation, where the teacher network guides the student network in learning normal patterns to effectively detect and localize defective regions within anomalous instances of a dataset collecting high-resolution images of industrial products.

In their proposed system, the teacher is a network pre-trained on the task of image classification on a large-scale image dataset. Aligned with the principles of the traditional teacher-student framework, the student network is trained to imitate the teacher's activations and learn the distribution of normal (anomaly-free) images by aligning its extracted features with those of the pre-trained counterpart (e.g. the teacher). To achieve this, STFPM leverages Feature Pyramid Matching (FPM), a technique

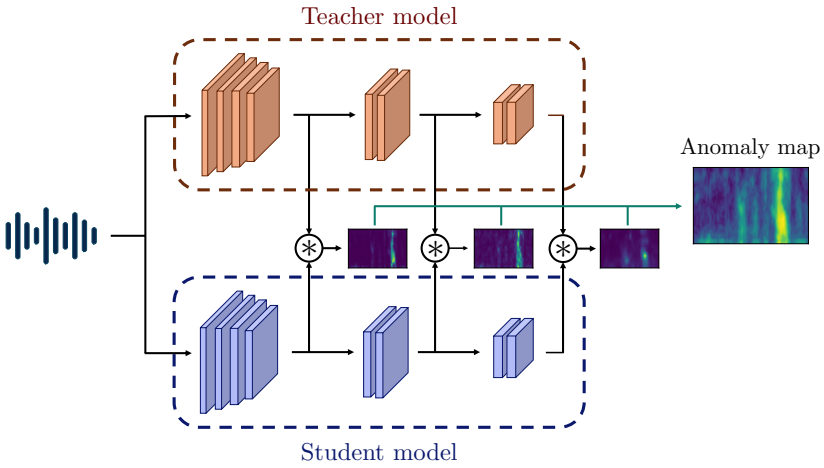


Figure 1: Pipeline of the proposed AD framework for speech deepfake detection.

that enables multi-scale feature comparison between the activations of the teacher and student networks, resulting in the ability to capture anomalous patterns at different levels of abstraction. By framing AD in this way, STFPM effectively learns the characteristics of normal data during training. Since the student network is optimized to mimic the teacher's feature representations on normal (anomaly-free) samples, it struggles to do the same when processing non-normal inputs, leading to noticeable deviations in feature maps. These deviations are then aggregated to generate anomaly maps, highlighting regions in the input image that are likely to contain defects.

In our proposed AD framework, we adapt the STFPM methodology to the domain of speech deepfake detection by replacing image-based inputs with acoustic representations extracted from the audio signal under analysis. Instead of relying on a teacher network pre-trained on an image classification task, we train our teacher model on the *speaker identification* task, enabling it to learn robust speech-specific representations from genuine speech data. The student network, in turn, is trained to replicate the teacher's feature activations through FPM [81], facilitating a structured comparison between the two networks.
During student training, the teacher network remains completely frozen, meaning its weights are not updated. Instead, it serves as a fixed reference model, using the parameters learned during its pre-training stage to extract feature representations from input spectrograms. These teacher-generated features act as the target embeddings that the student must approximate. By comparing the student's extracted features with those of the teacher at multiple network depths, the model learns to mimic the hierarchical representation of speech captured by the teacher.

Figure 1 illustrates the pipeline of the proposed system.
Since the teacher network is trained for speaker identification, its learned feature representations inherently capture the statistical distribution of genuine speech. The student network, learning to imitate these activations, aligns with the principles of One-Class (OC) anomaly detection, where deviations from the expected distribution signify the presence of anomalies within the analyzed signal. Additionally, FPM facilitates the detection of synthetic speech by capturing discrepancies between the hierarchical activations of the teacher and student networks. Lower layers of the network are responsible for encoding fine-grained, low-level acoustic features, while deeper layers extract more abstract, high-level patterns.

At inference, the two networks process both real and synthetic speech samples. The discrepancies between the normalized activations of the teacher and student models are then aggregated into what is referred to as the **anomaly map**, which serves as the primary anomaly indicator for both anomaly detection and anomaly localization. When analyzing real speech, we expect the student to closely replicate the teacher's activations, yielding low discrepancy values and an anomaly map with minimal deviations. Conversely, when processing deepfake speech, the student struggles to replicate the teacher's feature representations, resulting in higher discrepancy values. The resulting anomaly

map highlights time-frequency regions where the generation artifacts are most prominent, providing a valuable tool for both detecting and localizing synthetic speech anomalies.

3.2.2 Discrepancy Scaling

To further enhance the performance of our AD model, we incorporate DS [47], a normalization technique developed to standardize the discrepancy values between teacher and student activations. Since the activations differences vary across network layers and input samples, DS applies a layer-wise scaling strategy to ensure that discrepancies from different feature levels contribute proportionally to the final anomaly map and, consequently, the anomaly score. This prevents dominant layers from overshadowing subtle but meaningful variations in the data, leading to more consistent anomaly detection and localization across the processed dataset.
The scaling function is designed to emphasize anomalous regions while preserving the relative importance of different feature levels. Specifically, we apply DS in the attempt to maintain a consistent dynamic range across multiple layers and ensure that all meaningful anomalies are accounted for in the final anomaly map.
Normalization Process. DS involves a pre-processing step in which we calculate mean discrepancy and the standard deviation of the test data distribution across all layers. At inference time, these statistics are leveraged to normalize the anomaly scores layer by layer by applying the following equation:

$$A'_i = \frac{A_i - \mu_i}{\sigma_i + \epsilon}, \quad i = 0, 1, 2 \tag{1}$$

where:

- μ_i: precomputed mean of the discrepancy values at layer i, estimated from normal data.
- σ_i: precomputed standard deviation of the discrepancy values at layer i , capturing variability of normal activations.
- ϵ: a small scalar value added to prevent division by zero, thus ensuring stability in the computation.

This normalization technique amplifies subtle anomalies, enabling the model to detect fine-grained synthetic artifacts that might otherwise go unnoticed. Additionally, we employ DS as a calibration strategy to improve the detection performance of the model on out-of-distribution data, as we will show in Section 5.

4. Experimental Setup

In this section we illustrate implementation details of our setup. We begin by discussing the data sets chosen to conduct our study, along with the evaluation metrics used to evaluate our proposed method. We also provide technical specifics driving the training of both our teacher and student models.

We anticipate the discussion by stating that we trained and evaluated all our models on *windowed* audio samples, meaning that we only considered a short segment of the entire audio clip. We fixed the length of our signal window at 4 s to find a compromise between achieving high temporal resolution and ensuring sufficient contextual information for effective feature extraction and model learning. We ensured that all audio samples met the minimum required length (single window) by padding them if they were shorter. Padding was performed by repeating the signal until it reached the necessary length, ensuring continuity without introducing silence or artifacts. To introduce variability and robustness during the training and validation phases of our models, we windowed signals longer than the required length by extracting a fixed-length segment (single-window length) starting from a random point within the range $[0,\ len(\mathbf{x}) - len(\mathbf{w})]$, where $\mathbf{x}$ is the audio and $\mathbf{w}$ the window. As for the audio data used in the evaluation phase, we applied the same padding strategy, while windowing was performed by simply taking the first window-length samples of the processed track.

Moreover, we processed all audio data for both the training and evaluation steps at a sampling frequency of f_s =16 000 Hz.

4.1. Datasets

LJ-Speech [23]. It is a wide corpus of genuine speech audio recordings. Widely used in the field of synthetic speech research, it consists of 13 100 clips of a single female speaker reciting passages from non-fiction books, totaling approximately 24 hours of recorded speech. We consider this set to be the authentic (i.e. bona-fide) counterpart to all the single-speaker experiments. We do so since all datasets used to test our single-speaker model contain speech tracks from the same speaker, making it more challenging to discriminate between real and fake versions of the same voice.

ASVspoof19 [87]. Released for the homonymous challenge, ASVspoof19 is benchmark dataset specifically designed to develop effective Automatic Speaker Verification (ASV) models. It encompasses speech data from 78 different speakers (33 male and 45 female), providing a diverse collection of both bona-fide and synthetic speech. All bona-fide speech data are based on the VCTK corpus [82] while synthetic speech data were fabricated by 19 different synthetic speech generators, thus cover various spoofing attacks generated through Text-to-Speech (TTS), Voice Conversion (VC) and hybrid TTS/VC techniques. The Logical Access (LA) partition of this corpus, which is the one we employed to conduct this study, contains deepfakes further divided into training, development and evaluation subsets. Data within the LA partitions are distributed unevenly across the three subsets allowing for open-set evaluation.

TIMIT-TTS [66]. Based on the VidTIMIT corpus [69], TIMIT-TTS is a speech dataset containing counterfeit audio tracks only, for a total of almost 80 000 tracks generated using 12 different state-of-the-art (SOTA) TTS methods. This dataset is further subdivided into four partitions, each corresponding to a different post-processing operation applied by the authors to generate the spoofed speech tracks. For this study, we only employ the *clean* partition, for which no post-processing has been performed on audio samples after the TTS generation process. To conduct our experiments, we will only employ the *single_speaker* split of this partition, that includes all the tracks generated using TTS algorithms that support LJSpeech's speaker, for a total of 10 different SOTA generators. This split features 430 tracks for each of the generators.

MLAAD [46]. This large-scale corpus encompasses 378.0 hours of synthesized speech in 38 different languages and generated using 82 state-of-the-art TTS models. To conduct our experiments, we employ only English-language speech tracks (EU and US subsets) for our single-speaker scenario only, selecting speech tracks which voice matches the one of its authentic counterpart, namely LJSpeech.

Purde Speech Dataset [8]. Recently released, this corpus includes 25 000 spoofed tracks accounting for 52 hours of synthetic speech generated using 5 diffusion-model based voice cloning methods (ProDiff, DiffGAN-TTS, ElevenLabs, UnitSpeech and XTTS). This dataset spans across 11 different speakers, with 5000 synthetic speech signals generated using each of the five methods at a sampling frequency of f_s = 16 kHz. The real speech data were sourced from both the LJSpeech and the LibriSpeech corpora.

AISEC "In-The-Wild" [44]. Specifically designed to evaluate speech deepfake detection detectors in real-world conditions, this dataset includes 38 hours of audio clips (17 hours fake, 21 hours real) featuring English-speaking celebrities and politicians. The fake clips are created by segmenting 219 of publicly available video and audio deepfake content, while real clips are manually collected from sources like podcasts and speeches. Since the speakers talk absurdly and out-of-character, it is easy to verify that the audio files are really spoofed. We employ this dataset to conduct experiments in the multi-speaker scenario.

FakeOrReal [59]. This is a dataset specifically designed to facilitate research in the field of speech deepfake detection and is comprehensive of both real and synthetic speech, totaling over 111 000 utterances for the former category and over 87 000 utterances from 33 different speakers for the latter,

respectively. Authentic data were collected from a series of open source speech datasets, including platforms like TED talks and YouTube, while counterfeit data were generated using both open-source and commercial speech synthesis algorithms including Deep Voice 3[56], Google WaveNet[50] and Microsoft Azure-TTS. The authors released this dataset in 4 versions, namely *for-original, for-norm, for-2seconds and for-rerecorder*, which differ in the pre-processing applied to audio signals. To conduct our experiments in the multi-speaker scenario, we only employ the *for-original* version which includes a total of 195 541 utterances comprising of real and synthetic speech from both male and female speakers.

Generated Dataset. We created our own corpus of synthetic speech by employing 7 different neural vocoders, including HiFiGAN [32] and MelGAN [33] (in three different configurations), as well as the Griffin-Lim algorithm for waveform reconstruction. These techniques were used to synthesize tracks from samples constituting a partition of LJSpeech. As a result, this dataset consists of 8 subsets, each comprising 5000 spoofed speech samples. We used this dataset to conduct evaluation of our student model within the single-speaker scenario. Additionally, due to the direct correspondence between real and spoofed audio samples, this dataset served as the basis for investigating the Anomaly Localization (AL) capabilities of our model, thereby enhancing the explainability of the detection process. Section 4.1 provides a detailed overview of this corpus' composition.

Generators	**Num. Tracks**	**Num. Speakers**
HIFIGAN	5000	1
HN-SINC-NSF	5000	1
HN-SINC-NSF [hifi]	5000	1
WAVEGLOW	5000	1
MELGAN [bestnetG]	5000	1
MELGAN [linda_johnson]	5000	1
MELGAN [multi_speaker]	5000	1
GRIFFIN-LIM	5000	1
Total	40000	1

Table 1: Composition of our internal dataset.

Being our method One-Class (OC)-based, training of all our networks was performed on authentic data only. For the evaluation of the student network, for which we investigated robustness and generalization of the Anomaly Detection (AD) process and model explainability through the AL process, we employed both real and synthetic data.

4.2. System Configuration and Training

The model that we considered for both teacher and student networks is a modified version the ResNet-18 architecture, a Convolutional Neural Network (CNN) originally designed for image classification. To adapt it for audio processing, we modified the model to operate on visual representations of audio, namely spectrograms, instead of images. To achieve this, since spectrograms are single-channel representations the first convolutional layer of ResNet-18 is adjusted to accept one-channel inputs instead of three (i.e. the RGB channels in images). The original model's weights are averaged across channels to initialize the modified layer while maintaining the network's pre-trained structure. Additionally, the final fully connected layer is replaced with a three-layer classifier, including dropout regularization and a ReLU activation function to refine feature extraction.

4.2.1 Teacher Network Training

To perform training of our teacher network, we employed the LibriSpeech dataset [52]. More specifically, we employed the *train-clean-360* subset, which contains speech from a total of 921 speakers.

From this subset, we retained up to 100 samples for each speaker, for ease of use. We further partitioned this data splitting it between training and validation subsets and ensuring a balanced distribution of speakers across the two (thus stratification by speaker identity). We put 70% of the selected tracks towards the training subset, whereas the remaining 30% was put towards the validation subset. To summarize, we trained our teacher network using 62962 utterances and validated it with 26984 utterances from a total of 921 different speakers. Again, this dataset will only be used to support the pre-training of our teacher network and will not be involved in assessing the performance of our proposed system system on the anomaly detection and localization tasks.

The architecture of our teacher network is based on our modified version of the ResNet-18 model, retaining all convolutional, residual and fully connected layers. Since the Student-Teacher Feature Pyramid Matching (STFPM) framework requires a well-trained teacher to guide the student model, we tailored this architecture for speaker identification, ensuring it learns discriminative speech representations. To this end, we structured the network as a classifier taking spectrograms as input. The convolutional layers capture both spectral and temporal patterns, while the fully connected layer produces high-level embeddings encapsulating critical speech characteristics. We therefore achieve a model well-suited for learning discriminative features required for speech classification.

We considered mel-spectrograms as the acoustic feature representation to conduct our study, as they closely mimic human auditory perception. This choice enhances the interpretability of the generated anomaly maps, making them more reflective of how anomalies might be perceived by listeners. Therefore, the model processes mel-spectrogram representations computed from pre-processed raw audio signals (see Section 4.1) with a sample rate of 16 kHz, 512 FFT points, a window length of 512, a hop length of 128, and 128 mel filters. Moreover, the extracted features are normalized before being fed into the network.

We modified the last layer of the network, namely the fully connected layer, to return a number of outputs equal to the number of speaker classes of our training set ($N = 921$) instead of 2.
We built the dataloader to handle potential class imbalances in the dataset. First, we ensured a balanced representation of speakers during training by applied stratified sampling. Moreover, to avoid learning bias toward speakers with more data, we designed the dataloader to sample audio patches with equal probability across all speaker classes regardless of variations in the number of recordings per class.
We perform training of our teacher network with the following specifics:

- Batch size: 64
- Epochs: 300
- Learning Rate: 10^{-4}
- Weight Decay: 10^{-4}
- Loss function: Cross Entropy
- Optimizer: AdamW

During each training epoch, the model updates parameters through gradient backpropagation and optimizer steps. Validation is conducted at the end of each epoch, evaluating both loss and accuracy values.

To improve stability and convergence of the training process, we make use of a Cosine Annealing scheduler, which progressively reduces the learning rate from a maximum value to a minimum value over a specified number of epochs following a cosine decay pattern. The parameter we used as reference to evaluate learning progress is the validation loss. We set the minimum learning rate at 0.000 01 and the maximum number of epochs at 100.

To prevent overfitting, we make use of the *Early Stopping* function, enabling interruption of the training process when no further improvement in the learning process is observed. We set the *patience* value at 15 epochs. Once again, the parameter we used as reference to evaluate learning progress is the validation loss: if a new best loss is achieved, the model is saved and the early stopping counter is reset. Otherwise, the early stopping counter increments. Training stops when this counter reaches the patience value threshold.

4.2.2 Student Network Training

To investigate robustness and generalization capabilities of our method, we conducted two types of analyses: single-speaker and multi-speaker experiments.

- **Single-Speaker:** in this configuration, we focus on a specific target voice for both training and test the detectors. This scenario simulates a Person-of-Interest (POI) use case, where the goal is to protect a specific individual from deepfake attacks by focusing on their unique voice characteristics. For this analysis, we considered the voice of Linda Johnson from the LJSpeech dataset [23] as real data. To minimize bias, we excluded tracks used in the evaluation phase from both training and validation. After this initial skimming, 8100 tracks remained, which we further split into training and validation sets following a 70/30 ratio. Finally, we trained our single-speaker student on 5670 utterances and validated it on 2340 utterances.
- **Multi-Speaker:** in this configuration, we performed training and validation of our student network using all bona-fide samples within the LA partition of the ASVSpoof19 dataset. More specifically, we used the *train* subset for training and the *dev* subset for validation respectively.

The training objective of our student model is to minimize the discrepancy between its feature maps and those of the teacher network at the three final convolutional layers. This is done using a feature-matching loss based on the Mean Squared Error (MSE) between the L2-normalized activations of the teacher and student models. During training, both networks process the same input and generate their own feature representations. The weights of the teacher remain frozen, while those of the student are updated. The loss value is computed by averaging the squared differences across feature maps. As a result, the validation loss for the student indirectly influences the training by encouraging better feature matching rather than direct class-based accuracy, thus facilitating the distillation process on which the teacher-student framework relies.

Aligning with the Feature Pyramid Matching (FPM) method, core principle of the STFPM framework, unlike the teacher, our student feature extractor employs a truncated version of our modified ResNet-18 architecture. The feature extractor processes mel-spectrogram representations computed following the same parameters described in Section 4.2.1. The final residual block, needed for classification but not for AD, is excluded from the feature extraction process. Instead, the student network retains only the convolutional backbone up to the Conv4_x layer, enabling feature extraction at multiple resolutions. The final embeddings produced by the student are feature maps extracted from the last 3 convolutional layers, namely Conv2_x, Conv3_x, Conv4_x .

These multi-scale representations allow the model to detect anomalous patterns within the feature maps at different levels of abstraction, leading to more precise anomaly detection and localization.
We build the train dataloader to avoid bias and improve variability during the training process by randomizing the order in which data is being processed every epoch. For both our single-speaker and multi-speaker student models, we retain the same training specifics used for the teacher model as well as the training and validation strategies, ensuring consistency across analyses. For clarity and convenience, we report again the hyperparameters:

- Batch size: 64
- Epochs: 300
- Learning Rate: 10^{-4}
- Weight Decay: 10^{-4}
- Loss function: Cross Entropy
- Optimizer: AdamW

4.3. Inference Phase

4.3.1 Evaluation Methodology

To assess the generalization capabilities of our student network, we conducted evaluations on a diverse selection of both single-speaker and multi-speaker speech datasets. This analysis aimed to determine how well our proposed system adapts to a broad range of artifacts introduced by different speech synthesis and manipulation techniques in the single-speaker scenario, as well as to variations in

speaker identity, which lead to diverse time-frequency patterns, in the multi-speaker configuration. We assessed the detection capabilities of our framework in both the single-speaker and multi-speaker scenarios, while the explainability of the detection process — reflected in AL — was investigated only in the single-speaker configuration, specifically on our generated dataset. This choice was a natural consequence of the exact correspondence between real and spoofed tracks, as we specifically tailored this dataset ensuring a precise match between each spoofed sample and its real counterpart.

To assess the performance of our **single-speaker** configuration, for the authentic speech counterpart we employ bona-fide utterances from the LJSpeech corpus. Specifically, we consider the 5000 utterances previously excluded from the training and validation phases. As for the spoofed speech counterpart, we employ the following collections of audio data: TIMIT-TTS, MLAAD, Purdue, ouw generated dataset.
To test the detection capabilities of our considered system on speech data from unseen speakers, we evaluated the performance of our **multi-speaker** configuration by employing both bona-fide and spoofed utterances from each of these datasets individually: ASVSpoof19, In-The-Wild, FakeOrReal, Purdue.

4.3.2 Anomaly Maps Generation and Analysis

To quantify how much the activations of the student deviate from those of the teacher, we compute the feature-wise squared difference between the normalized teacher embeddings at each resolution. In line with our proposed system leveraging Discrepancy Scaling (DS) to enhance the STFPM framework, we then standardize these differences using precomputed mean and standard deviation statistics. This ensures a consistent comparison across features extracted by the two networks. Finally, the layer-wise discrepancies are aggregated to form a dense anomaly map that highlights time-frequency regions where discrepancies are more pronounced, indicating potential anomalies. Additionally, to facilitate the comparison between the anomaly map and a ground truth map (which visualizes the discrepancies between the spectrograms of genuine and synthetic speech samples), we perform interpolation on the feature maps extracted from each network layer before aggregation. This ensures that the dimensions of the feature maps match with those of the original spectrogram.

To compute the soft scores, for each anomaly map, we sum all the values to obtain a single scalar score reflecting the overall discrepancy between student network's predictions and the teacher's. Since we treat the teacher's activations as the ground truth for identifying whether a track is authentic or spoofed, a higher soft score indicates a greater likelihood of the track containing anomalies.
As for AL, we evaluate how precisely our model pinpoints anomalies within the mel-spectrogram, therefore artifacts within the audio track, by performing a relatively straightforward comparison operation. For each spoofed track, the model generates an anomaly map, which highlights regions suspected of containing anomalies. We then compare this predicted anomaly map with the corresponding ground truth map, which we previously generated by taking the raw difference between the mel-spectrogram of the spoofed track and that of the corresponding real track.

4.3.3 Discrepancy Scaling Values Computation

To implement the DS method and enhance AD and AL capabilities of the model, we computed mean and standard deviation values of the discrepancies between teacher and student's activations. These values were then used to perform normalization of the anomaly maps generating our anomaly scores.
Single-Speaker Scenario: we computed the mean and standard deviation values using authentic speech from both the training and validation subsets of the single-speaker student model (see Section 4.2.2).
Multi-Speaker Scenario: we computed the mean and standard deviation of discrepancies between teacher and student activations using authentic speech from subsets which we specifically curated to align with each evaluation set, leading to more calibrated statistics. For ASVSpoof19, we used both *train* and *dev* subsets from the authentic speech data within the dataset; for Purdue some speech data that includes a limited amount of tracks from both the single-speaker LJSpeech corpus and the *train-clean-360* partition of LibriSpeech, which constitutes the multi-speaker portion of the collection; for In-The-Wild, we employed 1000 bona-fide speech samples within the dataset, we randomly selected

without taking into consideration balance between speakers; while for FakeOrReal, non-synthetic speech samples from the validation subset of the *for-original* partition of the dataset.

4.4. Evaluation Metrics

In this section, we present the metrics adopted to evaluate the results obtained in this study.
To assess the performance of our system, we rely on a standard evaluation metric widely used in detection and classification tasks, namely the **Receiver Operating Characteristic (ROC)**.
Therefore, we introduce the True Positive Rate (TPR) and the False Positive Rate (FPR), which play a fundamental role in understanding the ROC. The TPR, also known as *sensitivity* or *recall*, provides a measure of how many actual positive instances (e.g. spoofed speech) are correctly classified as positives. Conversely, the FPR provides a measure of how many actual negative instances (e.g. bona-fide speech samples) are incorrectly classified as positive (spoofed).
The ROC curve is therefore a graphical representation showing how the performance of a model varies on a classification task as its classification threshold changes, and it is constructed by plotting TPR (on the y-axis) against FPR (on the x-axis), effectively capturing the trade-off between detection sensitivity and the rate of false positives.
An optimal prediction method would yield a point situated at the upper-left corner of the ROC curve, corresponding to coordinates $(0, 1)$ on the plot, which represent perfect classification. Conversely, a random classifier would produce a point along the diagonal line running from the lower-left to the upper right corners. This diagonal divides the ROC space into two distinct regions: points above the diagonal indicate effective classification of instances, while those below it signify poor performance.
Through the ROC curve, is also possible to determine which is the optimal threshold for converting probability scores, representing the usual outputs of classification models and which we will refer to as "*soft scores*", into binary "*hard scores*". In mathematical terms, we can formally define the optimal threshold p as the one satisfying the following equation:

$$TPR(p) = 1 - FPR(p) \tag{2}$$

This threshold is particularly significant when working with imbalanced datasets, as the conventional threshold of 0.5 may not always yield the best classification performance.

Closely related to the ROC metric is the **Area Under the Curve (AUC)**, which provides a quantitative measure, in the form of a single scalar value, of the two-dimensional area under the ROC curve. This value offers an assessment of the model's overall performance across all possible classification thresholds. AUC ranges from 0 to 1, where a value of 1 signifies perfect classification, a value of 0 indicates perfect classification but with inverted predictions (i.e. bona-fide classified as spoofed and viceversa), and a value of 0.5 suggests no better performance than random guessing. Therefore, the closer the AUC is to 1 (or 100%), the more accurate the model is at correctly distinguishing and classifying the classes. The characteristic curve ROC-AUC is a commonly used to quantify the accuracy of an anomaly detection system for a given test set [40].

Lastly, we introduce the **Equal Error Rate (EER)** metric. Commonly used to assess classification performance of a model, EER is defined as the point where the FPR equals the False Negative Rate (FNR). In other words, it represents the point at which the proportion of bona-fide samples mistakenly classified as spoofed (False Positives (FPs)) is equal to the proportion of spoofed samples mistakenly classified as bona-fide False Negatives (FNs). A lower EER indicates a better-performing model, as it suggests a smaller likelihood of misclassifying instances from both classes. Specifically, a lower EER means that the model can effectively minimize both FPs and FNs, leading to more accurate classifications overall.

4.5. Baseline Systems

To validate the effectiveness of our proposed framework for the detection task, we compare its performance against two baseline systems, each representing a different approach to speech deepfake

Table 2: Detection performance of the considered systems in a single-speaker scenario (%).

	TIMIT-TTS		Purdue		MLAAD		Generated	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
RawNet2 [77]	**99.9**	**0.5**	70.6	35.1	62.2	43.4	51.5	49.8
OC baseline [29]	56.8	46.9	62.4	39.8	78.1	29.7	77.1	29.7
Ours (w/o DS)	99.1	2.4	**100.0**	**0.0**	**94.3**	**13.3**	99.9	1.5
Ours (with DS)	96.3	9.7	**100.0**	0.1	86.0	23.6	**99.9**	**1.2**

detection.

The first baseline we consider is RawNet2 [77], a widely used supervised learning model trained to differentiate between real and synthetic speech. This model has demonstrated state-of-the-art performance in detecting deepfakes by leveraging raw waveform-based feature extraction and deep learning techniques. The second baseline we consider is a ResNet-18 model trained following the OC approach introduced in [29]. While in the original implementation, the authors utilized both genuine and spoofed speech data for training and validation, we modify this setup by excluding any synthetic samples during training. This adjustment ensures a fair comparison with our method, whose learning process relies solely on real data.

These baselines were selected to highlight two key aspects: first, the improved generalization capability of a OC model compared to a supervised classifier, which can struggle with unseen attacks and second, the advantages of the proposed AD framework over an existing OC approach, demonstrating its ability to refine speech deepfake detection by framing it as an anomaly detection problem.

5. Results

In this section, we present and discuss the results achieved with our proposed framework. We evaluate its performance in speech deepfake detection, reframed as an anomaly detection task, using Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC), and Equal Error Rate (EER) metrics introduced in Section 4.4.

We present results in this order: first, we assess our framework's performance on the detection task across both considered scenarios — single-speaker and multi-speaker. This section also includes a comparison with the two baseline models. Next, we present results for the anomaly localization challenge, beginning by demonstrating the the feasibility of our approach when testing the model in a controlled scenario. Finally, we investigate the interpretability of the detection process by examining the anomaly maps generated using our dedicated dataset, specifically designed for this purpose.

5.1. Single-Speaker Scenario

As a first experiment, we evaluate the considered system in a single speaker scenario, simulating a Person-of-Interest (POI) case study. Our single-speaker evaluation is based on the hypothesis that modeling the voice characteristics of an individual speaker is easier than learning multiple voices simultaneously, given the One-Class (OC) nature of our model. Furthermore, since the teacher network is trained for speaker identification, we expect the student network to effectively capture speaker-specific features.

Table 2 presents results across the four single-speaker datasets chosen to perform evaluation of our single-speaker configuration, i.e. TIMIT-TTS, Purdue, MLAAD and our generated dataset, revealing three key insights: the impact of Discrepancy Scaling (DS) on detection performances, comparison with considered state-of-the-art systems and dataset-specific trends.
Impact of Discrepancy Scaling. We can observe that our method performs strongly both with

and without DS, consistently outperforming the considered OC baseline across all datasets. However, we also notice how DS introduces a trade-off: while it enhances generalization in multi-speaker scenarios (as we will later discuss), in the single-speaker setting it slightly reduces performance on certain datasets. The most notable variation can be observed on the MLAAD dataset. Specifically, we observe how applying this normalization lowers the AUC from 94.3% to 86.0% and increases EER from 13.3% to 23.6%. This suggests that DS may be responsible for suppressing important speaker-specific anomaly cues. Conversely, on the generated dataset, DS provides a marginal benefit (EER slightly improves from 1.5% to 1.2%) revealing that its impact on the detection task may depend on dataset characteristics, thus on the technology used to synthesize data, rather than by its effect on speaker-specific anomaly cues alone.

Baseline Comparison. The OC baseline selected for our performance comparison consistently lags behind our approach, thus proving superior generalization capabilities of the Student-Teacher Feature Pyramid Matching (STFPM) approach over a standard OC method, across multiple manipulation techniques. Its AUC scores consistently remain below 80%, with the worst performance observed on the TIMIT-TTS dataset, suggesting it struggles to model speaker-specific feature effectively. As for the RawNet2 architecture, its state-of-the-art detection performance on TIMIT-TTS aligns with expectations, given its supervised nature. However, what is unexpected is its substantial performance drop on the Purdue and MLAAD corpora, indicating that its effectiveness may depend on the dataset characteristics rather than on its classification capabilities. Overall, our method achieves significantly better performance than the baselines' across all datasets, demonstrating greater adaptability to speaker-specific speech deepfake detection.

Dataset-Specific Trends.

- **TIMIT-TTS** emerges as the easiest dataset to handle, with all models – except the OC baseline – achieving near-perfect scores. This suggests that its synthetic speech contains highly detectable artifacts, making deepfake detection nearly straightforward.
- **Purdue** presents the greatest challenge for general-purpose deepfake detection among investigated datasets. This is evident from the significant performance drop compared to other corpora, reinforcing the idea that high-quality synthesis makes detection more difficult. However, our method achieves perfect classification, demonstrating its robustness in handling such high-quality synthetic data and effectively capturing even the subtlest of manipulations in speaker-specific features.
- **MLAAD** benefits the most from our method, improving significantly both over the OC baseline and the RawNet2, which poor performance on this set may suggest that artifacts within its synthetic speech data is almost imperceptible. If this is the case, our method proves the most effective (among those investigated) at detecting subtle discrepancies in speaker-specific features, showcasing its robustness in identifying even the most challenging anomalies.

5.2. Multi-Speaker Scenario

Given the strong performance in the single-speaker setting, we extend our evaluation to a multi-speaker scenario. This evaluation assesses the generalization ability of the considered models when confronted with a diverse set of speakers. Unlike the single-speaker scenario, where a model can focus on the acoustic features and patterns associated to a specific voice, a multi-speaker setup requires the model to differentiate between genuine and synthetic speech across various identities. This introduces additional variability in timbre, intonation, and prosody, making the task significantly more challenging.
Table 3 presents the detection performance of the considered systems across four multi-speaker datasets: ASVSpoof19, In-The-Wild, FakeOrReal and Purdue. Again, we proceed by analyzing results by highlighting the following key aspects: impact of DS, baseline comparisons, and dataset-specific trends.

Impact of Discrepancy Scaling. The introduction of DS in the multi-speaker scenario brings notable improvements across all investigated datasets. This indicates that DS significantly enhances the

Table 3: Detection performance of the considered systems in a multi-speaker scenario (%).

	ASVspoof 2019		In-the-Wild		FakeOrReal		Purdue	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
RawNet2 [77]	91.9	10.6	63.9	40.9	82.1	25.1	56.3	44.9
OC baseline [29]	**93.2**	**14.8**	79.0	28.9	**99.0**	**5.5**	66.6	35.9
Ours (w/o DS)	87.1	22.2	95.1	11.6	92.8	18.0	61.8	40.2
Ours (with DS)	92.7	15.6	**98.2**	**5.7**	97.4	9.1	**75.7**	**30.6**

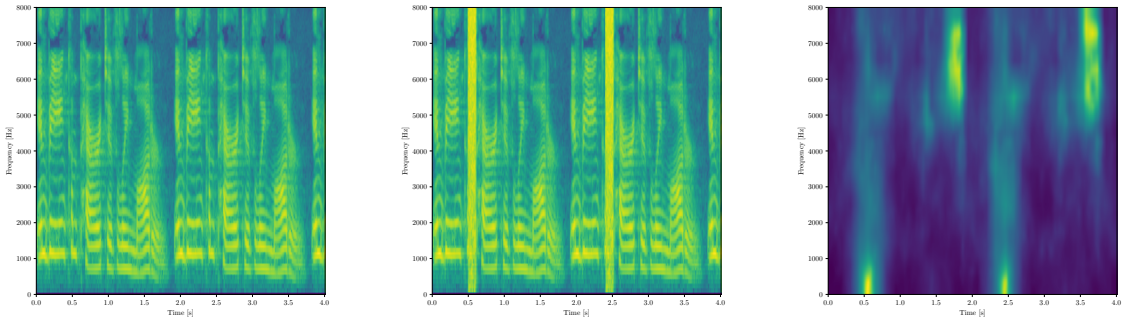
model's generalization capabilities, enabling it to detect anomalies across a wider range of speech-specific features and making it more robust to diverse speech patterns. Unlike the single-speaker scenario, where the detection process may heavily rely on speaker-specific cues, DS helps improve anomaly detection even when speaker-specific features are less prominent, demonstrating its effectiveness in more complex scenarios and in the presence of challenging deepfake data.

Baseline Comparison. Comparing our approach to the two baselines, RawNet2 and the OC baseline, we observe that our method performs strongly across all datasets, showcasing its excellent generalization capabilities. Again, RawNet2 performs well on its training dataset (ASVSpoof19) while struggles to distinguish between real and synthetic speech on the others. In contrast, our method with DS shows consistent robustness across all dataset, highlighting its ability to handle diverse speakers and challenging deepfake data on a dataset specifically built to simulate real-world scenarios. This can observed by comparing system's performance on the Purdue dataset, which includes speech deepfakes generated using the latest advanced technologies: our approach outperforms both baselines, showcasing its ability to detect even subtle anomalies in high-fidelity synthetic data. This further emphasizes the adaptability and effectiveness of our method, not just in controlled or simpler environments, but also in the presence of more complex, cutting-edge deepfake technologies. Moreover, while we expected the OC baseline to perform well on its training dataset ASVSpoof19, our method still delivers competitive results. In summary, although our approach does not outperform the baselines across all datasets, it consistently performs very close to the state-of-the-art methods, demonstrating its potential in handling complex and diverse speech anomalies.

Dataset-Specific Trends.

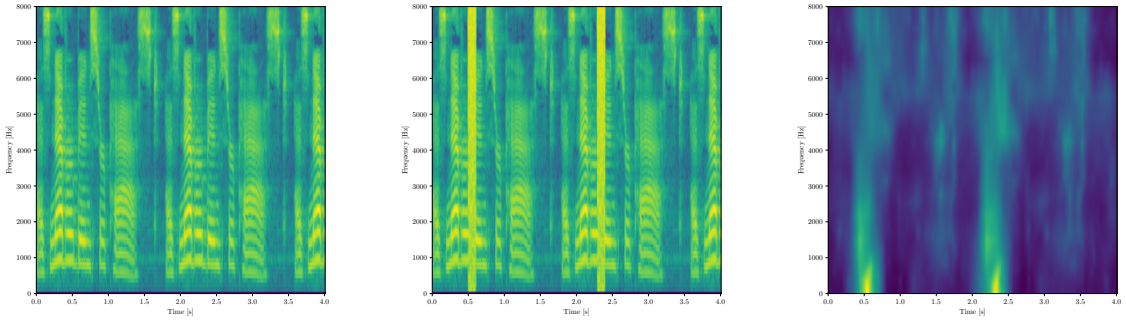
- **In-The-Wild** emerges as the most challenging dataset for all models, with RawNet2 notably struggling to classify data. This highlights how complex real-world scenarios, with diverse speakers and varying deepfake techniques, make detection more difficult, especially for supervised learning-based approaches. Our method, however, with the addition of DS, shows remarkable robustness, achieving a significantly higher AUC and lower EER compared to both baselines. This emphasizes our framework's ability to classify challenging speech data effectively as well as its superior generalization to unseen data, making its detection process in complex multi-speaker environments more robust than that of state-of-the-art (SOTA) affirmed methods.
- **Purdue** is characterized by high-fidelity speech deepfakes generated using the latest advanced technologies. This dataset proves to be the most challenging for detecting subtle anomalies in synthetic speech. Nevertheless, our method outperforms both baselines, showcasing its ability to effectively handle even the most sophisticated deepfakes, achieving superior performance over RawNet2 and the OC baseline.
- **ASVSpoof19**, on which the OC baseline was trained and tested, shows relatively high performance for the baseline, but our approach performs competitively. While the OC baseline yields strong results, particularly in this dataset, our method performs just below it, demonstrating its potential for generalization across multiple datasets, and confirming its competitive edge across a wide variety of deepfake detection tasks.

In summary, the AUCs clearly illustrate that discrepancy scaling consistently enhances performance across all datasets, allowing our method to achieve superior performance in most cases. While there are instances where the OC baseline performs well (e.g., on ASVSpoof19), our method with DS gen-



(a) Real Track (b) Real Track + Perturbation (c) Predicted Anomaly Map

Figure 2: Time-domain noise injection – Track "LJ001-0002.wav"



(a) Real Track (b) Real Track + Perturbation (c) Predicted Anomaly Map

Figure 3: Time-domain noise injection – Track "LJ001-0008.wav"

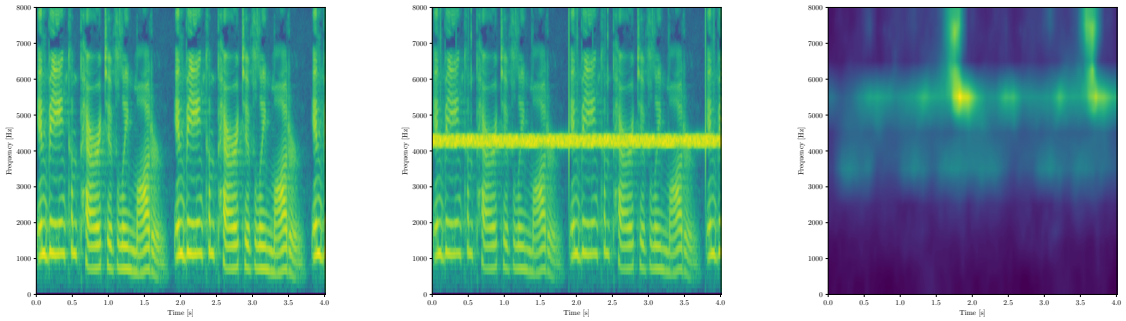
erally outperforms the baselines, especially in more challenging dataset scenarios like In-The-Wild and Purdue, confirming the superior robustness and generalization capabilities of our approach.

Overall, our model achieves state-of-the-art performance in both single and multi-speaker scenarios, demonstrating its robustness and adaptability to various generation techniques and diverse speaker and speech-specific features. Moreover, our initial hypotheses are confirmed, as the model generally performs better in the single-speaker scenario. However, it also shows excellent performance in multi-speaker scenarios, indicating its versatility in handling more complex conditions.

5.3. Localization Capability Assessment

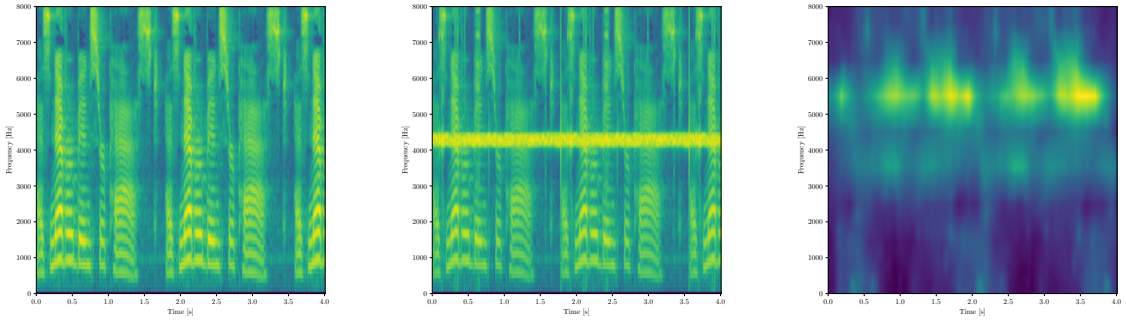
To investigate the feasibility of anomaly localization, we conduct a preliminary analysis by introducing controlled perturbations through the injection of noise in either the time domain or frequency domain directly into the raw audio. This approach enables us to assess whether the model can accurately detect and localize the artificially induced, more prominent artifacts. Below, we present visualizations of the original spectrogram, the spectrogram with noise injected into the raw audio, and the model's anomaly prediction for two arbitrary representative track within our dataset.

Time-domain perturbations. Figure 2 and Figure 3 showcase instances where noise was injected at a specific time interval. Despite the differences in features within the targeted time range due to the varying nature of different speech tracks, we observe how the model effectively detects anomalies. In Figure 3, we notice that, in some instances, the identified region appears slightly shifted in time. This behavior may stem from the time resolution of the mel-spectrogram representation or minor smoothing and normalization operations introduced by the model, such as interpolation of the final



(a) Real Track (b) Real Track + Perturbation (c) Predicted Anomaly Map

Figure 4: Frequency-domain noise injection – Track "LJ001-0002.wav"



(a) Real Track (b) Real Track + Perturbation (c) Predicted Anomaly Map

Figure 5: Frequency-domain noise injection – Track "LJ001-0008.wav"

anomaly maps to enable consistent comparison with the original mel-spectrogram and ground truth representations, respectively.

Frequency-domain perturbations. Figures 4 and 5 illustrate cases where noise was applied to specific frequency bands. The model successfully identifies these anomalies, although in some instances, the detection shows slight shifts in the spectral domain. Again, we hypothesize that this behavior may be influenced by the frequency resolution of the mel spectrogram or minor operations, such as spectral interpolation, applied during the processing of the anomaly maps to enable consistent comparison with the original mel-spectrogram and ground truth representations, respectively.
Additionally, we observed that silence is often detected as an anomaly. Figures 2c and 4c highlight instances where silent regions in the spectrogram are distinctly emphasized in the corresponding anomaly map, thus identified as anomalous cases. This suggests that the model perceives the absence of spectral content as irregular, possibly due to spectro-temporal artifacts introduced by generative technologies that subtly alter silent sections. This behavior aligns with demonstrations that silence in speech audio samples may influence the detection process of potential deepfakes, as portrayed in recent studies [45, 64].
Overall, by leveraging a controlled single-speaker scenario, we confirm that our model effectively detects and localizes anomalies across different perturbation scenarios. The observed slight shifts in time or frequency suggest potential opportunities for refining our framework, such as adopting higher-resolution spectro-temporal representations or further tuning the model's sensitivity to spectral variations. AUC-ROC scores computed for the employed dataset support the success of localization in these controlled scenarios; however, we refrain from reporting these scores, as we consider them less relevant to this discussion.

This preliminary analysis gives us confidence in the localization capabilities of our model, prompting us to continue exploring the anomaly localization challenge to achieve a deeper understanding of our system's internal mechanisms, thereby enhancing the explainability of the detection process.

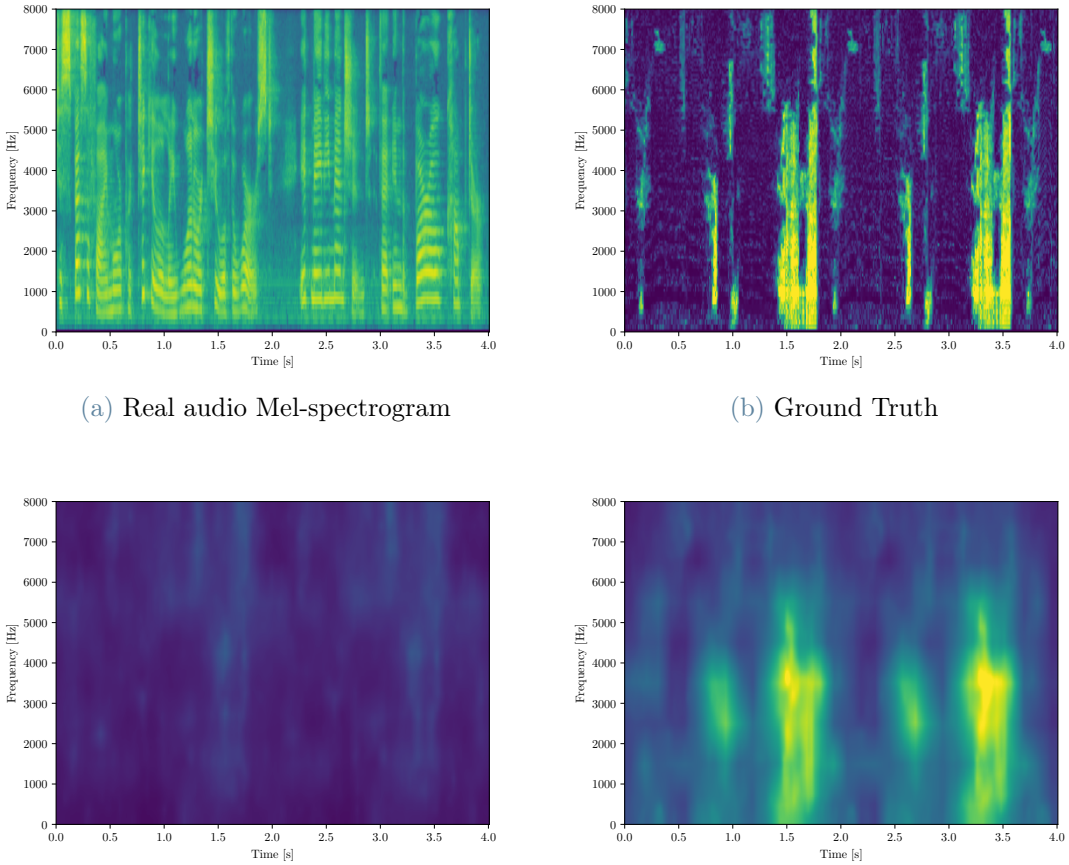
5.4. Interpretability Analysis

As anticipated in Section 4.1, to assess our framework's capabilities in pinpointing time-frequency anomalous regions and we conducted our evaluation exclusively on a tailored single-speaker dataset, i.e. our generated dataset. This choice is motivated by two key factors: (1) Our initial hypothesis that a single-speaker scenario would yield better overall performance, as the model can focus on speaker-specific characteristics, making it easier to characterize a single speaker's features rather than those of multiple speakers. (2) The dataset's design, which we specifically tailored to ensure a one-to-one correspondence between real and spoofed tracks allowing for a more consistent comparison between ground truth masks and anomaly maps.
To achieve a comprehensive evaluation, we conducted analyses at multiple levels, aiming to provide a deeper understanding of how the model operates—both in terms of its ability to detect and localize anomalies and in terms of the internal processes that contribute to this task. This exploration not only enhances our confidence in the model's performance but also provides valuable insights into its decision-making process. Ultimately, this understanding paves the way for further improvements and refinements of our proposed framework.

Anomaly Maps Comparison The first analysis aimed to assess performance on the localization task by comparing anomaly maps obtained from different subsets of the full dataset, thereby *investigating how the model reacts to artifacts produced by various generation techniques*. This was also done to investigate how well the model generalizes when presented with the same track produced with different synthesis processes. By comparing anomaly maps generated for tracks produced using different synthesis methods, we aim to understand whether the model can effectively detect anomalies regardless of the generation process, as well as to evaluate any potential variations in its performance based on the synthesis method used.
We observed that our model reacts differently to different generation techniques. Figures 6d, 7d and 8d showcase noticeable differences in localization precision across various vocoders. Our examinations revealed that the model performs well at localizing artifacts introduced by the WaveGlow[57] vocoder, as demonstrated by Figure 6. In contrast, the precision of localization is somewhat lower when applied to tracks generated with HiFiGAN and the MelGAN vocoders, as illustrated in 7, 8 respectively. Despite these differences, the model still manages to detect anomalies across all tracks, as confirmed by the corresponding ROC curves (Figure 12). This highlights the model's robustness in anomaly detection, even when localization precision varies depending on the synthesis method.

Then, we proceed by exploring *how the natural differences in the spectral content of different speech samples impact the detection and localization capabilities*, also to investigate if there is consistency in the detection of artifacts of a specific vocoder. This analysis helps us understand how different types of speech data, such as the inherent acoustic properties and variations in speaker characteristics, influence the model's performance in identifying and localizing anomalies. Specifically, we examine how the model reacts to variations in features across different tracks, and whether certain vocoders produce more distinguishable anomalies than others.

In this regard, we observed consistencies in the anomaly patterns detected across different tracks produced by the same generator. This is particularly evident when comparing tracks generated using the WaveGlow and MelGAN frameworks. This consistency suggests that specific vocoders may introduce artifacts at particular regions of the spectrogram. These recurring anomalies might result in biased detection and localization, as the model may tend to focus more on these regions that are uniquely impacted by the vocoder. For instance, the anomaly map for WaveGlow (Figures 6d, 9d) demonstrates clearer, more localized anomalies in the lower frequency range, whereas the anomaly map for MelGAN shows anomalies concentrated in higher frequency bands, as shown in Figures 8d, 11d. Such consistent behavior indicates that some vocoders produce more distinguishable and detectable anomalies, which could limit the model's generalization ability when exposed to tracks generated by



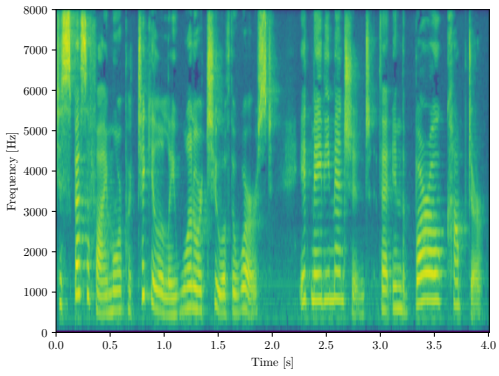
(a) Real audio Mel-spectrogram

(b) Ground Truth

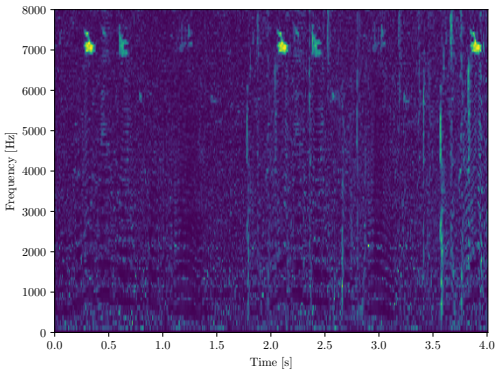
(c) Real audio anomaly map

(d) Fake audio anomaly map

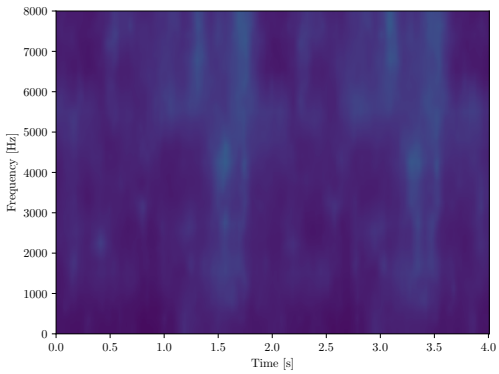
Figure 6: Anomaly localization results for a real audio track and its fake counterpart – Track: "LJ001-0008.wav", Vocoder: WAVEGLOW



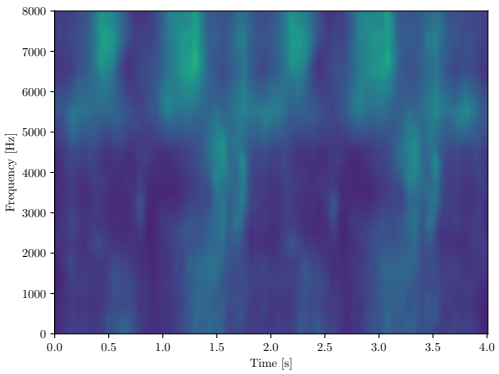
(a) Real audio Mel-spectrogram



(b) Ground Truth

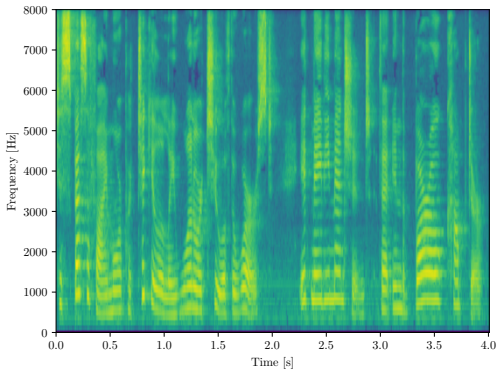


(c) Real audio anomaly map

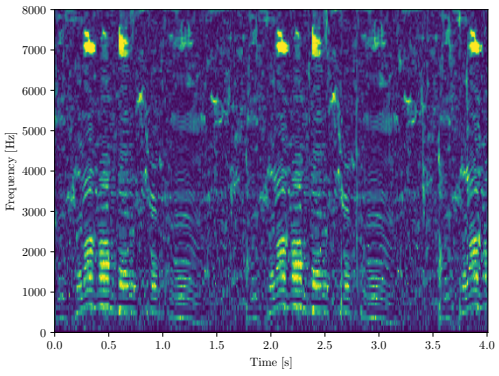


(d) Fake audio anomaly map

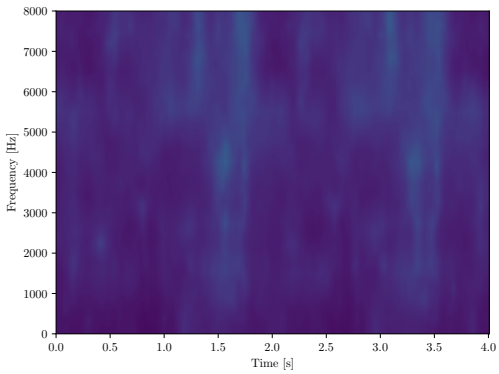
Figure 7: Anomaly localization results for a real audio track and its fake counterpart – Track: "LJ001-0008.wav", Vocoder: HIFIGAN



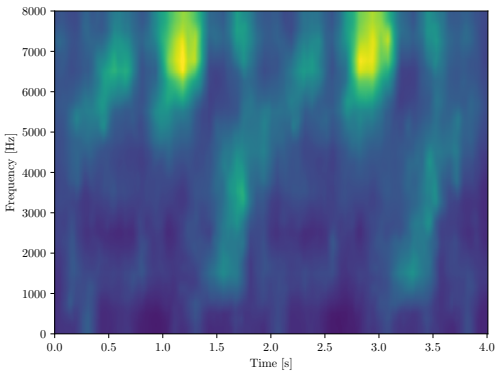
(a) Real audio Mel-spectrogram



(b) Ground Truth

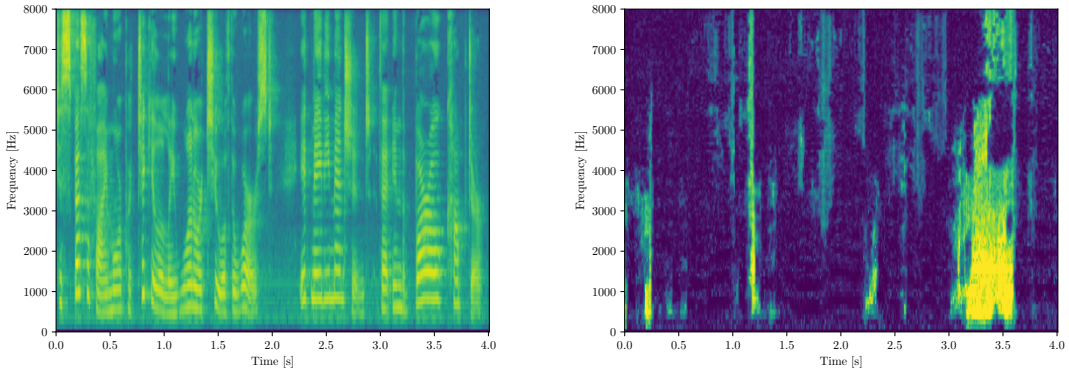


(c) Real audio anomaly map



(d) Fake audio anomaly map

Figure 8: Anomaly localization results for a real audio track and its fake counterpart – Track: "LJ001-0008.wav", Vocoder: MELGAN [linda_johnson]



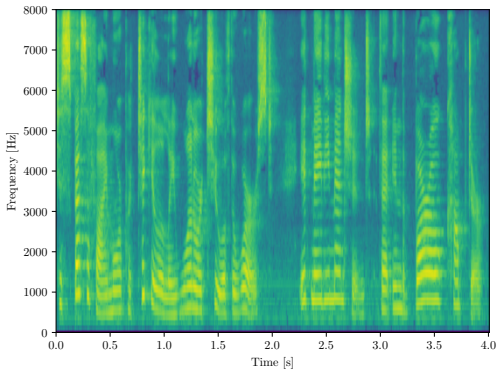
(a) Real audio Mel-spectrogram

(b) Ground Truth

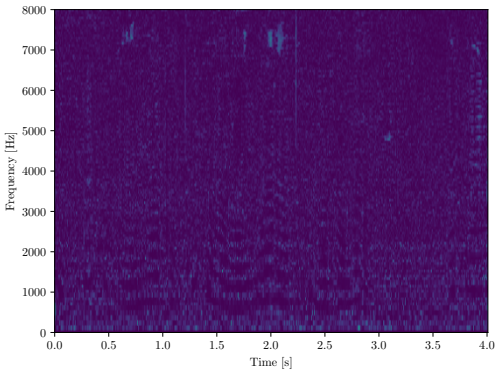
(c) Real audio anomaly map

(d) Fake audio anomaly map

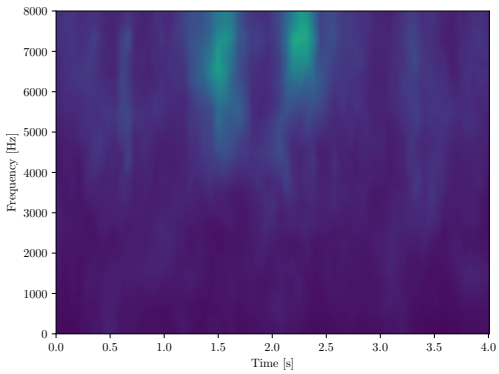
Figure 9: Anomaly localization results for a real audio track and its fake counterpart – Track: "LJ004-0131.wav", Vocoder: WAVEGLOW



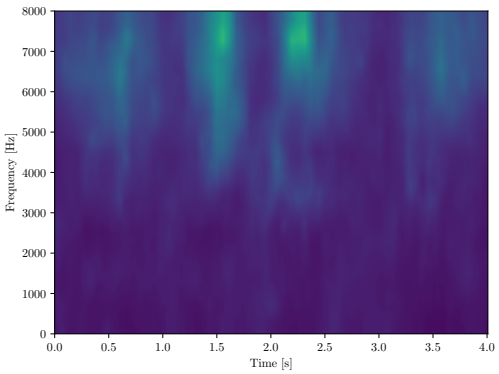
(a) Real audio Mel-spectrogram



(b) Ground Truth

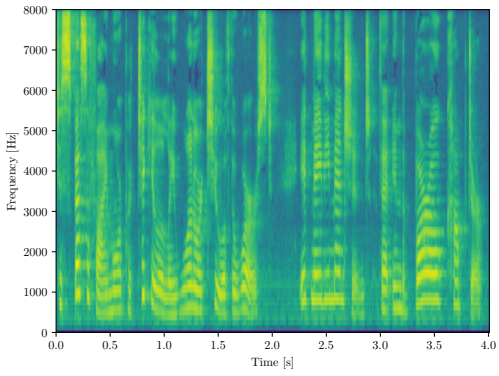


(c) Real audio anomaly map

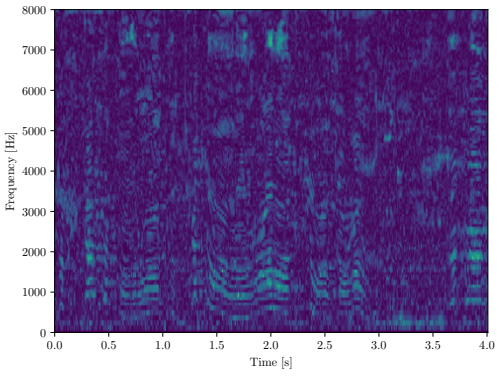


(d) Fake audio anomaly map

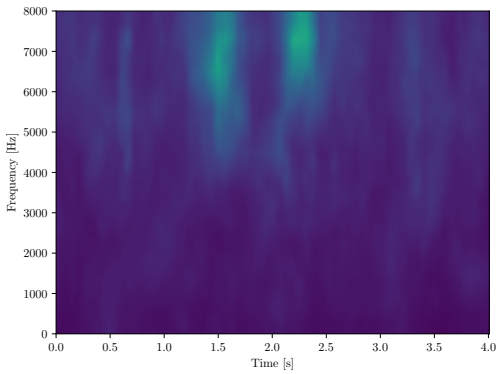
Figure 10: Anomaly localization results for a real audio track and its fake counterpart – Track: "LJ004-0131.wav", Vocoder: HIFIGAN



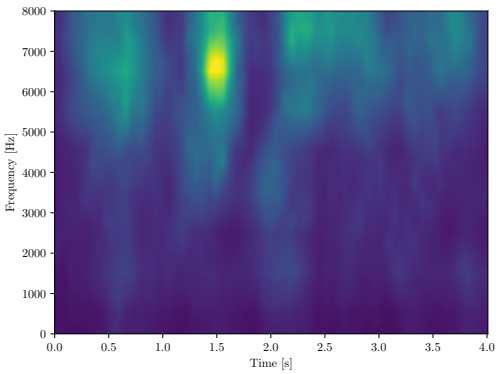
(a) Real audio Mel-spectrogram



(b) Ground Truth



(c) Real audio anomaly map



(d) Fake audio anomaly map

Figure 11: Anomaly localization results for a real audio track and its fake counterpart – Track: "LJ004-0131.wav", Vocoder: MELGAN [linda_johnson]

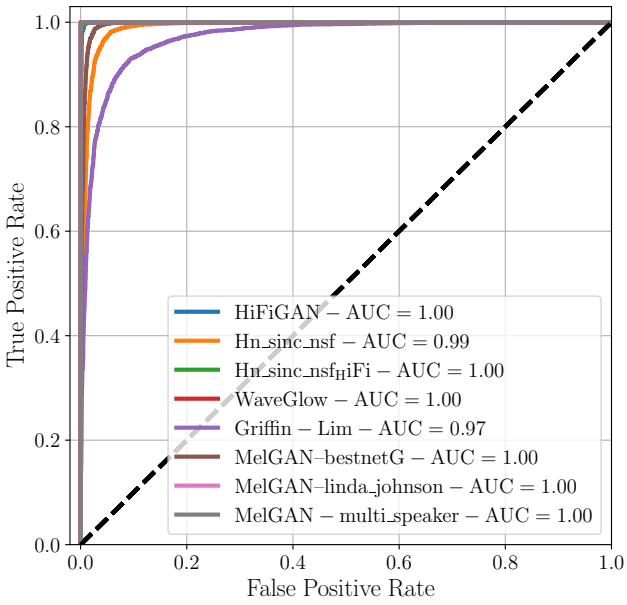


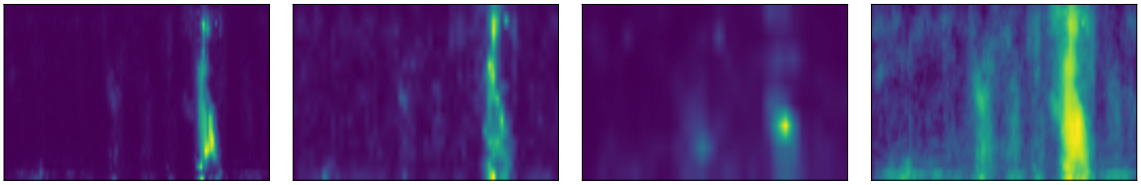
Figure 12: ROC-AUC curves showing detection performance for each subset of the investigated dataset.

different synthesis methods.

Finally, we can observe how the presence of silent regions within the audio track affects the detection and localization performance of the model. This aspect is particularly evident in Figures 10c and 10d, where darker regions within the spectrogram—corresponding to silence in the audio waveform—are consistently highlighted in both the genuine and spoofed track predictions. This suggests that the model interprets silent segments as anomalous, potentially due to the lack of spectral information in these regions, which differs significantly from typical voiced segments. Moreover, this behavior is consistent across multiple vocoders, as demonstrated by the real and spoofed predictions of HiFi-GAN and MelGAN (Figures 10c, 10d, 11c, and 11d), where silence-induced anomalies appear regardless of the synthesis technique used. These findings highlight the need for further investigation into how silence is perceived by the model and whether additional pre-processing techniques or post-processing refinements could mitigate this effect.

Feature Layer-Wise Comparison In the second analysis, we focus on the explainability of the model by examining how the features extracted at different layers of the network contribute to the final anomaly map. This is done for two out of the eight considered generative systems. By comparing the features and anomaly maps across different layers, we gain a deeper understanding of the model's internal processing and how each layer impacts the localization and detection of anomalies. This analysis aims to provide enhanced explainability of the detection process, offering insights into how the hierarchical structure of the model contributes to the identification of anomalies. Through this, we aim to shed light on the relationship between the extracted features at various layers and their influence on the model's final anomaly detection output.

After analyzing anomaly maps extracted at different layers, we observed that certain layers detected anomalies of larger size compared to others. This analysis was conducted across all vocoders and multiple tracks; however, for clarity, we report visual results for WaveGlow on an arbitrary track in Figure 13. Here, we present demonstrative examples of how the size of detected anomalies varies across layers, offering insights into the model's internal processing and its influence on anomaly localization.



(a) Layer 4 – Conv_2x (b) Layer 5 – Conv_3x (c) Layer 6 – Conv_4x (d) Final Anomaly Map

Figure 13: Feature Maps extracted at different layers and final Anomaly Map – Vocoder: WaveGlow, Track:"LJ005-0296.wav"

By investigating this aspect further, we found that the last feature extraction layer, `conv4_X`, consistently presented activations with a broader dynamic range. This observation held true across both track and vocoder variations, suggesting that deeper layers might capture more distinctive anomaly-related patterns. These finding warrant further investigation, as thid aspect could potentially lead to improved and less biased detection and localization performance, offering a more reliable and robust approach to anomaly detection in audio.

6. Conclusions and Future Works

In this work, we proposed a novel one-class anomaly detection framework for speech deepfake detection, addressing the generalization and interpretability limitations of traditional supervised approaches. Our system, based on the Student-Teacher Feature Pyramid Matching (STFPM) architecture, is trained exclusively on authentic speech to characterize its statistical distribution and detect anomalies attributable to synthetic content. Experimental results demonstrate that our method outperforms both supervised classifiers (RawNet2) and an alternative One-Class (OC) model, particularly in scenarios where training and evaluation data differ in terms of speakers or synthesis methods.
In the single-speaker scenario, our approach achieves high detection performance, effectively capturing fine-grained anomalies introduced by vocoders and synthesis models. The model successfully generalizes to unseen manipulations, detecting artifacts that might go unnoticed by traditional supervised methods. In the multi-speaker setting, while performance remains strong, we observe a slight degradation due to the increased variability in speaker characteristics. Despite this, our method maintains a significant advantage over supervised classifiers, which struggle with unseen speakers due to their reliance on explicit class labels during training.
Furthermore, our model effectively localizes artifacts within the time-frequency representation of audio signals, allowing for a more detailed interpretation of synthetic anomalies. Notably, we find that our system is particularly adept at detecting and localizing injected noise, such as frequency and time-domain distortions, highlighting its robustness to diverse synthetic manipulations. However, one challenge we identified is the influence of silent regions on anomaly predictions. Our results suggest that the model tends to highlight silence as anomalous, likely due to the absence of spectral content in these segments.
Future work will focus on refining the Discrepancy Scaling (DS) technique and exploring its applications in diverse real-world forensic scenarios. A key area of investigation will be a deeper study of how silence affects both anomaly detection and localization, as well as methods to mitigate potential biases introduced by silent regions. Additionally, we plan to analyze feature extraction across different network layers to determine which contribute most to effective anomaly detection. Beyond this, we aim to explore alternative acoustic features, such as different types of spectrogram representations (e.g., CQT, log-spectrograms) and chromagrams, to assess whether they provide complementary information that enhances deepfake detection.
Furthermore, we will investigate integrating additional modalities, such as prosody features and speaker-specific characteristics, to improve detection accuracy, particularly in cases where voice synthesis technologies target specific individuals. Extending our framework to multi-speaker detection across diverse languages will also be crucial in ensuring robustness in real-world forensic applications. These research directions hold the potential to significantly advance speech deepfake detection, making it more accurate, interpretable, and applicable to forensic voice analysis.

References

[1] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, 11(2):71–76, 1990.

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

[3] Alican Akman and Björn W Schuller. Audio explainable artificial intelligence: A review. *Intelligent Computing*, 2:0074, 2024.

[4] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[5] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, et al. Deepfake media forensics: Status and future challenges. *Journal of Imaging*, 11(3):73, 2025.

[6] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*. PMLR, 2017.

[7] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 1, 2018.

[8] Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J Delp. Are Recent Deepfake Speech Generators Detectable? In *ACM Workshop on Information Hiding and Multimedia Security*, 2024.

[9] Fadi Biadsy, Ron J Weiss, Pedro J Moreno, Dimitri Kanevsky, and Ye Jia. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *arXiv preprint arXiv:1904.04169*, 2019.

[10] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*, 2019.

[11] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. Synthetic speech detection through short-term and long-term prediction traces. *EURASIP Journal on Information Security*, 2021:1–14, 2021.

[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[13] Georgia Channing, Juil Sock, Ronald Clark, Philip Torr, and Christian Schroeder de Witt. Toward robust real-world audio deepfake detection: Closing the explainability gap. *arXiv preprint arXiv:2410.07436*, 2024.

[14] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265, 2021.

[15] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2022.

[16] Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. Open challenges in synthetic speech detection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022.

[17] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):954–964, 2010.

[18] Siwen Ding, You Zhang, and Zhiyao Duan. Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[19] Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. Audio Deepfake Detection With Self-Supervised Wavlm And Multi-Fusion Attentive Classifier. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[20] Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S Almadhor, Zunera Jalil, and Rouba Borghol. Deepfake audio detection via MFCC features using machine learning. *IEEE Access*, 10:134018–134028, 2022.

[21] Mihály Héder. Explainable ai: A brief history of the concept. *Ercim News*, (134):9–10, 2023.

[22] R Hemavathi and R Kumaraswamy. Voice conversion spoofing detection by exploring artifacts estimates. *Multimedia tools and applications*, 80(15):23561–23580, 2021.

[23] Keith Ito and Linda Johnson. The LJ Speech Dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[24] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[25] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[26] Stamatis Karnouskos. Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, 1(3):138–147, 2020.

[27] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. On the relation of trust and explainability: Why to engineer for trustworthiness. In *2021 IEEE 29th international requirements engineering conference workshops (REW)*, pages 169–175. IEEE, 2021.

[28] Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki, and Junichi Yamagishi. Rakugo speech synthesis using segment-to-segment neural transduction and style tokens—toward speech synthesis for entertaining audiences. In *Proc. 10th ISCA Speech Synth. Workshop*, pages 111–116, 2019.

[29] Hyun Myung Kim, Kangwook Jang, and Hoirin Kim. One-class learning with adaptive centroid shift for audio deepfake detection. *arXiv preprint arXiv:2406.16716*, 2024.

[30] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, 2021.

[31] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of imaging*, 4(2):36, 2018.

[32] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[33] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[34] Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[35] Jing Li, Yanhua Long, Yijie Li, and Dongxing Xu. Advanced rawnet2 with attention-based channel masking for synthetic speech detection. In *Proc. INTERSPEECH*, 2023.

[36] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 2025.

[37] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[38] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

[39] Daniele Mari, Davide Salvi, Paolo Bestagini, and Simone Milani. All-for-one and one-for-all: Deep learning-based feature fusion for synthetic speech detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2023.

[40] Roy A Maxion and Rachel R Roberts. Proper use of roc curves in intrusion/anomaly detection. *School of Computing Science Technical Report Series*, 2004.

[41] Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869, 2023.

[42] Zied Mnasri, Stefano Rovetta, and Francesco Masulli. Anomalous sound event detection: A survey of machine learning based methods and applications. *Multimedia Tools and Applications*, 81(4):5537–5586, 2022.

[43] Seyed Hamidreza Mohammadi and Alexander Kain. An overview of voice conversion systems. *Speech Communication*, 88:65–82, 2017.

[44] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

[45] Nicolas M Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is silver, silence is golden: What do asvspoof-trained models really learn? *arXiv preprint arXiv:2106.12914*, 2021.

[46] Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2024.

[47] Juha Mylläri and Jukka K Nurminen. Discrepancy scaling for fast unsupervised anomaly localization. In *IEEE Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023.

[48] Viola Negroni, Davide Salvi, Paolo Bestagini, and Stefano Tubaro. Analyzing the impact of splicing artifacts in partially fake speech signals. *arXiv preprint arXiv:2408.13784*, 2024.

[49] Viola Negroni, Davide Salvi, Alessandro Ilic Mezza, Paolo Bestagini, and Stefano Tubaro. Leveraging mixture of experts for improved speech deepfake detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[50] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[51] Monisankha Pal, Dipjyoti Paul, and Goutam Saha. Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language*, 48:31–50, 2018.

[52] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[53] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[54] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

[55] Marco Pasini. Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. *arXiv preprint arXiv:1910.03713*, 2019.

[56] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.

[57] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[58] Daniel Ramotsoela, Adnan Abu-Mahfouz, and Gerhard Hancke. A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study. *Sensors*, 18(8):2491, 2018.

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

[60] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[61] Karen McGregor Richmond, Satya M Muddamsetty, Thomas Gammeltoft-Hansen, Henrik Palmer Olsen, and Thomas B Moeslund. Explainable ai and law: An evidential survey. *Digital Society*, 3(1):1, 2024.

[62] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[63] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. 2015.

[64] Davide Salvi, Temesgen Semu Balcha, Paolo Bestagini, and Stefano Tubaro. Listening between the lines: Synthetic speech detection disregarding verbal content. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 883–887. IEEE, 2024.

[65] Davide Salvi, Paolo Bestagini, and Stefano Tubaro. Towards frequency band explainability in synthetic speech detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 620–624. IEEE, 2023.

[66] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C. Stamm, and Stefano Tubaro. TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection. *IEEE access*, 11:50851–50866, 2023.

[67] Davide Salvi, Viola Negroni, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Freeze and learn: Continual learning with selective freezing for speech deepfake detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[68] Davide Salvi, Amit Kumar Singh Yadav, Kratika Bhagtani, Viola Negroni, Paolo Bestagini, and Edward J Delp. Comparative analysis of asr methods for speech deepfake detection. *arXiv preprint arXiv:2411.17349*, 2024.

[69] Conrad Sanderson. The VidTIMIT database. Technical report, IDIAP, 2002.

[70] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

[71] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018.

[72] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157, 2020.

[73] Daeun Song, Nayoung Lee, Jiwon Kim, and Eunjung Choi. Anomaly detection of deepfake audio based on real audio using generative adversarial network model. *IEEE Access*, 2024.

[74] Yannis Stylianou, Olivier Cappe, and Eric Moulines. Statistical methods for voice quality transformation. In *Proc. Eurospeech 1995*.

[75] Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. AI-Synthesized Voice Detection Using Neural Vocoder Artifacts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2023.

[76] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*, 2019.

[77] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[78] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.

[79] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245, 2024.

[80] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

[81] Jorma Valjakka, Juha Mylläri, Lalli Myllyaho, Juhani Kivimäki, and Jukka K Nurminen. Anomaly localization in audio via feature pyramid matching. In *IEEE Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023.

[82] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE.

[83] Tomasz Walczyna and Zbigniew Piotrowski. Overview of voice conversion methods based on deep learning. *Applied sciences*, 13(5):3100, 2023.

[84] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021.

[85] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021.

[86] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1207–1216, 2020.

[87] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

[88] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[89] Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning. *arXiv preprint arXiv:2404.15854*, 2024.

[90] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[91] Yujie Yang, Haochen Qin, Hang Zhou, Chengcheng Wang, Tianyu Guo, Kai Han, and Yunhe Wang. A robust audio deepfake detection system via multi-view feature. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[92] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

[93] Hui Yin, Amir Aryani, Stephen Petrie, Aishwarya Nambissan, Aland Astudillo, and Shengyuan Cao. A rapid review of clustering algorithms. *arXiv preprint arXiv:2401.07389*, 2024.

[94] Dongsuk Yook, Geonhee Han, Hyung-Pil Chang, and In-Chul Yoo. Cyclediffusion: Voice conversion using cycle-consistent diffusion models. *Applied Sciences*, 14(20):9595, 2024.

[95] Khalid Zaman, Islam JAM Samiul, Melike Sah, Cem Direkoglu, Shogo Okada, and Masashi Unoki. Hybrid transformer architectures with diverse audio features for deepfake speech classification. *IEEE Access*, 2024.

[96] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2023.

[97] Chunlei Zhang, Chengzhu Yu, and John HL Hansen. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):684–694, 2017.

[98] You Zhang, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021.

[99] Zhenyu Zhang, Xiaowei Yi, and Xianfeng Zhao. Fake speech detection using residual network with transformer encoder. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*.

[100] Yu-Jun Zheng, Xiao-Han Zhou, Wei-Guo Sheng, Yu Xue, and Sheng-Yong Chen. Generative adversarial network based telecom fraud detection at the receiving bank. *Neural Networks*, 102:78–86, 2018.

Abstract in lingua italiana

Negli ultimi anni, l'ascesa di modelli generativi basati sull'intelligenza artificiale ha permesso la creazione di deepfake vocali altamente realistici—segnali audio sintetici in grado di imitare la voce di un determinato parlatore con una notevole precisione—sollevando preoccupazioni cruciali per la sicurezza. Sebbene questi sviluppi aprano la strada a nuove possibilità in vari ambiti applicativi, sollevano anche gravi preoccupazioni, in particolare nei settori dell'autenticazione, dell'audio forense e della disinformazione.
I metodi esistenti per rilevare i deepfake vocali si basano principalmente sull'apprendimento supervisionato da parte di reti neurali, il quale presenta tuttavia due limitazioni fondamentali: la generalizzazione a tecniche di sintesi non viste e l'interpretabilità dei processi decisionali del modello. La prima limitazione impedisce ai metodi di rilevamento di adattarsi facilmente a scenari reali, dove emergono frequentemente nuove e diverse tecniche di generazione di deepfake. La seconda mina la fiducia in questi sistemi, specialmente nelle applicazioni forensi, dove la trasparenza e l'interpretabilità sono cruciali per la validazione e l'accettazione legale.
La necessità di superare queste limitazioni ha spinto le ricerche recenti verso nuove strategie per il rilevamento dei deepfake vocali. Queste strategie includono approcci basati su un'unica classe di contenuti in fase di allenamento delle reti (one-class), per affrontare i problemi di generalizzazione, cosí come tecniche di localizzazione delle anomalie per migliorarne l'interpretabilità, permettendo una rilevazione più robusta e spiegabile dei deepfake vocali.
Per affrontare questi problemi, questa tesi introduce un nuovo framework di detection one-class, riformulando il problema della deepfake detection come un problema di detection di anomalie. Il nostro approccio sfrutta un modello addestrato esclusivamente su parlato reale, imparando a caratterizzarne la distribuzione e distinguere poi efficacemente le voci autentiche da quelle sintetiche, che presentano -di contro- una distribuzione diversa.
Un elemento distintivo del nostro framework è la sua capacità di produrre delle mappe delle anomalie rilevate, evidenziando aree di parlato che sono sospette sia nel dominio del tempo che in quello della frequenza. Ciò avviene grazie a un sistema di cosiddetto STFPM rafforzato da DS per migliorare le capacità di generalizzazione su distribuzioni di dati diverse e non viste.
Le valutazioni approfondite da noi intraprese indicano che il metodo proposto é in grado di superare i metodi dello stato del'arte presi in considerazione, rivelandosi efficace sia nel rilevare i deepfake, che nel localizzare gli artefatti di sintesi all'interno di una data traccia sintetica sotto analisi. Il nostro lavoro evidenzia il potenziale delle strategie di deepfake detection basate sulla detection di anomalie.

Parole chiave: Analisi Forense Multimediale; Audio Forense; Audio Deepfake; Interpretabilità; Anomaly Detection

Acknowledgements

Un sincero ringraziamento va all'Image and Sound Processing Lab (ISPL) e al mio relatore, prof. Paolo Bestagini, per avermi dato l'opportunità di lavorare a questo progetto, che si è rivelato essere ciò che di più affine a me io abbia sentito durante il mio percorso di studi. Un riconoscimento speciale va a Davide e Viola, mie stelle polari in questo percorso di tesi, che da sempre mi incuteva un po' di timore. Il vostro aiuto e il vostro supporto vi hanno resi indispensabili. Siete stati di grande ispirazione e vi stimo profondamente.
Un grazie, inoltre, a tutti coloro che, all'interno di questo spazio condiviso tra ISPL e AntLab, hanno reso gli ultimi mesi dedicati alla realizzazione di questo obiettivo più leggeri di quanto non lo siano stati, invece, i precedenti anni di studio. La vostra accoglienza e il continuo supporto hanno reso quest'esperienza una che ricorderò con il sorriso.
In ultimo, ma non per importanza, un grazie profondo a tutti coloro che, standomi vicino e prendendosi il tempo per ascoltarmi, hanno reso l'esperienza universitaria un percorso di inestimabile crescita, sia personale che professionale. Siete stati di incredibile supporto e conforto.