POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Music And Acoustic Engineering

Author: Manfredi Pletti

Advisor: Prof. Paolo Bestagini

Co-advisor: Viola Negroni

Academic year: 2024-2025

1. Introduction

In recent years, the rapid evolution of generative artificial intelligence has democratized the creation of highly realistic synthetic speech, more commonly known as deepfake audio. Technologies such as Text-To-Speech and Voice Conversion have transitioned from generating robotic voices to creating speech that is almost indistinguishable from human speech. While these technological advances offer notable benefits in areas such as accessibility and entertainment, they also pose critical security risks, making the human voice a vulnerable attack vector. In particular, malicious actors can rely on these generative tools to bypass biometric authentication systems or to fabricate fake news in order to manipulate public opinion, thereby undermining people's trust in digital media [1].

To counter these threats, deepfake speech detectors have been developed, mainly based on Deep Neural Networks. However, despite their remarkable performance on standard benchmarks, current systems operate primarily as opaque "black boxes". While recent research has started addressing the issue through Explainable AI (XAI), for example by designing architectures that link detection to specific phonetic features [4], or by demonstrating that models tend to concentrate their attention on extremely low and high frequency bands [5], the internal decision-making process of standard architectures remains largely unexplored. This opacity raises concerns about whether these models rely on authentic forensic traces or simply overfit specific artifacts in the datasets on which they were trained, compromising their reliability in real-world scenarios.

The main objective of this thesis is to systematically investigate the spectral dependencies of Convolutional Neural Networks (LCNN and ResNet) applied to fake speech detection. Moving beyond standard performance metrics, we aim to obtain information on the relationships between a model's spectral attention and its cross-domain generalization abilities, ultimately proposing an intervention strategy to study how generalization changes as spectral attention changes.

2. Proposed Method

To investigate the spectral dependencies of deepfake speech detectors, we designed a three-step diagnostic framework. Rather than focusing exclusively on maximizing accuracy metrics, this

methodology is specifically structured to observe and quantify how the internal behavior of the model changes depending on the spectral characteristics of the training data.

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 $\mathcal{D}$.

Quantification & Validation First, to understand which spectral regions drive the model's decisions, we adapt the Relative Contribution Quantification (RCQ) framework [3]. We employ Guided Grad-CAM [6] to generate high-resolution local explanations, combining detailed pixel-level gradients with class-discriminative heatmaps. The activation map M_{attr} is computed as:

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

To provide contextual explanations, we incorporate Semantic Segmentation (M_{sem}). We use Fast Context-based Pitch Estimation (Torch-FCPE) to track the fundamental frequency (f_0). Based on this, we generate binary masks to 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.
To aggregate these local explanations globally across the test set, we compute the RCQ_f for each frequency bin f and semantic segment s, which represents the percentage deviation from the global mean importance (μ_{global}):

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

To validate whether these attention profiles actually reflect a causal relationship, we use two types of stress tests. The first is **Frequency Swapping**, an adversarial intervention where specific frequency bands are injected into a target spectrogram of the opposite class to measure Sufficiency (Fake Injection) and Vulnerability (Real Injection). The second is the Bandwidth Stress Test, where we apply a band-pass filter equivalent to that applied in GSM technology (300 Hz-3400 Hz) to the samples, simulating a real-world scenario where the signal is degraded, to observe model survival rates when extreme frequencies are removed.

Intervention To study the plasticity of these models, we introduce Stratified Spectral Mixing (SSM) as a data augmentation tool for investigative purposes. Existing augmentation methods like SpecMix [2] 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 this logic by generating binary masks M composed exclusively of horizontal frequency bands. By generating masks M composed of horizontal frequency bands and stochastically mixing segments of different classes, we force the model to process conflicting and fragmented spectral data.
For a target sample x_A and a source sample x_B, the mixed spectrogram $\tilde{x}$ is synthesized as:

$$\tilde{x} = M \odot x_A + (1 - M) \odot x_B \tag{3}$$

By forcing the network to process conflicting and fragmented spectral data (e.g., low frequencies from real samples mixed with high frequencies from deepfakes), we force the model to explore the full spectrum. Furthermore, since the mixed spectrogram contains information from potentially different classes, we train the model using Soft Cross-Entropy Loss with a weighted target label:

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \tag{4}$$

where λ represents the proportion of the frequency bins belonging to file A. This soft-labeling strategy teaches the model to quantify the degree of fakeness based on the proportion of spectral components, rather than making a strict binary decision.

3. Experimental Setup

To ensure a comprehensive evaluation of model behavior and cross-domain generalization, we have structured a standardized data processing pipeline where we convert audio waveforms into frequency-wise normalized Log-Power Spectrograms.

Data Processing Pipeline The raw audio waveform is subject to a series of transformations before being fed into the models:

- **Voice Activity Detection (VAD):** To eliminate non-informative silent segments, all audio files are processed using Silero VAD.
- **Duration Standardization:** We enforce a standard duration of $T = 3.0$ seconds (looping if necessary).
- **Feature Extraction:** The waveform is converted into a time-frequency representation using STFT to obtain a Log-Power Spectrogram.
- **Frequency-wise Normalization:** We apply instance-level normalization independently for each frequency bin, standardizing the energy distribution across frequency bands.

Datasets The training, selection, and baseline evaluation of the model were performed using the ASVspoof 2019 Logical Access (LA) dataset [7], the standard community benchmark for this task featuring TTS and VC attacks. To accurately verify the generalization capabilities and impact of SSM augmentation, we performed cross-dataset evaluations using four out-of-domain datasets:

- **In-The-Wild:** 38 hours of audio sourced from social media, featuring uncontrolled acoustic environments and lossy compression codecs.
- **FakeOrReal:** 195,000 utterances balancing variety in generative models, with loudness normalization applied.
- **MLAAD (English subset):** A massive corpus comprising synthetic speech generated by 82 different TTS models.
- **ASVSpoof 5 (Track 1):** The latest challenge iteration, from which a balanced test set of 7,760 files per class was extracted under clean conditions.

Model Architectures To understand whether the observed spectral dependencies depend on the model structure or on the training data, we trained and evaluated three different CNN architectures:

- **LCNN:** A lightweight architecture that uses Max-Feature-Map activation layers, which act as a competitive feature selector to isolate sparse artifacts from the background signal.
- **ResNet-18:** A widely used residual network, characterized by very aggressive spatial downsampling in its initial layers.
- **ResNet-18 No-Stride:** A variant of ResNet modified to empirically evaluate whether spatial downsampling threatens robustness. By removing the initial stride and max pooling operation, the network processes the input at its original resolution in the early stages, preserving fine-grained details.

Training Protocol & Loss Functions We adopted a standardized training protocol using the Adam optimizer. To ensure sample variability, we applied stochastic data augmentation techniques to the waveform, including RawBoost, Codec Compression, Reverberation, and Noise Injection. Then, depending on the model type, we employed two different loss functions:

- **Focal Loss:** Used for the baseline models to focus the learning process on hard to detect examples.
- **Soft Cross-Entropy Loss:** Used for the SSM-trained models to handle the continuous soft labels ($\tilde{y} \in [0, 1]$) generated by the SSM process.

4. Results

The series of experiments shows how architectural choices and spectral dependencies influence model behavior, and how these models adapt when their spectral attention is forced to change through SSM.

4.1. Explainability Analysis

Through global RCQ analysis, we mapped the distribution of attention across frequency bands. Analysis of the LCNN architecture shows a prominent "U-shape" (Figure 1). The model intrinsically focuses on the band extremes (below

1 KHz and near the Nyquist frequency), placing significantly less emphasis on mid-range frequencies, which often show negative relative importance.
Furthermore, when evaluating the standard ResNet-18 architecture, we observed a phenomenon specific to this architecture: a high-frequency sawtooth oscillation in the attention profile, which we attributed to spatial downsampling in the initial layers (Figure 2a). By removing this stride (ResNet-18 No-Stride), we effectively smoothed the profile while also preserving spectral details (Figure 2b).

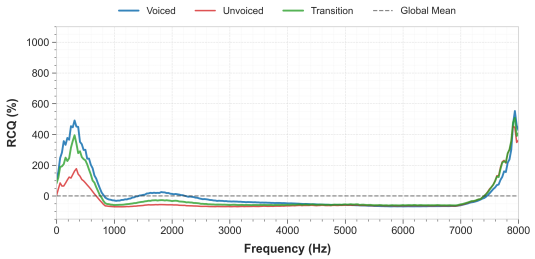
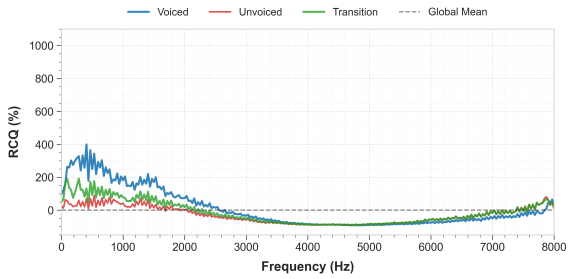
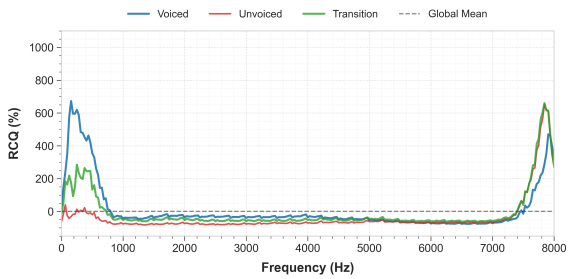


Figure 1: Aggregate RCQ Profile for the LCNN architecture, computed on True Positive samples from the ASVspoof 2019 LA evaluation set.



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



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

Figure 2: Comparison of RCQ profiles (True Positives) between Standard and No-Stride architectures.

4.2. SSM Intervention Study

To investigate the plasticity of these models, we calculated the RCQ profiles on the variants driven by SSM. Instead of treating SSM simply as a tool to increase performance, we used it to observe whether the models could reconfigure their spectral attention. As shown in Figure 3, disrupting vertical spectral coherence significantly flattened the U-shape of the LCNN profiles. Reliance on extreme bands decreased, and the profile generally moved toward the global mean, indicating more distributed attention across frequencies.

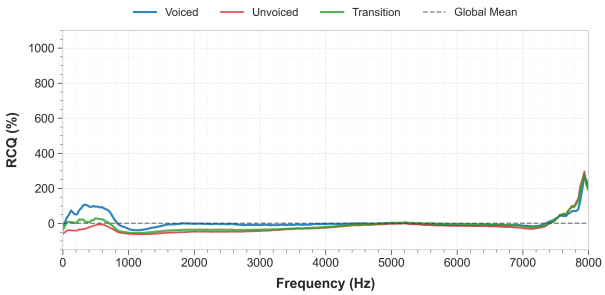


Figure 3: LCNN, True Positives computed on ASVspoof 2019 LA evaluation set, after SSM intervention. Note that this plot and the previous share the same Y-axis scale to facilitate a direct comparison.

After observing this shift in attention, we tested how generalization changed among the different datasets considered in this thesis. As shown in Table 1, models that were forced to learn from fragmented spectral information from SSM generally show an improvement in Equal Error Rate on Out-of-Domain datasets. In particular, ResNet-18 No-Stride shows significant changes when SSM is combined with spectral resolution preservation (e.g., dropping from 33.03% to 21.76% on the MLAAD dataset).

4.3. Robustness Evaluation

To verify whether the attention profiles reflect causality in the model's decisions, we performed a Band-wise Sensitivity Analysis via Frequency Swapping. As shown in Figure 4 for the LCNN architecture, the baseline model's decisions rely heavily on spectral extremes. The SSM-trained model (in red) shows a more balanced profile, effectively mitigating the over-reliance on band extremes by improving sensitivity in the mid-

Table 1: 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**

Table 2: 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 fake (TP) and real (TN) samples.

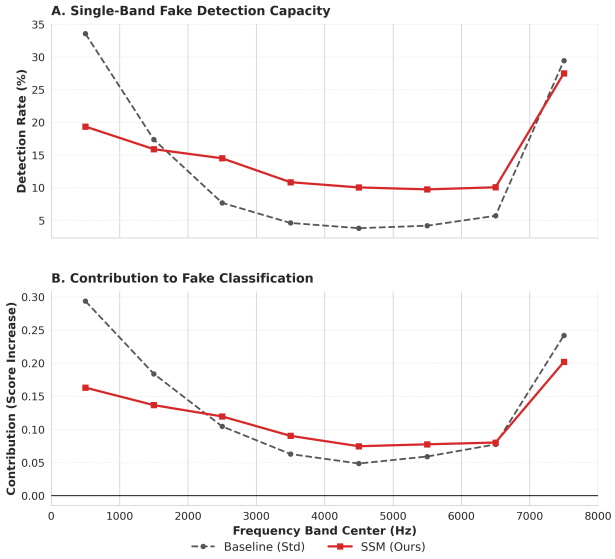
Architecture	Variant	Fake data resilience (TP)		Real data resilience (TN)	
		Survival Rate	Drift (ΔS)	Survival Rate	Drift (ΔS)
LCNN	*Baseline*	88.81%	4.62%	**86.23%**	19.40%
	SSM	**97.39%**	**-1.32%**	80.99%	**18.29%**
ResNet-18 Std	*Baseline*	**94.75%**	**-0.60%**	66.77%	27.96%
	SSM	82.96%	12.36%	**86.33%**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

band.

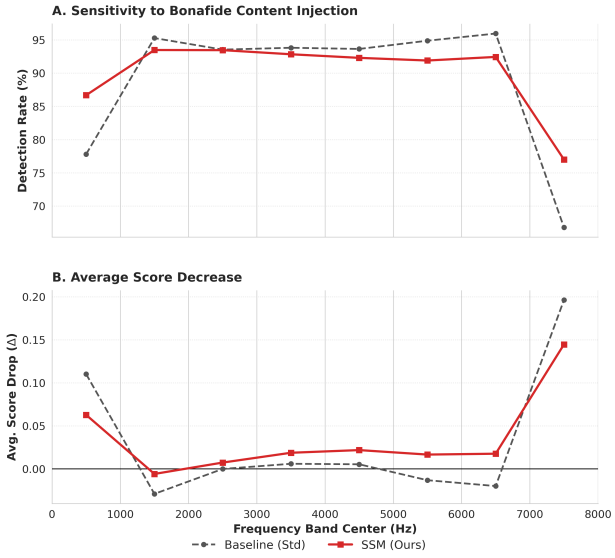
To contextualize these results, we studied how the models react when the frequency bands they rely on most are removed, simulating the degradation that obviously occurs during a GSM phone call (300-3400 Hz). Table 2 shows that the baseline ResNet collapses on genuine GSM samples (Survival Rate $\approx 50 - 66\%$). Clearly, the model has learned a strong dependence between the presence of high frequencies and authentic speech. In contrast, the SSM-trained versions maintain more stable behavior in this case (real data Survival Rate for ResNet-18 No-Stride rises to 98.29%). This demonstrates that actively modifying a model's spectral attention is a practical way to study and understand its behavior in a real-world scenario.

5. Conclusions

This thesis presented an exploratory study on the spectral dependencies of Convolutional Neural Networks applied to the detection of deepfake speech. By introducing a diagnostic framework based on Relative Contribution Quantification (RCQ), we observed a characteristic "U-shaped" attention profile, indicating a strong reliance on low- and high-frequency components. Through Stratified Spectral Mixing (SSM), we investigated the plasticity of these models, demonstrating that actively disrupting vertical spectral coherence encourages a reconfiguration of their spectral focus. This intervention allowed us to observe that a more distributed allocation of attention directly impacts cross-domain generalization and significantly alters the models' behavior under limited-bandwidth conditions, such as GSM telephony. Ultimately, these findings underscore the importance of shifting from opaque performance metrics to a transparent understanding of model behavior, paving the way for future investigations into time-domain architectures and diverse acoustic environments.



(a) Sufficiency Analysis (LCNN).



(b) Vulnerability Analysis (LCNN).

Figure 4: Band-wise Sensitivity Analysis for LCNN. The SSM intervention (Red) attenuates the over-reliance on the extreme low and high bands compared to the Baseline (Black), slightly improving sensitivity in the mid-band.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.