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

Xuechen Liu, Wanying Ge, Xin Wang, Junichi Yamagishi

IEEE IJCB 2025, Osaka, Japan

2025.09.10

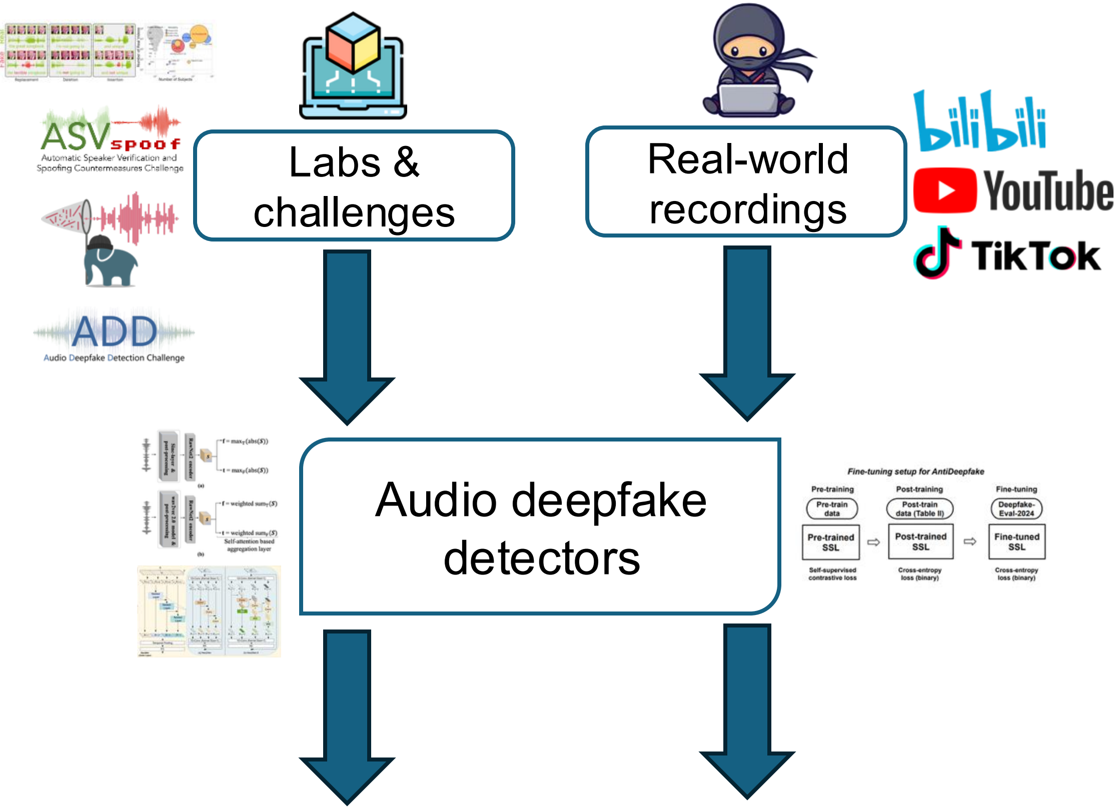
NII Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

IEEE Biometrics Council

IEEE

Our Objective

- The boost of social media and video/streaming platforms post new challenges to audio Deepfake detection
- Reigning dataset and their resulting deepfake detectors are promising on lab conditions and even challenges
- But they are mostly trained and benchmarked on **short, (largely) clean, and single speaker** audio, and they fail on real-world audios with **longer duration, noisy, and multi-speaker** audio
- We propose LENS-DF, a data complication pipeline, and investigate the adaptability and robustness of audio Deepfake detectors against various realistic factors



		min t-DCF	EER (%)
LA	[42]	0.2066	0.82
	T23	0.2176	1.32
	[44]	0.2780	3.54
	[45]	0.2882	4.66
	B03	0.3445	9.26
	[43]	0.3590	7.18
	B04	0.4257	9.50

447k[10]	N/A	**0.82 (1.00)**	2.85 (3.69)
447k	N/A	1.13 (1.36)	3.37 (4.09)
447k	N/A	0.93 (1.40)	3.56 (5.07)
511k	N/A	1.61 (1.90)	1.89 (2.12)
511k	N/A	1.73 (1.95)	1.65 (1.91)
511k	3	1.66 (1.87)	1.54 (1.98)
511k	5	1.88 (2.00)	**1.49 (1.78)**

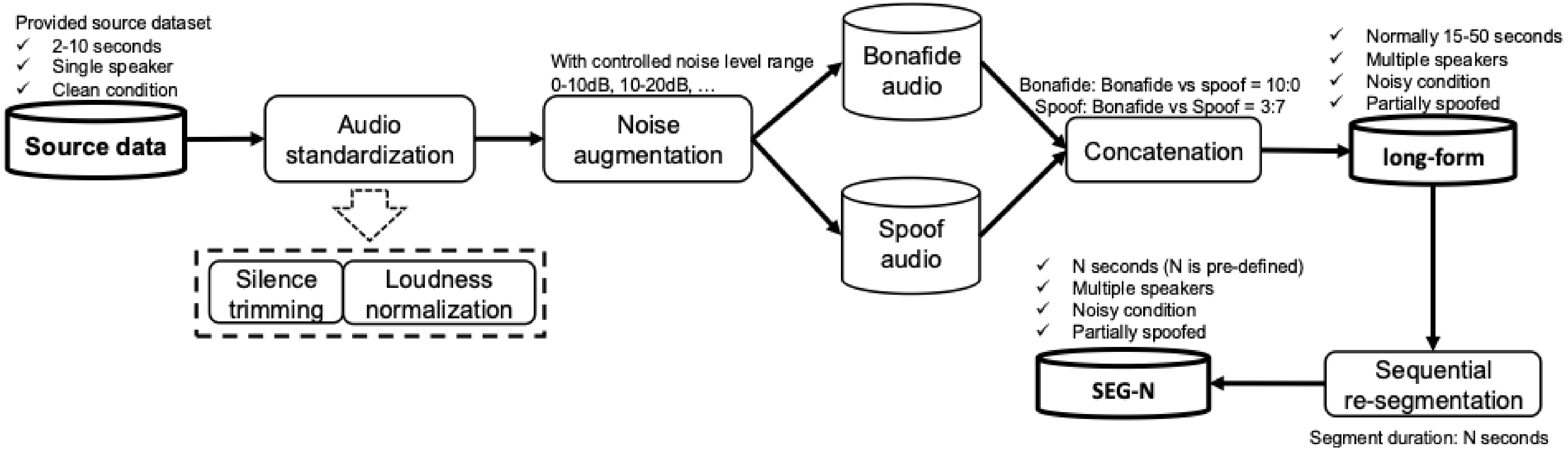
Deepfake-Eval-2024
Audio Test
34.08 / 47.72
33.33 / 31.97
33.38 / 30.05
32.80 / 31.38
27.70 / 27.55
26.76 / 26.17
27.77 / **25.78**

(Originally Sourced from YouTube)

[1] T. Liu, D. Troung, R. Das, K. Lee, and H. Li, "Nes2Net: A Lightweight Nested Architecture for Foundation Model Driven Speech Anti-spoofing", arxiv:2504.05657, 2025.
[2] X. Liu et al, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.
[3] 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.
[4] N. A. Chandra et al, "Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," arXiv:2503.02857, 2025.
[5] H. Tak, et al, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation", The Speaker Odyssey Workshop, 2022.
[6] Z. Cai et al, "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In Proceedings of the 32nd ACM International Conference on Multimedia, 2024.

NII

Data complication pipeline



- Loudness normalization follows ITU P.56 standard, we acquire a toolkit called sv56 to implement this
- Noise augmentation is based on MUSAN, a noise dataset that contains various background noises (speech, noise, music), with controllable SNR range
- Randomized concatenation followed by sequential re-segmentation (offset ignored)
- We generates long-form and SEG-N variants for training/evaluation, detection/temporal localization

[1] 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.

Detection & localization paradigm

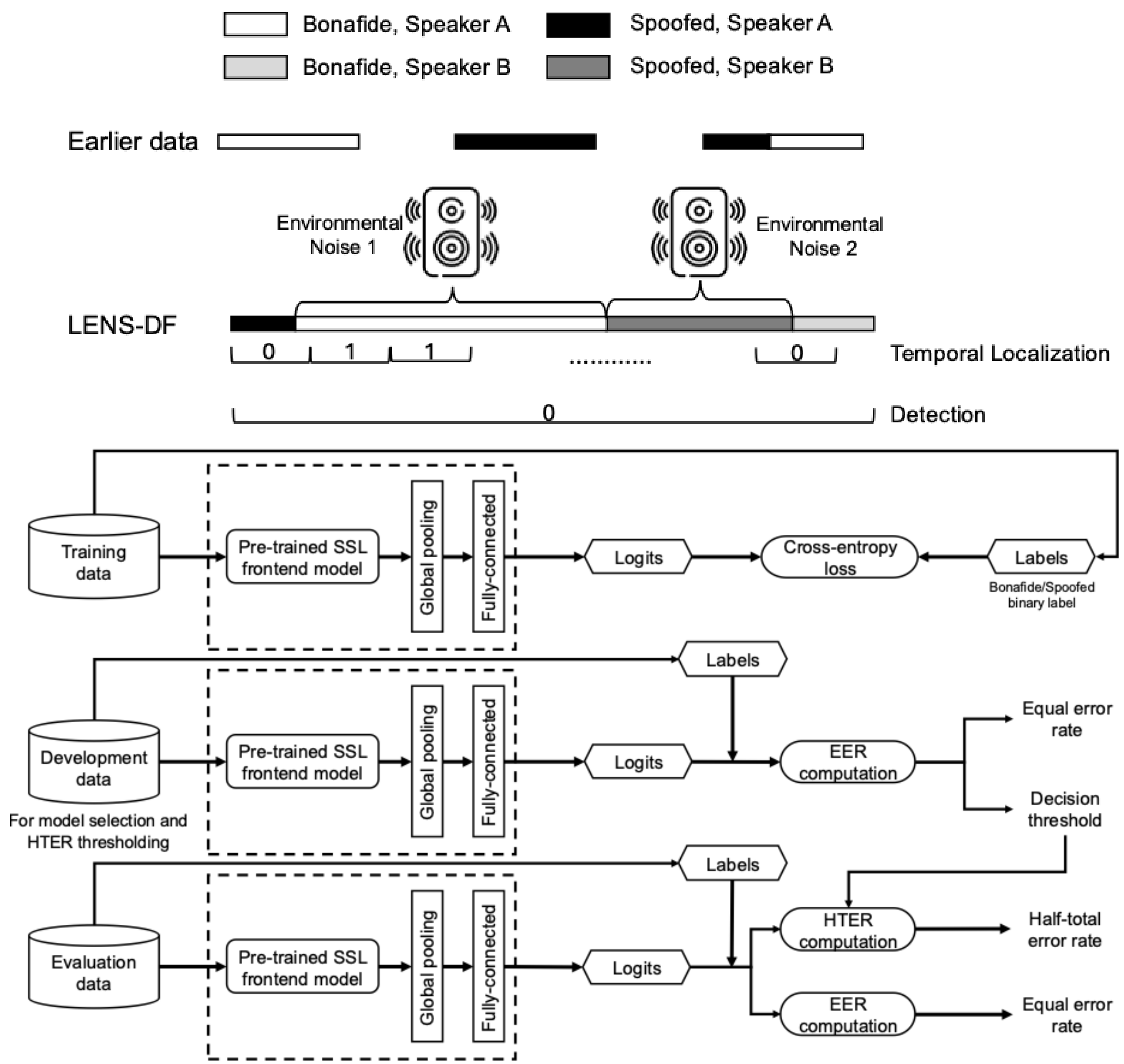
- We follow the original protocol of **ASVspoof 2019 LA** to partition the data to training, development and evaluation

- Theoretically we can generate ♾️ amount of data, while here we constraint the number for effective experimenting

- The training is done on normal audio deepfake detection paradigm, with pre-trained SSL frontend

- The development set is for model selection during training and deciding threshold to compute HTER during evaluation

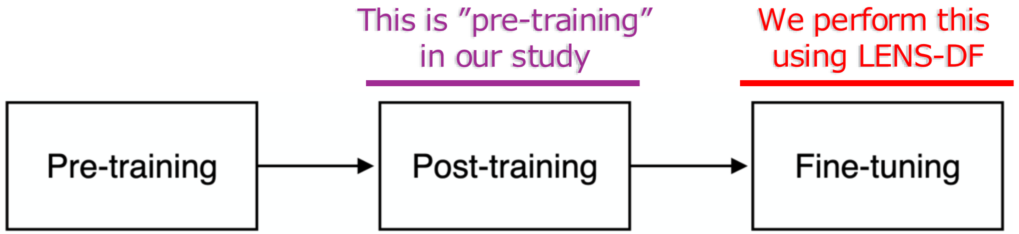
	long		SEG-4	
Partition	**Bonafide**	**Spoofed**	**Bonafide**	**Spoofed**
Train	2,580	22,800	17,857	129,805
Dev	1,000	1,000	5,132	5,640
Eval	1,000	1,000	4,984	5,663



Experimental setup

AntiDeepfake { ~74,000 hours, >100 languages

- Our audio Deepfake detector: AntiDeepfake, a large-scale model zoo with various model resources and massive training
- The models started from **pre-trained** model from Hugging Face, and has been **post-trained** on ~74K hours of specialized data in total (~56K real, ~18K fake), combining more than 100 languages
- We found applying online data augmentation does not necessarily bring better performance, so we included both strategies (NDA: no RawBoost during post-training)
- We **fine-tune** the model using generated training partition of LENS-DF



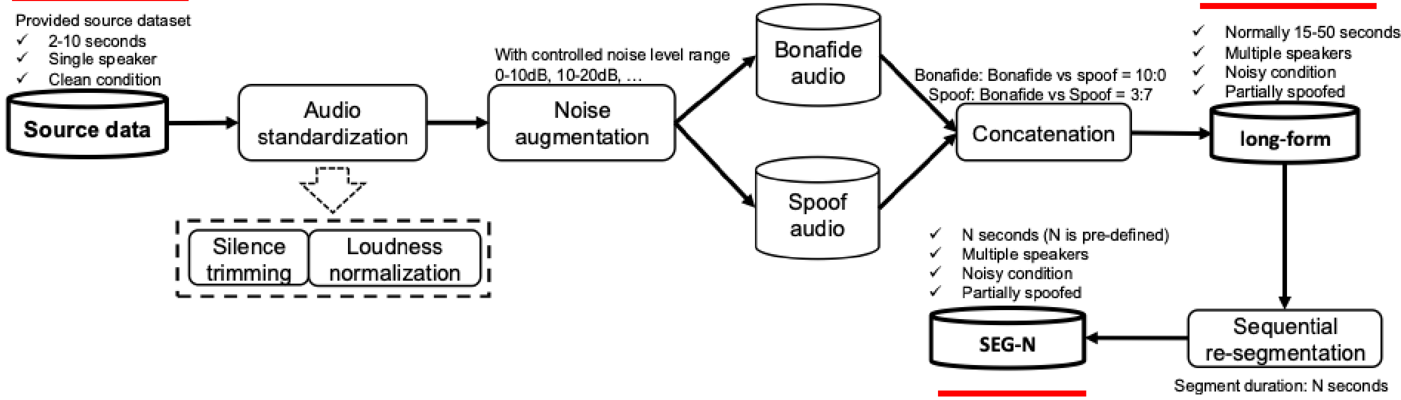
🤗 Model	Params	RawBoost	ADD2023	DEEP-VOICE	FakeOrReal	FakeOrReal-Norm	In-the-Wild	Deepfake-Eval-2024
HuBERT-XL-NDA	964M	×	35.34	14.87	3.67	15.52	17.99	47.72
W2V-Small-NDA	95M	×	19.41	16.22	1.05	6.47	4.65	31.97
W2V-Large-NDA	317M	×	12.67	5.01	0.80	1.44	2.25	30.05
MMS-300M-NDA	317M	×	11.22	3.04	0.46	2.71	2.00	31.38
MMS-1B-NDA	965M	×	9.46	2.27	0.89	1.10	1.86	27.55
XLS-R-1B-NDA	965M	×	6.58	2.96	3.16	10.91	1.36	26.17
XLS-R-2B-NDA	2.2B	×	6.84	2.63	1.18	1.73	1.31	25.78
HuBERT-XL	964M	✓	18.90	5.67	2.49	3.17	5.23	34.08
W2V-Small	95M	✓	13.02	9.80	21.94	17.85	4.24	33.33
W2V-Large	317M	✓	13.25	4.53	0.63	0.97	1.91	33.38
MMS-300M	317M	✓	7.93	2.27	1.35	5.92	2.90	32.80
MMS-1B	965M	✓	9.06	2.56	1.22	1.73	1.82	27.70
XLS-R-1B	965M	✓	5.39	2.52	5.74	12.14	1.35	26.76
XLS-R-2B	2.2B	✓	4.67	2.30	2.62	1.65	1.23	27.77

We use MMS-300M-NDA & MMS-1B-NDA

[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.

Results

- Three evaluation conditions
  - 19LA: Original 19LA evaluation data, clean
  - Long: Generated
  - SEG-4: Generated, re-segmented



- Conventional short, clean datasets are inadequate for detection on complex, realistic audio conditions. And using complex data for training helps

- Temporal localization requires further improvement even with enhanced training data

- RawBoost is helpful, not that much though

From MMS-300M-NDA

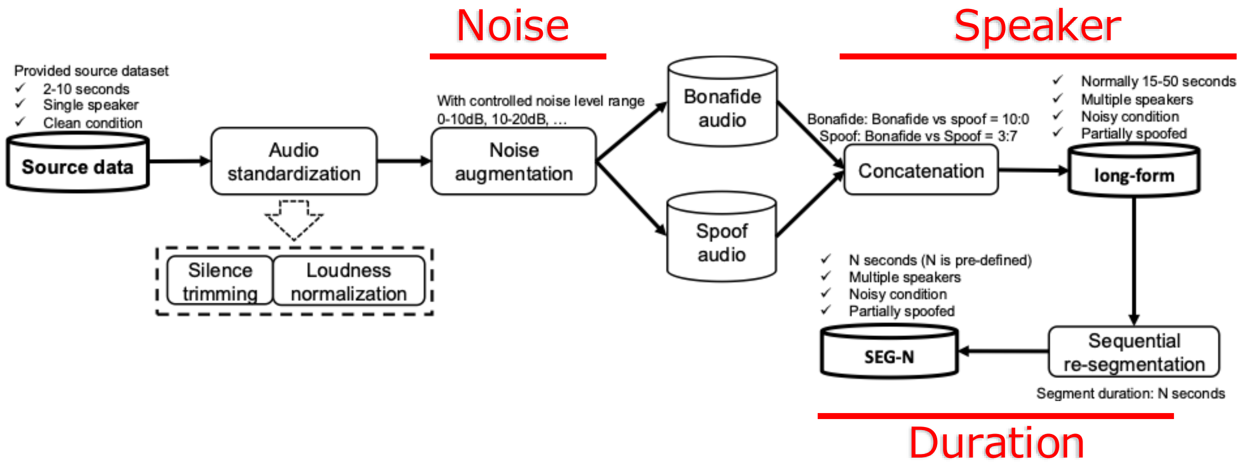
Eval ➡	Detection		Detection (Long)		Localization (SEG-4)	
Training ⬇	EER (%)	HTER (%)	EER (%)	HTER (%)	EER (%)	HTER (%)
19LA	**0.15**	**0.52**	2.90	4.05	21.12	21.09
Long	7.45	5.32	1.30	**1.40**	17.81	17.26
SEG-4	4.92	4.66	1.00	8.40	14.62	14.08
SEG-4 (w/RawBoost)	8.31	6.92	**0.60**	3.80	**13.68**	**13.52**
XLS-R-300M (earlier work)	0.19	0.94	15.70	17.60	30.41	27.30

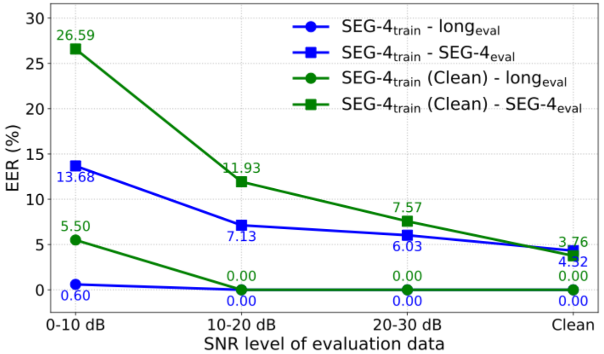
NII

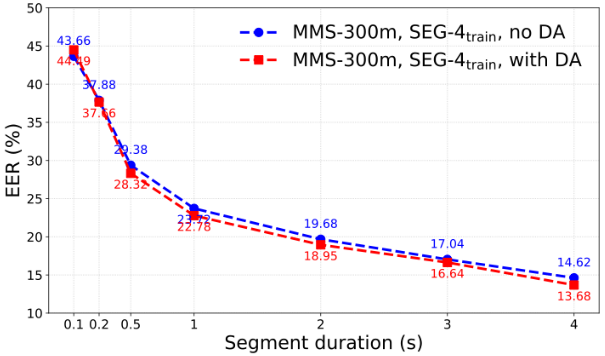
[1] Z . Cai, K. Stefanov, A. Dhall, and M. Hayat. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pages 1–10, 2022.
[2] H. Tak,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

Results

- **We perform additional analysis on noise, duration and speaker presence by varying the pipeline**
- Noise: As expected, noisy condition will create difficulties especially for localization, and that is invariant to training and evaluation variants
- Duration: Longer segments can improve temporal localization performance
- Speaker presence: Multiple vs. single speakers may cause short-cut learning so not doing well on multi-speaker cases
- Those additional artefacts may have distracted the model decision process, making the model more towards classifying something else







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

Summary

- We have proposed LENS-DF, a comprehensive data complication pipeline that real-world challenges in audio deepfake detection

- We acquire state-of-the-art audio Deepfake detectors and benchmark their adaptability against the more complicated data with more realistic distracting factors

- Training with LENS-DF improves detection performance under such more complicated conditions, including several factors that often occurs in the real-world data

- Future work will focus on more advanced model and training for temporal localization, and studying other speaker-related factors such as language

AntiDeepfake (Github)

LENS-DF data generation (Github)

Thanks for Listening!

Special thanks to all other Yamagishi Lab members and Dr. Huy H. Nguyen for helping and advising

For more queries, please visit poster #29 or email ***xuecliu@nii.ac.jp***

NII Inter-University Research Institute Corporation / Research Organization of Information and Systems National Institute of Informatics

IEEE Biometrics Council

IEEE