

Robust Localization of Partially Fake Speech: Metrics and Out-of-Domain Evaluation

Hieu-Thi Luong^{*}, Inbal Rimon[†], Haim Permuter[†], Kong Aik Lee[‡] and Eng Siong Chng^{*}

^{*} Nanyang Technological University, Singapore

hieuthi.luong@ntu.edu.sg, aseschng@ntu.edu.sg

[†] Ben Gurion University, Be'er Sheva, Israel

inbalri@post.bgu.ac.il, haimp@bgu.ac.il

[‡] The Hong Kong Polytechnic University, Hong Kong SAR, China

kong-aik.lee@polyu.edu.hk

Abstract—Partial audio deepfake localization poses unique challenges and remain underexplored compared to full-utterance spoofing detection. While recent methods report strong in-domain performance, their real-world utility remains unclear. In this analysis, we critically examine the limitations of current evaluation practices, particularly the widespread use of Equal Error Rate (EER), which often obscures generalization and deployment readiness. We propose reframing the localization task as a sequential anomaly detection problem and advocate for the use of threshold-dependent metrics such as accuracy, precision, recall, and F1-score, which better reflect real-world behavior. Specifically, we analyze the performance of the open-source Coarse-to-Fine Proposal Refinement Framework (CFPRF), which achieves a 20-ms EER of 7.61% on the in-domain PartialSpoof evaluation set, but 43.25% and 27.59% on the LlamaPartialSpoof and Half-Truth out-of-domain test sets. Interestingly, our reproduced version of the same model performs worse on in-domain data (9.84%) but better on the out-of-domain sets (41.72% and 14.98%, respectively). This highlights the risks of over-optimizing for in-domain EER, which can lead to models that perform poorly in real-world scenarios. It also suggests that while deep learning models can be effective on in-domain data, they generalize poorly to out-of-domain scenarios, failing to detect novel synthetic samples and misclassifying unfamiliar bona fide audio. Finally, we observe that adding more bona fide or fully synthetic utterances to the training data often degrades performance, whereas adding partially fake utterances improves it.

I. INTRODUCTION

While full-utterance spoofing has been widely studied [1]–[3], partial audio deepfakes, where only segments of an utterance are manipulated, are more challenging due to their subtlety [4], [5]. These localized forgeries can evade traditional spoofing detection systems that rely on global audio characteristics, making them particularly threatening in scenarios such as voice-controlled authentication and disinformation generation [6]. The performance evaluation in partial deepfake scenarios remains inconsistent across the literature [5], [7]. Zhang et al. [4] used segment-based Equal Error Rate (EER)

[†]Dr. Hieu-Thi Luong is funded by RIE2025 NRF International Partnership Funding Initiative. This research is supported by the National Research Foundation, Singapore, under the AI Singapore Programme (AISG Award No.: AISG2-TC-2023-011-SGIL). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

as the primary metric, while Cai et al. [7] adopted Average Recall (AR) and Average Precision (AP). Although these metrics are threshold-independent, which simplifies evaluation, they can be difficult to interpret in the context of real-world deployment. Current localization models report outstanding results on in-domain datasets [6], [8]. For example, Cai et al. [9] achieved a 20-ms segment-based EER of 0.064% on the ADD 2023 Challenge Track 2 development set [5], and 1.74% on the PartialSpoof evaluation set [4]. However, cross-domain evaluations reveal a significant drop in performance, 15.2% EER on PartialSpoof and 45.34% on ADD, indicating that the models struggle to generalize beyond their training domain [10]. These results raise concerns about whether these systems are sufficiently robust and reliable for real-world use.

To create a shift in research towards more robust models [11]–[13], we recently released the LlamaPartialSpoof dataset [6], which contains high-quality fully and partially fake speech designed for out-of-domain evaluation. In this work, we discuss the shortcomings of the current evaluation methodology [14], and propose a simpler framing that better assesses system robustness that helps bridge the gap between research and deployment. The contributions of this paper are as follows: (1) Metric Redefinition: We argue that the metrics used in recent research are problematic and can be misleading, particularly on out-of-domain data. To address this, we propose reframing the localization task as a standard anomaly detection problem and advocate the use of conventional metrics such as accuracy, precision, recall, and F1; (2) Generalization Limitations of SSL Models: We show that current self-supervised learning (SSL)-based models fail to generalize to out-of-domain data. These models struggle not only to detect novel synthetic speech but also to misclassify out-of-domain bona fide audio; (3) Impact of Training Data: Finally, we show that increasing the volume of training data does not necessarily improve performance and, in some cases, may even degrade it.

II. PARTIALLY FAKE SPEECH LOCALIZATION

A. Problem Definition

Given an input audio of N frames, each of duration d seconds, the task of partially fake localization aims to produce

TABLE I: Partially Fake Speech Datasets for In-Domain and Out-of-Domain Training and Evaluation

Dataset	Language	Number of Utterance		
		Bonafide	Full Fake	Partial Fake
PartialSpoof	English			
- train		2580	0	22,800
- dev		2548	0	22,296
- eval		7,355	0	63,882
LlamaPartialSpoof	English	10,573	33,461	32,194
- (subset) train		2,391	13,991	13,498
- (subset) test		8,182	19,470	18,696
Half-Truth	Chinese			
- test		0	0	9072

an output sequence $\tilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_N\}$ that best matches the ground truth labels $Y = \{y_1, y_2, \dots, y_N\}$. Each frame label can be either 0, denoting a bona fide segment, or 1, denoting a fake segment. Note that this labeling scheme is the inverse of that used in several previous studies [4], [8]. We adopt this formulation to better align the task with standard practices in anomaly detection [15] and to reduce the risk of misinterpreting the evaluation results. In this work, we label segments that contain any amount of fake audio as fake, instead of requiring a majority.

B. Datasets and Generalization

Compared to audio deepfake detection, the task of localizing small fake segments within longer utterances is less explored, partly due to the scarcity of publicly available datasets. This is largely because creating such datasets requires greater expertise. We trained our systems on the train set of PartialSpoof [4] and used its eval set for in-domain evaluation. The main focus of this work is to investigate whether localization systems can generalize to unseen scenarios, which is a key indicator of their readiness for real-world deployment. To assess out-of-domain generalization, we additionally evaluated our models on two additional datasets: LlamaPartialSpoof [6] and Half-Truth [16]. Table I shows the number of utterances included in each dataset. Notably, the LlamaPartialSpoof dataset contains bona fide samples, as well as both fully fake and partially fake utterances, whereas the test set of the Half-Truth dataset includes only partially fake utterances.

C. Fake Speech Localization Models

In this paper, we use two SSL-based fake speech localization systems to examine various aspects of evaluating and improving robustness in out-of-domain scenarios. Specifically, we utilize the Multi-resolution Model (MRM)¹, introduced alongside the PartialSpoof dataset [4], and the Coarse-to-Fine Proposal Refinement Framework (CFPRF)² [8]. MRM processes input at multiple temporal resolutions, ranging from 20 ms to 640 ms, using a hierarchical architecture that enables the model to learn temporal patterns at varying scales. CFPRF employs a two-stage architecture: a 20-ms Frame-level Detection Network (FDN) first identifies coarse candidate regions based on inconsistency cues between real and fake frames,

¹<https://github.com/hieuthi/MultiResoModel-Simple>

²<https://github.com/ItzJuny/CFPRF>

TABLE II: Evaluation results of the CFPRF model using genuine and fake scores calculated on the PartialSpoof evaluation set at the 20-ms resolution.

Score	EER	Threshold at EER			
		Acc.	Pre.	Rec.	F1
genuine score (reported)	7.41	92.58	95.23	92.59	93.89
1 – genuine score (ours)	7.61	92.39	88.77	92.39	90.54

followed by a Proposal Refinement Network (PRN) that fine-tunes boundary predictions. CFPRF has reported state-of-the-art results on in-domain localization tasks. These models were trained using the default setting of their respectable setups.

D. Evaluation Metrics

Current approaches to deepfake audio detection typically follow the anti-spoofing evaluation methodology, relying on metrics such as EER or tandem Detection Cost Function (t-DCF) [17] to quickly assess model performance. However, these metrics postpone the selection of a decision threshold and reduce the interpretability of results, making it more difficult to evaluate system performance in practical scenarios. While closely related, fake speech detection is a broader and simpler task compared to speaker verification spoofing, as it does not necessarily involve the imitation of a specific target speaker. In the case of partially fake speech, the problem can be naturally framed as a sequential anomaly detection task. This allows for the use of straightforward and interpretable metrics such as accuracy, precision, recall, and F1-score.

Table II shows the evaluation results of the CFPRF model released by the authors [8]. The first row reports results based on the model’s output score (interpreted as a genuine score), while the second row uses 1 – score (interpreted as a fake score). In these two scenarios, the meaning of EER and accuracy remains the same, while precision, recall, and F1 differ depending on which class is defined as the positive class. This accounts for the significant differences observed in the precision results, beyond small mismatches that may arise from differences in segment labeling. This highlights how the current evaluation framing can misrepresent model performance and lead to confusion. For example, readers might expect the fake speech localization system to achieve 95.23% precision, when its actual precision is only 88.77%.

Zhang et al. [18] advocated for the use of Range-based EER, which is essentially a segment-based EER computed over very short segments. This metric enables finer-grained and fairer comparisons between different setups. However, in this work, we focus on the robustness of the model rather than its finer-grained precision, and therefore leave such evaluations for future work. Throughout the rest of this paper, all evaluation results are calculated with the fake segment defined as the positive class, in alignment with standard anomaly detection practices. To support standardized evaluation, we have open-sourced our library of partial spoofing metrics³.

³<https://github.com/hieuthi/partialspoof-metrics>

TABLE III: EERs at different resolutions.

(a) PartialSpoof (in-domain)							
Model	Evaluation resolution (s)						Utt.
	0.02	0.04	0.08	0.16	0.32	0.64	
MRM	13.72	14.46	15.29	11.60	9.63	7.24	1.48
CFPRF	7.61	7.36	6.84	6.04	5.24	4.80	1.72
reCFPRF	9.84	9.47	8.73	7.42	5.89	4.88	1.65

(b) LlamaPartialSpoof (out-of-domain)							
Model	Evaluation resolution (s)						Utt.
	0.02	0.04	0.08	0.16	0.32	0.64	
MRM	46.29	45.93	45.13	43.43	40.79	37.04	24.50
CFPRF	43.25	42.85	42.15	40.95	38.95	36.04	31.50
reCFPRF	41.72	41.37	40.77	39.77	37.78	34.10	24.81

(c) Half-Truth (out-of-domain)							
Model	Evaluation resolution (s)						Utt.
	0.02	0.04	0.08	0.16	0.32	0.64	
MRM	46.48	46.29	45.50	43.39	41.83	42.93	-
CFPRF	27.59	27.98	28.41	29.21	31.03	33.31	-
reCFPRF	14.98	14.86	14.65	14.62	14.97	16.18	-

III. EXPERIMENTS

A. Detecting fake segment at different resolutions

In the first experiment, we explored the trade-off between precisely localizing fake segments and simply detecting their presence. In other words, we evaluated the performance across different segment durations. Detecting shorter segments is more challenging and requires a more precise system; however, in real-world applications, such precision may be less valuable than overall robustness. We selected the MRM model for our experiments because it produces predictions at multiple resolutions. Although the CFPRF model achieved a lower EER at 20-ms resolution [8], it does not natively output scores for longer segments. To obtain scores at coarser resolutions, we downsampled the CFPRF score sequence by a factor of n and aggregated adjacent segments using the maximum value, a simple heuristic that has been shown to be effective [18].

Table III shows the segment-based EER at various resolutions on several in-domain and out-of-domain evaluation sets. It includes the MRM model, the original CFPRF model provided by the authors, and our reproduced version (reCFPRF), all trained on the training set of the PartialSpoof dataset. As expected, performance generally improves with larger segment sizes, suggesting that increasing segment length is a valid trade-off for enhancing robustness. Interestingly, our reproduced model, reCFPRF, achieved higher EER on in-domain data but lower EER on both out-of-domain sets⁴ compared to the original CFPRF. This suggests that the original model may have been overfitted to its training data.

B. Threshold, Recall, Precision, and F1

The problem with using EER to evaluate performance across multiple datasets is that each result corresponds to different decision thresholds. In other words, models that achieve strong EER scores across different evaluation sets may still perform poorly in deployment which requires a single threshold. Since

⁴Half-Truth does not have utterance-based EER since its test set does not contain bona fide samples

TABLE IV: Evaluation results at different resolutions at a specific threshold and recall values.

(a) PartialSpoof (in-domain)							
Model	Reso.	Threshold=0.5000			Recall=95%		
		Pre.	Rec.	F1	Pre.	Thres.	F1
MRM	0.02 s	96.73	70.28	81.41	53.95	0.0048	68.86
	0.04 s	97.21	71.36	82.30	51.65	0.0024	66.97
	0.08 s	97.25	74.57	84.41	46.15	-0.0232	62.13
	0.16 s	97.57	79.05	87.34	62.55	0.0000	75.55
	0.32 s	98.06	84.09	90.54	80.53	0.0116	87.19
	0.64 s	98.74	89.23	93.75	90.85	0.0172	92.89
Utt.	99.19	99.52	99.35	99.96	0.9264	97.43	
CFPRF	0.02 s	94.93	86.90	90.74	81.35	0.0212	87.65
	0.04 s	95.07	87.76	91.27	83.66	0.0304	88.97
	0.08 s	95.29	89.37	92.24	87.44	0.0556	91.07
	0.16 s	95.75	91.76	93.71	92.30	0.1536	93.63
	0.32 s	96.58	94.29	95.42	96.14	0.4148	95.57
	0.64 s	97.78	96.33	97.05	98.18	0.6396	96.57
Utt.	99.45	99.42	99.44	99.89	0.9488	97.40	
reCFPRF	0.02 s	96.43	81.60	88.40	64.28	-0.0192	76.69
	0.04 s	96.51	82.80	89.13	67.82	-0.0120	79.15
	0.08 s	96.60	84.92	90.38	74.53	0.0004	83.53
	0.16 s	96.86	88.05	92.24	85.84	0.0396	90.19
	0.32 s	97.38	91.60	94.40	94.69	0.2340	94.85
	0.64 s	98.24	94.71	96.45	98.16	0.4752	96.55
Utt.	99.58	99.17	99.37	99.89	0.8160	97.39	

(b) LlamaPartialSpoof (out-of-domain)							
Model	Reso.	Threshold=0.5000			Recall=90%		
		Pre.	Rec.	F1	Pre.	Thres.	F1
MRM	0.02 s	71.76	44.42	54.87	67.98	0.0072	77.66
	0.04 s	72.71	43.52	54.45	68.46	0.0040	77.89
	0.08 s	74.31	44.84	55.93	69.46	-0.0132	78.49
	0.16 s	76.98	48.24	59.31	71.48	-0.0012	79.88
	0.32 s	80.54	54.23	64.82	74.14	0.0048	81.34
	0.64 s	84.28	64.20	72.88	78.12	0.0116	83.70
Utt.	88.35	98.24	93.03	92.01	0.8652	91.00	
CFPRF	0.02 s	72.47	68.54	70.45	69.68	0.0092	78.59
	0.04 s	72.92	69.64	71.24	70.17	0.0144	78.86
	0.08 s	73.69	71.58	72.62	71.03	0.0204	79.45
	0.16 s	74.95	74.77	74.86	72.50	0.0304	80.31
	0.32 s	76.81	79.49	78.13	74.82	0.0544	81.71
	0.64 s	79.44	85.49	82.35	78.28	0.1896	83.73
Utt.	86.30	99.65	92.50	89.92	0.9852	90.00	
reCFPRF	0.02 s	75.75	50.68	60.73	69.63	-0.0348	78.56
	0.04 s	76.33	52.31	62.08	70.14	-0.0288	78.86
	0.08 s	77.30	55.00	64.27	71.06	-0.0216	79.44
	0.16 s	78.68	59.51	67.77	72.66	-0.0124	80.41
	0.32 s	80.45	66.56	72.85	75.25	0.0012	81.99
	0.64 s	82.59	75.92	79.11	78.93	0.0404	84.11
Utt.	86.51	98.67	92.19	91.08	0.9464	90.54	

(c) Half-Truth (out-of-domain)							
Model	Reso.	Threshold=0.5000			Recall=90%		
		Pre.	Rec.	F1	Pre.	Thres.	F1
MRM	0.02 s	18.73	96.67	31.38	19.38	0.9664	31.90
	0.04 s	19.22	96.14	32.04	19.79	0.9440	32.44
	0.08 s	19.96	96.40	33.07	20.62	0.9644	33.56
	0.16 s	21.52	97.28	35.24	22.42	0.9604	35.91
	0.32 s	24.56	98.68	39.33	25.69	0.9720	39.98
	0.64 s	29.83	99.76	45.93	31.04	0.9656	46.17
CFPRF	0.02 s	24.55	99.28	39.36	32.51	0.9444	47.78
	0.04 s	24.83	99.40	39.74	33.27	0.9536	48.60
	0.08 s	25.28	99.57	40.32	34.58	0.9648	49.96
	0.16 s	26.41	99.72	41.76	36.48	0.9756	51.93
	0.32 s	28.83	99.92	44.75	39.54	0.9852	54.98
	0.64 s	33.33	100.00	50.00	44.23	0.9928	59.40
reCFPRF	0.02 s	35.27	95.57	51.52	48.87	0.9784	63.34
	0.04 s	35.37	95.89	51.68	50.42	1.0012	64.63
	0.08 s	35.59	96.52	52.00	52.98	1.0296	66.70
	0.16 s	36.48	97.52	53.10	56.23	1.0536	69.23
	0.32 s	38.36	98.67	55.25	60.81	1.0744	72.58
	0.64 s	41.54	99.32	58.58	65.43	1.0904	75.82

all of our models output score value approximately between 0 and 1 (uncapped, as logits were used), we can use a neutral threshold of 0.5 for evaluation. Table III reports the Precision,

TABLE V: Effects of waveform augmentation on generalization ability of the localization model at the 20-ms resolution

(a) PartialSpoof (in-domain)						
Model	EER	Threshold at EER				
		Thres.	Acc.	Pre.	Rec.	F1
MRM	13.72	0.0124	86.24	80.20	86.46	83.21
MRM+rb	15.27	0.0924	84.73	78.34	84.70	81.40
MRM+ms	12.86	0.0084	87.11	81.39	87.26	84.22
CFPRF (*)	7.61	0.0880	92.39	88.77	92.39	90.54
reCFPRF	9.84	0.0504	90.15	85.62	90.16	87.83
reCFPRF+rb	9.52	0.1892	90.48	86.09	90.48	88.23
reCFPRF+ms	11.92	0.0800	88.07	82.76	88.11	85.35

(b) LlamaPartialSpoof (out-of-domain)						
Model	EER	Threshold at EER				
		Thres.	Acc.	Pre.	Rec.	F1
MRM	46.29	0.0548	53.71	71.36	53.71	61.29
MRM+rb	47.88	0.0764	52.12	70.04	52.11	59.76
MRM+ms	44.70	0.0196	55.33	72.64	55.40	62.86
CFPRF (*)	43.25	0.8828	56.75	73.81	56.76	64.17
reCFPRF	41.72	0.2444	58.28	75.00	58.28	65.59
reCFPRF+rb	42.59	0.0844	56.41	73.55	56.39	63.84
reCFPRF+ms	42.23	0.1052	57.77	74.60	57.78	65.12

(c) Half-Truth (out-of-domain)						
Model	EER	Threshold at EER				
		Thres.	Acc.	Pre.	Rec.	F1
MRM	46.48	0.9908	53.44	20.65	53.64	29.82
MRM+rb	38.40	0.9728	61.57	26.61	61.65	37.17
MRM+ms	38.51	0.9956	62.32	26.79	60.19	37.07
CFPRF (*)	27.59	0.9704	72.34	37.19	72.52	49.16
reCFPRF	14.98	1.0420	84.97	56.10	85.12	67.63
reCFPRF+rb	13.44	0.9128	86.56	59.29	86.55	70.37
reCFPRF+ms	12.00	0.8280	88.01	62.42	87.99	73.03

(*) Since it is a training from scratch experiments, the pretrained CFPRF did not include waveform augmentation results.

Recall, and F1 of the baseline models at different resolutions. Interestingly, while the PartialSpoof and LlamaPartialSpoof datasets exhibit higher precision than recall, the Half-Truth test set shows the opposite trend.

Moreover, since deepfake detection typically operates under the assumption that a false alarm is far less costly than a missed detection [19], we evaluate models at a high recall level, 95% for in-domain and 90% for out-of-domain data, and compare their precision at these points. For PartialSpoof evaluated at 20-ms resolution and a 95% recall level, MRM achieves 53.95% precision, while CFPRF and reCFPRF reach 81.35% and 64.28%, respectively. However, on LlamaPartialSpoof at 90% recall, the differences in precision among the models are smaller. On the Half-Truth dataset, reCFPRF performs better than CFPRF under the same recall condition. It is also important to note that the thresholds required to achieve a specific recall value vary across datasets. In real-world applications, one may use a neutral threshold (e.g., 0.5) or select a threshold based on a specific dataset. However, in the latter case, an optimal threshold for in-domain evaluation may not generalize well to out-of-domain scenarios, and vice versa. We will discuss this issue further in the next section.

C. EER Threshold and Effect of Waveform Augmentation

Since both tested models take raw waveforms as input, we investigated whether waveform augmentation methods effective in detection tasks can also benefit localization systems.

TABLE VI: Evaluation results of the models tuned on Llama-PartialSpoof train subset. Gray cells indicate results that is better than the original checkpoints, while bold values indicates the best within the same system type.

(a) EER					
Model	PS	LPS (test)			HAD
		both	full	partial	
MRM	13.72	44.54	40.98	32.76	46.48
+lps-train-bonafide	17.04	48.40	48.25	34.34	30.73
+lps-train-full	8.22	46.55	43.49	46.10	54.13
+lps-train-partial	9.30	11.39	10.70	9.75	9.62
+lps-train-all	9.03	8.20	6.04	10.31	7.55
CFPRF	7.61	41.47	36.22	35.01	27.59
+lps-train-bonafide	12.14	47.29	45.98	35.77	9.53
+lps-train-full	7.83	42.66	41.01	47.40	44.77
+lps-train-partial	8.11	7.62	7.19	6.66	4.74
+lps-train-all	7.62	5.96	2.89	8.72	3.52

(b) Accuracy (Threshold=0.5)					
Model	PS	LPS (test)			HAD
		both	full	partial	
MRM	87.34	52.04	47.73	66.98	22.02
+lps-train-bonafide	83.96	38.69	31.07	65.45	51.52
+lps-train-full	92.34	65.06	71.25	38.96	18.46
+lps-train-partial	90.36	89.17	89.27	89.12	74.39
+lps-train-all	90.45	92.37	94.62	88.99	84.49
CFPRF	93.00	61.19	62.86	61.27	43.59
+lps-train-bonafide	90.29	41.26	34.63	66.00	90.69
+lps-train-full	93.03	64.98	71.09	38.68	18.60
+lps-train-partial	93.33	92.85	93.20	92.07	92.21
+lps-train-all	93.59	93.43	95.13	88.71	94.86

Specifically, we applied the third algorithm of RawBoost [20] and the band-shaped MaskedSpec [21] as augmentation methods, as these were reported by the authors to be the most effective in their respective works. Table V presents the EER and other evaluation results, calculated at the EER threshold, for models trained with additional waveform augmentation. Interestingly, MaskedSpec (+ms) enhances the performance of the MRM model across both in-domain and out-of-domain scenarios, while RawBoost (+rb) shows inconsistent effectiveness. For CFPRF, the effects of augmentation are inconsistent across datasets. Overall, we conclude that while waveform augmentation can enhance in-domain performance, its effectiveness in out-of-domain settings is not guaranteed.

Figure 1 shows the score distributions of the three baseline models along with their EER thresholds. Among them, CFPRF stands out, as the large gap between its in-domain and out-of-domain EER thresholds suggests that it is highly overfitted to in-domain data. In practice, using the in-domain EER threshold in a production setting may result in suboptimal performance in real-world scenarios. Moreover, we observe that all models misclassified both unseen bona fide and synthetic samples. These findings highlight a key limitation of machine learning-based approaches to fake speech detection: their strong dependence on the quality and diversity of the training data.

D. Tuning with emerging data

Previous sections have shown that current SSL-based localization methods perform well on in-domain data but generalize poorly to out-of-domain scenarios. While it is often assumed that training on large-scale and diverse datasets can resolve

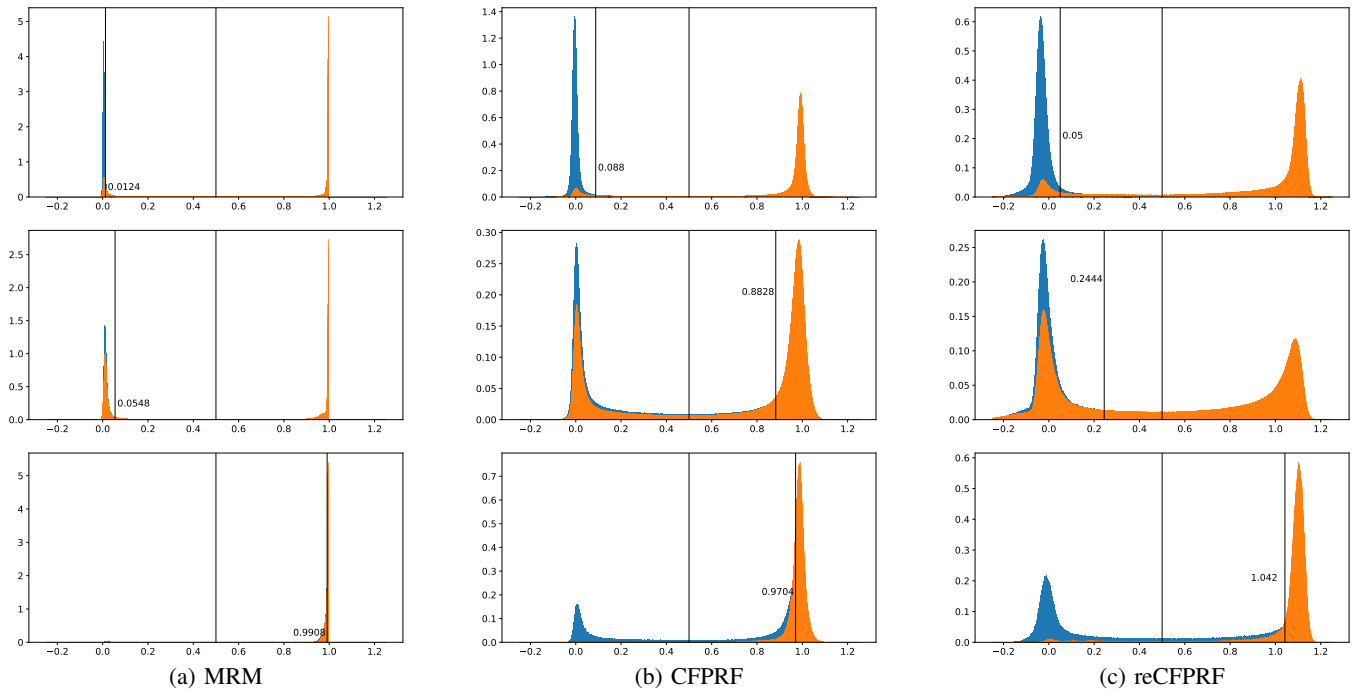


Fig. 1: Distributions of 20-ms segment score produced by the detections models. Blues are bona fide segments while oranges are fakes. Higher scores mean they are more likely to be fake. Upper row is PartialSpooF (in-domain), middle row is LlamaPartialSpooF (out-of-domain) and lower row is Half-Truth (out-of-domain). Vertical lines indicate either 0.5 threshold or EER threshold.

generalization issues, we test this hypothesis by simulating a scenario in which new data becomes available for tuning. Specifically, we selected approximately 20 speakers from the LlamaPartialSpooF (LPS) dataset as the new train subset for fine-tuning, with the remaining speakers used as the new test set, as shown in Table I. Starting from the checkpoints of the two baseline systems trained on PartialSpooF (PS), we fine-tuned the models using a combination of the PS train data and either part or all of the LPS train subset.

Table VI presents the EER and the accuracy of the baseline and fine-tuned models on the PS eval set, the new LPS test subset (including separate evaluations on fully and partially fake utterances), and the Half-Truth Audio Deepfake (HAD) test set. Interestingly, adding bona fide or fully synthetic utterances to the training data often worsens performance across most metrics, contradicting the common belief. It is worth noting that adding bona fide utterances does improve the performance on the HAD test set, while adding fully synthetic utterances improves the performance of the MRM model on the PS eval set, even though they are largely unrelated. More importantly, adding partial spooF utterances from the LPS train set leads to significant improvements across most test conditions, suggesting that these models learn to distinguish between real and fake segments based on temporal inconsistencies within a single utterance. As expected, models fine-tuned with all three types of samples produce the most consistent results overall. This suggests that simply increasing the amount of training data does not guarantee improved performance. A thoughtful data creation and selection strategy

may be necessary to develop reliable localization systems.

IV. CONCLUSIONS

In this work, we revisited the evaluation setup for partially fake speech localization and identified key gaps between academic benchmarks and real-world deployment, both in terms of metrics and dataset diversity. We argued that metrics like EER, while useful for benchmarking, are insufficient for interpreting real-world system behavior. To address this, we reframe the localization task as a sequential anomaly detection problem and advocated for the use of more interpretable, threshold-dependent metrics. We further showed that current SSL-based models, despite strong in-domain performance, generalize poorly to out-of-domain data. These models not only fail to detect novel synthetic samples but also misclassify unfamiliar bona fide audio, highlighting their reliance on domain-specific patterns. For future work, we recommend always including an out-of-domain evaluation set, such as LlamaPartialSpooF, and focusing on areas such as domain adaptation [22], [23] and generalizable training methods [13], [24]. Such a shift is critical for advancing the field toward robust solutions that withstand increasingly sophisticated synthetic speech [25] and adversarial attacks [26] in real-world conditions.

REFERENCES

- [1] Z. Wu, J. Yamagishi, T. Kinnunen, *et al.*, “AsvspooF: The automatic speaker verification spooFing and countermeasures challenge,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

- [2] N. M. Müller, P. Kawa, W. H. Choong, *et al.*, “Mlaad: The multi-language audio anti-spoofing dataset,” in *Proc. IJCNN*, IEEE, 2024, pp. 1–7.
- [3] X. Wang, H. Delgado, H. Tak, *et al.*, “Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech,” *Computer Speech & Language*, p. 101 825, 2025.
- [4] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, “The partialspooft database and countermeasures for the detection of short fake speech segments embedded in an utterance,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 813–825, 2023.
- [5] J. Yi, J. Tao, R. Fu, *et al.*, “Add 2023: The second audio deepfake detection challenge,” *arXiv preprint arXiv:2305.13774*, 2023.
- [6] H.-T. Luong, H. Li, L. Zhang, K. A. Lee, and E. S. Chng, “Llamapartialspooft: An llm-driven fake speech dataset simulating disinformation generation,” in *Proc. ICASSP*, IEEE, 2025, pp. 1–5.
- [7] Z. Cai, S. Ghosh, A. P. Adatia, *et al.*, “Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7414–7423.
- [8] J. Wu, W. Lu, X. Luo, R. Yang, Q. Wang, and X. Cao, “Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7395–7403.
- [9] Z. Cai and M. Li, “Integrating frame-level boundary detection and deepfake detection for locating manipulated regions in partially spoofed audio forgery attacks,” *Computer Speech & Language*, vol. 85, p. 101 597, 2024.
- [10] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, “Does audio deepfake detection generalize?” In *Proc. Interspeech*, 2022, pp. 2783–2787. DOI: 10.21437/Interspeech.2022-108.
- [11] Y. Yang, H. Qin, H. Zhou, *et al.*, “A robust audio deepfake detection system via multi-view feature,” in *Proc. ICASSP*, IEEE, 2024, pp. 13 131–13 135.
- [12] Y. Zhu, S. Koppiseti, T. Tran, and G. Bharaj, “Slim: Style-linguistics mismatch model for generalized audio deepfake detection,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 67 901–67 928, 2024.
- [13] T.-P. Doan, H. Dinh-Xuan, T. Ryu, *et al.*, “Trident of poseidon: A generalized approach for detecting deepfake voices,” in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 2222–2235.
- [14] O. Pascu, D. Oneata, H. Cucu, and N. M. Müller, “Easy, interpretable, effective: Opensmile for voice deepfake detection,” *arXiv preprint arXiv:2408.15775*, 2024.
- [15] E. Coletta, D. Salvi, V. Negroni, D. U. Leonzio, and P. Bestagini, “Anomaly detection and localization for speech deepfakes via feature pyramid matching,” *arXiv preprint arXiv:2503.18032*, 2025.
- [16] J. Yi, Y. Bai, J. Tao, *et al.*, “Half-Truth: A Partially Fake Audio Detection Dataset,” in *Proc. Interspeech*, 2021, pp. 1654–1658. DOI: 10.21437/Interspeech.2021-930.
- [17] T. Kinnunen, H. Delgado, N. Evans, *et al.*, “Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2195–2210, 2020.
- [18] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, “Range-based equal error rate for spoof localization,” in *Proc. Interspeech*, 2023, pp. 3212–3216.
- [19] X. Wang, H. Delgado, H. Tak, *et al.*, “Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale,” *arXiv preprint arXiv:2408.08739*, 2024.
- [20] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, “Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing,” in *Proc. ICASSP*, IEEE, 2022, pp. 6382–6386.
- [21] I. Rimon, O. Gal, and H. Permuter, “Unmasking deepfakes: Leveraging augmentations and features variability for deepfake speech detection,” *arXiv preprint arXiv:2501.05545*, 2025.
- [22] Y. Li, M. Zhang, M. Ren, *et al.*, “Cross-domain audio deepfake detection: Dataset and analysis,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 4977–4983.
- [23] X. Chen, W. Lu, R. Zhang, *et al.*, “Continual unsupervised domain adaptation for audio deepfake detection,” in *Proc. ICASSP*, IEEE, 2025, pp. 1–5.
- [24] J. Lu, Y. Zhang, Z. Li, Z. Shang, W. Wang, and P. Zhang, “Leveraging distance information for generalized spoofing speech detection,” *Computer Speech & Language*, p. 101 804, 2025.
- [25] A. Vyas, B. Shi, M. Le, *et al.*, “Audiobox: Unified audio generation with natural language prompts,” *arXiv preprint arXiv:2312.15821*, 2023.
- [26] N. Müller, P. Kawa, W.-H. Choong, *et al.*, “Replay attacks against audio deepfake detection,” *arXiv preprint arXiv:2505.14862*, 2025.