LENS-DF:
Deepfake Detection and Temporal Localization for Long-Form Noisy Speech

Xuechen Liu Wanying Ge Xin Wang Junichi Yamagishi
National Institute of Informatics, 2-1-2 Hitotsubashi, Tokyo, 101-8430 Japan
{xuecliu, gewanying, wangxin, jyamagis}@nii.ac.jp

Abstract

*This study introduces **LENS-DF**, a novel and comprehensive recipe for training and evaluating audio deepfake detection and temporal localization under complicated and realistic audio conditions. The generation part of the recipe outputs audios from the input dataset with several critical characteristics, such as longer duration, noisy conditions, and containing multiple speakers, in a controllable fashion. The corresponding detection and localization protocol uses models. We conduct experiments based on self-supervised learning front-end and simple back-end. The results indicate that models trained using data generated with LENS-DF consistently outperform those trained via conventional recipes, demonstrating the effectiveness and usefulness of LENS-DF for robust audio deepfake detection and localization. We also conduct ablation studies on the variations introduced, investigating their impact on and relevance to realistic challenges in the field*[1].

1. Introduction

The exponential growth of audio deepfake technology via various speech synthesis methods, driven by advanced deep learning techniques for realistic generation, has given rise to serious societal concern [3, 14, 10]. Deepfake audio clips can spread convincingly fake information, threatening information security. Researchers respond to such challenges by developing various detection methods, leading to adequate deepfake detectors — usually binary classifiers—that can spot synthetic speech. Advancing beyond conventional classifiers, recent detection models [12, 33, 13] leveraging deep neural networks (DNNs) have shown promising performance.

In parallel to the advances in spoofing detectors, datasets that are used to train and evaluate them have also advanced.

[1]This study is supported by the New Energy and Industrial Technology Development Organization (NEDO, JPNP22007), and JST AIP Acceleration Research (JPMJCR24U3). This study was partially carried out using the TSUBAME4.0 supercomputer at the Institute of Science Tokyo.

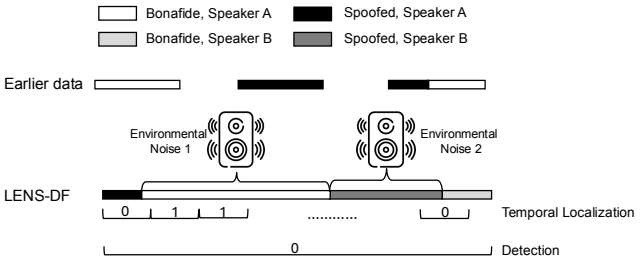

Figure 1. Conceptual differences between earlier datasets and datasets generated for this study. Note that the boxes in the figure represent audio samples. Earlier datasets typically feature short (2–10 seconds), single-speaker audio under clean, stable conditions. Dataset generated with LENS-DF includes longer clips (normally 15-50 seconds), environmental noise, and multiple speakers per sample. Bottom shows the conceptual task definition with how the annotation is conducted (1 denotes bonafide samples, and 0 denotes spoofed ones).

Mainly driven by the challenge series such as ASVspoof [34, 22, 18] and Audio Deepfake Detection (ADD) [37], these datasets have become complex. Later research has tackled partial spoofing [40], different compression, including neural-based codecs; and added background noise. Studies have also expanded to languages beyond English [20] and incorporated video alongside audio [4].

Although recent DNN-based deepfake detectors show promising performance, their real-world applicability remains limited because they have been predominantly trained and evaluated under laboratory-controlled audio conditions. State-of-the-art detectors have been trained on typically clean, short segments, ranging from under 2 to approximately 10 seconds [13, 22, 27]. Real-world issues and factors related to audio deepfakes, such as complex acoustic conditions, extended audio durations, and multi-speaker scenarios, remain underexplored. These issues and factors present additional challenges for audio deepfake detection, highlighting the disparity between the current research and real-world conditions and motivating more representative studies. Addressing these challenges requires new training and evaluation data and a comprehensive methodology cov-

ering data generation, training, and evaluation.

To address the current under-exploration in the research field, we introduce LENS-DF (**L**onger duration, **E**nhanced multi-speaker, **N**oisy **S**peech for audio **D**eep**F**ake detection). As shown in Figure 1, it contains several interesting characteristics such as longer duration, noisy conditions, and multi-speaker presence compared with early datasets and recipes. We further test whether state-of-the-art detectors can feasibly generalize in this scenario, measuring performance on both detection and localization. As illustrated at the bottom of the figure, throughout the paper, detection refers to a binary decision at the sample level. Localization, on the other hand, refers to predicting whether a specific region of the audio sample is spoofed.

- We introduce LENS-DF, a simple and comprehensive data generation recipe for robust audio deepfake detection that simulates multiple real-world conditions, , including extended audio duration, noise conditions, and multiple speakers within a single audio sample. We use ASVspoof 2019 logical access (LA) [22] for this study and generate variants of LENS-DF for training and evaluation[2].
- We conduct an investigation into current state-of-the-art neural deepfake detection models based on self-supervised learning (SSL) by fine-tuning them with the training data generated with LENS-DF, examining their adaptability and performance in audio deepfake detection and localization under complex conditions. We also benchmark the performance against various realistic factors and systematically analyze their impact on robustness and reliability.

2. Related Work

2.1. Audio Deepfake Datasets

Beyond advances in speech synthesis techniques, such as *text-to-speech* and *voice conversion* (VC) for spoofing attacks, recent datasets in audio deepfake research have increasingly emphasized realistic non-speech audio characteristics. ASVspoof 2021 [18] was among the first to address this critical aspect by incorporating audio-compression techniques, including transmission and media codecs, applied to initially clean source audio [22]. Conversely, a different approach was taken with datasets such as In-the-Wild [21] and DeepFakeEval [8] in that audio samples are directly collected from real-world recordings already containing natural artifacts and noise. Another realism-enhancing factor is the inclusion of multiple languages beyond English, as addressed with the Multi-Language Audio Anti-Spoof Dataset [20]. Previous studies (e.g., [19]) began exploring spoofing detection under complex conditions and with longer audio durations. However, these initial explorations typically focused on evaluating pre-trained models without SSL front-ends, and the impact of real-world factors has yet to be thoroughly investigated, aspects that this study aims to comprehensively address.

[2]Data generation code and sample data can be found at `https://github.com/nii-yamagishilab/LENS-DF-DataGen`

2.2. Audio Deepfake Detection and Localization

Audio deepfake detection has been an active research area for decades. Early approaches primarily relied on statistical models, such as *Gaussian mixture models* [30], combined with various acoustic features [24, 29]. Advances have seen a shift toward DNN-based methods, including architectures such as RawNet2 [32, 27] and AASIST [13]. Data-augmentation techniques, such as RawBoost [26], have significantly enhanced these models, achieving promising performance under controlled, lab-like evaluation conditions. The evolution of deepfake detectors has increasingly embraced Transformer architectures; for example, [7] integrates a Transformer encoder with a 1-D ResNet to effectively detect partially spoofed audio clips, while RawFormer [17] combines convolutional and Transformer layers to further refine detection capabilities. Leveraging large-scale pre-trained SSL speech models as front-ends has also proven highly effective. Extracted features from these SSL models can be seamlessly used by downstream classifiers, ranging from simple structures such as global average pooling (GAP) and fully connected (FC) layers [31], to more sophisticated models such as AASIST [28].

In parallel to the widely-known detection task, the task of *temporal localization* focuses on locating manipulated segments within an audio utterance, shifting the objective from utterance-level classification toward fine-grained, segment-level detection. Benchmarks such as PartialSpoof [40] and ADD [37] have supported research in this area. Further, a content-driven partially spoofed audio dataset was developed, in which specific parts of the speech are altered to change the sentential meaning [6, 4]. DNN-based methods dominate temporal-localization studies; for instance, an earlier study used hybrid convolutional and recurrent neural networks to effectively capture spectro-temporal patterns and contextual information at the frame level [41]. Transformer-based methods have been increasingly used [17, 42], along with ensemble frameworks combining boundary detection and classification enhanced by SSL front-ends [5]. Approaches using temporal convolution paired with contrastive learning have also emerged, further advancing localization performance [35]. However, previous studies primarily evaluated these models using relatively short audio samples under controlled conditions. There has been no exploration of spoof localization for long-duration audio with complex, realistic situations, which is a gap this study aims to address.

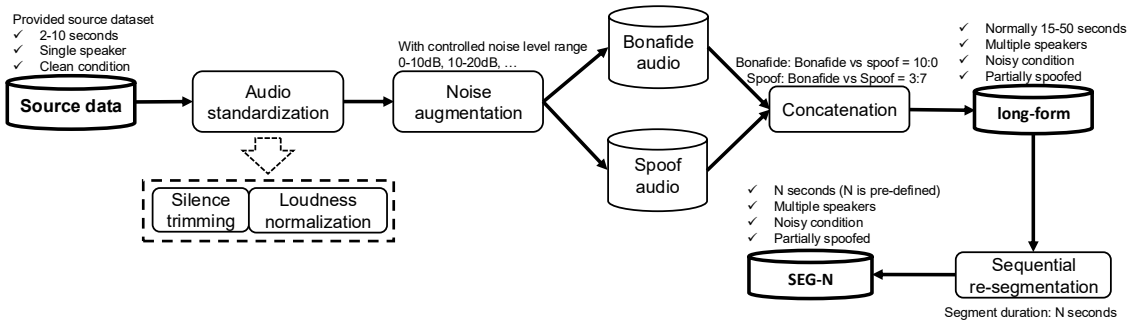

Figure 2. Data-generation pipeline of LENS-DF. Datasets marked as bold-lined cylinders are used to train and evaluate deepfake detectors in this study.

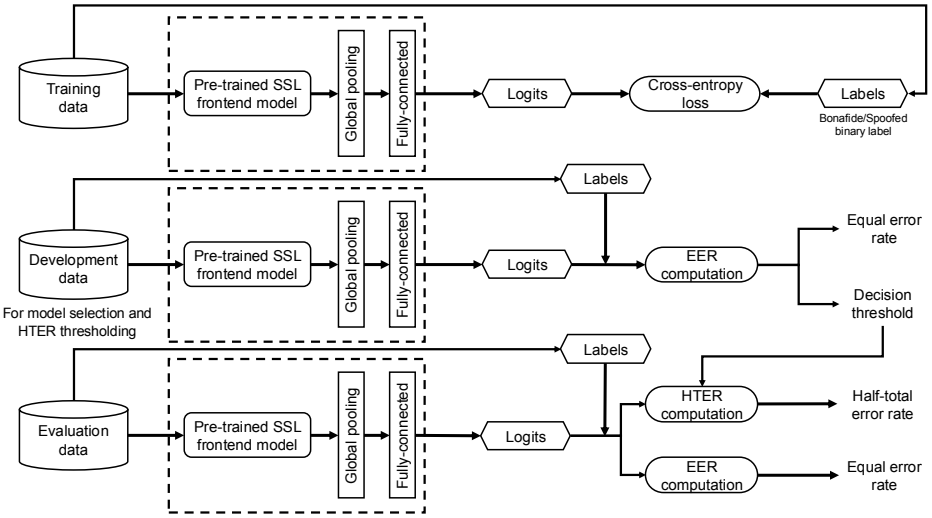

Figure 3. Detection and localization paradigm for LENS-DF. Area within dashed lines shows model architecture used in this study.

3. LENS-DF

In this section, we describe the data generation and annotation pipeline and the corresponding detection and localization paradigm of LENS-DF.

3.1. Source Data: ASVspoof 2019 LA

As an exemplar source dataset to demonstrate LENS-DF, we use ASVspoof 2019 LA [22]. This dataset is widely used for speech anti-spoofing, including recordings from 107 speakers and spoofed utterances generated using 17 different TTS and VC algorithms (or their hybrids). This diversity of speakers and attack methods provides a rich foundation for creating partial deepfakes. All clips in the dataset are labeled as bonafide or spoofed, and it is essential to note that this dataset does not contain any partially spoofed audio — each spoofed file is entirely synthetic. The original audio lengths range from 2 to 10 s, and the audio-recording condition is controlled from its source — the VCTK dataset [36], with a single speaker in each audio sample and clean acoustic condition. This makes this dataset suitable for this study. The original training, development, and evaluation partitions of the dataset[3] [are maintained throughout this paper — were maintained throughout this study?]. The audio throughout this study was mono-channel, with a sampling rate 16KHz.

3.2. Generation

The data-generation pipeline, depicted in Fig. 2, encompasses several systematic steps designed to produce realistic, standardized, multi-speaker audio under controlled noise conditions[4]. While the pipeline shares some procedural similarities with [19], it significantly differs from the latter in applying more controlled and systematic techniques. The detailed steps are as follows.

Audio standardization Similar to [19], initial audio standardization consists of two pre-processing stages: 1) silence trimming, where leading and trailing silence is removed us-

[3] https://www.asvspoof.org/index2019.html

[4] Note that different from earlier studies, the term "generation" for this study does not mean speech synthesis via generative models.

ing LibROSA[5], and 2) loudness normalization, involving randomized audio-volume normalization based on the ITU P.56 standard[6]. Importantly, these steps do not alter the original labels (bonafide/spoofed) of the data.

Noise augmentation Noise augmentation uses background noise from the MUSAN dataset [25], categorized into human chatter (`babble`), music (`music`), and general environmental sounds (`noise`). Each audio segment is randomly assigned either no noise or exactly one noise type, with noise intensity uniformly randomized between 0 and 10 dB to reflect realistic scenarios in which background noise is typically lower than the primary speech signal. Unlike [19], the proposed noise augmentation in this study does not include room impulse response [16] and augmentations using media codecs, such as mp3, to maintain control over noise characteristics.

Audio concatenation Following noise augmentation, short audio segments are concatenated to create long-form audio clips, each containing exactly 10 segments. Specifically, fully bonafide long-form audio consists exclusively of bonafide segments. In contrast, partially spoofed audio contains three bonafide and seven spoofed segments in randomized order, based on pilot experiments and [19]. To facilitate comparative analysis, the number of generated bonafide and spoofed long-form samples match that in ASVspoof 2019 LA (2,580 bonafide, 22,800 spoofed). Development and evaluation subsets contain 1,000 samples per class, with audio lengths typically 15 to 50 seconds, featuring multiple speakers and varied noise conditions. Labels are assigned on the basis of any spoofed content within the concatenation. We label the data generated from this step as `long` in the remainder of this paper.

Audio re-segmentation We introduce an additional re-segmentation step absent in prior studies [19]. Each long-form audio is uniformly segmented into shorter clips of N seconds (default $N = 4$), with any residual audio at the end discarded. The motivation for this re-segmentation step is to practically increase the number of training samples under the current training paradigm and enhance the model's capability to detect localized spoofing artifacts, which may occur in specific segmental and temporal regions. Table 1 shows the resulting number of audio samples. Labels follow the same rule as in the last step. The data generated from this step are labeled `SEG-N` throughout the remainder of this paper.

[5] `https://librosa.org/doc/latest/index.html`
[6] `https://github.com/openitu/STL`

Table 1. Number of generated speech samples from ASVspoof 2019 LA, using LENS-DF data generation pipeline. `SEG-4` samples are *4-s* and from `long`.

	`long`		`SEG-4`	
Partition	**Bonafide**	**Spoofed**	**Bonafide**	**Spoofed**
Train	2,580	22,800	17,857	129,805
Dev	1,000	1,000	5,132	5,640
Eval	1,000	1,000	4,984	5,663

3.3. Detection & Localization

In conjunction with our generation pipeline, we implement the training and evaluation procedures depicted in Figure 3. Our reference model uses an SSL front-end to generate frame-level audio representations, subsequently processed by a straightforward back-end classifier composed of global average pooling (GAP) and a fully connected (FC) layer, producing posterior logits for binary classification. During training, model selection is guided by logits computed from the development set on the basis of the equal error rate (EER).

Perhaps it is essential to highlight that our inference strategy differs from conventional video and audio-visual deepfake localization approaches [6, 11], with which the model produces a boundary-matching confidence map. We conduct localization by creating a "confidence" per fixed-length segment. Each audio segment, obtained by explicitly re-segmenting long audio recordings with annotations, is independently classified as either bonafide or spoofed. This approach means that longer audio inputs are directly processed with a model trained on shorter fixed-length segments, and each segment receives a separate binary classification without any fusion or aggregation across segments. This aligns with earlier audio-specific deepfake detection and localization practices [41, 40, 37], directly yielding segment-level localization decisions.

The optimal threshold identified from the development-set EER is used to calculate the half total equal error rate (HTER) [9], another widely used metric in biometric evaluations alongside the EER. The EER is computed directly from the logits at evaluation time, while the HTER uses the predefined threshold from the development set. The conditions of the development set matched those of the training set throughout our experiments. Both metrics were used for the two tasks presented in this study.

4. Experimental Design

4.1. Training

SSL front-end On the basis of preliminary experiments, we selected two pre-trained SSL models from the *massive multilingual speech* (MMS) framework [23]. Both selected models leverage the Wav2Vec 2.0 architecture [2] and have been pre-trained extensively on speech data covering over

Table 2. Details of two SSL models used for this study. Information is obtained from [23].

Model	Backbone	Training data	Model size #params.
MMS-1B	Wav2Vec 2.0 [2]	55K hours 1K+ languages	∼965 million
MMS-300M	Wav2Vec 2.0 [2]	55K hours 1K+ languages	∼317 million

1,400 languages. Specifically, we used two MMS variants: one with around 1 billion parameters (MMS-1B[7]) and the other with around 300 million parameters (MMS-300M[8]), with their respective pre-trained checkpoints publicly accessible. The basic information of these two models is listed in Table 2.

Training details As shown in Figure 3, we fine-tune the SSL front-end together with the back-end using a cross-entropy loss. For comparison, we experiment with three different training sets for fine-tuning the SSL models: $19LA_{train}$, the original ASVspoof 2019 LA training dataset; $\texttt{long}_{train}$, the training partition of our generated data (see Table 1); and $\texttt{SEG-4}_{train}$, the re-segmented version of $\texttt{long}_{train}$. To improve training efficiency, we use dynamic batching[9], enabling input audio durations to vary while remaining within a similar numerical range. This approach also minimizes the need for excessive zero padding. We also constrain the total audio duration per mini-batch to 100 s, ensuring workable memory usage. The temporal pooling via the GAP layer is done at the whole audio length without any further fixed-length chunking. All models are optimized using the Adam optimizer [15]. The learning rate is linearly increased to 1×10^{-7} over the first 80k steps then linearly decayed to zero by 800k steps, at which point training terminates. We stop training if either 100 epochs or 800k steps are reached, whichever criterion is met first. Each model is trained and evaluated on a single NVIDIA TESLA A100 GPU card.

4.2. Evaluation

Data We evaluate both detection and segment-level temporal localization performance on the LENS-DF evaluation set, which comprises our generated long-form audio, as detailed in Section 3. We cover the source evaluation data, denoted as "19LA" in the next section, and the two generated variants via LENS-DF: (i) `long`, which refers to the generated long-form audio; and (ii) `SEG-4`, which refers to the re-segmented samples (see Section 3.2). The length is set to 4 s. We mark the training and evaluation sets using subscripts "$_{\text{train}}$" and "$_{\text{eval}}$", respectively. We assess performance under both matched-condition and cross-condition settings, where matched conditions use the same variant (including source evaluation data) for training and evaluation. At the same time, cross-conditions involve different variants or sources between the two.

Metrics As introduced in Section 3.3, we use the EER and HTER as the primary evaluation metrics. Following prior research, we treat localization as detection applied to very short segments, computing both metrics in the same manner for segment-based decisions. The HTER threshold is determined using the development set corresponding to the evaluation set to ensure a fair comparison of model performance. For instance, if the evaluation condition is `long`, the development set will also be from `long`.

5. Results and Discussion

The detection and localization results on the LENS-DF evaluation set are presented in Table 3. We have several key findings, which are detailed below.

Conventional datasets with short, clean speech are clearly not suitable for training models for detecting long and complex spoofed audios. Models trained on conventional short, clean speech (ASVspoof 2019 LA) do not perform well when evaluated on $\texttt{long}_{eval}$. While MMS-300M achieved the best detection performance on the matched $19LA_{eval}$, its performance significantly degraded under more complex conditions such as $\texttt{long}_{eval}$ (from 0.15 to 2.9% EER, from 0.52% to 4.05% HTER). Conversely, training on $\texttt{long}_{train}$ and its re-segmented variant $\texttt{SEG-4}_{train}$ improves the performance for both types of models, with MMS-1B being the second-best considering both metrics (0.9% EER, 1.65% HTER) and MMS-300M reaching the best EER with the latter training data (0.6%). These findings highlight the importance of the incorporated complexity required to achieve robust spoofing detection. The complexity in `long` includes several variations introduced in a compounding manner. Thus, isolating specific factors contributing to improvement is important.

Model trained on conventional datasets with short, clean speech needs further effort for localizing long-form deepfake speech in noisy, complex environments. We now move on to the $\texttt{SEG-4}_{eval}$ results. Localization using models trained on $19LA_{train}$ resulted in unsatisfying performance, indicating that models trained solely on short and clean audio without multi-speaker presence struggle with temporal localization. Expectedly, models trained explicitly on the long-form or re-segmented datasets demonstrated improved temporal localization performance, with the best result (MMS-300M trained on $\texttt{SEG-4}_{train}$) achieving 14.62% EER and 14.08% HTER.

Detection models trained on complex data might not perform localization effectively. Although LENS-DF variants (`long` and `SEG-4`) notably improve detection of

[7] https://huggingface.co/facebook/mms-1b

[8] https://huggingface.co/facebook/mms-300M

[9] https://speechbrain.readthedocs.io/en/latest/tutorials/advanced/dynamic-batching.html

Table 3. Statistical results on LENS-DF evaluation set. Note that models in this table do not have RawBoost [26] applied. Numbers in **bold** and underlined indicate best and second-best results under each evaluation condition, respectively.

		Detection, $19LA_{eval}$		**Detection**, $long_{eval}$		**Localization**, $SEG{-}4_{eval}$	
SSL Model	**Training Data**	**EER (%)**	**HTER (%)**	**EER (%)**	**HTER (%)**	**EER (%)**	**HTER (%)**
MMS-1B	$19LA_{train}$	0.39	1.97	6.40	9.30	22.10	23.04
MMS-300M	$19LA_{train}$	**0.15**	**0.52**	2.90	4.05	21.12	21.09
MMS-1B	$long_{train}$	5.15	4.97	0.90	1.65	19.20	17.86
MMS-300M	$long_{train}$	7.45	5.32	1.30	**1.40**	17.81	17.26
MMS-1B	$SEG{-}4_{train}$	4.62	7.98	**0.60**	4.20	15.14	14.34
MMS-300M	$SEG{-}4_{train}$	4.92	4.66	1.00	8.40	**14.62**	**14.08**

Table 4. Statistical results on LENS-DF evaluation set, with RawBoost [26] applied as online data-augmentation method during training. The last row is model released by [28], which has been pre-trained on 19LA with AASIST back-end and RawBoost applied. **bold** and underlined indicate best and second-best results under each evaluation condition, respectively.

			Detection, $19LA_{eval}$		**Detection**, $long_{eval}$		**Localization**, $SEG{-}4_{eval}$	
SSL Model	**Training Data**	**RawBoost**	**EER (%)**	**HTER (%)**	**EER (%)**	**HTER (%)**	**EER (%)**	**HTER (%)**
MMS-300M	$19LA_{train}$	Yes	0.46	7.80	9.60	14.80	26.32	27.75
	$long_{train}$	Yes	12.06	7.99	0.90	**1.90**	18.89	18.36
	$SEG{-}4_{train}$	Yes	8.31	6.92	**0.60**	3.80	**13.68**	**13.52**
	$SEG{-}4_{train}$ (Clean)	Yes	6.85	7.68	5.90	12.30	26.66	26.65
	$SEG{-}4_{train}$ (Clean)	No	3.26	8.36	8.90	9.75	29.94	31.41
XLS-R-300M	$19LA_{train}$	Yes	**0.19**	**0.94**	15.70	17.60	30.41	27.30

spoofed speech under complex conditions, localization performance remains inadequate for practical use, with the lowest EER and HTER observed at 14.62% and 14.08%, respectively. Although training with $SEG{-}4_{train}$ provided more shorter samples and greater exposure to spoofed content and results in notable improvements compared with $long_{train}$, its performance may not be considered feasibly satisfying. These findings highlight the need for distinct training strategies and datasets tailored explicitly. It also indicates the significant impact of the realistic and complex factors introduced in the data-generation pipeline. We discuss these factors via multiple self-contained and detailed ablation studies covered in Section 6.

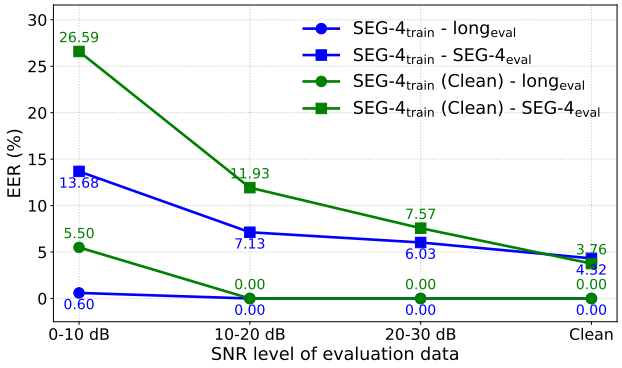

Figure 4. Detection and localization performance against different noise level ranges of LENS-DF evaluation. The difference in colors and shapes corresponds to different training and evaluation conditions, respectively.

6. Ablation Study

In this section, we investigate several realistic factors created in the pipeline illustrated in Figure 3.2. Specifically, we examine the impact of noise, duration, and presence of multiple speakers within single audio sample.

6.1. The effect of noise augmentation

We first study the impact of our controlled noise augmentation during both training and evaluation. Two specific aspects are addressed separately: 1) the comparative effectiveness of our noise augmentation strategy introduced in Section 3.2 versus earlier established methods; and 2) the impact of varying noise levels during evaluation, measured using the signal-to-noise ratio (SNR).

For the first aspect, we employ RawBoost [26], a validated data augmentation method known for controlled additions such as room impulse response [16] and digital compression using signal processing and filtering techniques. We integrate this augmentation during the data loading phase of training, using a publicly available implementation[10]. To systematically evaluate the efficacy of RawBoost in our experimental setup, we first create a variant of $SEG{-}4$ by removing the noise augmentation step illustrated in Figure 2. The variant audio samples retain longer duration and multi-speaker presence in such cases. . We label this variant as $SEG{-}4$ (Clean) and use it to serve training and evaluation phases. We then examine the performance

[10] https://github.com/TakHemlata/RawBoost-antispoofing

impact of applying RawBoost on SEG−4$_{\text{train}}$ (Clean), allowing for a direct and fair comparison against the noise-augmentation strategy we introduced in our study. For additional comparative reference, we evaluate a baseline model comprising an XLS-R [1] front-end with approximately 300 million parameters and an AASIST back-end [13]. This baseline, labeled as XLS-R-300M, was trained using randomly chunked or padded audio segments of 4 seconds in duration. The pre-training is carried out on the 19LA$_{\text{train}}$ dataset with RawBoost applied, achieving a reported 0.82% EER on the ASVspoof 2021 LA evaluation set [28][11].

Results are presented in Table 4. In comparison to findings from Table 3, for more complex training scenarios (such as those utilizing long or SEG−4 data), RawBoost provides moderate improvements, especially in the long$_{\text{eval}}$ scenario. This gain comes at a considerable performance penalty for the 19LA detection task (where EER deteriorates from 7.45% to 12.06%). Localization improvements under these conditions are modest, with EER improving slightly from 14.62% to 13.68%. Removing our noise augmentation from SEG−4$_{\text{train}}$ notably degrades performance on the two complex evaluation sets, with moderate improvement at 19LA$_{\text{eval}}$ in terms of EER. Applying RawBoost recovers such degradation for long$_{\text{eval}}$ detection but does not for the other two sets. These results collectively suggest that RawBoost demonstrates limited effectiveness compared with LENS-DF due to its highly controlled setup.

Regarding the second aspect, we examine the impact of varying evaluation noise levels, considering practical conditions with non-negative SNRs. Given that our training data is augmented within an SNR range of 0 to 10dB, we evaluate model performance across evaluation sets with noise levels ranging from 0 to 30 dB at selected intervals. For this analysis, we choose the MMS-300M model trained on SEG−4$_{\text{train}}$ with RawBoost, as it yielded the best EER and HTER on SEG−4$_{\text{eval}}$.

The resulting detection and localization performances across these varying SNR conditions on the LENS-DF dataset are illustrated in Figure 4. Our analysis reveals a clear trend: reducing noise levels enhances detection and localization accuracy. Under noise-free conditions, detection accuracy approaches near-perfect performance (0% EER), and localization accuracy also significantly improves, achieving around 5% EER. Conversely, performance deteriorates sharply at higher noise levels, approaching speech signal amplitude. These findings highlight the importance of maintaining a complex yet controlled acoustic environment for reliable detection and temporal localization. Future work may involve exploring more adaptive noise augmentation methods that better align with complex real-world scenarios.

[11]Checkpoint available at https://github.com/TakHemlata/SSL_Anti-spoofing

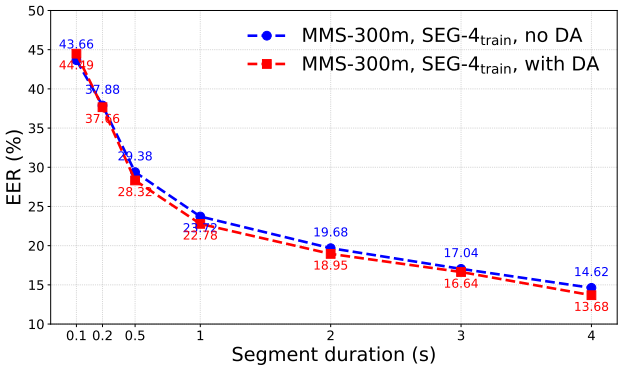

Figure 5. Localization performance against different durations of LENS-DF evaluation set in EER(%). "DA" corresponds to the application of RawBoost addressed in Section 6.1.

Table 5. Detection and localization performance measured using EER (%) under single- and multi-speaker conditions. multi. in this table is same as long$_{\text{train}}$ in earlier sections.

Train / Eval cond.	single		multi.	
	Detection, long$_{\text{eval}}$	**Localization,** SEG−4$_{\text{eval}}$	**Detection,** long$_{\text{eval}}$	**Localization,** SEG−4$_{\text{eval}}$
single.	7.90	16.10	**1.30**	19.56
multi.	11.10	17.37	**0.60**	**13.68**

6.2. The effect of duration on localization

In Section 5, we explored two different tasks under two distinct segment durations (long-form audio and 4-second segments), observing significant performance differences. We conduct experiments to systematically study this effect by varying the duration (N, measured in seconds) of segments re-extracted from the long portion of the LENS-DF evaluation set. Specifically, we vary N from 0.01 (typical hop size of audio framing) up to 4, selecting several intermediate durations for comparison. We use the MMS-300M trained on SEG−4 audio, with and without RawBoost applied, as they achieved the best performance on SEG−4_eval among all trained models.

The resulting performance is shown in Figure 5. Extremely short segments yield insufficient localization performance. However, performance consistently improves as segment duration increases. This indicates that longer segments facilitate better temporal localization under current evaluation conditions and setup. Future work may focus on developing architectures and training strategies specifically optimized for ultra-short audio segments.

6.3. The effect of speaker presence

We then evaluate whether multiple speakers in long-form audio segments can affect detection and localization performance. Similar to Section 6.1, we modify the original data generation pipeline by limiting post-augmentation concatenation to segments from a single speaker (as illustrated in Figure 1, where only speaker A or B is used).

Table 6. Statistical performance of models trained on LENS-DF and LAV-DF, measured by precision and recall metrics. Time resolution was 0.04 seconds.

Data	AP@0.25	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
LAV-DF	99.32	29.28	0.16	0.00	27.77	26.27	24.63	23.73
LENS-DF (ours)	100.00	33.50	0.00	0.00	38.56	37.52	36.61	35.70

This variant preserves the long-form structure and diverse noise conditions but includes only one speaker per audio. To enable this single-speaker setup, we leverage the speaker correspondence between bonafide and spoofed samples in ASVspoof 2019 LA. We refer to the original multi-speaker variant as `multi.` and the new single-speaker variant as `single`. For comparison, we additionally train a model on the `single` speaker data using the same MMS-300M setup as in Section 6.1.

The results are presented in Table 5. While both models perform well for detection under multiple-speaker conditions (1.3% and 0.6% EER), performance notably drops for single-speaker cases (7.9% and 11.1% EER). Both models struggle with the $\texttt{SEG-4}_{\text{eval}}$ regardless of speaker configuration, reflected by the lowest EER being larger than 13%. These findings suggest that when only one speaker is present, the model may over-rely on speaker identity rather than spoofing artifacts, which leads to shortcut learning and degraded performance, especially with unseen or multi-speaker inputs. Hence, speaker information may be an unreliable cue in this context, and future work may focus on alternative features and strategies to improve robustness.

6.4. Comparative results on LAV-DF

Finally, we compare our dataset with existing work on audio-visual deepfake localization, represented by LAV-DF [6]. We use the data protocol and adopt the evaluation metric from [6, 11] for a sensibly fair comparison.

It is worth noting that our model hypothesis differs fundamentally from the previous ones incorporating visual information. The audio-visual deepfake detectors trained in those works predict variable-length segments in time, whereas the audio deepfake models in this study work on fixed-length windows. This mismatch affects not just what the model outputs but also how one matches predictions to ground truth. Intersection-over-union (IoU) is typically computed over continuous frame segments and assigns each ground-truth event to its best-matching prediction. The fixed-window approach in audio tasks generates overlapping predictions requiring IoU computation over discrete chunks. Therefore, the matching logic will also change: instead of one prediction per event, we aggregate point-based model predictions with a specific time resolution and predefined IoU threshold. Prediction scores for these fixed segments are binarized for classification, and IoU is computed directly from these binary predictions and ground-truth labels per segment. Subsequently, average precision is calculated across varying IoU thresholds, which resembles earlier practices in audio deepfake localization [39, 38]. The time resolution is set to 0.04 s, which is the same as that used in a previous study [6][12]. There is no trimming or padding during our evaluation. We evaluate using MMS-300M trained on $\texttt{SEG-4}_{\text{train}}$ without RawBoost.

As shown in Table 6, model performance on LENS-DF is comparable to that on LAV-DF in AP, despite LAV-DF being more mismatched to our model. Additionally, our model achieves better AR on LENS-DF, likely due to a more balanced representation of bonafide and spoofed data than on LAV-DF. This observation aligns with previous findings [19], indicating that overly abundant bonafide data may obscure spoofed regions, negatively impacting spoof detection. These results underscore the effectiveness of LENS-DF and the necessity of adapting evaluation metrics specifically tailored to the characteristics of audio-based detection models. Future research could further explore tailored training methods and evaluation frameworks to address better temporal localization challenges in audio-visual tasks represented by LAV-DF and others [4].

7. Conclusion and Future Work

In this study, we have introduced LENS-DF, a comprehensive approach addressing realistic complexities in audio deepfake detection and localization, such as extended audio duration, varying noise conditions, and multi-speaker scenarios within single audio samples. We have extended previous work on audio deepfake data generation in a controllable manner to facilitate both training and evaluation. Moreover, we have confirmed through systematic experiments and ablation studies that current state-of-the-art SSL models trained and, while effective under laboratory conditions, demonstrate performance degradation in detection and temporal localization under complex conditions. This degradation can be mitigated by explicitly incorporating additional complexities into the training. Meanwhile, robust localization may require more effort in complex, real-world scenarios, even after systematically exploring multiple contributing factors. Future work will focus on advanced model architectures and training methodologies for temporal localization. Adapting to multi-speaker scenarios, language variations, and noise can further bridge the gap between experimental settings and practical applications.

[12] In [6], the audio features used in [6] is log spectrogram with 0.01-second resolution, and the scores are downsampled 4 times.

References

[1] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli. XLS-R: Self-supervised cross-lingual speech representation Learning at Scale. In *Proc. Interspeech*, pages 2278–2282, 2022.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proc. NeurIPS 2020*, 33:12449–12460, 2020.

[3] T. Brewster. Huge bank fraud uses deep fake voice tech to steal millions. `https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions`, 2021. Accessed: 2024-06-26.

[4] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 7414–7423, 2024.

[5] 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*, 85:101597, 2024.

[6] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10, 2022.

[7] Z. Cai, W. Wang, and M. Li. Waveform boundary detection for partially spoofed audio. In *Proc. ICASSP*, pages 1–5, 2023.

[8] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, J. Choi, A. Kim, and O. Etzioni. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024, 2025.

[9] I. Chingovska, A. R. d. Anjos, and S. Marcel. Biometrics evaluation under spoofing attacks. *IEEE Transactions on Information Forensics and Security*, 9(12):2264–2276, 2014.

[10] Cointelegraph. Security researchers unveil deepfake artificial intelligence(ai) audio attack hijacks live conversations. `https://cointelegraph.com/news/security-researchers-unveil-deepfake-artificial-intelligenceai-audio-attack-hijacks-live-conversations`, 2024. Accessed: 2024-06-26.

[11] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4367, 2021.

[12] G. Hua, A. B. J. Teoh, and H. Zhang. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters*, 28:1265–1269, 2021.

[13] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proc. ICASSP*, pages 6367–6371, 2022.

[14] F. Karimi. 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. `https://amp.cnn.com/cnn/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html`, 2023. Accessed: 2025-04-09.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *Proc. ICASSP*, pages 5220–5224, 2017.

[17] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang. Leveraging positional-related local-global dependency for synthetic speech detection. In *Proc. ICASSP*, pages 1–5, 2023.

[18] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

[19] X. Liu, X. Wang, and J. Yamagishi. A preliminary case study on long-form in-the-wild audio spoofing detection. In *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2024.

[20] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger. MLAAD: The multi-language audio anti-spoofing dataset. *International Joint Conference on Neural Networks (IJCNN)*, 2024.

[21] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger. Does audio deepfake detection generalize? In *Proc. Interspeech*, pages 2783–2787, 2022.

[22] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee. ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):252–265, 2021.

[23] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

[24] M. Sahidullah, T. Kinnunen, and C. Hanilçi. A comparison of features for synthetic speech detection. In *Proc. Interspeech*, pages 2087–2091, 2015.

[25] D. Snyder, G. Chen, and D. Povey. MUSAN: A music, speech, and noise corpus, 2015.

[26] 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*, pages 6382–6386, 2022.

[27] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher. End-to-end anti-spoofing with RawNet2. In *Proc. ICASSP*, pages 6369–6373, 2021.

[28] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 112–119, 2022.

[29] M. Todisco, H. Delgado, and N. Evans. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516–535, 2017.

[30] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi. Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion. In *Interspeech 2018*, pages 77–81, 2018.

[31] X. Wang and J. Yamagishi. Investigating self-supervised front ends for speech spoofing countermeasures. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 100–106, 2022.

[32] J. weon Jung, S. bin Kim, H. jin Shim, J. ho Kim, and H.-J. Yu. Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms. In *Proc. Interspeech*, pages 1496–1500, 2020.

[33] Z. Wu, R. K. Das, J. Yang, and H. Li. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In *Proc. Interspeech*, pages 1101–1105, 2020.

[34] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Proc. Interspeech*, pages 2037–2041, 2015.

[35] Y. Xie, H. Cheng, Y. Wang, and L. Ye. An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection. In *Proc. ICASSP*, pages 966–970, 2024.

[36] J. Yamagishi, C. Veaux, and K. MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). 2019.

[37] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li. ADD 2022: The first audio deep synthesis detection challenge. In *Proc. ICASSP*, pages 9216–9220, 2022.

[38] B. Zhang and T. Sim. Localizing fake segments in speech. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3224–3230, 2022.

[39] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi. Range-based equal error rate for spoof localization. In *Interspeech 2023*, pages 3212–3216, 2023.

[40] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi. The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:813–825, 2023.

[41] L. Zhang, X. Wang, E. Cooper, and J. Yamagishi. Multi-task learning in utterance-level and segmental-level spoof detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 9–15, 2021.

[42] J. Zhong, B. Li, and J. Yi. Enhancing partially spoofed audio localization with boundary-aware attention mechanism. In *Proc. Interspeech*, pages 4838–4842, 2024.