POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Explainable Speech Deepfake Detection: an Investigation into Model Behavior and Generalization

Tesi di Laurea Magistrale in
Music And Acoustic Engineering

Manfredi Pletti, 216473

Abstract: The rapid evolution of generative artificial intelligence has democratized the production of highly realistic synthetic speech, often referred to as deepfake audio, posing significant challenges to biometric security systems. Although current deepfake detection systems achieve high scores on standard benchmarks, they operate as opaque "black boxes", leaving their internal decision-making processes largely unexplored. This thesis presents an exploratory study on the spectral dependencies of Convolutional Neural Networks (LCNN and ResNet) applied to the speech deepfake detection domain. We introduce a diagnostic framework based on an adaptation of Relative Contribution Quantification (RCQ) to generate global attention profiles in the frequency domain. Our analysis confirms a characteristic "U-shaped" attention profile, in which models favor low- and high-frequency components of the speech signal, placing less emphasis on the mid-frequency range. To investigate the plasticity of these attention profiles, we introduce Stratified Spectral Mixing (SSM), a data augmentation strategy designed in this context as an investigative tool to disrupt vertical spectral coherence. Experimental results show that this intervention encourages models to reconfigure their spectral focus, leading to a more distributed allocation of attention across frequency bands. This shift is accompanied by improved generalization to previously unseen attack algorithms and greater robustness under limited-bandwidth conditions, such as GSM telephony. Overall, these findings demonstrate that actively modifying a model's spectral focus provides a practical means of examining how its detection behavior evolves.

Advisor:
Prof. Paolo Bestagini

Co-advisors:
Viola Negroni

Academic year:
2024-2025

Key-words: Deepfake Detection, Explainable AI, Audio Forensics, Spectral Analysis

1. Introduction

In recent years, the digital media production sector has witnessed a paradigm shift driven by the rapid and constant evolution of generative artificial intelligence. This technological advancement

has not only improved existing tools but has profoundly redefined the nature of how audiovisual content is created, manipulated, and consumed. We have moved from an era where digital synthesis was a complex, labor-intensive process reserved for highly skilled experts who knew how to use specialized software to a new reality where powerful generative algorithms can autonomously produce high-quality media. This transition marks a departure from traditional manual editing towards data-driven generation, where machines learn to mimic reality with unprecedented accuracy [11]. As a result, the barrier to entry for producing realistic synthetic content has been drastically lowered, democratizing access to capabilities that were previously reserved exclusively for high-budget production studios. This accessibility, in addition to fueling creativity, has simultaneously blurred the line between authentic and artificially generated media, creating a scenario where the origin of digital content is increasingly difficult to verify.

In the specific context of the audio domain, this evolution is represented by two main categories of algorithms: Text-To-Speech (TTS) and Voice Conversion (VC). The former refers to systems capable of synthesizing speech directly from textual input, effectively enabling machines to "read" text with a human-like voice. The latter involves transforming a source recording to match the vocal identity of a different speaker, while preserving the original linguistic content. These tools have moved from generating robotic voices to creating speech that is virtually indistinguishable from the human voice.

The democratization of these technologies brings with it great benefits that extend far beyond mere technological innovation. In the field of accessibility and health, generative audio serves as a tool for individuals with speech impairments or reading difficulties [56], effectively restoring their ability to communicate or consume written content. In addition, the entertainment and creative industries are leveraging these advances to revolutionize post-production processes, enabling realistic dubbing for cinema and personalized voices that preserve original expressiveness. Finally, at a daily life level, these algorithms have become ubiquitous in human-computer interaction, being integrated into services such as GPS navigation and voice assistants.

However, this technological democratization introduces major security risks, turning the human voice into a vulnerable attack vector [1, 48]. The most immediate threats target biometric security systems: since voice authentication can be used for secure access to services, buildings, and banking channels, Automatic Speaker Verification systems face unprecedented risks. Generative models can produce spoofed signals capable of bypassing these logical access controls, allowing unauthorized entities to breach protected systems by stealing a user's biometric identity. In addition to these violations, the potential for malicious use extends deeply into the realm of criminal activity and fraud. The ability to clone a target's voice has paved the way for sophisticated social engineering attacks, identity theft, and unauthorized commercial transactions. One particular application involves fraudsters cloning the voices of victims' family members, such as children, to stage fake kidnappings and demand ransoms. On a societal level, these tools facilitate misinformation and political manipulation. By generating arbitrary words that a target speaker never said, malicious actors can craft fake news and propaganda designed to manipulate public opinion or interfere with political elections [29, 49]. This ability to put false words in the mouths of public figures undermines their credibility and reduces trust in digital media. Finally, the proliferation of deepfakes poses a crucial challenge for forensic science and the legal system. The difficulty in distinguishing between synthetic and genuine recordings compromises the admissibility of audio evidence in court. This problem is further exacerbated by less invasive manipulation techniques such as audio splicing, where synthetic segments are interspersed with genuine segments to alter the semantic meaning of a spoken recording while evading detection by standard countermeasures.

To mitigate these threats, the research community has engaged in an "arms race" against generative models, focusing in particular on the development of deepfake speech detection systems. State-of-the-art synthetic speech detectors, mainly based on deep neural networks nowadays, achieve remarkable results on standard benchmarks, reaching impressive detection performance in controlled environments. Despite these significant achievements, there is still a significant gap between "in-lab" and "in-the-wild" performance [23]. Indeed, detectors that perform flawlessly on clean, academic datasets often suffer from performance drops when exposed to audio data that has undergone unseen compression codecs, environmental noise, or novel algorithms.

This vulnerability is critically emphasized by the asymmetry in development speeds: generative models evolve at a pace that significantly exceeds forensic detection models. This weakness suggests

that although current models are effective in recognizing specific artifacts in training data, they often struggle to generalize to unknown attacks.
Another significant limitation lies in the intrinsic opacity of current state-of-the-art neural detectors, broadly defined as the "black-box" nature of Deep Learning. Although these neural networks are exceptionally capable of mapping audio input to a detection score, their internal decision-making mechanism is largely inaccessible to human interpretation. Unlike traditional forensic algorithms based on hand-crafted features, where detection rules are explicitly defined, modern deep models learn complex, non-linear representations often based on millions of internal parameters. As a result, it is difficult to determine precisely which acoustic clues or spectral artifacts are exploited by the model to classify an utterance as fake.
In situations where security comes first, knowing that a model works is not enough; it is necessary to understand how it reaches a decision. Blind trust in performance metrics can be misleading, as models often learn "shortcuts" (such as background noise or silence patterns specific to a dataset) [7] rather than actual traces left by synthesis systems. This is becoming an undeferrable topic to address, especially in the synthetic speech detection landscape, where the application and use cases of detectors extend to high-stakes domains (e.g., biometric authentication, forensic investigations, and misinformation mitigation), thus inherently demanding transparency to ensure reliability. Without interpretability, it is not possible to distinguish between a valid detector and one that simply overfits the dataset with which it was trained.
This is where Explainable AI (XAI) comes into play. Explainable Artificial Intelligence represents a paradigm designed to bridge the gap between high-performance "black-box" models and the human need for transparent decision-making. In the specific context of digital forensics, XAI ceases to be an auxiliary tool and becomes a fundamental requirement: knowing that a model classifies an audio sample as fake is insufficient; a forensic analyst must understand why a decision was made to ensure that the verdict is legally defensible and reproducible. While the field of computer vision has achieved a robust suite of interpretability methods, the application of XAI in the specific landscape of synthetic speech remains comparatively fragmented and exploratory. Current literature still predominantly prioritizes minimizing performance metrics on standard benchmarks, often neglecting the diagnostic analysis of the internal features that drive classification. This lack of understanding causes a critical vulnerability: without determining whether a detector relies on genuine forensic traces or simply overfits training artifacts, its reliability in real-world scenarios remains unverified. Consequently, integrating interpretability into the detection pipeline is essential to move from opaque prediction systems to trustworthy forensic tools.
Guided by these motivations, the primary objective of this thesis is to systematically investigate the spectral dependencies of speech deepfake detectors. Moving beyond standard performance metrics, we adopt a mechanistic approach to explore how Convolutional Neural Networks weight different frequency bands during the classification process. To this end, we first introduce a diagnostic framework based on an adaptation of Relative Contribution Quantification (RCQ) for time-frequency representation. This metric allows us to derive global attention profiles, quantifying the extent to which standard architectures rely on specific ranges while potentially ignoring others. Next, to test the nature of these dependencies, we employ Stratified Spectral Mixing (SSM), a data augmentation strategy designed to destroy vertical spectral coherence during training. SSM is used as an investigative tool to study the plasticity of the model's attention mechanism. By forcing the network to process conflicting spectral information, we analyze how different architectures (LCNN and ResNet) reconfigure their attention patterns, obtaining information on the relationship between spectral focus and generalization capabilities on unseen data.
This thesis is structured as follows: Section 2 reviews the technological background, covering synthetic speech generation, deepfake detection architectures, and Explainable AI (XAI) in audio forensics. Section 3 defines the problem and presents the proposed methodology, including the adapted Relative Contribution Quantification (RCQ) framework and the Stratified Spectral Mixing (SSM) strategy. Section 4 outlines the experimental setup, detailing datasets, preprocessing, model training (LCNN and ResNet), and evaluation metrics. Section 5 presents and discusses the results, analyzing spectral attention profiles, intervention effects, and robustness tests. Finally, Section 6 summarizes the main findings and suggests directions for future research in interpretable audio forensics.

2. Background and Related Work

To understand the vulnerabilities of deepfake detection systems, it is first necessary to understand the nature of the threat. This section provides an overview of the technologies used to generate synthetic speech, the architectures employed for detection, and the theoretical reasons why models tend to prioritize specific frequency regions. Finally, we examine the data augmentation and explainability techniques relevant to our investigation.

2.1. Synthetic Speech Generation

The term "deepfake" in the audio domain refers to speech signals generated or manipulated by deep neural networks to imitate human vocal characteristics with high fidelity. Current state-of-the-art technologies can be broadly generalized into two main paradigms: Text-To-Speech (TTS) and Voice Conversion (VC). TTS systems aim to synthesize speech directly from textual input, generating both linguistic content and prosodic features such as intonation and rhythm from scratch. In contrast, VC operates on an existing audio waveform, transforming the source signal to match the vocal identity of a target speaker while preserving the original linguistic content. While early approaches relied on concatenative synthesis, modern pipelines are predominantly data-driven. The transition of these systems has moved from statistical parametric methods to sequence-to-sequence neural networks, and finally to fully end-to-end architectures[39, 47].

2.1.1 Acoustic Modeling Stage

In the standard architecture of modern speech synthesis systems, the generation process is conceptually divided into a front-end and a back-end. The first crucial stage is Acoustic Modeling, which serves the purpose of mapping the linguistic input into an intermediate acoustic representation.

Statistical Parametric Speech Synthesis (SPSS) Historically, this task was performed using Hidden Markov Models (HMM). In this framework, speech is generated by predicting parameters (such as fundamental frequency and spectral envelope) from linguistic features using decision trees. Although flexible, HMM-based synthesis often resulted in muffled sounds and a buzzy quality due to the over-smoothing of spectral parameters and the simplified assumptions of the statistical models [57].

Hybrid HMM-DNN Approaches The first integration of Deep Learning into this pipeline maintained the parametric approach but replaced the decision trees with Deep Neural Networks (DNNs) to map linguistic features to acoustic parameters. These hybrid HMM-DNN systems significantly improved the mapping accuracy compared to traditional HMMs, but they still relied on complex, hand-crafted linguistic features and separate vocoding stages, limiting the potential naturalness of the output.

Sequence-to-Sequence (Seq2Seq) Models The advent of end-to-end Deep Learning fundamentally revolutionized this stage by introducing architectures that eliminate the need for manual feature engineering. Sequence-to-sequence architectures, such as Tacotron 2 [38], utilize attention mechanisms to align text sequences directly with acoustic frames (e.g., mel-spectrograms). Concurrently, non-autoregressive transformer-based models such as FastSpeech [33] increased generation speed by predicting spectrograms in parallel. Unlike their statistical predecessors, these neural architectures learn complex nonlinear mappings between character sequences and spectrograms, effectively capturing long-term dependencies and subtle nuances of human expression.

2.1.2 Vocoding Stage

The final stage of the synthesis pipeline is the Vocoder, the component responsible for converting intermediate acoustic features, such as mel-spectrograms predicted by the acoustic model, into

audio waveforms. The datasets analyzed in this thesis represent the technological evolution of this field, starting from traditional signal processing methods such as the Griffin-Lim algorithm [8] and parametric vocoders such as WORLD [22]. These systems reconstruct speech by estimating phase information or modeling vocal parameters (e.g., F_0 and spectral envelope), but often introduce audible buzzing or metallic artifacts due to simplified mathematical assumptions. A significant paradigm shift has occurred with autoregressive neural vocoders such as WaveNet [50], which generate waveforms one sample at a time. Despite achieving a high degree of naturalness, these models often rely on μ-law companding to reduce computational complexity, introducing a specific type of quantization noise that serves as a distinctive fingerprint for detection systems. More recently, this field has been dominated by non-autoregressive GAN-based architectures, such as MelGAN [17] and HiFi-GAN [16]. Although these models introduce state-of-the-art real-time synthesis, the use of transposed convolution to fill resolution gaps often leads to checkerboard artifacts [28] and artificial spectral periodicity. Understanding these different artifacts is crucial for forensics, as a detection model trained on the robotic inconsistencies of a traditional vocoder may fail to generalize when exposed to the more subtle, high-frequency textures typical of modern neural generators.

2.1.3 Fully End-to-End Architectures

While the two-stage pipeline (Acoustic Model + Vocoder) remains the standard for many applications, it suffers from the "mismatch problem": the vocoder is trained on ground-truth spectrograms but, during inference, receives predicted spectrograms from the acoustic model, which contain prediction errors. To address this, recent research has moved towards fully End-to-End (E2E) architectures that merge acoustic modeling and vocoding into a single network.
Models such as VITS (Conditional Variational Autoencoder with Adversarial Learning) [13] directly generate waveforms from text without explicit intermediate spectrogram supervision. By jointly optimizing the entire generation path, these systems reduce error propagation and produce highly realistic speech. However, they also introduce new challenges for detection, as the absence of a fixed intermediate representation (like a mel-spectrogram) can result in more subtle and complex artifacts that are harder to categorize than standard vocoder traces.

2.2. Synthetic Speech Detection

This section aims to provide readers with an overview of the current state of the art in synthetic speech detection, summarizing the main detection architectures.

2.2.1 Spectrogram-based CNN architectures

A first family of deepfake detectors is based on Convolutional Neural Networks designed to process time-frequency representations of audio signals. These models treat audio spectrograms as single-channel images, exploiting the ability of convolutional layers to automatically extract local and global patterns from spectral energy distributions. In this category, two of the standard architectures are **Light CNN (LCNN)** and **Residual Networks (ResNet)**.
Originally proposed for facial recognition, LCNN [54] has proven to be exceptionally effective in audio forensics due to the use of the Max-Feature-Map (MFM) activation function [18]. Unlike standard activation functions, MFM acts as a competitive feature selector, retaining only the most dominant activations among groups of feature maps. This mechanism is particularly well suited to the detection of speech deepfakes, as it helps the model isolate sparse synthesis artifacts from the complex background signal of natural speech.
ResNet architecture, on the other hand, uses residual connections to facilitate the training of deeper networks by allowing the gradient to flow through skip connections [9]. In the audio context, ResNet-18 is frequently used to capture high-level features from log-magnitude or power spectrograms. These models are commonly employed either as standalone architectures or combined in ensemble configurations to improve robustness and generalization performance [26, 27].

However, while highly performant, the standard design of these networks often includes aggressive spatial downsampling in the early layers. A key aspect of our investigation will be determining how this reduction in resolution interacts with the model's ability to focus on specific frequency bands. In this thesis, we focus exclusively on these architectures because they allow us to directly map classification decisions in the frequency domain, giving us a clear diagnostic path for analyzing spectral dependencies.

2.2.2 SincNet-based architectures

Another family of detection methods is based on the SincNet architecture [31]. SincNet was originally proposed to operate directly on raw waveforms by using parametrized band-pass filters, allowing the network to learn a task-specific filterbank optimized for speech processing.
A first prominent extension of this idea is RawNet2 [44], which combines SincNet-based filterbanks with residual convolutional blocks. The extracted representations are then passed through gated recurrent units (GRUs) to model temporal dependencies before final classification. To better capture joint time–frequency relationships, RawGAT-ST [44] replaces the GRU-based decoder with a spectro-temporal graph attention network (ST-GAT), enabling more flexible modeling of structured dependencies.
Building on this direction, AASIST [10] introduces parallel spectral and temporal branches that are fused through a heterogeneous stacking graph attention mechanism, explicitly targeting artifacts that manifest across both domains.
Overall, these approaches constitute an alternative design paradigm to conventional convolutional architectures, achieving performance generally comparable to the previously discussed methods.

2.2.3 Emerging architectures

More recently, research has increasingly shifted toward detection systems built upon large self-supervised learning (SSL) pre-trained encoders [36], most notably wav2vec 2.0 [4] and its multilingual extension, XLS-R [3]. One of the first works to adopt this paradigm was XLSR-AASIST [45], which replaced the SincNet front-end of AASIST with embeddings extracted from a pre-trained XLS-R model.
This approach was later refined in XLSR-SLS [58], where representations from all transformer layers were aggregated through a learnable layer-wise weighted sum. This design is motivated by the assumption that deepfake artifacts may emerge at different levels of abstraction within the network. More recently, XLSR-Mamba [55] substituted the AASIST back-end with a state space sequence encoder, further increasing modeling capacity.
Although these SSL-based architectures achieve remarkable detection performance, their decision-making process remains largely opaque. This limitation is even more pronounced than in earlier models, as these networks are deeper, more complex, and pre-trained on massive external corpora. Research on their interpretability is still ongoing, and they remain far from ideal for real-world applications where explainability (XAI) is a strict requirement.
In this thesis, we instead focus on architectures operating on time–frequency representations (spectrograms) processed by Convolutional Neural Networks (CNNs). This choice is driven by our objective to explicitly analyze how detection models weight different frequency bands. Spectrogram-based models enable a direct mapping between model responses and the frequency domain, thereby facilitating a more transparent diagnostic analysis. In particular, we examine two widely adopted architectures for this task: Light CNN (LCNN) and ResNet-18.

2.3. Explainable AI in Synthetic Speech Detection

A primary obstacle in the development of speech deepfake detectors is their lack of explainability. Despite achieving high accuracy, modern systems based on deep neural networks often operate as black boxes, providing detection scores without a clear justification for their internal logic. This lack of transparency represents a significant limitation in high-stakes environments, such as

forensic investigations or financial transactions, where human-understandable reasoning is critical to ensuring that a verdict is legally defensible and trustworthy.
Historically, traditional detection methods based on MFCC or CQCC offer a high level of interpretability since decisions can be directly linked to specific acoustic properties. However, these approaches do not achieve the performance levels obtained by end-to-end models. To bridge this gap, recent research has explored post-hoc attribution techniques, such as SHAP, LIME, and Grad-CAM, to retroactively identify the temporal and/or spectral regions that most influenced a model's predictions.
Alternatively, some frameworks seek to achieve explainability by design through architectural constraints. For example, the SFAT-Net series of models, a type of multi-task transformer, links detection to phonetic features [25], while other strategies leverage emotional or prosodic inconsistencies [2]. Other approaches adopt a fine-grained phoneme-level analysis, explicitly aligning detection cues with individual phonetic units in order to obtain a more localized and interpretable characterization of synthetic artifacts [35].
A fundamental study for the research conducted in this thesis is the work of Salvi et al. [34], which introduces a diagnostic approach to frequency band explainability. Using both a posteriori interpretations (LIME) and active interpretations through band-limited training, the authors demonstrate that forensic traces are not uniform across the spectrum. Instead, they find that synthesis artifacts are often concentrated in specific regions, mainly in the low frequencies (below 1600 Hz) and in the extremely high frequencies.
Although these results provide a clear starting point, interpretability remains a fragmented field in forensic audio. Most state-of-the-art models continue to prioritize performance over diagnostic clarity, leaving the underlying causes of spectral dependence largely unexplored. This investigation aims to build on these findings to analyze how standard architectures distribute their attention across frequency bands.

3. Proposed Methodology

3.1. Problem Formulation

The synthetic speech detection problem can be formally defined as follows. Let us consider a discrete-time input speech signal x sampled at a frequency f_s and associated with a class $y \in \{0, 1\}$, where 0 denotes that the signal is authentic while 1 indicates that it is has been synthetically generated. The goal of this task is to develop a speech deepfake detector $\mathcal{D}$ that estimates the class of the signal x as $\hat{y} \in [0, 1]$, where $\hat{y}$ is the likelihood that the signal $\mathbf{x}$ is fake. The aim of this thesis is to examine the relationship between the characteristics of x and the resulting output $\hat{y}$ of $\mathcal{D}$.

3.2. Considered Approach

Current state-of-the-art systems for deepfake speech detection operate as "black boxes" [30]. While their performance on standard datasets is impressive, the internal decision-making mechanism remains largely unknown. This lack of in-depth knowledge poses a significant risk: without understanding the specific features a model uses, it is impossible to guarantee its robustness against unseen attacks or manipulations [5].
In this thesis, we address this problem by proposing an investigative framework designed to analyze the spectral dependencies of detection models. As suggested by recent literature [34], Convolutional Neural Networks trained on audio spectrograms do not process the frequency spectrum uniformly. On the contrary, they tend to develop strong dependencies on specific frequency ranges, such as high or low-frequency components. While this behavior may achieve high performance on clean data, it risks compromising the robustness of the detector in real-world scenarios, where the specific bands on which the model relies may be corrupted or altered. Consequently, our proposed intervention aims to emphasize generalization by encouraging the model to exploit information spread across the entire spectrum, rather than relying solely on narrow spectral regions.

However, a comprehensive study in this direction across multiple datasets and architectures is currently lacking. Consequently, the objective of this thesis is twofold: validate the existence and nature of these spectral dependencies (*goal 1*): explore potential mitigations to emphasize robustness against them (*goal 2*).
To achieve these results, we designed a three-step pipeline:

1. **Quantification** (targeting *goal 1*): To analyze the spectral focus of the models, we adapt the Relative Contribution Quantification (RCQ) [20] framework. This technique allows us to aggregate local, instance-based explanations [37] into a global metric, effectively quantifying the model's attention profile across the frequency spectrum.
2. **Validation** (targeting *goal 1*): Since attention maps indicate correlation rather than causation, we apply adversarial stress tests via *Frequency Swapping*. By modifying specific spectral bands, we verify that the observed attention values are effectively the driver of the model's decisions
3. **Intervention** (targeting *goal 2*): Finally, to mitigate the identified dependencies, we introduce a data augmentation strategy, *Stratified Spectral Mixing (SSM)*. This technique is based on methods like SpecMix [12], but it is designed to disrupt only the vertical spectral coherence in training, encouraging the model to learn more distributed features.

3.2.1 Quantification

To investigate the spectral dependencies of deepfake detectors, we must quantify the contribution of each time-frequency bin to the model's decisions. Since standard saliency mapping techniques are instance-based, we adopt a two-step approach: first, we generate high-resolution explanations for each individual sample using Guided Grad-CAM; then, we aggregate these explanations into a global metric using an adaptation of the RCQ framework.

Guided Grad-CAM
Standard visualization techniques often face a trade-off between resolution and class specificity. Grad-CAM [37] applied to deeper activation maps of the network excels at localizing discriminative regions but produces low-resolution heatmaps due to spatial downsampling in these latter convolutional layers. In contrast, Guided Backpropagation [42] provides pixel-level detail by modifying the backpropagation of ReLU activations, but it can be noisy and less class-discriminative. In order to detect both the high-resolution details of the synthetic artifacts and the semantic concentration of the model, we employ Guided Grad-CAM. This method combines the two approaches via element-wise multiplication:

$$M_{attr} = M_{GradCAM}^{upsampled} \odot M_{GuidedBackprop} \tag{1}$$

Global RCQ
While M_{attr} provides a detailed explanation for a single spectrogram, it only offers a local perspective. Identifying systematic spectral dependencies requires aggregating these maps into a global metric. Our adaptation of the RCQ framework processes the evaluation set in such a way as to calculate importance scores for each frequency bin, depending on the audio content. The process consists of two steps for each input sample x:

- *Semantic Segmentation (M_{sem}):* First, we divide the signal into sections based on content. We use Fast Context-based Pitch Estimation (TorchFCPE) [21] to track the fundamental frequency (f_0). Based on this detection, we generate binary masks that separate the signal into Voiced regions (where $f_0 > 0$), Unvoiced regions, and Transition regions. This allows us to determine whether the model focuses on the harmonic structure of the voice or on noisy components.
- *Saliency Map Calculation (M_{attr}):* We generate the saliency map M_{attr} using the Guided Grad-CAM method defined above. In particular, to accurately represent the model's decision process, we calculate the gradients with respect to the predicted class $\hat{y}$ rather than the ground truth label y. This ensures that in cases of errors (false positives or false negatives), we visualize the features that misled the model.

The global RCQ profile is derived from the aggregation of these saliency maps across the entire test set. For each frequency bin f, detection category $c \in \{TP, TN, FP, FN\}$, and segment type $s \in \{Voiced, Unvoiced, Trans\}$, we compute the mean importance $\mu_{f,c,s}$ accumulated over the time dimension T and across all N_{category} samples. The RCQ_f is defined as the percentage deviation of that bin from the global mean importance (μ_{global}):

$$RCQ_{f,c,s} = \frac{\mu_{f,c,s} - \mu_{global}}{\mu_{global}} \times 100 \tag{2}$$

A value of $RCQ_f > 0$ indicates that the frequency band f contributes more than the average to the model's decision. By plotting RCQ_f against the frequency axis, we obtain a "spectral attention profile", revealing whether a model relies on specific bands (e.g., high-frequency artifacts) while ignoring others.

3.2.2 Validation

RCQ profiles highlight correlations between frequencies and model attention, but they do not prove causation. To verify whether the highlighted frequency bands actually drive the model's decision process, we introduce an adversarial stress test based on *Frequency Swapping*. This method allows us to answer the question: "How much would the model change its prediction if the content of this specific band were different?"

Experimental Protocol
We analyze the spectrum by discretizing it into fixed-width bands of 1000 Hz (e.g., 0-1 kHz, 1-2 kHz, up to Nyquist). For each band B, we generate a dataset of $N = 5000$ hybrid samples. The hybrid sample S_{hybrid} is created by taking a frequency band B from a source spectrogram S_{src} onto a target spectrogram S_{tgt} belonging to the opposite class, using a binary mask M_B:

$$S_{hybrid} = M_B \odot S_{src} + (1 - M_B) \odot S_{tgt} \tag{3}$$

To ensure the validity of the test, S_{src} and S_{tgt} are selected from the *correctly classified samples*. This ensures that only the spectral manipulation is responsible for any change in the model's prediction.

Testing Modes
To confirm the causal relationship between spectral bands and the decisions made by the model, we apply this swapping procedure in two different test modes. For both modes, we quantify the impact using the *Fake Detection Rate (FDR)* and *Score Sensitivity (ΔS)* metrics, formally defined in Section 4.5.

1. *Fake Injection (Sufficiency Test):* We inject a frequency band from a Deepfake sample into a Bonafide file. The goal is to verify if a specific spectral band carries enough artifacts to independently trigger a Fake detection (high FDR) or significantly raise the anomaly score (high ΔS).
2. *Real Injection (Vulnerability Test):* We inject a frequency band from a Bonafide sample into a Deepfake file. This tests whether the presence of genuine spectral cues in a specific band is sufficient to "repair" the deepfake and lower the detection score. A significant drop in the score implies the model relies heavily on artifacts in that band.

The "Frankenstein Effect"
An effect observed during the *Real Injection* is what we call the *Frankenstein Effect*. Intuitively, an injection of a Bonafide band into a Deepfake should lower the score (the file becomes "more real"). However, sometimes the opposite is observed: the fakeness score increases. This suggests that the model detects the *Spectral Coherence*: it penalizes the spectral discontinuity more than it rewards the presence of bonafide features.

Bandwidth Stress Test
While the Frequency Swapping experiment provides a quantification of the influence of individual

spectral bands on classification, it is an artificial manipulation that does not represent real-world scenarios. To integrate this analysis with a plausible use-case, we introduce the *Bandwidth Stress Test*.
Telephony networks often drastically reduce the audio bandwidth, cutting off the high and low frequencies. To assess the model's robustness in such environments, we filter the test set using the standard *GSM Band* (300 Hz - 3400 Hz) [6]. We then measure the *Survival Rate (SR)*, defined as the percentage of correctly detected samples that maintain the correct classification after filtering. This test reveals whether the model has learned robust cues in the voice band or if it relies solely on high and low frequency artifacts.

3.2.3 Intervention

The third and final stage of our framework involves an intervention. If the hypothesis of spectral bias holds, forcing the model to learn from spectrally fragmented data should decrease the reliance on specific bands. To achieve this, we propose *Stratified Spectral Mixing (SSM)*, a data augmentation strategy designed to break vertical spectral coherence.

Conceptual Design
Existing augmentation methods like SpecMix [12] operate by cutting and pasting random time-frequency rectangles. While effective for regularization, this approach does not specifically target the vertical structure of the spectrograms. SSM modifies the SpecMix logic by generating masks composed exclusively of *horizontal frequency bands*.

- *Why Horizontal Bands?:* Deepfake artifacts typically manifest as patterns over specific frequency ranges. Standard training often leads model to focus on the most prominent artifacts (e.g., exclusively in high frequencies), neglecting the rest of the spectrum. By preserving the horizontal integrity of the bands, SSM maintains these local artifacts but randomizes the global context in which they appear.
- *Goal:* The objective is to reduce the model's tendency to rely excessively on the most immediately discriminative spectral cues. By training on samples composed of conflicting spectral information (e.g., Bonafide low and mid frequencies mixed with Deepfake high frequencies) we force the model to look beyond the most obvious features. This makes the network explore the full spectrum and learn to identify cues in frequency bands it would typically ignore.

Mixing Protocol and Soft Labels
SSM Augmentation is applied stochastically during the training phase. For each input sample x_A (Target), the transformation is triggered with a probability of $p = 0.75$. When applied, the algorithm proceeds as follows:

1. *Source Selection (x_B):* We select another sample x_B from the dataset. To prevent class imbalance during augmentation, x_B is sampled with a balanced strategy (50% probability Real, 50% Fake), regardless of the original class distribution.
2. *Mask Generation:* We randomly select a number of cut points k in the range $\{1, 2, 3\}$. These points partition the spectrum into $N = k + 1$ horizontal bands. Then we generate a binary mask M of the same dimensions as the spectrogram. For each band, we randomly assign a value of 0 or 1. This creates a spectral stratification where $M = 1$ indicates regions to keep from x_A, and $M = 0$ regions to inject from x_B.
3. *Synthesis:* The final spectrogram $\tilde{x}$ is constructed by mixing the two samples according to the mask:
$$\tilde{x} = M \odot x_A + (1 - M) \odot x_B \tag{4}$$
Note that, before mixing, the samples are cropped to the minimum duration between x_A and x_B to ensure alignment.

Since the mixed spectrogram contains spectral information from potentially different classes, we compute the target label $\tilde{y}$ as the weighted average of the source labels, proportional to the spectral area:

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \tag{5}$$

where λ represents the proportion of the frequency bins belonging to the file A. This soft-labeling strategy teaches the model to quantify the degree of fakeness based on the proportion of spectral components.

4. Experimental Setup

In this section, we detail the datasets, data processing pipeline, model architectures, and training parameters. The experimental environment is implemented in PyTorch.

4.1. Datasets

We employed the ASVSpoof 2019 Logical Access (LA) dataset for training, model selection and evaluation. Additionally, to verify the generalization capabilities of our models and the results of the SSM augmentation, we performed cross-dataset evaluation on four different out-of-domain datasets, characterized by different recording conditions and synthesis algorithm.

- **ASVspoof 2019 LA:** Based on the VCTK corpus [51], this dataset serves as the standard community benchmark for synthetic speech detection. It comprises recordings from 107 speakers (46 male and 61 female), sampled at 16 kHz and captured in a hemi-anechoic chamber to ensure high-quality, noise-free conditions. The Logical Access partition specifically targets text-to-speech (TTS) and voice conversion (VC) attacks, excluding physical replay scenarios [47].
  The training set contains 2,580 bona fide and 22,800 spoofed utterances generated by six algorithms (A01–A06). The development set includes 2,548 bona fide and 22,296 spoofed samples produced by the same attack algorithms as in training. The evaluation set consists of 7,355 bona fide and 63,882 spoofed utterances generated by 13 different systems (A07–A19), 11 of which are unseen during training, enabling the assessment of robustness against previously unknown synthesis methods.
- **In-The-Wild**: To simulate real-world scenarios, we employed the In-The-Wild dataset [23]. Unlike controlled studio recordings, this corpus consists of approximately 38 hours of audio (17 hours fake, 21 hours real) collected from social media and video-sharing platforms. It features speech from 58 celebrities and politicians, characterized by uncontrolled acoustic environments, background noise, and diverse lossy compression codecs, making it a challenging benchmark for robustness.
- **FakeOrReal:** This large-scale dataset [32] is designed to balance variety in generative models. It contains approximately 195,000 utterances, divided between bonafide speech collected from open sources (e.g., TED Talks, YouTube) and synthetic speech generated by 7 state-of-the-art TTS systems, including Deep Voice 3 and Google WaveNet. In this thesis, we utilized the `for-norm` version, which applies loudness normalization to the audio files. This preprocessing step is crucial to mitigate potential biases arising from the different recording volumes and conditions of the diverse source materials.
- **MLAAD (English subset):** The Multi-Language Audio Anti-spoofing Dataset (MLAAD) [24] is a massive corpus comprising 378 hours of synthetic speech in 38 languages generated by 82 different TTS models. For our experiments, we selected the English subset to maintain consistency with the other datasets. The bonafide counterpart is sourced from the M-AILABS dataset [41], while the spoofed samples are drawn from MLAADv8. We applied a random 70/15/15 split.
- **ASVspoof 5 (Track 1):** We also evaluated our models on the latest version of the challenge, ASVspoof 5 [53]. We focused on Track 1, filtering the test set to select only samples corresponding to the clean condition (codec '-'). We put together a balanced test set by randomly sampling 7,760 files for the bonafide class and for each of the specific attack systems (A17, A19, A21, A22, A24, A25, A26, A28, A29), ensuring a uniform distribution across the attack types.

4.2. Data Processing Pipeline

We structured a standardized data processing pipeline to ensure consistency across all experiments. The specific parameters used in this pipeline are detailed in Table 1.
The raw audio waveform undergoes a series of transformations before being fed into the model:

1. **VAD:** To eliminate non-informative and possibly biasing silent segments, all audio files are processed through Voice Activity Detection (VAD) using *Silero VAD* [46].
2. **Duration Standardization:** We enforce a standard duration of $T = 3.0$ seconds. Audio shorter than T is looped, while longer audio is cropped.
3. **Feature Extraction:** We convert the waveform into a time-frequency representation using Short-Time Fourier Transform (STFT). Then, the power and the logarithm of the STFT is computed to obtain a Log Power Spectrogram.
4. **Frequency-wise Normalization:** We apply instance-level normalization independently for each frequency bin. We compute the mean μ_f and standard deviation σ_f along the time axis for each frequency f, and standardize the bin as $X_{f,t} = (X_{f,t} - \mu_f)/\sigma_f$. This step, conceptually similar to Cepstral Mean and Variance Normalization (CMVN) [52], is crucial to normalize the energy distribution across frequency bands.

Parameter	**Value**
Sampling Rate	16,000 Hz
FFT Size (N_{FFT})	512
Hop Length	128
Window Function	Hann
Feature Type	Log-Power Spectrogram
Normalization	Instance-level, Frequency-wise
Input Dimensions	257×376 (Freq $\times$ Time)

Table 1: Parameters configuration for the data processing pipeline.

4.3. Model Architectures

To investigate whether the observed spectral dependencies originate from the data or from the model structure, we evaluate three distinct architectures. All networks are initialized from scratch and modified to accept 1-channel input (spectrograms) and 2 output classes (Bonafide vs. Spoof).

Light Convolutional Neural Network (LCNN)
The LCNN architecture [54] is widely adopted in the speech deepfake detection domain for its efficiency and state-of-the-art performances on logical access tasks. Our implementation follows the 9-layer configuration described in [18]. The defining characteristic of LCNN is the use of Max-Feature-Map (MFM) activations instead of standard Relu. MFM introduces a selection mechanism where, given a convolutional layer outputting $2N$ feature maps, it performs an element-wise maximum between two candidate groups:

$$MFM(x) = \max(x_{1:N}, x_{N+1:2N}) \tag{6}$$

This mechanism works as a feature selector: by suppressing low-activation features in favor of dominant ones, MFM effectively filters out components unrelated to the classification task. In the context of spoofed speech detection, this allows the model to separate sparse synthesis artifacts from the background signal.

ResNet-18
We employ a standard ResNet-18 [9] adapted for audio by modifying the first convolutional layer

to accept single-channel input. The standard architecture is characterized by an aggressive downsampling in its initial stages, consisting of:

- A 7×7 Convolution with *stride 2*.
- A 3×3 Max Pooling with *stride 2*.

This configuration results in a $4\times$ reduction in spatial resolution immediately at the first layer. While being powerful for computer vision, our analysis suggests that such an aggressive downsampling might harm the robustness in audio forensics, as fine-grained spectral details might be discarded before deeper layers can process them

ResNet-18 (No-Stride)
To empirically verify if the downsampling mentioned above causes a loss of information, we evaluate a modified variant denoted as *ResNet-18 No-Stride*. In this architecture, we preserve the full input dimension in the first layers by:

- Modifying the first convolutional layer to use a *stride of 1* (instead of 2).
- Removing the following Max Pooling operation.

This modification allows the network to process the input at is original resolution in the initial layers, keeping the fine-grained details without early spatial downsampling.

4.4. Training Protocol

We adopted a unified training protocol through all the experiments to assure comparability and reproducibility. All models were implemented, trained and evaluated using the Pytorch framework.

Optimization and Hyperparameters
We use the Adam optimizer [14] with a learning rate of 10^{-4} and a weight decay of 10^{-3} to prevent overfitting. The batch size is set to 256. We adopted a dynamic learning rate strategy: a *ReduceLROnPlateau* scheduler monitors the Equal Error Rate (EER) on the validation set and halves the learning rate if no improvement is seen for 5 consecutive epochs. The training runs for a maximum of 100 epochs, with an early stopping check triggered if the EER does not improve after 20 epochs.

Class Balancing
Since the ASVSpoof 2019 training set is highly imbalanced, with a majority of spoofing attacks, we employ a weighted random sampler during training. This ensures that each batch contains a balanced distribution (50/50) of Bonafide and Spoof samples.

Data Augmentation Pipeline
We applied existing data augmentation techniques during training to ensure sample variability. Each augmentation step is applied stochastically (with $p = 0.4$ for each transformation) to the waveform before feature extraction:

- *RawBoost:* A specialized augmentation for anti-spoofing that introduces noise, convolutive noise and distortion to simulate different real-life recording situations [43]
- Codec Compression: We introduce artifacts by applying lossy compression (MP3, Vorbis) at various bitrates (8-128 kbps), or reducing the bit-depth of the samples.
- Reverberation: We convolve the signal with real and simulated Room Impulse Response (RIRs) to add diverse environmental characteristics. [15]
- Noise Injection: We add background noises from the MUSAN dataset [40] with a random SNR between 5 and 30.

Note that for the SSM experiments, the Stratified Spectral Mixing is applied together with this augmentation pipeline, forcing the model to adapt both to channel distortion and spectral fragmentation.

Loss Functions
Depending on the training experiment (Standard vs. SSM), we employ two different loss functions:

- Focal Loss (baseline): For the baseline experiments (models trained without SSM), where targets remain binary, we minimize the Focal Loss ($\gamma = 2$) [19]. This loss functions scales the standard cross-entropy, down-weighting easily classified samples to focus learning on hard-to-detect examples.
- Soft Cross-Entropy (SSM): Since Stratified Spectral Mixing generates continuous soft labels $\tilde{y} \in [0, 1]$ as described in section 3.2.3, we minimize the Soft Cross-Entropy Loss:

$$\mathcal{L}_{soft} = -\frac{1}{N} \sum_{i=1}^{N} [\tilde{y}_i \log(p_i) + (1 - \tilde{y}_i) \log(1 - p_i)] \tag{7}$$

  where p_i is the predicted probability for the spoof class.

4.5. Evaluation Metrics

To provide a complete overview of both detection performance and spectral robustness, we employ the standard metric for binary classification and a set of custom metrics for our spectral analysis experiments. Since the detection task is a binary classification problem (where the positive class is typically defined as "Spoof"), we define the fundamental outcomes as follows:

- True Positive (TP): A synthetic sample correctly classified as Spoof.
- True Negative (TN): A genuine sample correctly classified as Bonafide.
- False Positive (FP): A genuine sample incorrectly classified as Spoof.
- False Negative (FN): A synthetic sample incorrectly classified as Bonafide.

Based on these quantities, we compute the following metrics:

Equal Error Rate (EER): the standard metric used in the ASVSpoof challenge and represents the operating point where False Acceptance Rate (FAR) equals the False Rejection Rate (FRR).

$$FAR = \frac{FP}{FP + TN}, \quad FRR = \frac{FN}{FN + TP} \tag{8}$$

It is threshold-independent and a lower value indicates better performance.

Fake Detection Rate (FDR): This metric (used in the Frequency Swapping experiments) measures the percentage of hybrid samples that are classified as "Spoof" by the model (output probability > 0.5).

$$FDR = \frac{N_{detected}}{N_{total}} \times 100 \tag{9}$$

Score Sensitivity (ΔS): Measures the average variation in the model output probability caused by spectral perturbations. We compute it in two different variants:

- Score increase (sufficiency): When injecting a Fake band into a Real sample (x_{real}), we measure how much the fakeness probability increases:

$$\Delta S_{impact} = P(spoof|\tilde{x}_{hybrid}) - P(spoof|x_{real}) \tag{10}$$

- Score drop (vulnerability): When injecting a Real band into a Fake sample (x_{fake}), we measure how much the fakeness probability decreases:

$$\Delta S_{drop} = P(spoof|x_{fake}) - P(spoof|\tilde{x}_{hybrid}) \tag{11}$$

Metrics for Bandwidth Stress Test: To evaluate robustness against the GSM bandwidth limitation, we introduce two specific metrics:

- **Survival Rate (SR):** The percentage of samples that were originally classified correctly and remain correctly classified after the application of the band-pass filter. A high SR indicates robustness to signal degradation.
- **Average Score Drift (ΔS_{drift}):** Quantifies how much the model's confidence shifts away from the correct class after filtering. It is defined based on the sample type:

- For *Spoofed* samples (TP):

$$\Delta S_{drift} = P(spoof|x_{original}) - P(spoof|x_{filtered}) \tag{12}$$

where a positive drift indicates a loss of confidence in the fake class (the score moves towards 0).

- For *Bonafide* samples (TN):

$$\Delta S_{drift} = P(spoof|x_{filtered}) - P(spoof|x_{original}) \tag{13}$$

where a positive drift indicates an increase in fakeness probability (the score moves towards 1).

5. Results

This section presents the outcomes of our experimental campaign.
The aim of this work is to analyze the spectral dependencies of deepfake detectors. Rather than limiting the study to conventional performance evaluation, we seek to understand how existing architectures rely on specific frequency bands, and whether actively manipulating these dependencies through our proposed SSM approach can improve generalization and robustness.
To systematically address these questions, the experimental campaign is structured into three stages: *Explainability Analysis*, *SSM Intervention Study*, and *Robustness Evaluation*.

5.1. Explainability Analysis

In this stage, we compute global RCQ profiles to characterize how each model distributes its attention across frequency bands. This analysis reveals the extent to which detectors rely on particular spectral components.

We first analyze the behavior of the LCNN architecture, considering it as the reference for this task. Figure 1 shows the aggregate RCQ profile for True Positive samples computed on the ASVSpoof 2019 LA evaluation set. The visualization reveals a prominent attention imbalance: the model exhibits a *U-shape* profile, concentrating its discriminative attention at the spectral extremes while reducing it in the center.

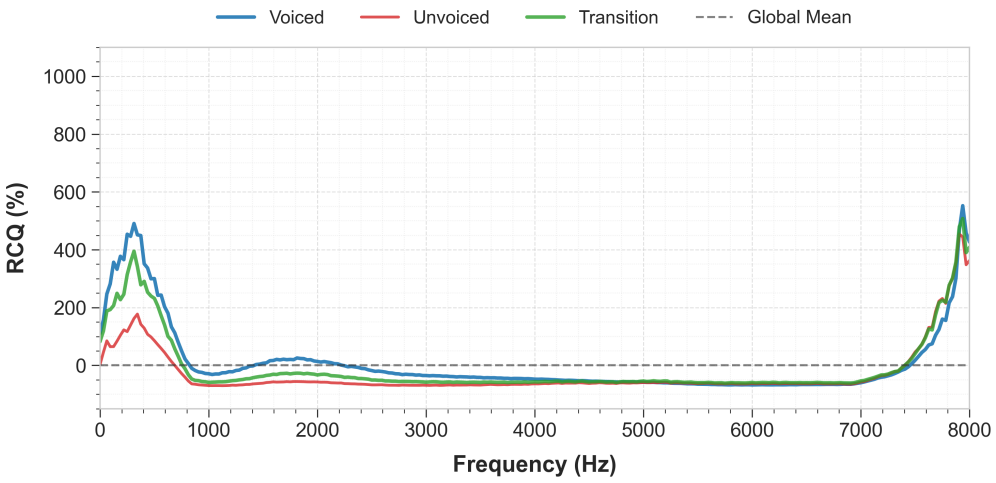


Figure 1: Aggregate RCQ Profile for the LCNN architecture, computed on True Positive samples from the ASVspoof 2019 LA evaluation set.

A significant region of interest appears to be the frequency region below 1KHz. The voiced component (blue line) shows a strong activation in this area, reaching relative importance values around 500%. This suggests the model is analyzing the fundamental frequency (f_0) and the first harmonic structures.
The highest contribution is observed near the Nyquist frequency, where the RCQ values peak above 550%. This aligns with known characteristics of vocoders, which often fail in reconstructing high-fidelity spectral elements in the upper bands, leaving distinct artifacts that the model exploits as primary detection clue.
Notably, the mid-frequency band shows negative RCQ values, indicating that the model is suppressing information from this extended frequency range.

Comparing the profiles of TP and FP on the source dataset ASVspoof 2019 offers an insight into the reliability of these features. While TP samples trigger high attention in both low and high bands, the FP show a different pattern (Figure 2).

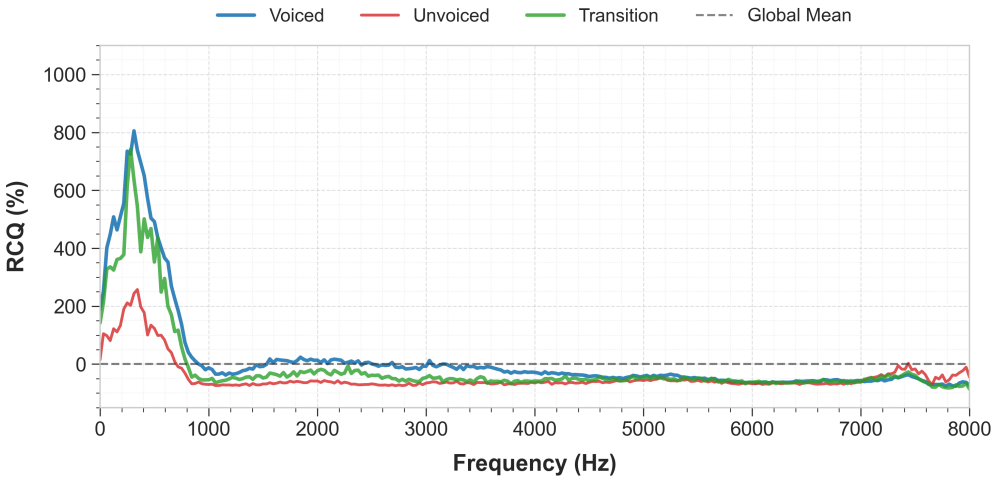
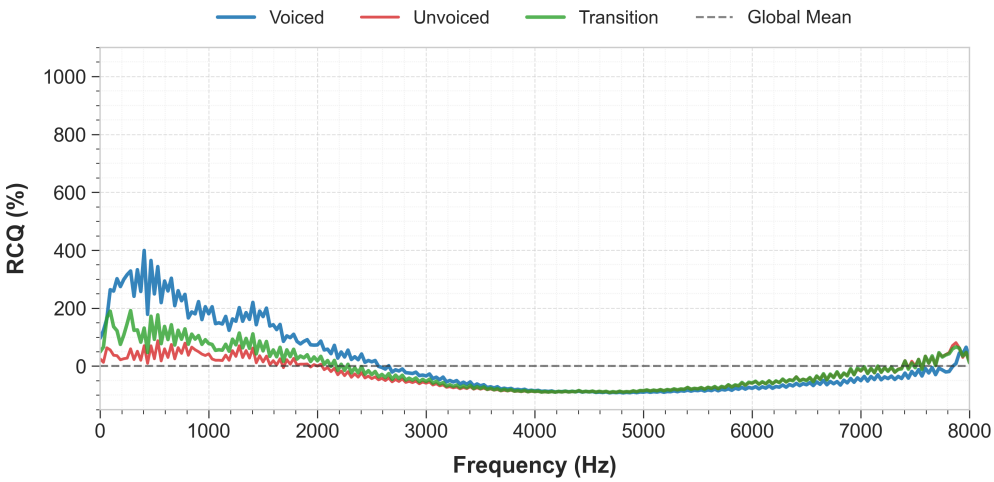


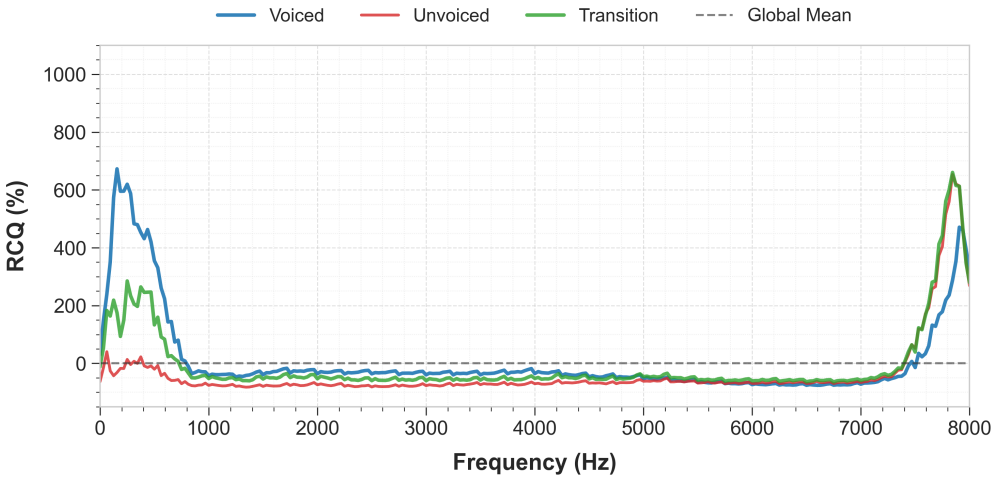
Figure 2: Aggregate RCQ Profile for the LCNN architecture, computed on False Positive samples from the ASVspoof 2019 LA evaluation set.

In these FP cases, the attention on the high-frequency band disappears, while the low-frequency peak is still prominent. This suggest that within the training data, high-frequency artifacts are a strong indicator: their absence usually means that the sample is bonafide. On the contrary, low-frequency features are error-prone: the model occasionally confuses natural irregularities of bonafide speech with deepfake artifacts, leading to false positives.
It is worth mentioning that in cross-dataset analysis (e.g., InTheWild), high-frequency attention often reappears in False Positives. This is likely because of environmental noise and compression artifacts being misinterpreted by the model as synthetic speech patterns.

While the "U-Shape" bias seems to be data-driven, we observed a secondary, model-specific artifact affecting the standard ResNet-18. Figure 3 compares the RCQ profiles computed on the ASVSpoof 2019 LA evaluation set of the Standard ResNet-18 (Top) against the modified ResNet-18 No-Stride (Bottom).



(a) Standard ResNet-18: Note the "Sawtooth" oscillation.



(b) ResNet-18 No-Stride: The profile is smoother

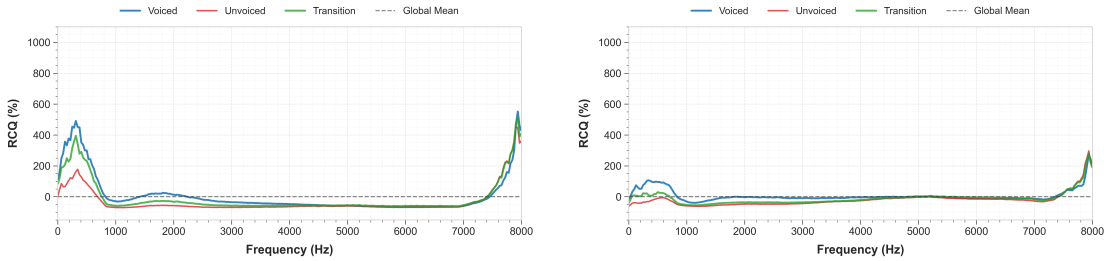
Figure 3: Comparison of RCQ profiles (True Positives) between Standard and No-Stride architectures. The Standard model shows high-frequency fluctuations caused by spatial downsampling, which are significantly mitigated in the No-Stride variant.

The Standard ResNet profile (Fig. 3a) is characterized by prominent, high-frequency oscillations across the entire spectrum with *sawtooth*-like effect. Adjacent frequency bins show drastically different importance scores, suggesting periodic "blind spots" in the model's perception.
This effect is caused by the initial downsampling (stride 2) in the standard ResNet architecture. This creates a mismatch between the continuous harmonic structure of the audio and the way gradients are propagated through the network. By removing the stride in the ResNet-18 No-Stride variant (Fig. 3b), we preserve the full input resolution.

5.2. SSM Intervention Study

In this stage, we investigate the effect of SSM on the attention patterns of the considered models. By disrupting vertical spectral coherence during training, we assess whether models adapt their frequency attention patterns and examine how such changes relate to cross-domain generalization.

Qualitative Analysis
To verify if the intervention alters the models' internal focus, we computed the RCQ profiles for the SSM-trained variants. Figure 4 shows a comparison between the LCNN Baseline (Left) and the SSM-trained variant (Right) on True Positive samples.



(a) LCNN Baseline: Note the "U-Shape" with high values on extreme bands.

(b) LCNN + SSM: The profile is significantly flattened.

Figure 4: Side-by-side comparison of RCQ profiles (LCNN, True Positives) before and after SSM intervention. While a residual "U-Shape" persists, SSM significantly reduces the magnitude of the spectral bias. Note that both plots share the same Y-axis scale to facilitate a direct comparison.

The visual comparison confirms a significant *dampening effect*. The most notable effect is the drastic reduction in peak intensity. The low frequency attention, which exceeded 500% in the baseline (Fig. 4a), drops to 100% in the SSM-trained model (Fig. 4b). Similarly, the high frequency prominence is highly reduced. While the mid-frequency band still does not become the primary focus, the depth of the valley is reduced. The profile moves towards the global mean, indicating that the model is no longer actively suppressing these frequencies as aggressively as it did in the baseline, leading to more balanced feature extraction.

Quantitative Impact
We now evaluate how this change in spectral attention translates into performance metrics. Table 2 compares the Equal Error Rate (EER) of the baseline models against their SSM-augmented counterparts for all three architectures.

Table 2: EER (%) comparison across five dataset. The best performance for each architecture is highlighted in **bold**.

Model Architecture	In-Domain	Out-of-Domain (OOD)			
	ASV19 LA	InTheWild	MLAAD	FakeOrReal	ASVSpoof 5
LCNN (Baseline)	16.86	**26.91**	35.99	21.11	27.36
LCNN + SSM	**14.24**	27.56	**34.87**	**11.48**	**19.33**
ResNet-18 Std (Baseline)	19.51	**32.25**	38.46	34.72	25.82
ResNet-18 Std + SSM	**16.51**	36.39	**38.05**	**33.17**	**21.24**
ResNet-18 NS (Baseline)	15.93	**21.57**	33.03	10.56	23.93
ResNet-18 NS + SSM	**14.09**	22.74	**21.76**	**06.32**	**19.55**

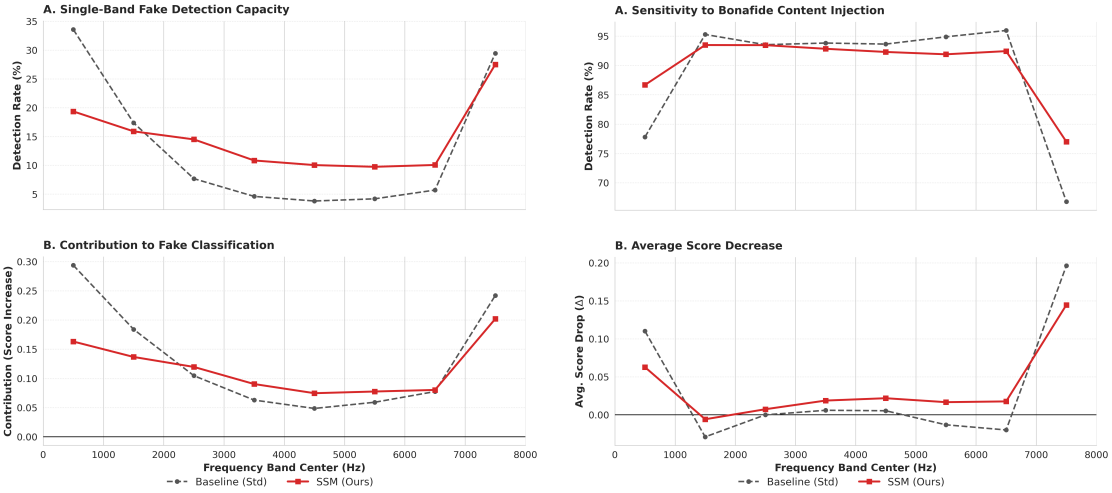
A notable outcome is that SSM consistently improves performance on the source dataset (ASVSpoof 2019 LA) across all three architectures. For instance, LCNN improves from 16.86% to 14.24%, and the ResNet-18 NS achieves the best overall score of 14.09%. Since the ASVSpoof 2019 Evaluation set contains attack algorithms not present in the training set, this improvement indicates that SSM successfully helps the model in generalizing to unseen algorithms.

The method improves particularly the model's classification capabilities on datasets characterized by unseen attack algorithms rather than just environmental noise. On ASVSpoof 5, SSM yields consistent improvements (e.g., ResNet-NS drops from 23.93% to 19.55%). Notably, on FakeOrReal, LCNN with SSM achieves a remarkable relative EER reduction of 46%.
While SSM is model-agnostic, the ResNet-18 NS benefits the most from the intervention. On the MLAAD dataset, while LCNN shows modest gains, the ResNet-18 No-Stride registers an impressive improvement, dropping from 33.03% to 21.76%. This suggests that removing the downsampling, combined with spectral mixing, unlocks the full potential of this network.
The only case in which SSM causes a consistent performance degradation is the InTheWild dataset (e.g., LCNN moves from 26.91% to 27.56%). Unlike the other datasets, InTheWild is characterized by uncontrolled environmental noise. A hypothesis could be that the vertical spectral coherence disruption might interfere with the model's ability to handle background noise. However, the degradation is marginal compared to the gains achieved on all other OOD datasets.

5.3. Robustness Evaluation

In this last stage, we evaluate whether SSM-based training yields measurable robustness improvements under realistic signal degradations.
To address this, we designed two complementary experiments. In the first (*band-wise sensitivity analysis*), we mechanically dissect the spectrum to verify whether the model has learned to use the mid-frequencies. In the second (*real-world scenario*), we simulate a real-world degradation scenario (telephony bandwidth limitation) to see if the attention shift translates into practical robustness.

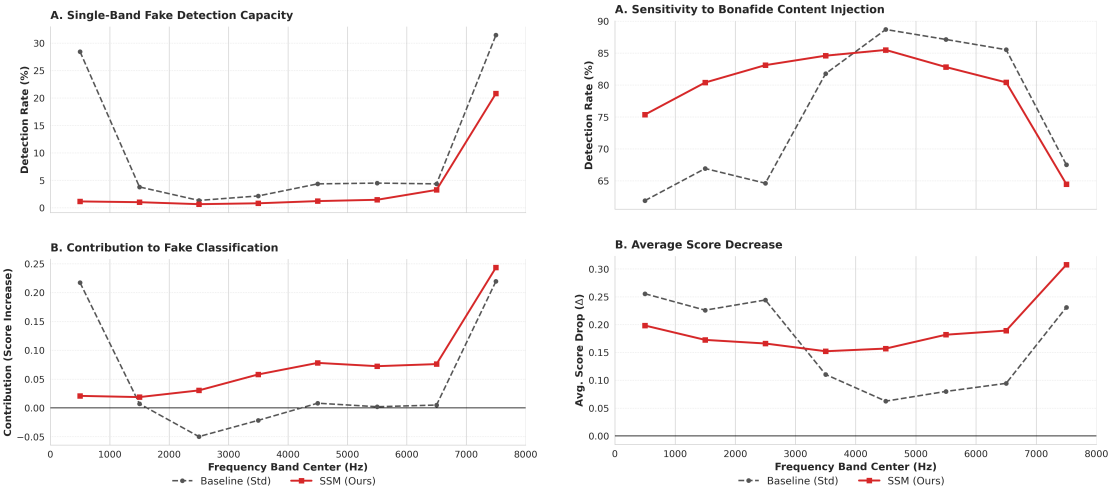
Band-wise Sensitivity Analysis
To study the link between spectral attention and detection performance, we applied the *adversarial frequency swapping* protocol defined in Section 3. We evaluated the three architectures using the *Sufficiency (Fake Injection)* and *Vulnerability (Real Injection)* modes. The spectrum was segmented into 8 bands of 1000 Hz each, utilizing a subset of 5,000 samples per band to ensure statistical significance.
Regarding the LCNN model (Figure 5), the results confirm the diagnosis from Phase 1. Both the Baseline and the SSM-trained models mirror the "U-Shape" profile observed in the RCQ analysis. However, a crucial difference emerges: the *SSM model (Red)* exhibits a more balanced profile. It effectively attenuates the over-reliance on the extreme low and high bands while slightly improving sensitivity in the mid-band. Notably, in the Vulnerability analysis (Fig. 5b), injecting real bands into a spoofed audio occasionally causes the "Fake" score to increase rather than decrease. This suggests that the model is detecting the spectral discontinuity introduced by the swapping process, penalizing the sample.



(a) Sufficiency Analysis (LCNN). (b) Vulnerability Analysis (LCNN).

Figure 5: Band-wise Sensitivity Analysis for LCNN. **(a)** Sufficiency: The Baseline (Black) relies heavily on extremes. **(b)** Vulnerability: Note the "Frankenstein Effect" where real injection increases the fake score, indicating a check for spectral coherence.

The Standard ResNet (Figure 6) exhibits a distinct behavior compared to LCNN. The SSM intervention (Red) appears to reduce sensitivity in the low-frequency range (< 2 kHz). However, contrary to the LCNN case, the high-frequency bands (> 6 kHz) remain the dominant factor for classification. This suggests that for this specific architecture, SSM preserves the model's dependency on high-frequency artifacts.

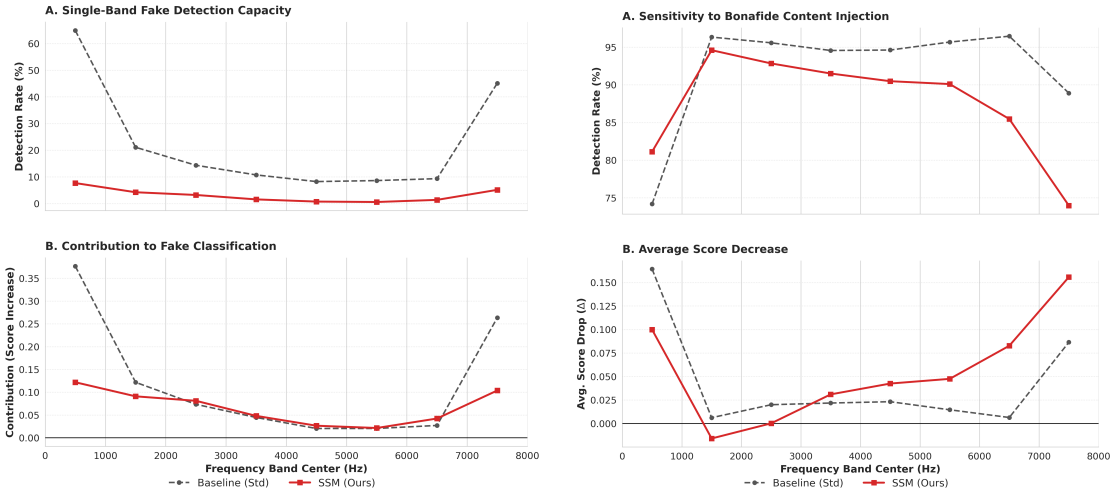


(a) Sufficiency Analysis (ResNet-18 Std). (b) Vulnerability Analysis (ResNet-18 Std).

Figure 6: Sensitivity Analysis for ResNet-18 Standard. The SSM intervention redistributes attention, dampening low-frequency reliance while maintaining high-frequency sensitivity.

The analysis of the ResNet-18 No-Stride (Figure 7) shows another response to the intervention. In the Sufficiency analysis (Fig. 7a), the SSM model retains the sensitivity in the mid-frequencies at the same level of the baseline, but significantly dampens the sensitivity at the spectral extremes (high and low bands). This results in a more distributed attention profile. In the Vulnerability analysis, it almost mirrors the Standard ResNet behavior by increasing sensitivity in the mid-high

range, showing that the SSM training activates the 3-6 kHz range.



(a) Sufficiency Analysis (ResNet-18 NS). (b) Vulnerability Analysis (ResNet-18 NS).

Figure 7: Sensitivity Analysis for ResNet-18 No-Stride. The combined effect of architecture and SSM creates a more balanced attention profile, reducing over-reliance on spectral edges.

This comparative analysis shows that the impact of SSM is not uniform. The behavioral change induced by frequency mixing is architecture-dependent. While LCNN tends to balance the entire spectrum, the ResNet models show a persistent sensitivity to high-frequency cues, with SSM primarily acting on the low-frequency bands. This suggests that the architecture plays a critical role in modulating the effects of spectral regularization.

Real-World Scenario

In this last experiment, we study whether the outcomes of the previous analysis translate into practical behavior in response to real-world bandwidth degradation.

We simulate a *GSM telephony scenario* to replicate the acoustic conditions of a standard cellular call. Traditional telephone protocols reduce the audio signal bandwidth (approximately 300 to 3400 Hz) to optimize transmission capacity while keeping the speech signal intelligible. This scenario is critical for forensic analysis as it eliminates high-frequency spectral content (> 3400 Hz), which, as we have seen, is fundamental for the discrimination implemented by these models to detect synthetic speech.

To isolate the effect of the filter, we adopt a *survival analysis* approach: we select a subset of samples correctly classified by the models (TP for the spoof class, TN for the bonafide one), and we apply a standard Butterworth band-pass filter (300-3400 Hz) to these samples.

Table 3: Results of the GSM Bandwidth Stress test comparing Baseline and SSM-trained models across different architectures. The table reports Survival Rate (%) and Score Drift (ΔS) for both Spoof (TP) and Bonafide (TN) samples.

Architecture	**Variant**	**Spoof Resilience (TP)**		**Bonafide Resilience (TN)**	
		Survival Rate	**Drift (ΔS)**	**Survival Rate**	**Drift (ΔS)**
LCNN	*Baseline*	88.81%	4.62%	**86.23%**	19.40%
	SSM	**97.39%**	**-1.32%**	80.99%	**18.29%**
ResNet-18 Std	*Baseline*	**94.75%**	**-0.60%**	66.77%	27.96%
	SSM	82.96%	12.36%	**86.33%**	**17.94%**
ResNet-18 NS	*Baseline*	**99.21%**	**-2.60%**	50.92%	29.86%
	SSM	83.83%	8.74%	**98.29%**	**7.05%**

The results in Table 3 show a clear contrast in how the different architectures we used react to band-pass filtering. While LCNN + SSM improve the detection of spoofed samples, the most significant impact was seen on the ResNet architecture, where there was a marked improvement in the classification of bona fide samples. More specifically, the baseline ResNet model incorrectly classifies a large portion of genuine GSM samples as fake (survival rate $\approx 50 - 66\%$), suggesting considerable vulnerability to bandwidth reduction. On the contrary, the same model driven by the SSM restores reliability in this scenario, increasing the bona fide survival rate to approximately $86 - 98\%$.

We attribute the divergence of the different architectures to their internal feature extraction mechanisms: Max-Feature-Map (MFM) for LCNN versus residual/ReLU layers for ResNet.

The ResNet architecture (both Standard and No-Stride) shows a collapse on Bonafide samples (Survival $\approx 50\%$). ResNets, with their residual connections and ReLU activations, tend to preserve global statistics throughout the depth of the network. The baseline evidently learned a fragile heuristic: "High-Frequency Energy = Real." When the GSM filter removes this energy, the aggregated signal at the Global Average Pooling layer drops significantly, pushing the classification toward "Fake". The high Score Drift (+29.86%) confirms this massive internal shift. The SSM intervention corrects this by forcing the ResNet to find discriminative features in other spectral bands. On ResNet No-Stride, SSM improves Bonafide Survival to 98.29% and drastically reduces drift. This proves that SSM successfully prevents the network from overfitting to the presence of high-frequencies as proof of authenticity.

The LCNN Baseline is naturally more robust on Bonafide samples (86.23%) compared to ResNet. This is attributed to the MFM activation layers. Unlike ReLU, which passes all positive activations, MFM performs a competitive selection (keeping only the maximum between two feature maps). This strategy filters out non-dominant high-frequency noise layer-by-layer before it reaches the final Global Average Pooling. Therefore, for LCNN, the SSM training has a different effect: it does not fix a broken Bonafide detector, but rather improves the model's ability to spot mid-frequency artifacts in Spoof samples (Survival rising from 88.81% to 97.39%).

6. Conclusions and future developments

This thesis explores the internal mechanisms of deepfake speech detection, focusing on how state-of-the-art models depend on different frequency regions rather than solely on performance metrics. Our goal was to understand whether reliance on specific spectral bands limits generalization.

Using Relative Contribution Quantification (RCQ), we observed a consistent "U-shaped" attention pattern in models such as LCNN and ResNet: strong dependence on low-frequency harmonics and high-frequency artifacts, with limited use of mid-band information. This imbalance is partly driven by architectural design, not just data.

Through Stratified Spectral Mixing (SSM), we showed that this dependency can be reshaped. Forcing models to attend to mid frequencies improved generalization on external datasets, confirming

that mid-band underutilization constrained robustness. Finally, stress tests revealed architecture-specific failure modes: ResNet models over-rely on high-frequency energy, while LCNN exhibits greater stability, highlighting the critical role of architectural choices in determining robustness.
Although this exploratory study successfully diagnosed the frequency dependencies of models trained on a standard dataset, these findings open new questions regarding the universality of the U-Shape pattern and the applicability of our methods across different domains. Since our investigation focused on models trained on the ASVSpoof 2019 LA dataset, future works should replicate the RCQ diagnosis by training models on diverse datasets, such as InTheWild, characterized by real-world noise, or the recently released ASVSpoof 5.
Another promising direction for future work is extending this analysis to time-domain architectures. In this thesis, we focused on models operating on frequency-domain representations derived from the STFT. However, recent approaches increasingly rely on raw waveform architectures, such as RawNet2 and Wav2Vec 2.0, which learn feature extraction filters directly from the time-domain signal. An important open question is whether the observed imbalance in spectral attention persists independently of the input representation, or whether raw-waveform models develop fundamentally different attention patterns.

References

[1] 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.

[2] Luigi Attorresi, Davide Salvi, Clara Borrelli, Paolo Bestagini, and Stefano Tubaro. Combining Automatic Speaker Verification and Prosody Analysis for Synthetic Speech Detection, October 2022.

[3] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv*, abs/2111.09296, 2021.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[5] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction.

[6] ETSI. Digital cellular telecommunications system (phase 2+); full rate speech; transcoding (gsm 06.10 version 8.1.1 release 1999). Standard ETSI EN 300 961 V8.1.1, European Telecommunications Standards Institute, 2000.

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. 2(11):665–673.

[8] D. Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, April 1984.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal

graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

[11] Stamatis Karnouskos. Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society*, 1(3):138–147, September 2020.

[12] Gwantae Kim, David K. Han, and Hanseok Ko. SpecMix : A Mixed Sample Data Augmentation method for Training withTime-Frequency Domain Features, August 2021.

[13] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, June 2021.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

[15] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224, March 2017.

[16] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020.

[17] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[18] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. STC Antispoofing Systems for the ASVspoof2019 Challenge, April 2019.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[20] Tianchi Liu, Lin Zhang, Rohan Kumar Das, Yi Ma, Ruijie Tao, and Haizhou Li. How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio?, June 2024.

[21] Yuxin Luo, Ruoyi Zhang, Lu-Chuan Liu, Tianyu Li, and Hangyu Liu. Fcpe: A fast context-based pitch estimation model, 2025.

[22] Masanori MORISE, Fumiya YOKOMORI, and Kenji Ozawa. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D:1877–1884, July 2016.

[23] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does Audio Deepfake Detection Generalize?, August 2024.

[24] 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, June 2024.

[25] Viola Negroni, Luca Cuccovillo, Paolo Bestagini, Patrick Aichroth, and Stefano Tubaro. *Multi-Tast Transformer for Explainable Speech Deepfake Detection via Formant Modeling.* January 2026.

[26] Viola Negroni, Davide Salvi, Alessandro Ilic Mezza, Paolo Bestagini, and Stefano Tubaro. Attention-based mixture of experts for robust speech deepfake detection. *arXiv preprint arXiv:2509.17585*, 2025.

[27] Viola Negroni, Davide Salvi, Alessandro Ilic Mezza, Paolo Bestagini, and Stefano Tubaro. Leveraging mixture of experts for improved speech deepfake detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[28] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and Checkerboard Artifacts. *Distill*, 1(10):e3, October 2016.

[29] Simon Parkin. The rise of the deepfake and the threat to democracy. http://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

[30] Hanwei Qian, Lingling Xia, Ruihao Ge, Yiming Fan, Qun Wang, and Zhengjun Jing. From Black Boxes to Glass Boxes: Explainable AI for Trustworthy Deepfake Forensics. *Cryptography*, 9(4):61, December 2025.

[31] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE, 2018.

[32] 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, October 2019.

[33] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, Robust and Controllable Text to Speech. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] 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, September 2023.

[35] Davide Salvi, Viola Negroni, Sara Mandelli, Paolo Bestagini, and Stefano Tubaro. Phoneme-level analysis for person-of-interest speech deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1586–1595, 2025.

[36] Davide Salvi, Amit Kumar Singh Yadav, Kratika Bhagtani, Viola Negronil, Paolo Bestagini, and Edward J Delp. Comparative analysis of asr methods for speech deepfake detection. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pages 329–333. IEEE, 2024.

[37] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017.

[38] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, April 2018.

[39] 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, 2021.

[40] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus, October 2015.

[41] Imdat Solak and M.AI Labs. The m-ailabs speech dataset, 2019. Accessed: 2024-01-28.

[42] Jost Tobias Springenberg, Alexey Dosovitskiy, T. Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. *CoRR*, December 2014.

[43] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386, May 2022.

[44] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with RawNet2. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[45] 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.

[46] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2024.

[47] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Hector Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection, April 2019.

[48] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6(1):2056305120903408, 2020.

[49] Cristian Vaccari and Andrew Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1):2056305120903408, January 2020.

[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, September 2016.

[51] 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)*, pages 1–4, November 2013.

[52] Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147, August 1998.

[53] Xin Wang, Hector Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale, August 2024.

[54] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A Light CNN for Deep Face Representation With Noisy Labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, November 2018.

[55] Yang Xiao and Rohan Kumar Das. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 2025.

[56] Junichi Yamagishi, Christophe Veaux, Simon King, and Steve Renals. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1):1–5, 2012.

[57] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966, May 2013.

[58] Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6765–6773, 2024.

Abstract in lingua italiana

La rapida evoluzione dell'Intelligenza Artificial generativa ha democratizzato la produzione di parlato sintetico realistico, introducendo sfide importanti alla sicurezza biometrica. Nonostante gli attuali rilevatori di deepfake ottengano risultati elevati sui benchmark standard, operano come "black boxes" opache, lasciando i loro processi decisionali interni largamente inesplorati. Questa tesi presenta uno studio esplorativo sulle dipendenze spettrali delle Reti Neurali Convoluzionali (LCNN e ResNet) applicati all'audio forense. Introduciamo un framework diagnostico basato su un adattamento della Relative Contribution Quantification (RCQ) per generare dei profili di attenzione globale nel dominio della frequenza. La nostra analisi conferma un caratteristico profilo di attenzione a "U", in cui i modelli privilegiano i componenti a bassa e ad alta frequenza, ponendo minore enfasi sulla gamma delle frequenze medie. Per indagare la plasticità di questi profili di attenzione, impieghiamo lo Stratified Spectral Mixing (SSM), una strategia di data augmentation utilizzata in questo contesto come strumento investigativo per rompere la coerenza spettrale verticale. I risultati sperimentali dimostrano che questo intervento spinge i modelli a riconfigurare il loro focus spettrale, promuovendo un'attenzione piú distribuita tra le diverse bande di frequenza. Questo cambiamento si accompagna ad una migliore generalizzazione su algortmi di attacco non visti e ad una maggiore robustezza in scenari a banda limitata (telefonia GSM). In definitiva, questo lavoro evidenzia come, modificando attivamente il focus spettrale dei modelli, sia possibile studiare come cambia il comportamento di rilevamento.

Parole chiave: Rilevamento Deepfake, Intelligenza Artificiale Spiegabile, Audio Forense, Analisi Spettrale