Oral #4-3,
Poster #2-29

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

Xuechen Liu, Wanying Ge, Xin Wang, Junichi Yamagishi @ National Institute of Informatics, Japan

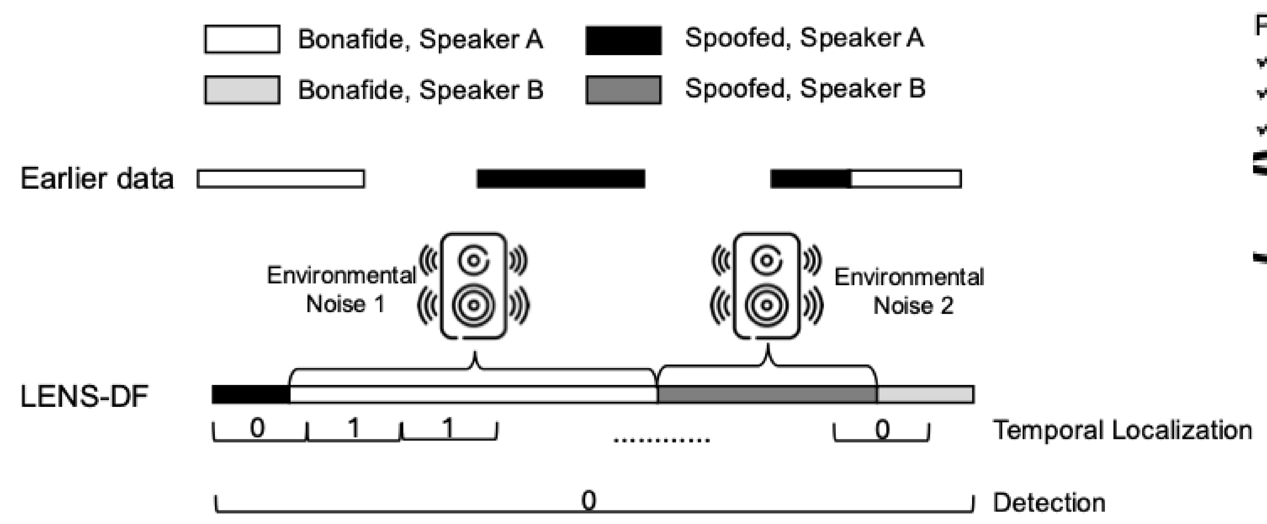
Main Contributions

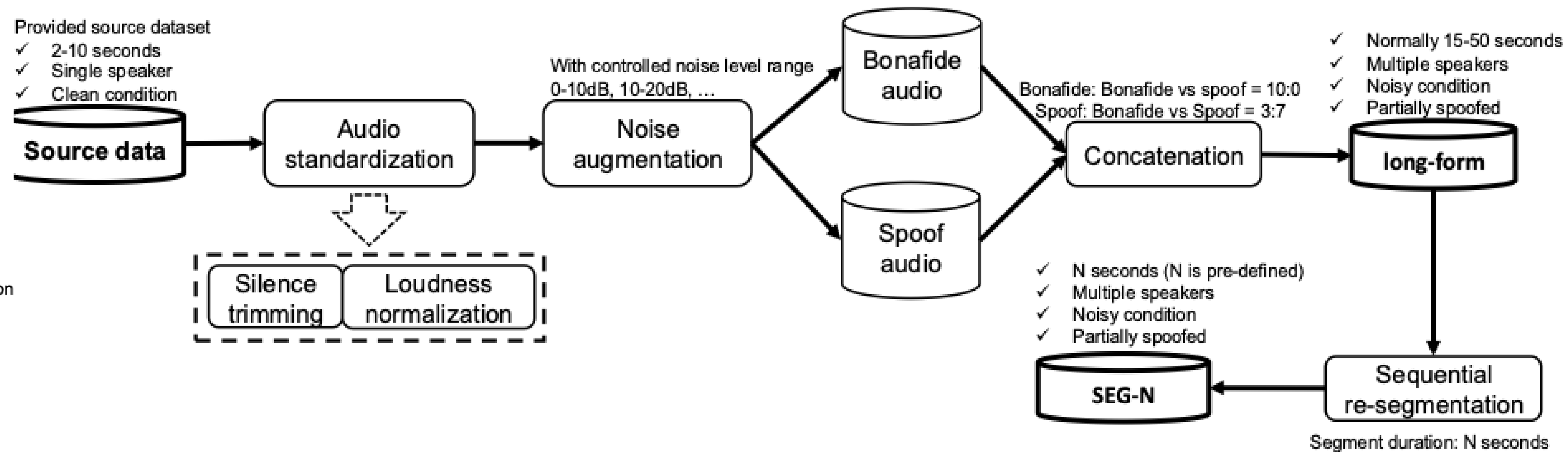
- We propose a controllable, comprehensive data complication framework that includes several artefacts that reflects real-world, in-the-wild Deepfake audios
- We benchmark the adaptability and robustness of the state-of-the-art self-supervised learning-based audio Deepfake detectors against those realistic variations
- We perform ablation analysis on those artefacts' impact on model robustness and reliability

The Conceptual Novelty

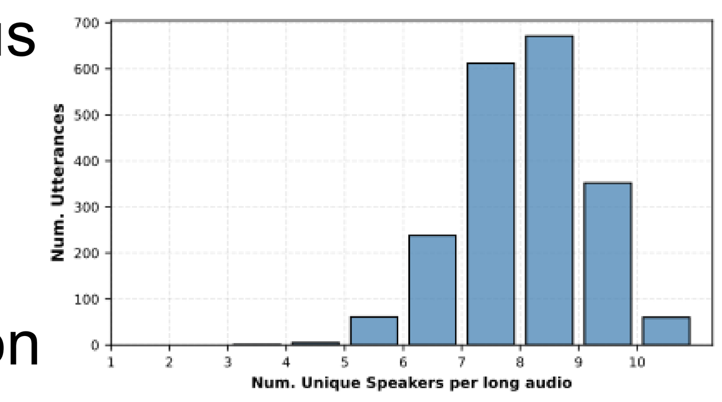
- We are not completely generating "real-world" audio, since the definition is hard and itself unrealistic
- Instead, we include multiple realistic variations/artefacts into the generated audio: longer duration, multiple speakers (mix of Bonafide and spoof), and noises
- The detection is performed per long audio, while the temporal localization is formed here as detection per short segment individually

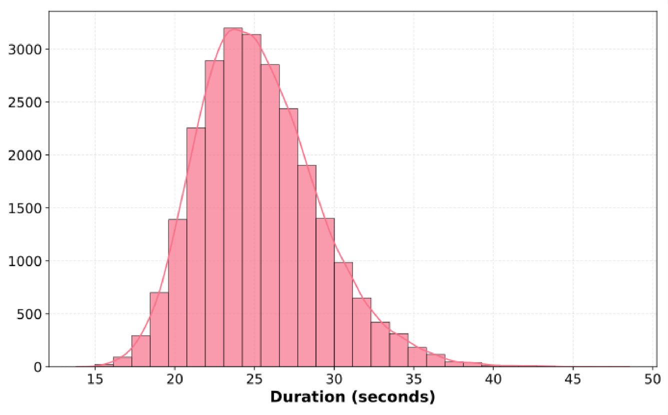
LENS-DF Data Complication Framework





- Loudness normalization: ITU P.56
- Noise augmentation is based on MUSAN, a noise dataset that contains various sound sources
- Randomized concatenation is done, followed by sequential re-segmentation
- We ignore the offset (last segment length < N seconds) during re-segmentation

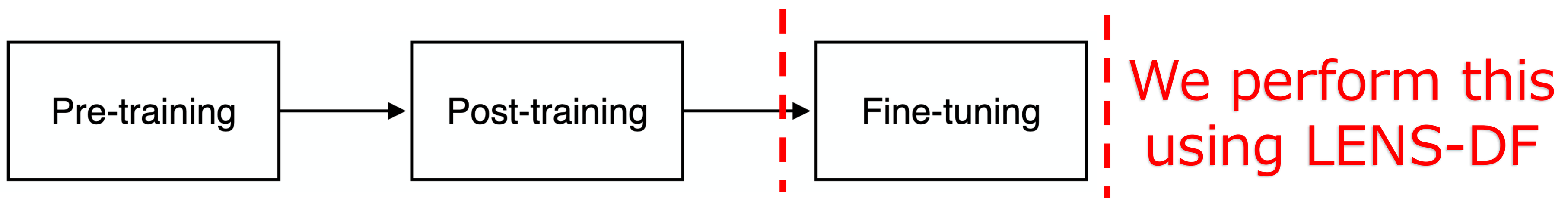




Models & Experimental Setup

- Models: MMS-300M and MMS-1B from AntiDeepfake
- Evaluation metric: Equal error rate (EER) & Half-Total error Rate (HTER, threshold tuned by development set)

AntiDeepfake { ~74,000 hours >100 languages



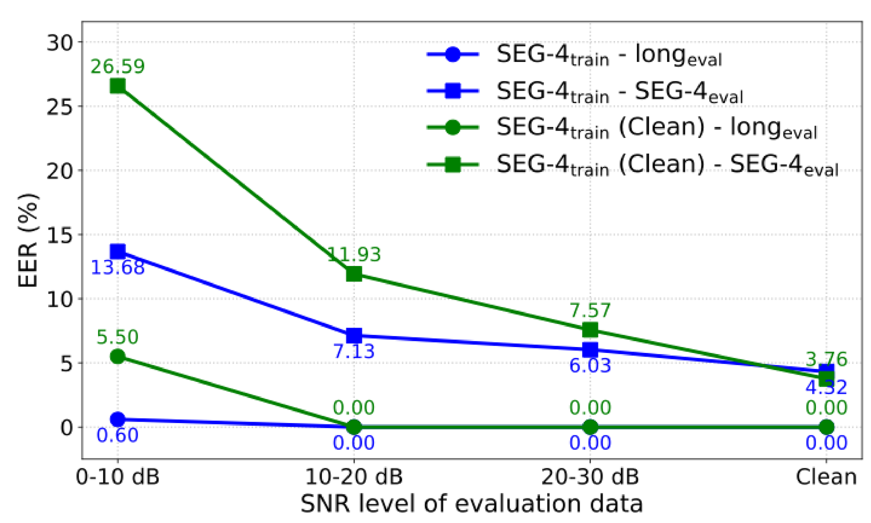
Main Results

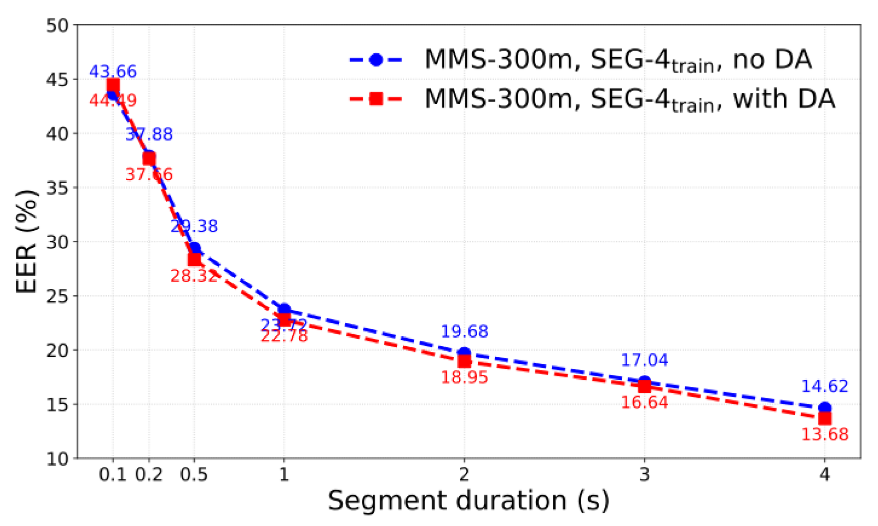
		Detection, $19LA_{eval}$		Detection, $long_{eval}$		Localization, $SEG-4_{eval}$	
SSL Model	**Training Data**	**EER (%)**	**HTER (%)**	**EER (%)**	**HTER (%)**	**EER (%)**	**HTER (%)**
MMS-1B	$19LA_{train}$	0.39	1.97	6.40	9.30	22.10	23.04
MMS-300M	$19LA_{train}$	**0.15**	**0.52**	2.90	4.05	21.12	21.09
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**

			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

- Acquiring LENS-DF generated data for training substantially improves adaptability on complex data
- Temporal localization remains as a challenging task even with LENS-DF
- RawBoost addition improves the performance of temporal localization, but not much

Ablation Study





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

- Higher noise level in SNR and shorter segments leads to bad performance
- Single speaker cannot generalize on multi-speaker cases
- Those realistic conditions & artefacts may unravel task switching/distraction of the model, which is for future work

AntiDeepfake (Github)

LENS-DF data generation (Github)

[1] W. Ge, X. Wang, X. Liu, and J. Yamagishi, "Post-training for Deepfake Speech Detection", IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Honolulu, Hawaii, USA, 2025.
[2] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Computer Speech & Language, vol. 64, p. 101114, 2020.
[3] 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.