POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching

Laurea Magistrale in Music and Acoustic Engineering

Author: Emma Coletta

Advisor: Prof. Paolo Bestagini

Co-advisor: Davide Salvi, Viola Negroni

Academic year: 2023-2024

1. Introduction

In recent years, the rapid advancement of AI-driven generative systems has made the creation of high-quality synthetic content easier than ever. While the democratization of generative tools brings numerous benefits, it also introduces serious social risks. Malicious actors could exploit these technologies to produce deceptive synthetic media, enabling identity theft, financial fraud, and disinformation campaigns. Among these threats, speech deepfakes represent a particularly concerning issue in the audio domain. These are synthetically generated speech samples that mimic the voice of a target speaker, making them say arbitrary utterances. Speech deepfakes have already been used in fraud, blackmail, and other security-related threats, prompting the multimedia forensics community to actively develop deepfake detection systems to preserve the integrity of digital communications.

Diverse strategies have been proposed for speech deepfake detection, spanning from the analysis of low-level acoustic features to that of higher-level semantic aspects. These approaches leverage diverse frameworks, including traditional machine learning techniques, advanced DL architectures, and pre-trained models. Most existing detection methods follow a supervised learning strategy, where models are trained on labeled datasets containing both real and synthetic speech samples. However, this approach typically presents two critical challenges: generalization and interpretability. Generalization refers to the model's ability to detect deepfakes generated by speech synthesis methods not encountered during training. Interpretability concerns the "black box" nature of the detection systems, which typically provide little insight into the rationale underlying their predictions. These two limitations significantly reduce the practical effectiveness of deepfake detectors in real-world scenarios.

Recent research has explored novel strategies to address these limitations. One promising direction involves One-Class (OC) classification systems to improve generalization. These models are trained exclusively on authentic speech data, learning its distribution so that any sample deviating from it is classified as synthetic. On the other hand, efforts to enhance interpretability have focused on identifying the specific aspects of the audio signal that drive model predictions, improving the transparency of the detection pro-

cess.
In this paper, we propose a novel method that addresses these challenges through an interpretable, OC speech deepfake detector. Specifically, we reframe the detection problem as an Anomaly Detection (AD) task, training a OC model exclusively on real speech samples. This enables the model to learn the statistical distribution of genuine speech and classify out-of-distribution samples as fake, preventing it from being fooled by unseen synthesis methods. Our approach builds upon the Student-Teacher Feature Pyramid Matching (STFPM) framework proposed in [3], enhanced with Discrepancy Scaling (DS) [2], and is capable of detecting and localizing speech anomalies in both time and frequency domains. To evaluate our approach, we test it across diverse scenarios, comparing its performance against multiple baselines. Our results indicate its superiority and validate the effectiveness of framing speech deepfake detection as an AD problem.

2. Proposed Method

In this work, we address the problem of speech deepfake detection and propose an AD framework to tackle it. Our approach is based on the hypothesis that synthetic speech can be conceptualized as authentic speech embedded with anomalies, which arise from imperfections in the speech generative process. These imperfections manifest as subtle artifacts that an analyst can leverage to discriminate between synthetic and authentic signals. Traditional supervised learning systems explicitly learn the characteristics of these artifacts by training on labeled datasets containing both real and synthetic samples. Then, at inference time, they use their acquired knowledge to differentiate between the two classes. Although effective in controlled scenarios, this approach has significant limitations, particularly in its ability to generalize detection performance over speech generation methods that are unseen during training. Different generative models introduce distinct artifacts, and detectors trained on a limited set of synthesis techniques may fail to recognize deepfakes produced by other methods. To overcome this limitation, the proposed framework adopts a OC learning paradigm, considering only authentic speech data during training. The goal is to characterize the statistical distribution of genuine speech so that, during inference, any deviation from this distribution is flagged as an anomaly, indicating potential deepfake content. By focusing only on authentic speech, the model's performance becomes independent of any speech synthesis method, enhancing its generalization capabilities. Moreover, our AD framework enhances interpretability by generating anomaly maps that localize these deviations in both time and frequency domains, providing valuable forensic insights into the input speech signal.

2.1. Problem Formulation

The speech deepfake detection problem can be formally defined as follows. Let us consider a discrete-time speech signal $\mathbf{x}$ associated with a class $y \in \{0, 1\}$, where 0 denotes that the signal is authentic and 1 indicates that it has been synthetically generated. The goal of this task is to develop a detector $\mathcal{D}$ capable of estimating the class of the signal $\mathbf{x}$ as $\hat{y}$, where $\hat{y}$ is the likelihood of the signal $\mathbf{x}$ being fake.

2.2. Proposed System

We address the speech deepfake detection task by adopting a One-Class (OC) Anomaly Detection (AD) strategy. In this framework, the detector $\mathcal{D}$ is trained exclusively on authentic speech samples to learn their underlying statistical distribution. During inference, the model evaluates how much a given speech signal deviates from this learned distribution. Signals that closely align with the distribution of genuine speech are assigned a low $\hat{y}$ value, indicating authenticity. Conversely, signals exhibiting significant deviations are assigned a higher $\hat{y}$ value, suggesting the presence of deepfake content.
To implement the proposed framework, we adopt a Student-Teacher architecture inspired by the STFPM method introduced in [4]. As illustrated in Figure 1, the system consists of two identical CNN-based networks, i.e., the teacher and the student, which process acoustic features extracted from an input speech signal. The teacher network is pre-trained on the speaker identification task, enabling it to learn robust, speech-specific representations. The student network is then trained to replicate the teacher's activations through Feature Pyramid Matching

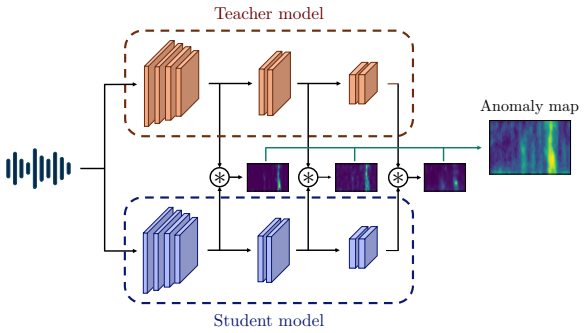


Figure 1: Pipeline of the proposed One-Class Anomaly Detection framework for speech deepfake detection.

(FPM) [3]. This training process encourages the student to mimic the teacher's feature representations. However, since the two models are trained exclusively on authentic data, the student learns to approximate the teacher's behavior only for real speech samples. During inference, when presented with synthetic speech, the student will struggle to replicate the teacher's feature maps, resulting in discrepancies between their activations. These discrepancies are quantified as anomaly scores, which we use to detect deepfakes.

To compute these scores, we leverage FPM, which compares the activations of the teacher and student networks at multiple designated layers. This multi-level comparison captures anomalies at different scales: lower layers encode fine-grained, low-level acoustic details, while higher layers capture broader, context-rich information. The discrepancy between the normalized activations of the teacher and student models serves as the anomaly indicator. Ideally, the anomaly values will be low when the system receives real speech, and high when the input is synthetic. Finally, these discrepancies can be visualized through an anomaly map, highlighting time-frequency regions where generation artifacts are most prominent. Regions with higher anomaly values indicate areas in the signal where synthetic speech deviates most significantly from real speech, providing an interpretable way to analyze deepfake detection.

To further enhance the performance of our AD model, we incorporate Discrepancy Scaling (DS) [2]. This technique standardizes the discrepancy values between the teacher and student networks during inference based on statistical properties of the test data. The normalization process involves computing the mean and standard deviation of the test distribution using a small calibration set. DS is applied layer by layer to both the teacher and student networks, amplifying subtle anomalies and enabling the detection of fine-grained synthetic artifacts. In our analysis, we use DS as a calibration strategy to enhance detection performances on out-of-distribution data, as shown in our experimental results in Section 4.

3. Experimental Setup

In this section, we describe the evaluation setup used in our experiments. We begin by detailing the architecture and the strategies used to train and evaluate the teacher and student networks. Then, we introduce the baseline systems used to validate our findings. Finally, we present the datasets we employed in our experimental campaign.

3.1. Training and Evaluation Strategies

The model that we considered for both teacher and student networks is a modified version of ResNet18, adapted to process 2D acoustic features. It incorporates a dropout layer for regularization, a fully connected layer with 256 units, and a ReLU activation function. We considered mel-spectrograms as the acoustic feature representation, as they closely mimic human auditory perception. This choice enhances the interpretability of the generated anomaly maps, making them more reflective of how anomalies might be perceived by listeners. The networks are trained and tested on audio segments of 4 s. We trained the teacher network on the speaker identification task using the LibriSpeech dataset. We utilize the *train-clean-360* subset, which contains speech from 921 speakers. This strategy is inspired by [3], where the backbone was pre-trained on the image classification task. In our implementation, we decided to consider a pre-training that better aligns with the speech domain. The network has been trained for 300 epochs with an early stopping of 15 epochs, considering a batch size of 64 samples and a learning rate of 10^{-4}. We assumed Cross Entropy as the loss function, with AdamW as optimizer and Cosine Annealing as learning rate scheduler.

The student network shares the same architecture of the teacher model, except for the final classification layer, which is removed since it is not needed for AD. It is trained to minimize the discrepancy between its feature maps and those of the teacher network at the three final convolutional layers. This is done using a feature-matching loss, which computes the Mean Squared Error (MSE) between the L2-normalized activations of the teacher and student models. During training, both networks process the same input and generate their respective feature representations. The weights of the teacher remain frozen, while those of the student are updated. The loss value is computed by averaging the squared differences across feature maps. All the other hyperparameters are identical to those used for training the teacher. At inference time, the discrepancy between the teacher and student activations is used to generate an anomaly map. The average value of this map serves as the anomaly score for binary classification.

To enhance the anomaly detection capabilities of the proposed framework, we integrate Discrepancy Scaling (DS). DS computes the mean and standard deviation of the discrepancies between the teacher and student activations on a small calibration dataset of real speech tracks. Then, during inference, these statistics are used to normalize the anomaly scores at each layer, improving the sensitivity to synthetic artifacts and enhancing the model's generalization capabilities on out-of-distribution data.

3.2. Baseline systems

To validate the effectiveness of our proposed framework, we compare its performance against two baselines. The first one is RawNet2, a widely used supervised learning model designed to differentiate between real and synthetic speech samples. The second baseline is the same ResNet18 model considered above, trained following the OC approach introduced in [1]. Unlike the original implementation, we train this model using a strict OC setup, excluding synthetic samples during training. This ensures a fair comparison with our proposed method, which also relies exclusively on real data. These baselines were selected to highlight two key aspects: the improved generalization capability of a OC model compared to a traditional supervised classifier, and the advantages of the proposed AD framework over an existing OC approach.

3.3. Datasets

To assess the generalization capabilities of the considered systems, we conducted experiments across multiple speech deepfake datasets, with all audio sampled at 16 kHz.

We performed two distinct types of analyses: single and multi-speaker experiments. In single-speaker experiments, we focus on a specific target voice for both training and testing the detectors. This scenario simulates a Person-of-Interest (POI) use case, where the goal is to protect a specific individual from deepfake attacks by leveraging their unique voice characteristics. For this analysis, we considered the voice of Linda Johnson from LJSpeech as real data source. The synthetic datasets used for evaluation include TIMIT-TTS, Purdue speech dataset, and MLAAD, focusing on their subsets which contain deepfake speech generated to mimic the target voice. In this scenario, the OC detectors were trained on a partition of LJSpeech excluded from the evaluation, while RawNet2 was trained on the same LJSpeech partition combined with a subset of TIMIT-TTS.

In multi-speaker experiments, we tested generalization across diverse voices using ASVspoof 2019, In-the-Wild, FakeOrReal, and the complete Purdue speech dataset. In this setting, models were trained on the *train* and *dev* partitions of ASVspoof 2019. For OC detectors, only authentic speech data was used during training, while RawNet2 was trained on both real and fake samples.

4. Results

In this section, we analyze and discuss the performance of our proposed framework for speech deepfake detection, using Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC), and Equal Error Rate (EER). Additionally, we assess the interpretability of our approach by examining the anomaly maps it produces.

Single-Speaker Scenario. As a first experiment, we evaluate the system in a single-speaker scenario, simulating a POI use case. We do this

Table 1: Single-speaker detection performance (%). Best values are bold, second-best are italic.

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

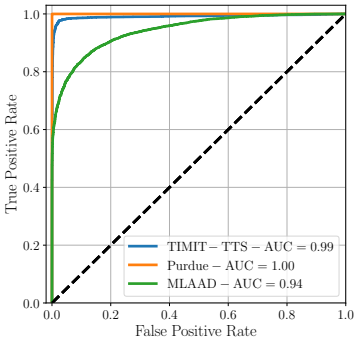


Figure 2: ROC curves of the proposed framework without DS evaluated in a single-speaker scenario.

as an initial analysis because we hypothesize that characterizing the voice of a single target speaker may be easier for a OC model, compared to modeling multiple voices simultaneously. Table 1 shows the complete results of this analysis. Our proposed framework, both with and without DS, outperforms the baselines across most datasets. On the TIMIT-TTS dataset, RawNet2 achieves the best results, which is expected given its supervised training on in-domain data. However, the model performance degrades significantly on the other datasets. On the other hand, the OC baseline shows reasonable performance on MLAAD, but performs the worst overall, demonstrating its limitations in a POI scenario. Our proposed framework without DS achieves the best performance, with perfect detection on Purdue and strong results on both TIMIT-TTS and MLAAD. The ROC curves and anomaly score distributions for this configuration are shown in Fig. 2 and Fig. 3, respectively. In this scenario, applying DS slightly reduces performance on some datasets but still maintains high detection accuracy.

Multi-Speaker Scenario. Given the strong performance in the single-speaker setting, we extend our evaluation to a multi-speaker scenario, where detecting deepfakes becomes more challenging due to the increased variability in

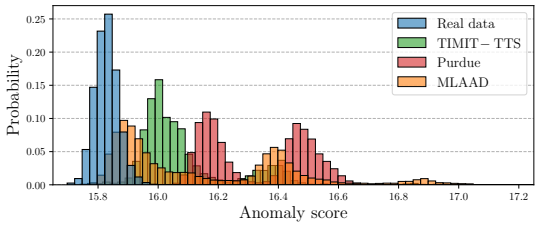


Figure 3: Anomaly score distributions of the proposed framework without DS evaluated in a single-speaker scenario.

Table 2: Multi-speaker detection performance (%). Best values are bold, second-best are italic.

	ASVspoof 2019		In-the-Wild		FakeOrReal		Purdue	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
RawNet2	**93.2**	**10.6**	63.9	40.9	82.1	25.1	56.3	44.9
OC baseline	91.9	*14.8*	79.0	28.9	**99.0**	**5.5**	*66.6*	*35.9*
Ours (w/o DS)	88.1	21.2	*95.1*	*11.3*	92.8	18.0	63.8	38.2
Ours (with DS)	*92.7*	15.6	**98.2**	**5.7**	*97.4*	*9.1*	**75.7**	**30.6**

real speech. Table 2 presents the results of this analysis. Again, RawNet2 performs best on its training dataset (ASVspoof 2019) but struggles significantly on unseen datasets, confirming the limitations of supervised learning methods. The OC baseline performs better than in the single-speaker case, particularly on FakeOrReal, indicating that characterizing multiple voices together can improve detection compared to focusing on a single target voice. The best-performing model is our proposed framework with DS, achieving the best detection results on all datasets except ASVspoof 2019. The ROC curves for this model are shown in Figure 4. These results demonstrate the effectiveness of DS in enhancing the model's ability to detect deepfakes, particularly in datasets where the real speech distribution differs significantly from the training data.

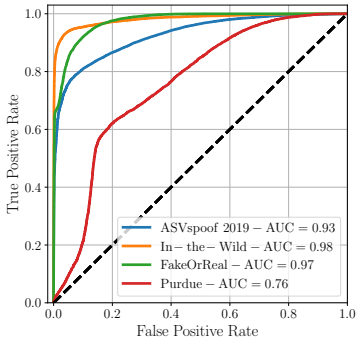
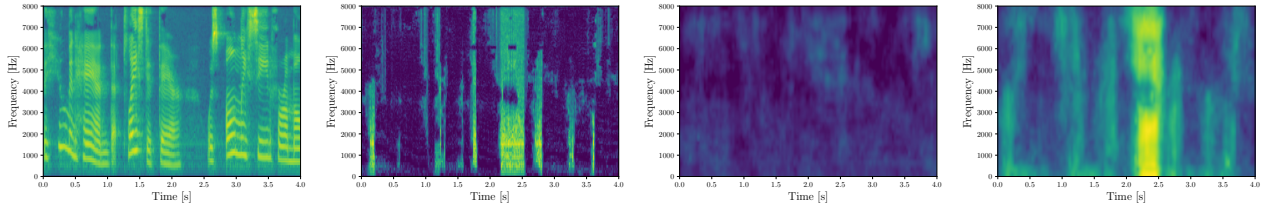


Figure 4: ROC curves of the proposed framework evaluated in a multi-speaker scenario.



(a) Real audio Mel-spec. (b) Ground truth (c) Real anomaly map (d) Fake anomaly map

Figure 5: Anomaly localization results for a real audio track and its fake counterpart.

Interpretability Analysis. As a final experiment, we assess the interpretability of our framework by analyzing the anomaly maps it produces. To do this, we take a real audio track from LJSpeech, compute its mel-spectrogram, and re-synthesize it using the Waveglow vocoder. Following a well-established approach, we consider this re-synthesized track as synthetic, even though it is derived from real speech features. This method ensures that the real and fake tracks are perfectly time-aligned, allowing us to compute their difference and generate a ground truth anomaly map. We then use our framework to estimate the predicted anomaly map. Figure 5 shows the results of this analysis, including the mel-spectrogram of the original audio, the ground truth anomaly map, and the anomaly maps produced by the system for both the real and fake tracks. The anomaly map for the authentic track is nearly empty, indicating no detected anomalies, while the map for the fake track closely matches the ground truth, highlighting the system's ability to localize synthetic artifacts.

5. Conclusions

In this work, we proposed a novel One-Class (OC) Anomaly Detection (AD) framework for speech deepfake detection, addressing the generalization and interpretability limitations of traditional supervised approaches. Our system, based on the STFPM architecture, is trained exclusively on authentic speech to characterize its statistical distribution and detect anomalies attributable to synthetic content. Experimental results show that our method outperforms both supervised classifiers (RawNet2) and an alternative OC model. By incorporating DS, we further enhance the system's generalization capabilities on out-of-distribution data. Additionally, the framework generates anomaly maps that localize artifacts in time and frequency, improving the interpretability of the results. Our findings validate the effectiveness of framing speech deepfake detection as an AD problem, offering superior generalization over traditional supervised methods. Future work will refine DS and explore applications to diverse real-world forensic scenarios.

References

[1] Hyun Myung Kim, Kangwook Jang, and Hoirin Kim. One-class learning with adaptive centroid shift for audio deepfake detection. *Interspeech*, 2024.

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

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

[4] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *The British Machine Vision Conference (BMVC)*, 2021.