Minor Manipulations, Major Threat: An Overview of Partially Fake Speech

Lin Zhang@JHU, Xin Wang@NII, Erica Cooper@NICT,
Nicholas Evans@EURECOM, Junichi Yamagishi@NII

Nov. 20th, 2025

Slides by Lin Zhang
Johns Hopkins University

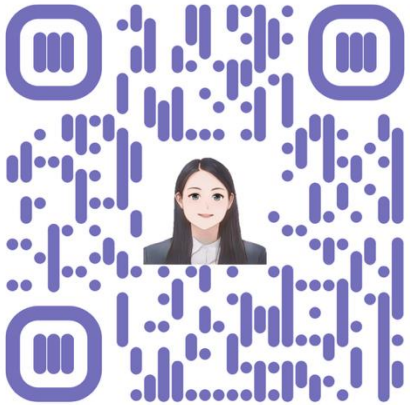
© 2025, Lin Zhang. All rights reserved.

This work is licensed under the Creative Commons Attribution 3.0 license.
See http://creativecommons.org/ for details.

Self-Introduction

Lin Zhang
琳　张

LinkedIn



Homepage

EDUCATION

Ph.D | The Graduate University for Advanced Studies (SOKENDAI) — Oct. 2020 – Mar. 2024, Tokyo, Japan
Supervisor: Prof. Junichi Yamagishi; *Mentor*: Dr. Xin Wang, Dr. Erica Cooper
Focus: Anti-spoofing (Partial Spoof), Speaker Recognition
Thesis: "Whether, When, What": Detection, Localization, and Diarization of Partially Spoofed Audio

M.Sc | Tianjin University — Sep. 2017 – Jan. 2020, Tianjin, China
Supervisor: Prof. Kiyoshi Honda and Prof. Jianguo Wei
Focus: Speaker Recognition, Speech Production, Speech Recognition
Thesis: Study on Static Individual Characteristics for Speaker Recognition

WORK EXPERIENCE

Postdoctoral Fellow — April 2025 – present
Johns Hopkins University (Dr. Nicholas Andrews, Dr. Matthew Wiesner, Dr. Leibny Paola Garcia, Prof. Sanjeev Khudanpur), Baltimore, United States
- Speech Security and Privacy

Research Worker — July 2024 – March 2025
Brno University of Technology (Prof. Lukáš Burget), Brno, Czech Republic
- Anti-spoofing, Speaker Verification, Speaker Diarization.

Research Assistant — Oct. 2020 – Mar. 2024
National Institute of Informatics (Prof. Junichi Yamagishi, Dr. Xin Wang, Dr. Erica Cooper), Tokyo, Japan
- Defined a new spoofing scenario – '**Partial Spoof**,'
- Built a database PartialSpoof with 8k+ downloading,
- Defined tasks and measurements for Partial Spoof,
- Proposed several benchmark countermeasures.

Visitor — May 2023 – Oct. 2023
Brno University of Technology (Prof. Lukáš Burget, Dr. Mireia Diez), Brno, Czech Republic
- *Best Paper* of Odyssey 2024
- Applied the variational information bottleneck to investigate the essential information required for EEND-EDA.

Research Assistant — Feb. 2020 – Aug. 2020
Duke Kunshan University (Prof. Ming Li), Jiangsu, China
- Data selection, knowledge-based feature extraction, model structure for an autism project.

Game time: Real vs. Fake

1

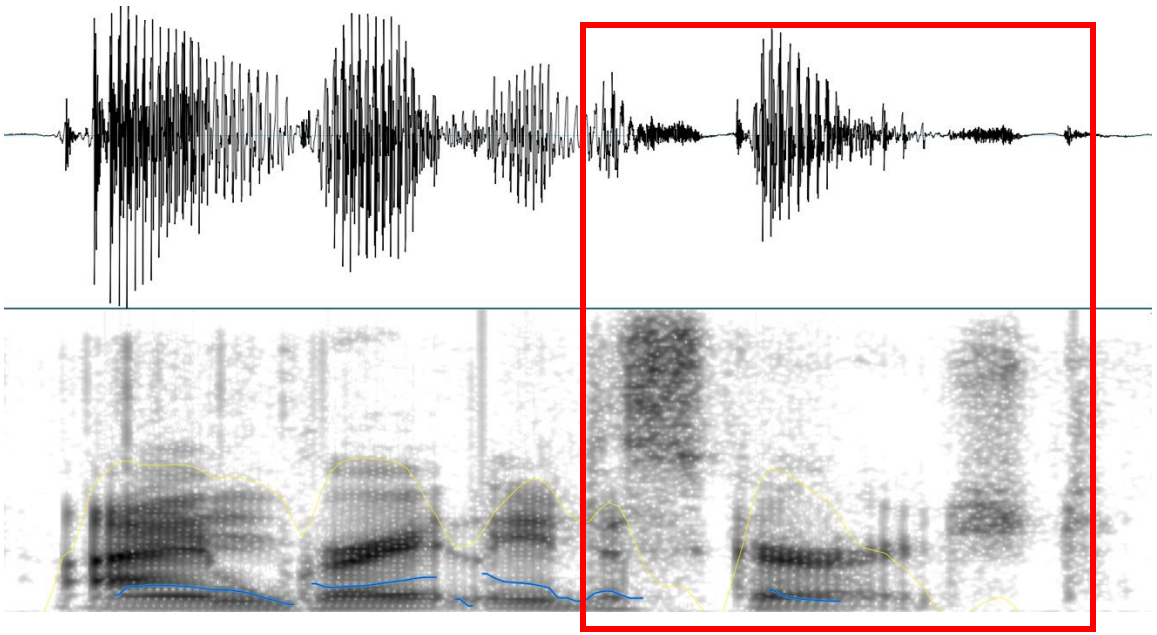
2

https://store.line.me

Game time: Real vs. Fake

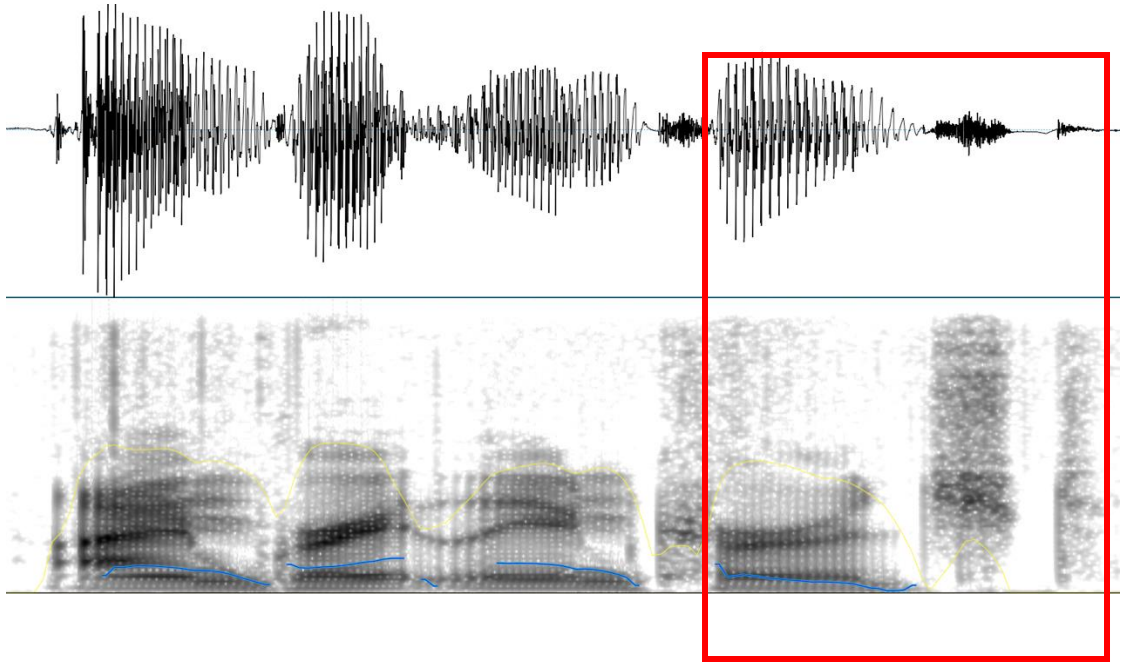
- **Is the man encouraged or discouraged?**

1



I am greatly discouraged.

2



I am greatly encouraged.

Small thing about this talk

The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance

Lin Zhang, *Student Member, IEEE*, Xin Wang, *Member, IEEE*, Erica Cooper, *Member, IEEE*, Nicholas Evans, *Member, IEEE*, and Junichi Yamagishi, *Senior Member, IEEE*

has been identified as being one of the IEEE Signal Processing Society's top 25 downloaded articles from Sept. 2023 - Sept. 2024 for IEEE/ACM Transactions on Audio, Speech, and Language Processing on IEEE Xplore®! -> IEEE SPS Webinar
Thanks for the invitation!


Xin Wang
NII


Erica Cooper
NII -> NICT


Nicholas Evans
EURECOM


Junichi Yamagishi
NII

paper


https://store.line.me

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10003971

Outline

0. Introduction

1. Database

Whether

2. Detection

When

3. Localization

What + When

4. Diarization

How

5. Analysis

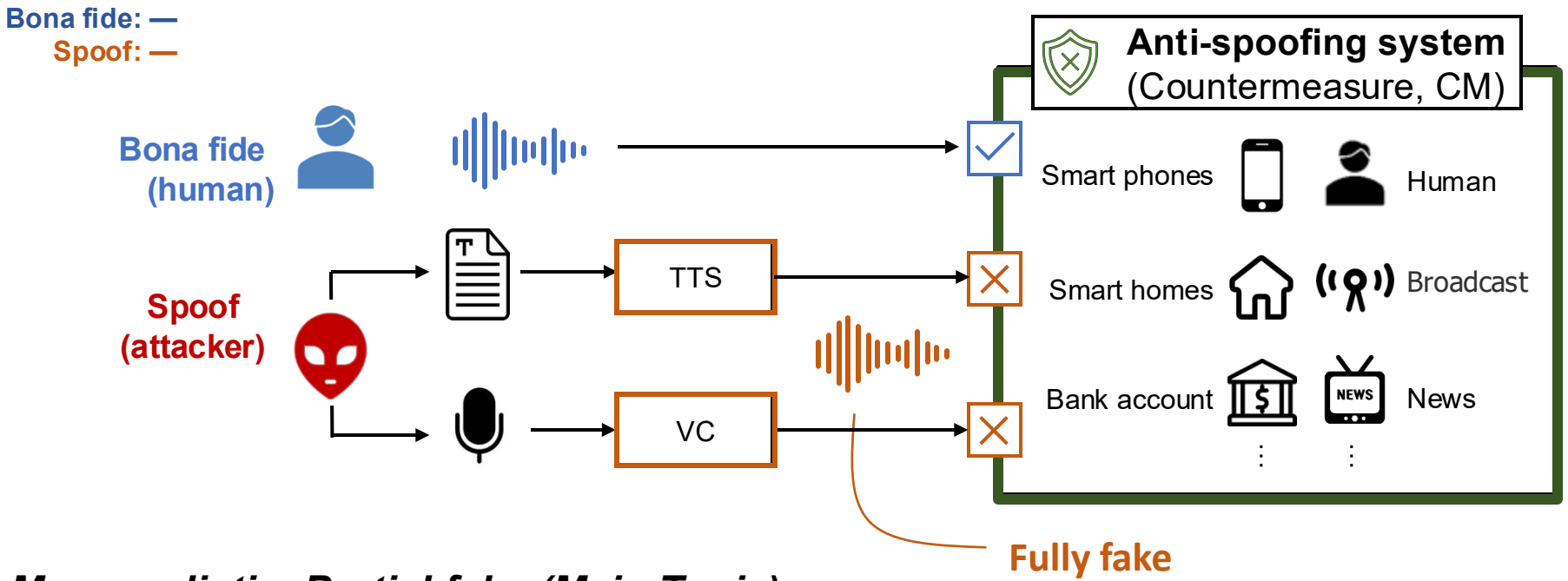
6. Summary & Open Challenges

0 Introduction

Commonly studied scenario: Fully fake

TTS: Text-to-speech

VC: Voice conversion



More realistic: Partial fake (Main Topic)

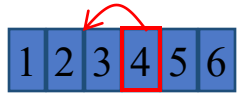
Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc j, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in Proc. Interspeech 2015, pp. 2037–2041.
T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in Proc. Interspeech, 2017, pp. 2–6.
A. Nautsch, X. Wang, etc, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021.
J. Yamagishi, X. Wang , etc, "ASVspoof2021: accelerating progress in spoofed and deep fake speech detection," in Proc. ASVspoof 2021 Workshop, 2021.

0 Introduction

Bona Fide (real)

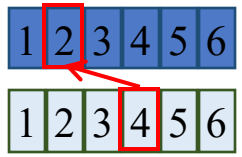
1 2 3 4 5 6 | 1 2 3 4 5 6 7

Tampering Forgery (copy-move)



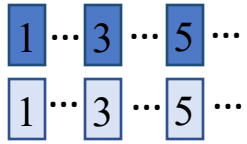
1 2 3 4 5 6 | 1 2 **4** 3 4 5 6

Tampering Forgery (splicing)



1 **4** 3 4 5

Unit-Selection



5 2 1 4 3 7

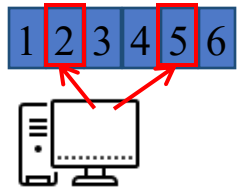
Confined to the manipulation of **existing** bona fide data

Fully Spoof

1 2 3 4 5 6 | 1 2 3 4 5 6

→ May not need to construct an entire utterance

Partial Spoof



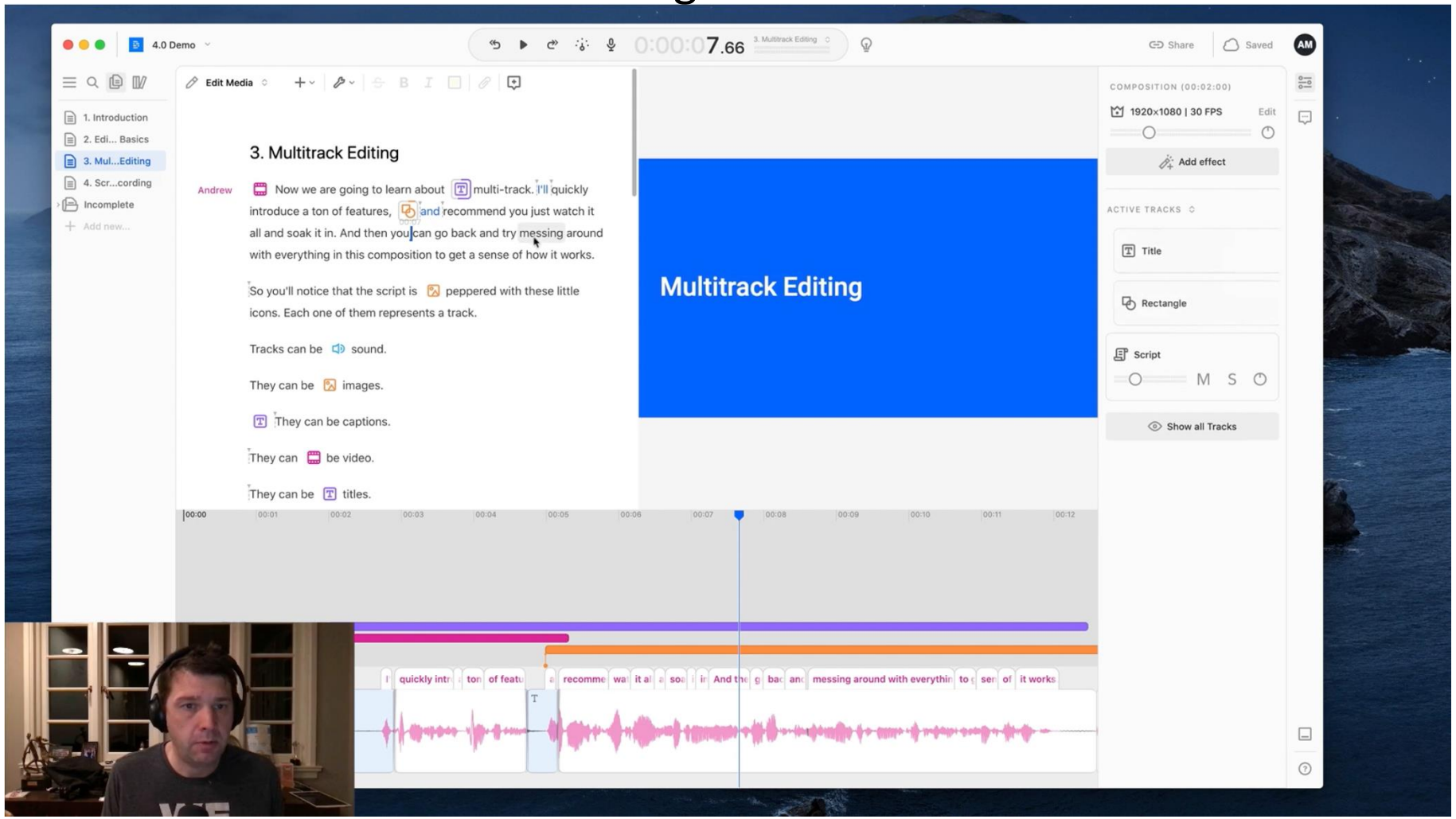
1 4 3 4 2 6 | 1 4 3 4 5 3

→ Utterances comprise audio segments that are spoken by real humans and synthesized by machines.

L. Zhang, "Whether, when, what: Detection, localization, and diarization of partially spoofed audio," Ph.D. dissertation, National Institute of Informatics, SOKENDAI, Japan, 2024.

0 Introduction

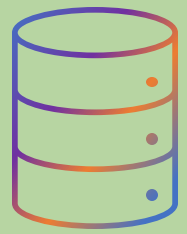
Speech Editing enables inexperienced users to modify specific segments of existing speech without re-recording the entire utterance.



Descript (demo): https://www.descript.com/overdub, 2023

Outline

0. Introduction



1. Database

Whether

2. Detection

When

3. Localization

What + When

4. Diarization

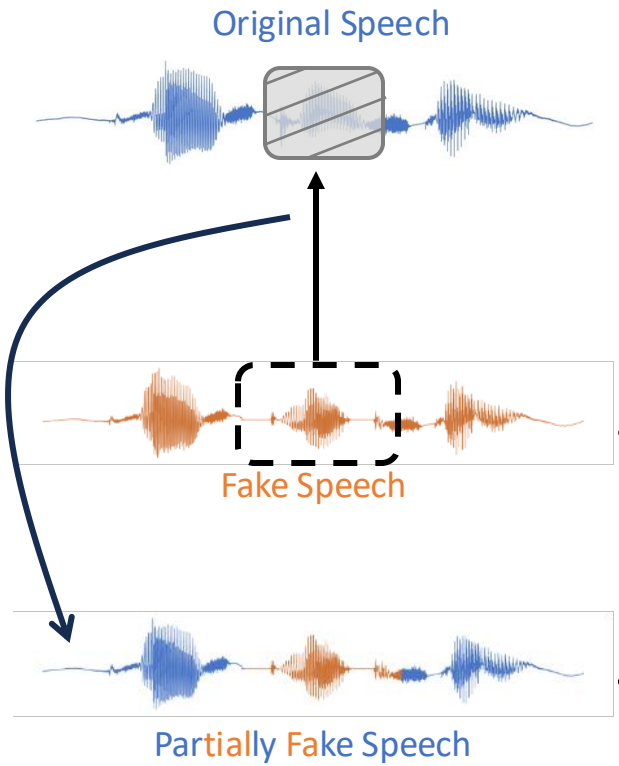
How

5. Analysis

6. Summary & Open Challenges

1 Databases

➢ How will attacker manipulate the audio?



1. Where to edit?
2. How to generate?
3. How to concatenate?

+ Different types

LLM: large language model
DSP: digital signal processing

1. Random: PartialSpoof [Zhang 2021] (VAD), Psynd [Zhang 2022], LENS-DF [Liu 2025]
2. Keyword–antonym substitution (maximal change in meaning):
   HAF [Yi 2021], ADD 2022-2023 [Yi 2022, Yi 2023], LAV-DF [Cai 2023],
3. LLM (modify meaning to an opposite direction)
   AV-Deepfake1M [Cai 2024], LlamaPartialSpoof [Luong 2024], AV-Deepfake1M++ [Cai 2025]

1. Reference database: PartialSpoof [Zhang 2021], LENS-DF [Liu 2025]
2. TTS + DSP: HAF [Yi 2021], ADD 2022-2023 [Yi 2022, Yi 2023],
   Psynd [Zhang 2022], AV-Deepfake1M [Cai 2024, 2025], LlamaPartialSpoof [Luong 2024]
3. In-filling: PartialEdit [Zhang 2025]

1. Naive Concatenate: LENS-DF [Liu 2025]
2. Signal processing: PartialSpoof [Zhang 2021], LlamaPartialSpoof [Luong 2024]
3. Pydub toolkit: HAF [Yi 2021], Psynd [Zhang 2022], ADD 2022-2023 [Yi 2022, Yi 2023]
3. In-filling: PartialEdit [Zhang 2025]

Long-form: LENS-DF [Liu 2025]
Audio–visual: LAV-DF [Cai 2023], AV-Deepfake1M, AV-Deepfake 1M++ [Cai 2024, 2025]
Competitions: ADD 2022 - 2023, AV-Deepfake1M, AV-Deepfake1M++

1 Databases

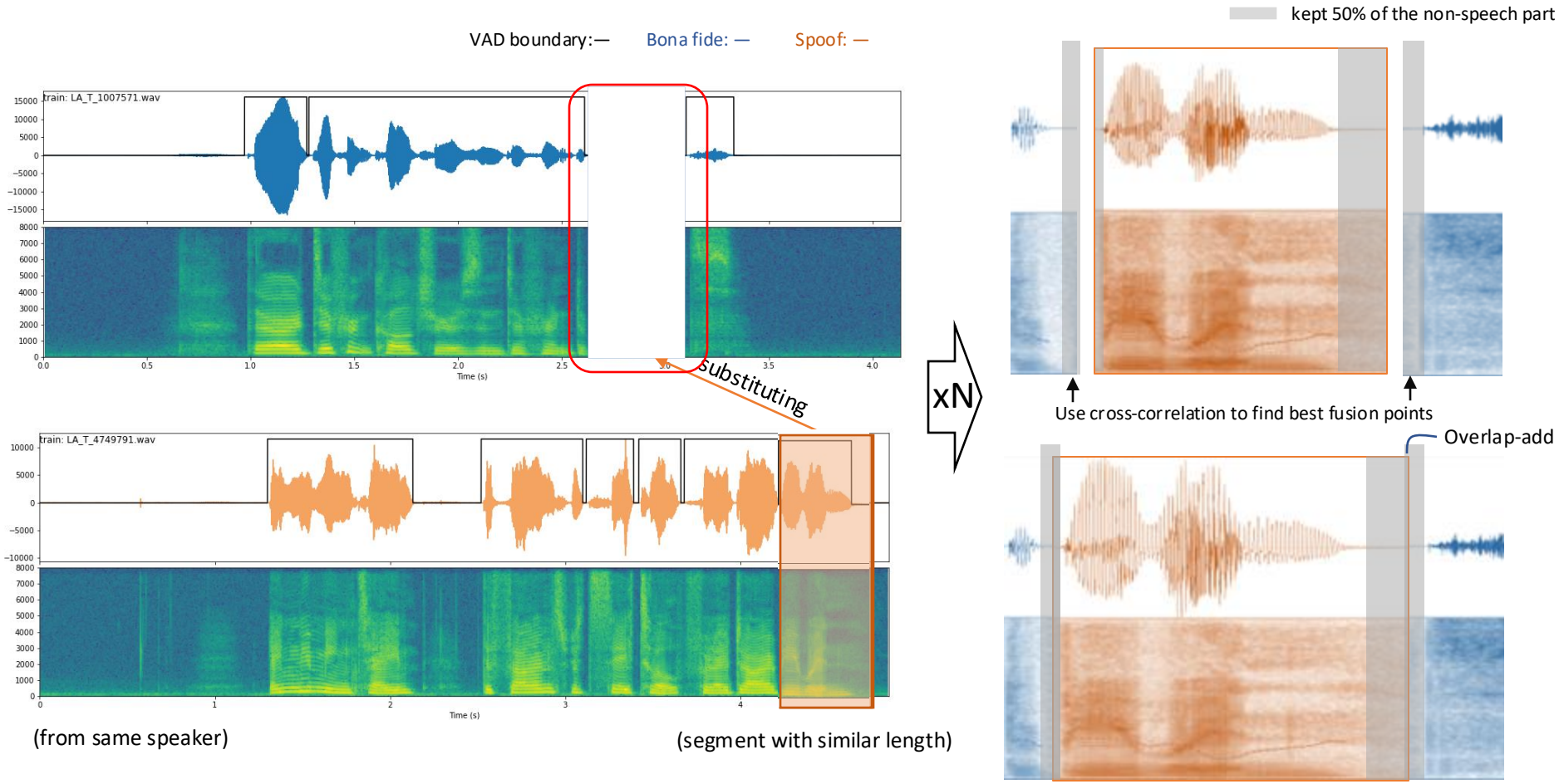
Database	Year	Utt. - bona	Utt. - spf	TTS/VC	Generation
PartialSpoof	2021	2,580/2,548/7,355	22,800/22,296/63,882	6/6/13	Random + ref. data + DSP
Half-Truth	2021	26,554/8,914/18,144	26,554/8,914/18,144	1	Keyword-antonym + TTS + DSP
ADD 2022	2022	23,897	127,414	unk.	Keyword-antonym + TTS + DSP
ADD 2023	2023	26,554/ 8,914	26,539/8,910	2	Keyword-antonym + TTS + DSP
Psynd	2022	-	1,963/94/79	1	Random + TTS + DSP
LAV-DF	2022	36,431	99,873	1	Keyword-antonym + TTS + concat.
AVDeepfake1M	2023	186,666/14,235	186,344/14,515	2	LLM + TTS + concat.
LlamaPartialSpoof	2024	10,573	32,194	6	LLM + TTS + DSP
LENS-DF	2025	2,580/1,000/1,000	22,800/1,000/1,000	6/6/13	Random + ref. data + concat.
PartialEdit	2025	44,070	43,358	3	LLM + infilling
AVDeepfake1M++	2025	297,389/20,220/310,327	801,828/57,106/564,285	4	LLM + TTS + DSP

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans. (2021) An Initial Investigation for Detecting Partially Spoofed Audio. Proc. Interspeech 2021, 4264-4268
Yi, J., Bai, Y., Tao, J., Ma, H., Tian, Z., Wang, C., Wang, T., Fu, R. (2021) Half-Truth: A Partially Fake Audio Detection Dataset. Proc. Interspeech 2021, 1654-1658
Bowen Zhang and Terence Sim. Localizing fake segments in speech. In Proc. ICPR 2022, pages 3224–3230. IEEE, 2022.
Cai, Z., Ghosh, S., Adatia, A.P., Hayat, M., Dhall, A., Gedeon, T. and Stefanov, K., 2024, October. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In Proc. *MM*
More please refer to the reference page.

1 Databases - PartialSpoof

Processing

Source database: ASVspoof 2019 LA database [Wang 2020] (Normalized)
https://www.asvspoof.org



Processing to construct PartialSpoof [Zhang 2021]

X. Wang, J. Yamagishi, M. Todisco, et al. ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech[J]. Computer Speech and Language, vol. 64, pp. 101--114, 2020

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans. (2021) An Initial Investigation for Detecting Partially Spoofed Audio. Proc. Interspeech 2021, 4264-4268

Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization

VoxCeleb2

Google Speech-to-Text service 2023

identify the tokens within D that should be replaced in order to achieve the maximum change in perceived sentiment.

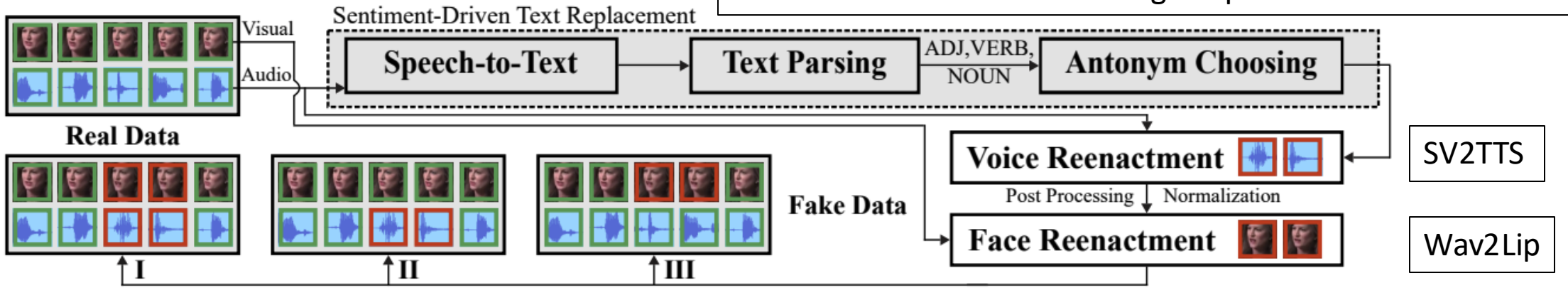


Figure 2. **Content-driven audio-visual manipulation for the creation of the LAV-DF dataset**. The real transcript is used to find the word tokens for replacement based on the largest change in perceived sentiment. Then the modified tokens are used as input for generating audio. Post-processing and normalization are applied to the generated audio to maintain loudness consistency in the temporal neighborhood. The generated audio is then used as input for facial reenactment. The green-edge audio and visual frames are real data, and red-edge are fake data. *In total, three categories of data are generated: Fake Audio and Fake Visual, Fake Audio and Real Visual and Real Audio and Fake Visual.*

Cai, Zhixi, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. "Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization." *Computer Vision and Image Understanding* 236 (2023): 103818.

AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset

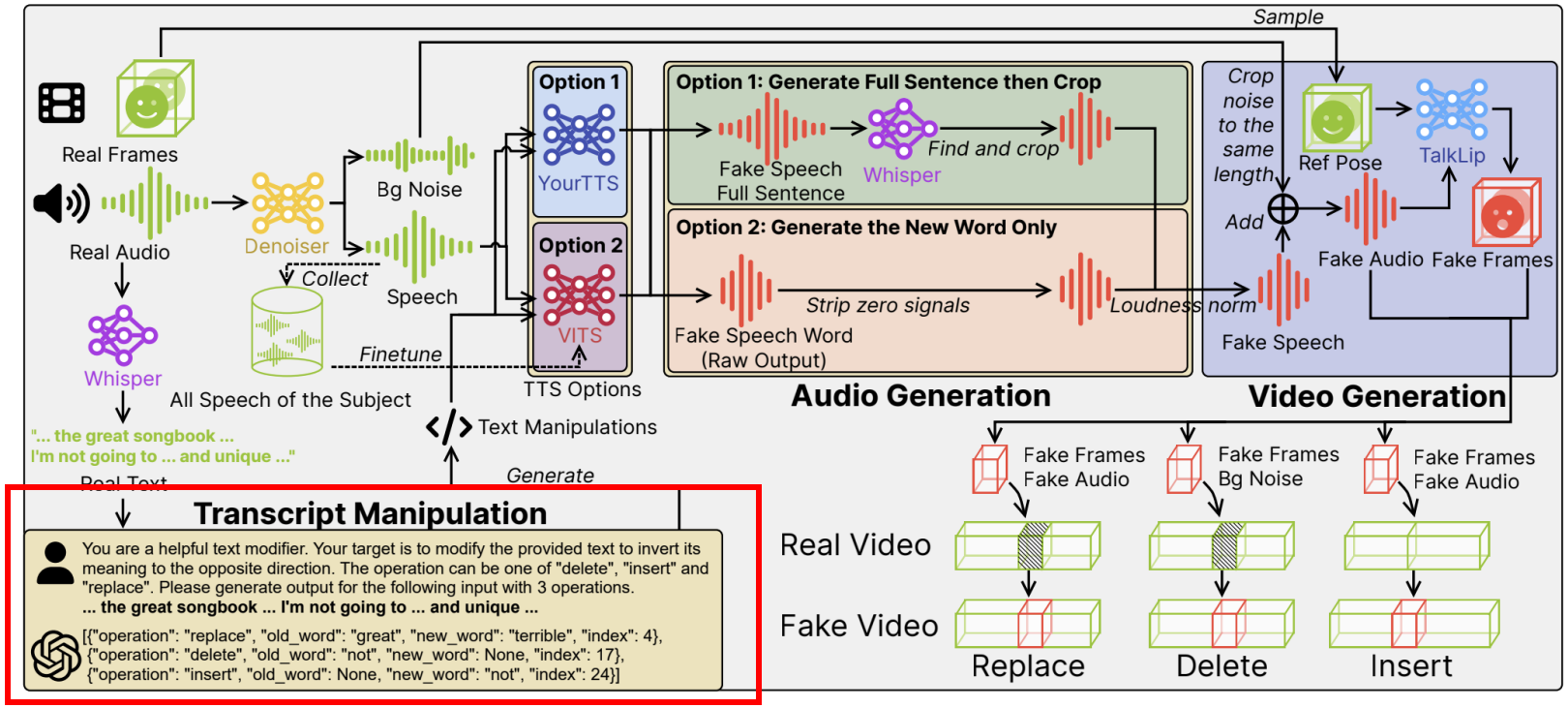


Figure 2: Data manipulation and generation pipeline. Overview of the proposed three-stage pipeline. Given a real video, the pre-processing consists of audio extraction via FFmpeg followed by Whisper-based transcript generation. In the first stage, transcript manipulation, the original transcript is modified through word-level insertions, deletions, and replacements. In the second stage, audio generation, based on the relevant transcript manipulation, the audio is generated in both speaker-dependent and independent fashion. In the final stage, video generation, based on the generated audio, the subject-dependant video is generated with smooth transitions in terms of lip-synchronization, pose, and other relevant attributes.

Cai, Zhixi, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset." In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7414-7423. 2024.

1 Databases – AV-Deepfake1M++

AV-Deepfake1M++: A Large-Scale Audio-Visual Deepfake Benchmark with Real-World Perturbations

Zhixi Cai
zhixi.cai@monash.edu
Monash University
Melbourne, Australia

Kartik Kuckreja
kartik.kuckreja@mbzuai.ac.ae
MBZUAI
Abu Dhabi, United Arab Emirates

Shreya Ghosh
shreya.ghosh@curtin.edu.au
Curtin University
Perth, Australia

et al.

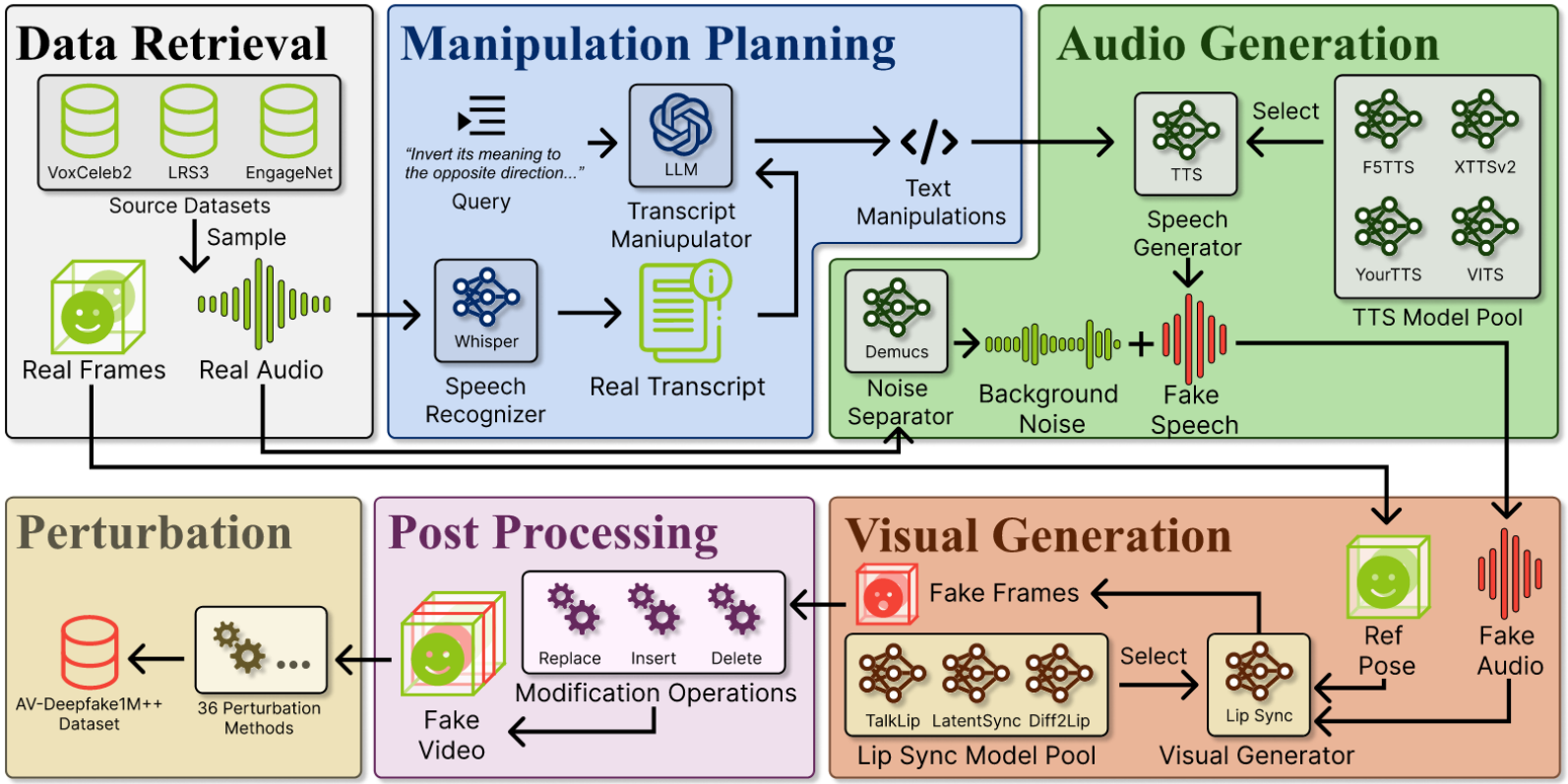


Figure 2: Data generation pipeline of AV-Deepfake1M++.

Cai, Z., Kuckreja, K., Ghosh, S., Chuchra, A., Khan, M.H., et. al. 2025. AV-Deepfake1M++: A Large-Scale Audio-Visual Deepfake Benchmark with Real-World Perturbations. *arXiv preprint arXiv:2507.20579*

LlamaPartialSpoof: An LLM-Driven Fake Speech Dataset Simulating Disinformation Generation

Hieu-Thi Luong, Haoyang Li*, Lin Zhang, Kong Aik Lee, Eng Siong Chng*

1) one or two words;
2) several words;
3) one or two nouns;
4) one or two pronouns;
5) several nouns or pronouns;
6) change the meaning to the opposite;
7) change the meaning to something else;
8) make it better;
9) make it simpler;
10) change the tone of the sentence.

TABLE I: Number of fully fake and partially fake utterances generated from each TTS model. Word Error Rates (WERs) were calculated on 5,000 sentences.

ID	Model	# of Utterances		WER (%)	
		Full	Partial	Full	Partial
TTS001	LJ JETS	5,577	5,387	2.28	3.59
TTS002	YourTTS	5,577	5,387	5.24	5.52
TTS003	XTTS V2	5,577	5,384	2.29	3.42
TTS004	GPT Sovits	5,577	5,387	4.13	4.23
TTS005	CosyVoice	5,576	5,286	3.23	3.50
TTS006	ElevenLab	5,577	5,387	**1.27**	**2.39**

Source Data: LibriTTS

TL;DR: We designed prompts based on attackers' motivation for LLM to create partially spoofed audio for evaluation.

Luong, Hieu-Thi, Haoyang Li, Lin Zhang, Kong Aik Lee, and Eng Siong Chng. "LlamaPartialSpoof: An LLM-driven fake speech dataset simulating disinformation generation." In ICASSP 2025, pp. 1-5.
Zen, Heiga, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. "LibriTTS: A corpus derived from librispeech for text-to-speech." in Proc. Interspeech 2019, 1526-1530

LlamaPartialSpoof: An LLM-Driven Fake Speech Dataset Simulating Disinformation Generation

Hieu-Thi Luong, Haoyang Li*, Lin Zhang, Kong Aik Lee, Eng Siong Chng*

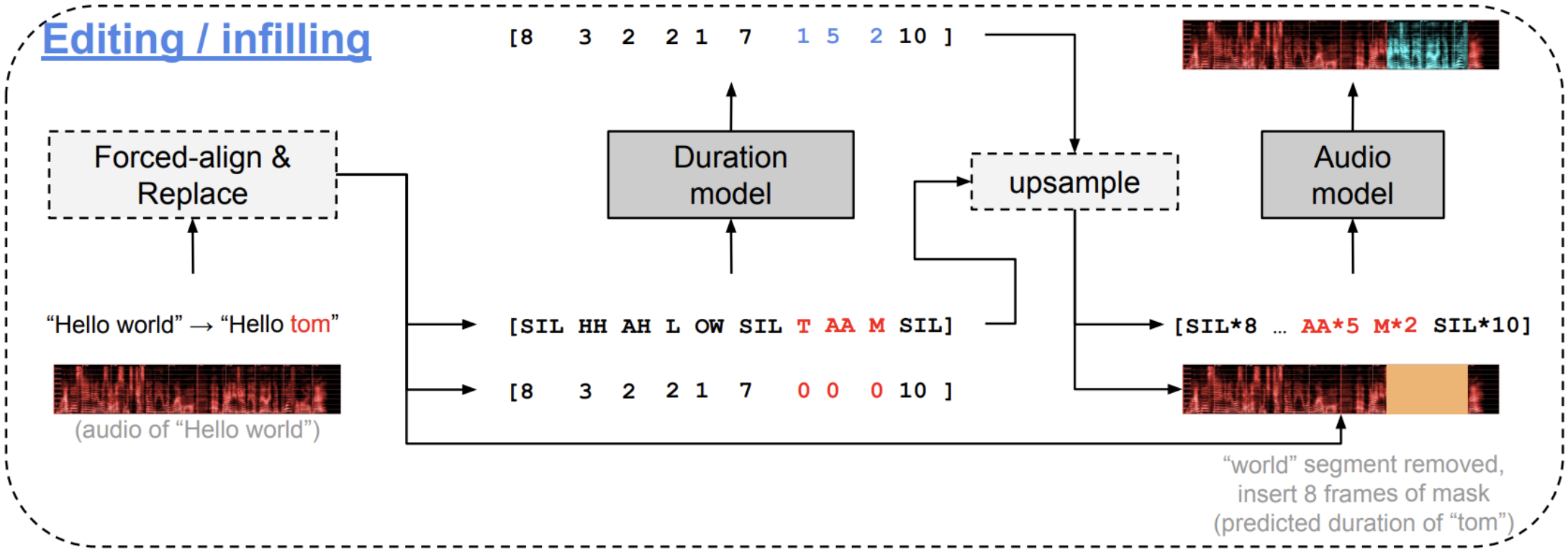
TABLE IV: Utterance-based EER (%) calculated on each TTS model subsets of LlamaPartialSpoof. The results were calculated on 5,000 sentences. Bold values indicate the lowest EER (in each column) of the particular evaluation subset.

Train Set	LJ JETS		YourTTS		XTTS V2		GPT-SoVITS		CosyVoice		ElevenLab		All	
	Full	Partial	Full	Partial	Full	Partial	Full	Partial	Full	Partial	Full	Partial	Full	Partial
A. ps-train	47.31	**17.14**	20.25	10.72	27.11	**14.90**	18.47	**13.50**	**30.62**	**16.92**	46.13	**18.37**	**32.35**	**15.33**
B. had-train	**44.93**	36.02	66.55	49.65	70.58	52.41	68.21	53.56	66.37	47.50	69.35	47.87	65.21	47.54
C. 1m-train	55.23	48.99	54.12	48.26	55.68	49.58	54.79	50.09	54.05	48.22	51.21	47.36	54.17	48.90
D. asv5-train	44.98	44.10	**0.63**	**9.03**	**24.25**	26.33	**6.60**	22.63	40.17	37.64	48.41	45.29	34.14	34.52
A+B+C	50.11	27.84	18.39	17.24	32.61	20.83	40.07	24.80	47.88	27.35	**38.75**	26.57	39.00	24.55

- Model A, trained on the PartialSpoof dataset, had the best performance on detecting partial subsets. Semantically ignorant data can still be used to train partially fake speech detection model.

Luong, Hieu-Thi, Haoyang Li, Lin Zhang, Kong Aik Lee, and Eng Siong Chng. "LlamaPartialSpoof: An LLM-driven fake speech dataset simulating disinformation generation." In *ICASSP 2025*, pp. 1-5.

Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale



Content editing in Voicebox

Le, Matthew, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson et al. "Voicebox: Text-guided multilingual universal speech generation at scale." *Advances in neural information processing systems* 36 (2023): 14005-14034.

PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing

You Zhang[*1], *Baotong Tian*[*1], *Lin Zhang*[2], *Zhiyao Duan*[1]

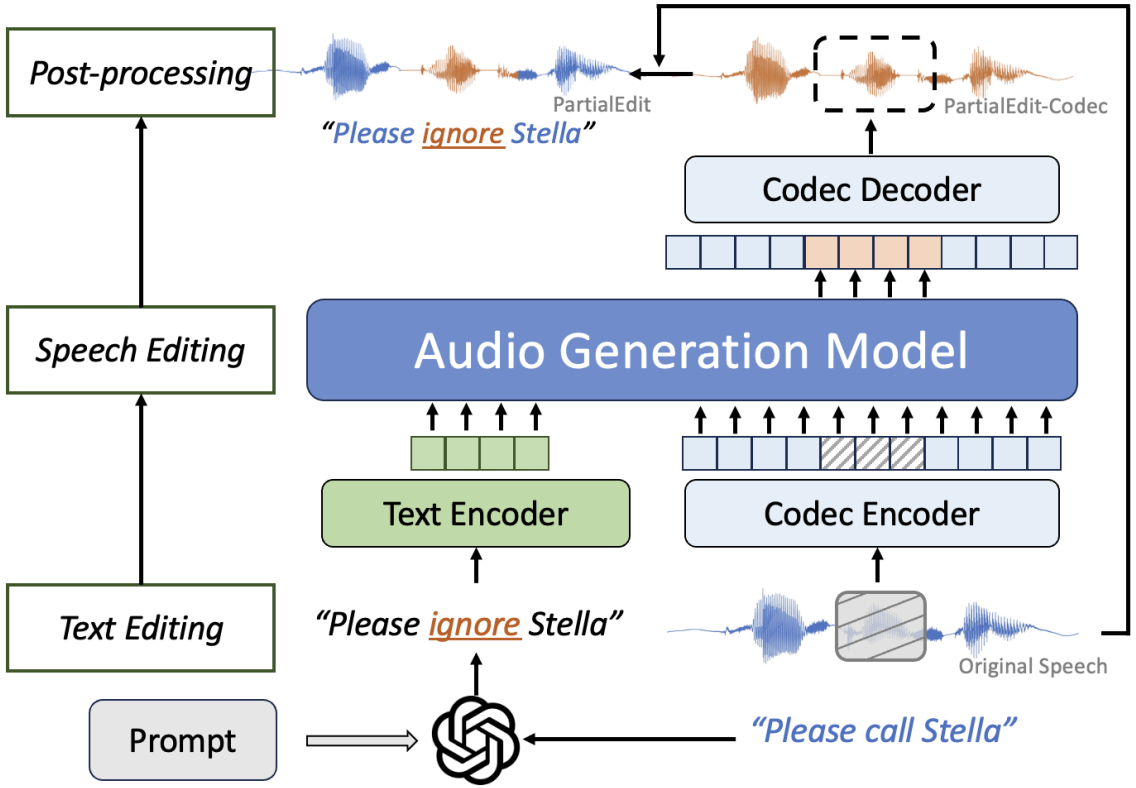


Table 1: *Duration (hours) and predicted mean opinion score (MOS) for PartialEdit and (PartialEdit-Codec). Duration report as train/dev/eval splits and shares between both versions.*

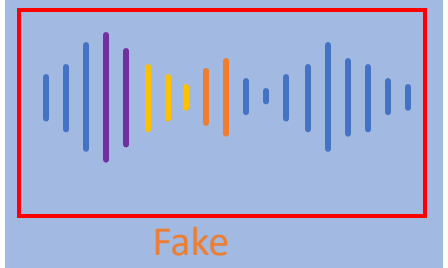
Subset	Duration (h)	MOS
VCTK [23]	7.80 / 8.18 / 25.13	3.88±0.28
VoiceCraft (E1)	8.28 / 8.06 / 27.79	3.80±0.32 (3.60±0.38)
SSR-Speech (E2)	7.82 / 7.64 / 26.26	3.83±0.30 (3.71±0.34)
Audiobox-Speech (E3)	7.94 / 7.96 / 25.69	3.90±0.32 (3.53±0.32)
Audiobox (E4)	8.14 / 7.96 / 26.44	3.90±0.33 (3.54±0.32)

Zhang, You, Baotong Tian, Lin Zhang, and Zhiyao Duan. "PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing." in Proc. *Interspeech2025*
Wang, Helin, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. "SSR-Speech: Towards Stable, Safe and Robust Zero-shot Text-based Speech Editing and Synthesis" in Proc. ICASSP, 2025
A.Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan et al., "Audiobox: Unified audio generation with natural language prompts," arXiv:2312.15821, 2023.
P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in Proc. ACL, 2024, pp. 12 442–12 462.

Outline

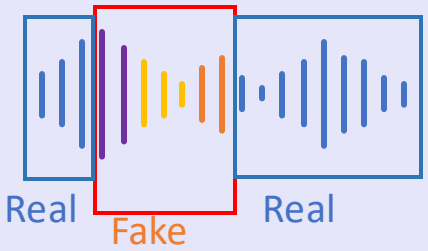
0. Introduction

1. Database



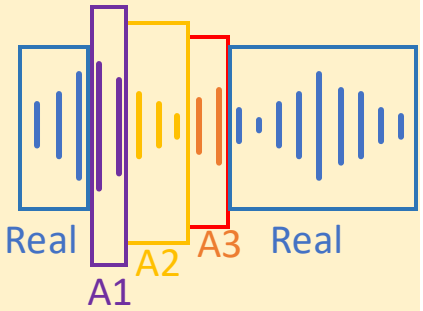
Whether

2. Detection



When

3. Localization



What + When

4. Diarization

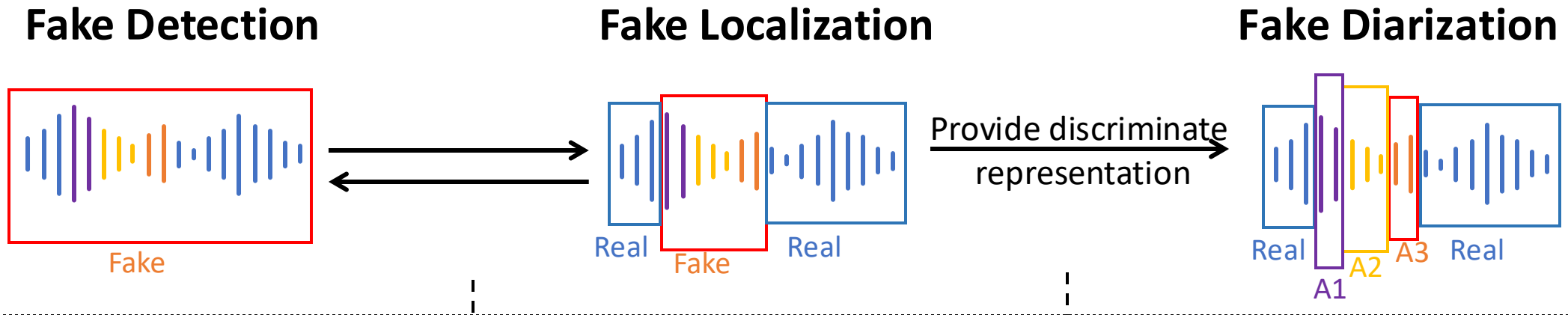
How

5. Analysis

6. Summary & Open Challenges

Outline

Tasks



	Fake Detection	Fake Localization	Fake Diarization
Equation	$\boldsymbol{x}_{1:T} \rightarrow c$ $c \in \{bonafide, spoof\}$	$\boldsymbol{x}_{1:T} \rightarrow c_{1:M}$ $c_m \in \{bonafide, spoof\}.$	$\boldsymbol{x}_{1:T} \rightarrow c^{\star}_{1:M}$ $c^{\star}_m \in \{bonafide, A_1, A_2, \cdots\}.$
Objective	Binary classification		Multi classification
Level of Analysis	Utterance level	Precise time domain (fine-grained segment level)	

Outline

0. Introduction

1. Database

Whether

2. Detection

When

3. Localization

What + When

4. Diarization

How

5. Analysis

6. Summary & Open Challenges

2 Fake Detection - Metric

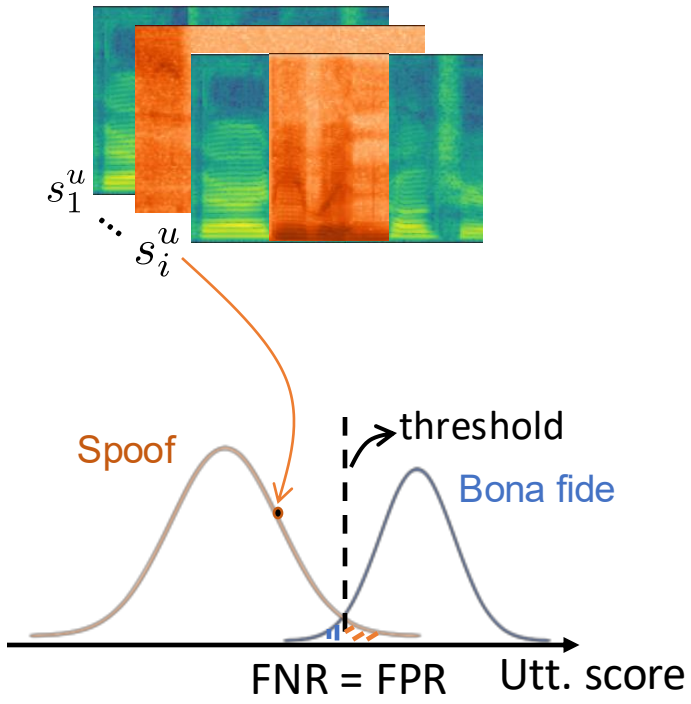
Equal Error Rate (**EER**)

(The error rate with a specific threshold where the FPR is closest to the FNR)

		Hypothesis	
		Positive (*spoof*)	**Negative** (*bona fide*)
Reference	**Positive** ($\mathcal{P}$)	TP	FN
	Negative ($\mathcal{N}$)	FP	TN

Fake Detection

- Utterance EER

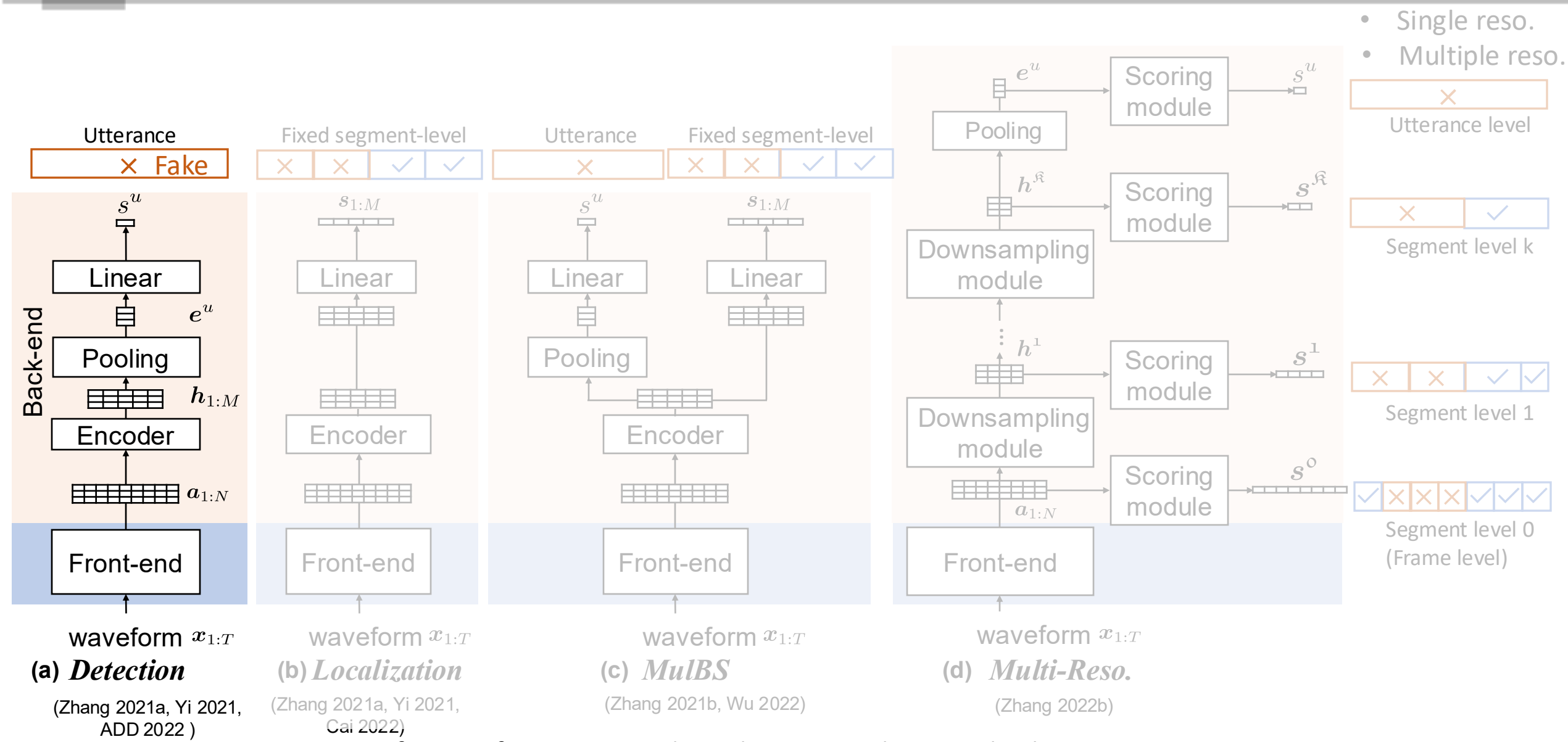


FNR: False Negative Rate; FPR: False Postive Rate

Point-based, range-based: N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," in Proc. NeurIPS 2018, 2018, p. 1924–1934.

Bona fide (1): ☑ Spoof (0): ☒

2 Fake Detection



CMs for Spoof Detection and Localization on the Partial Fake

2 Fake Detection - PartialSpoof

➢ Whether the input utterance is spoofed?

× Fake

- Spoof detection on the PS scenario is **more difficult** than on the fully spoofed scenario.
- **Fully-spoofed CM** appear to **lack generalization ability**, while CM trained on **the partially-spoofed** database is relatively **robust**.

Table 2. *EERs (%) of the cross-scenario study.*

Train	ASVspoof 2019 Dev.	ASVspoof 2019 Eval.	PartialSpoof Dev.	PartialSpoof Eval.
ASVspoof 2019	0.21	2.65	9.59	15.96
PartialSpoof	4.28	5.38	3.68	6.19

$$Spoof\_to\_Utterance\_Ratio = \frac{Spoof\_Duration}{Total\_Duration}$$

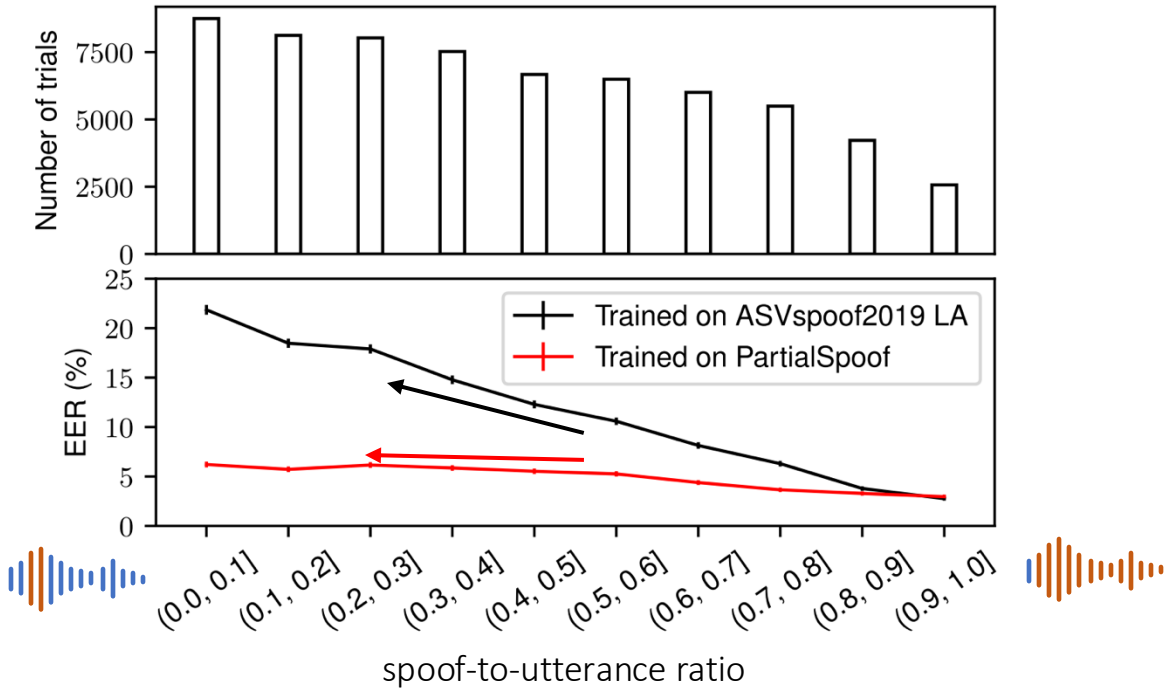


Figure 3. *Break-down results of the cross-scenario study. Top: Histogram of number of trials having different spoof segment ratio. Bottom: EERs for each of the quantized spoof*

- **Partially-spoofed CM** is more **robust** to changes in the spoof-to-utterance ratio.

LCNN-LSTM

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans. (2021) An Initial Investigation for Detecting Partially Spoofed Audio. Proc. Interspeech 2021, 4264-4268.

2 Fake Detection - PartialEdit

Table 2: *EERs (%) of deepfake detection on PartialEdit and existing databases. Each row demonstrates one system trained on a specific setting, and each column demonstrates its test results on various datasets. The same applies to the following tables.*

Data	Train \Test	I	II	III
I	PartialSpoof	2.55	12.95	23.72
II	PartialEdit-Codec	14.54	0.13	27.59
III	PartialEdit	23.06	0.41	2.14
IV	I + III	3.00	0.64	2.61

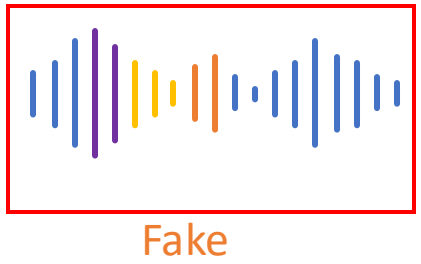
- CM trained on PartialSpoof can not handle PartialEdit, vice versa.
- Combining PartialSpoof & PartialEdit in training improves generalization, highlighting the need to address both traditional and modern deepfake techniques.
- Post-processing in PartialEdit with cutting and pasting are less detectable than those introduced by codec processing in PartialEdit-Codec

You Zhang, Baotong Tian, Lin Zhang, and Zhiyao Duan. "PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing." in Interspeech2025

2 Fake Detection

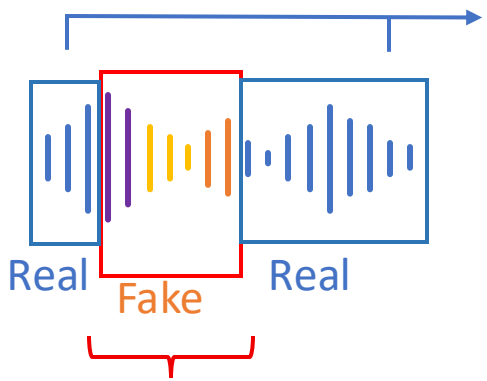
- **Fake Detection: <u>Whether</u> the input utterance is spoofed?**

  (Utterance-level)



- **Fake Localization: <u>When</u> do spoofs happen?**

  (segment-level)



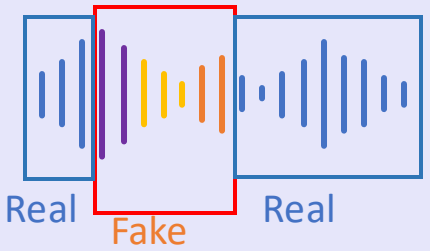
Outline

0. Introduction

1. Database

Whether

2. Detection



When

3. Localization

What + When

4. Diarization

How

5. Analysis

6. Summary & Open Challenges

3 Fake Localization - Metric

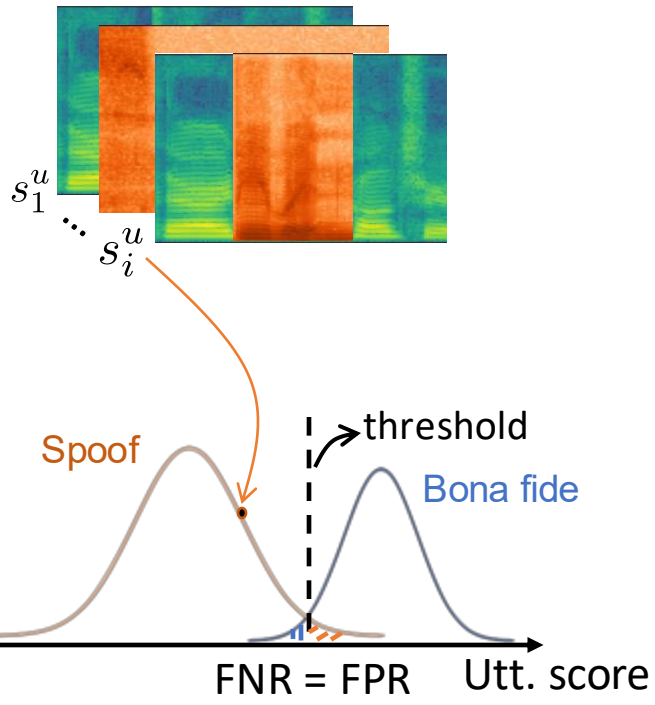
Equal Error Rate (**EER**)

(The error rate with a specific threshold where the FPR is closest to the FNR)

		Hypothesis: Positive (*spoof*)	Hypothesis: Negative (*bona fide*)
Reference	Positive ($\mathcal{P}$)	TP	FN
	Negative ($\mathcal{N}$)	FP	TN

Fake Detection

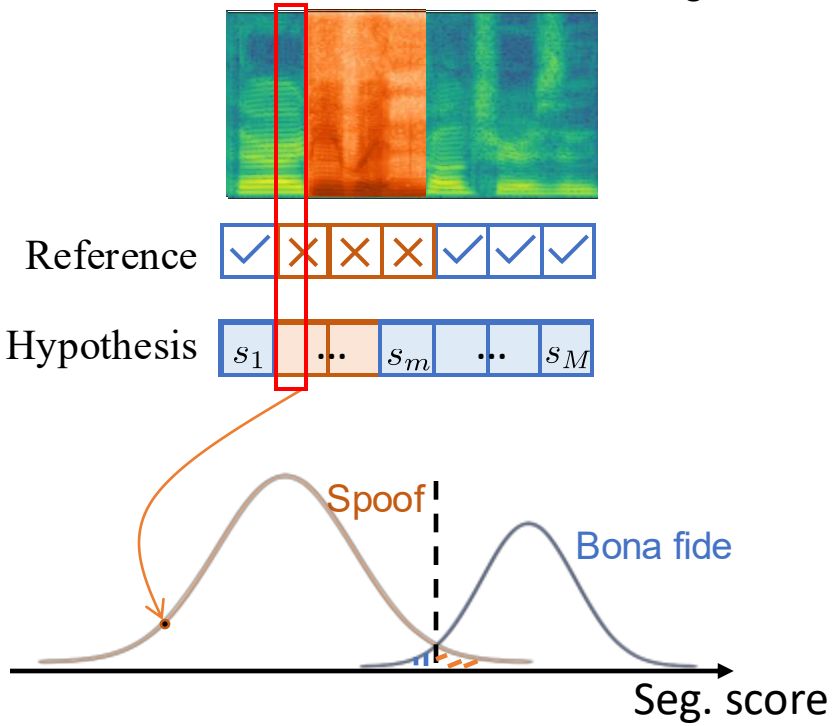
- Utterance EER



Fake Localization

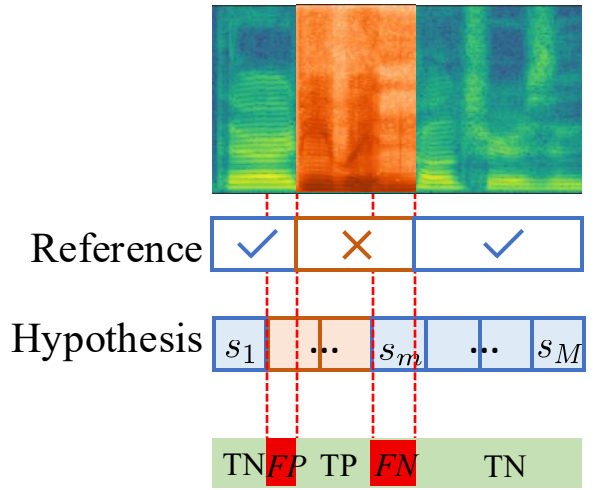
- Segment EER (Point-based)

  Number *of misclassified segments*



- Range-based EER [Zhang 2023]

  Duration *of misclassified regions*



FNR: False Negative Rate; FPR: False Postive Rate

Bona fide (1): ☑ Spoof (0): ☒

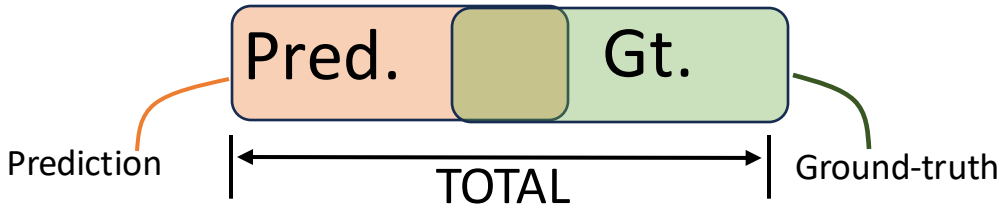
Point-based, range-based: N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," in Proc. NeurIPS 2018, 2018, p. 1924–1934.

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans and Junichi Yamagishi, (2023) Range-Based Equal Error Rate for Spoof Localization. Proc. INTERSPEECH 2023, 3212-3216, doi: 10.21437/Interspeech.2023-1214

3 Fake Localization - Metric

Other metrics

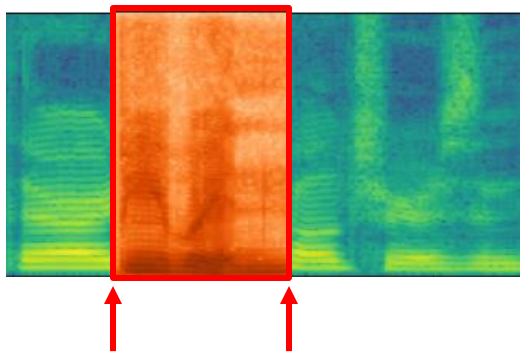
- Point-based: EER, Precision/Recall/F1 ADD [Yi 2022, Yi 2023], IoU
- Range-based:
  - RangeEER
  - Intersection over Union (IoU)



  - Average precision (AP) and average recall (AR). AV-Deepfake 1M [Cai 2023, 2025]
    - AP: measures the performance by averaging precision at different recall levels
    - AR: focuses on the recall ability at different confidence thresholds

3 Fake Localization

➢ **<u>When</u> do spoofs happen? ≈ Whether the segment(s) are spoofed?**



(1) Change points / boundaries + classification

Quantify

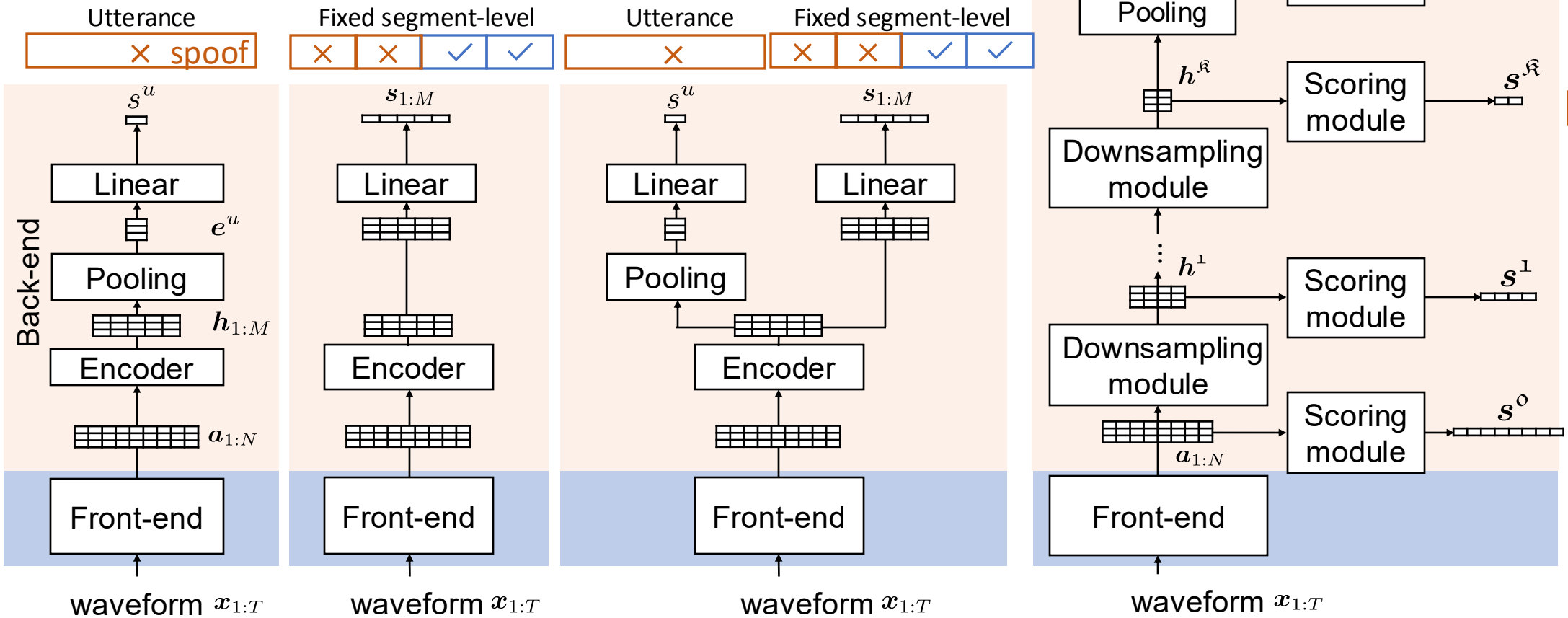
(2) Uniform segmentation

- Varying length of the segment created an additional variability into the representation and deteriorated the fidelity of the representations. [Park 2022]
- Multiple stages increase complexity.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317.

3 Fake Localization

➢ When do spoofs happen?



- Single reso.
- Multiple reso.

Utterance level

Segment level k

Segment level 1

Segment level 0 (Frame level)

(a) *Detection*

(Zhang 2021a, Yi 2021, ADD 2022)

(b) *Localization*

(Zhang 2021a, Yi 2021, Cai 2022)

(c) *MulBS*

(Zhang 2021b, Wu 2022, Cai 2024)

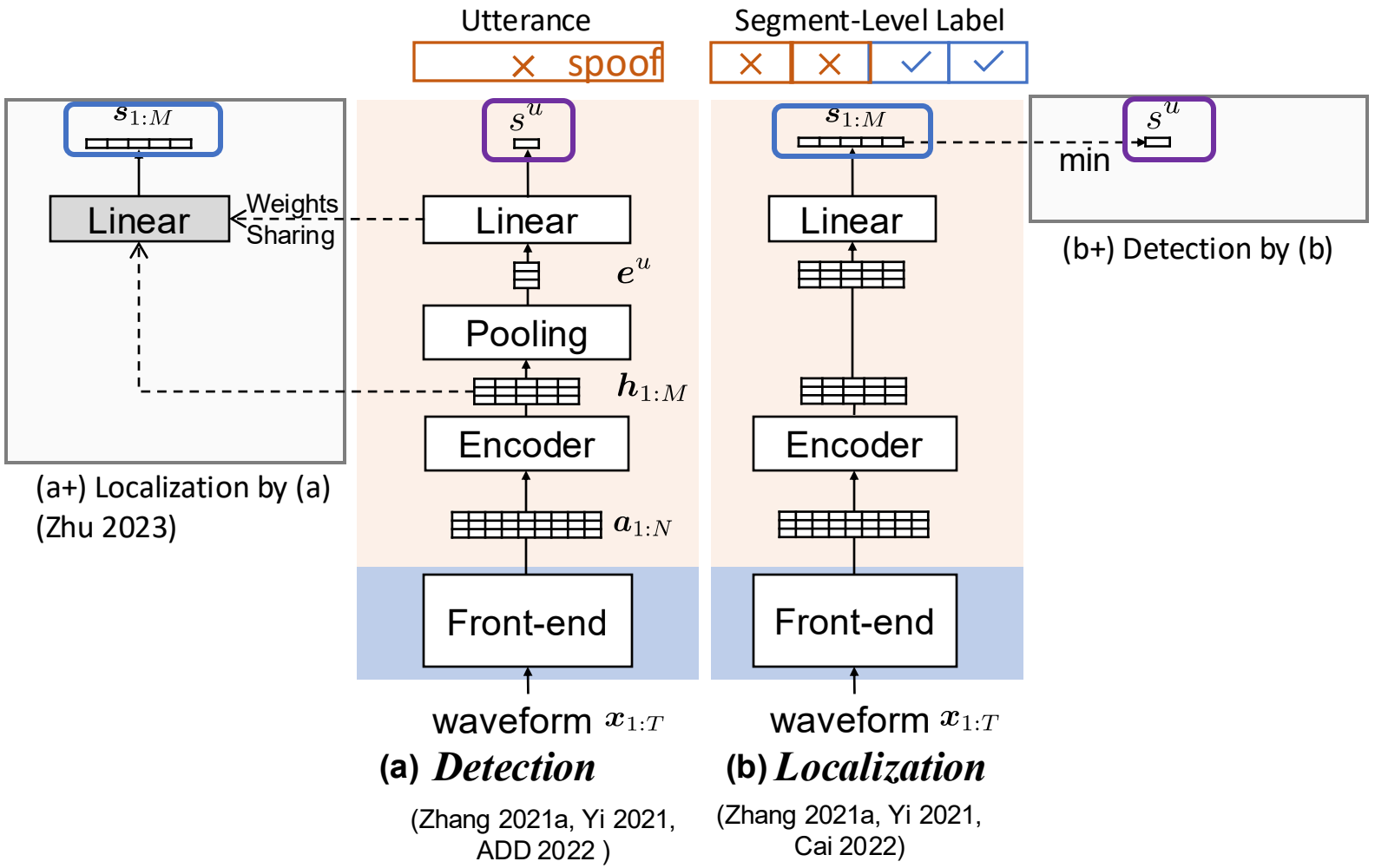
(d) *Multi-Reso.*

(Zhang 2022b)

CMs for Spoof Detection and Localization on the Partial Spoof

3 Fake Localization - PartialSpoof

Single-task trained spoof detection and localization models can predict each other's tasks **without additional labeling or training.**



RangeEER (%) for spoof localization.

Model	Train labels	Localization Dev.	Localization Eval.
Detection	Utterance	41.02	42.71
Localization	Segment	27.49	33.76

EER (%) for spoof detection.

Model	Train labels	Detection Dev.	Detection Eval.
Detection CM	Utterance	3.68	6.19
Localization CM	Segment	5.01	8.61

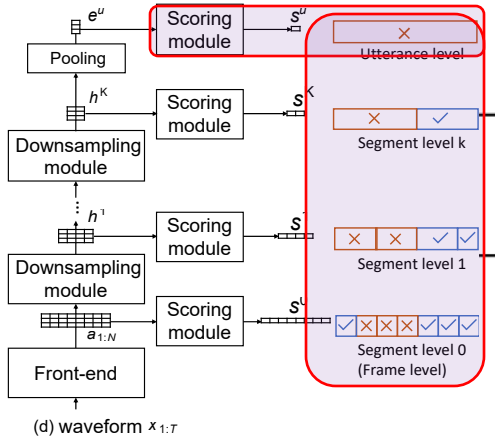
LCNN-LSTM

Yupeng Zhu, Yanxiang Chen, Zuxing Zhao, Xueliang Liu, and Jinlin Guo. 2023. Local Self-attention-based Hybrid Multiple Instance Learning for Partial Spoof Speech Detection. ACM Trans. Intell. Syst. Technol. 14, 5, Article 93 (October 2023), 18 pages
Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi. (2021) Multi-task Learning in Utterance-level and Segmental-level Spoof Detection. Proc. ASVspoof workshop 2021, 9-15, doi: 10.21437/ASVSPOOF.2021-2

3 Fake Localization - PartialSpoof

- **Database**: PartialSpoof
- **Model**: Previous slide

Performance of different CMs on the PartialSpoof evaluation set.

Sec.			Training Resolutions	Front-end	Back-end	Localization RangeEER(%)	Detection EER(%)
4.2			utt. 160 ms	LFCC	LCNN-BLSTM	42.71 33.76	6.19 8.61
5.2							
6.2			utt. 160 ms 160 ms, utt.	LFCC	SELCNN-BLSTM	42.28 33.56 33.81	6.33 7.69 5.90
7.2	*Single reso.* *Multi reso.*	d d	20 ms or utt. 20~640, utt.	w2v2-large (Baevski 2020)	5gmlp (Liu 2021)	**29.27** 30.40	0.64 **0.49**

- **Multi-resolution** CM can do detection and localization at different resolutions and **SSL-based** front-end is helpful.
- For spoof detection, training on the localization task with more fine-grained information can help improve the performance.
  - For spoof localization, training at the finer temporal resolution performs better.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS 2020, pp. 12449–12460

H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in Proc. NeurIPS 2021, pp. 9204–9215.

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans and Junichi Yamagishi, "The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 813-825, 2023, doi: 10.1109/TASLP.2022.3233236.

All experiments were repeated three times with different random seeds.

PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing

You Zhang[*1], *Baotong Tian*[*1], *Lin Zhang*[2], *Zhiyao Duan*[1]

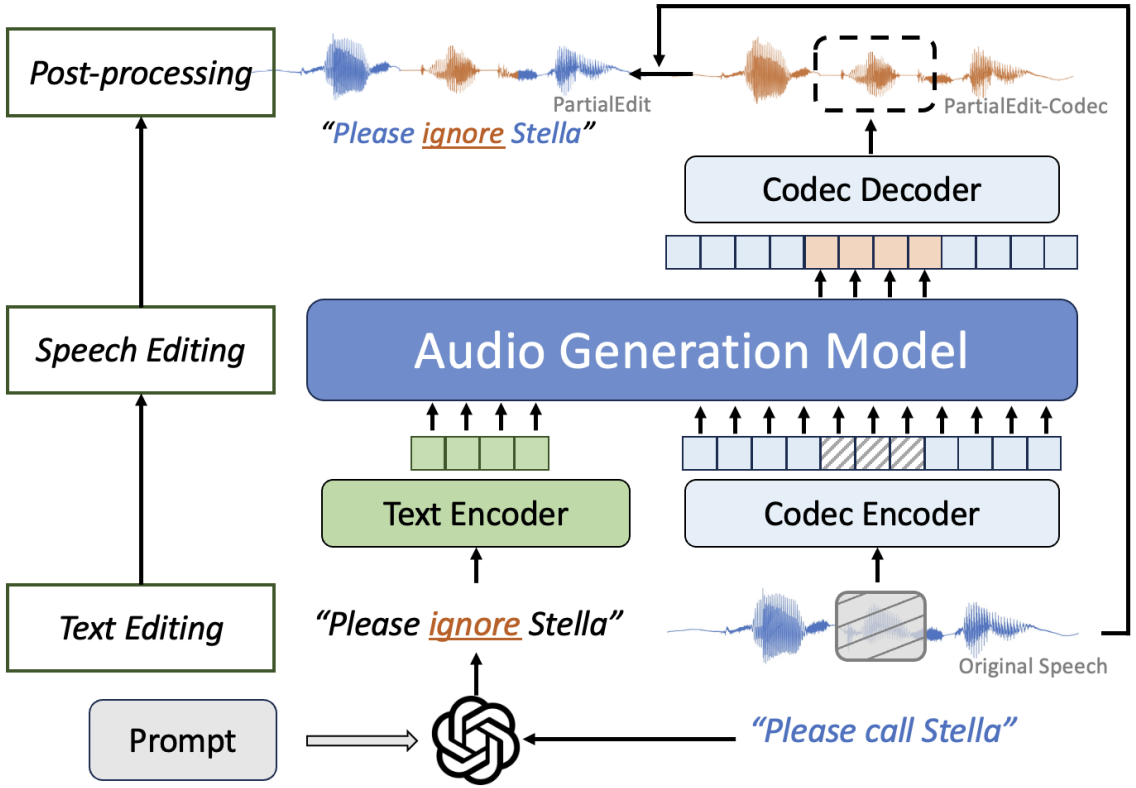


Table 1: *Duration (hours) and predicted mean opinion score (MOS) for PartialEdit and (PartialEdit-Codec). Duration report as train/dev/eval splits and shares between both versions.*

Subset	Duration (h)	MOS
VCTK [23]	7.80 / 8.18 / 25.13	3.88±0.28
VoiceCraft (E1)	8.28 / 8.06 / 27.79	3.80±0.32 (3.60±0.38)
SSR-Speech (E2)	7.82 / 7.64 / 26.26	3.83±0.30 (3.71±0.34)
Audiobox-Speech (E3)	7.94 / 7.96 / 25.69	3.90±0.32 (3.53±0.32)
Audiobox (E4)	8.14 / 7.96 / 26.44	3.90±0.33 (3.54±0.32)

Zhang, You, Baotong Tian, Lin Zhang, and Zhiyao Duan. "PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing." in Proc. *Interspeech2025*
Wang, Helin, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. "SSR-Speech: Towards Stable, Safe and Robust Zero-shot Text-based Speech Editing and Synthesis" in Proc. ICASSP, 2025
A.Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan et al., "Audiobox: Unified audio generation with natural language prompts," arXiv:2312.15821, 2023.
P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in Proc. ACL, 2024, pp. 12 442–12 462.

3 Fake Localization - PartialEdit

- **Database**: PartialEdit
- **Model**: BAM

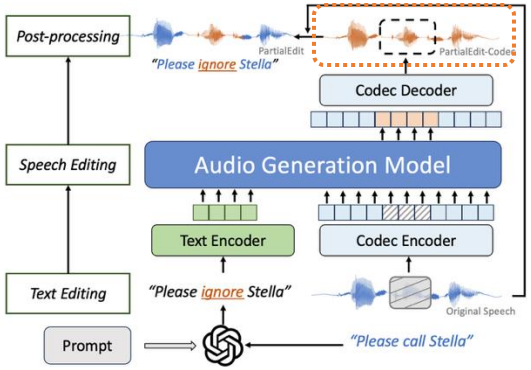
Table 3: *Frame-level EERs (%) of localization with cross-algorithm evaluation on different editing algorithms of PartialEdit.*

Train \Test		E1	E2	E3	E4	PartialEdit
VoiceCraft	(E1)	3.80	3.61	6.79	7.75	7.10
SSR-Speech	(E2)	6.50	3.57	9.17	9.33	9.51
Audiobox-Speech	(E3)	22.35	20.86	0.11	0.14	15.26
Audiobox	(E4)	16.32	15.32	0.17	0.11	11.77
PartialEdit		4.07	3.30	0.18	0.16	2.77

- CMs perform best when trained on data that match the test data.
  - The diversity of the training data plays a crucial role.

Zhang, You, Baotong Tian, Lin Zhang, and Zhiyao Duan. "PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing." in Proc. *Interspeech2025*
J. Zhong, B. Li, and J. Yi, "Enhancing partially spoofed audio localization with boundary-aware attention mechanism," in Proc. Interspeech, 2025.

3 Fake Localization - PartialEdit



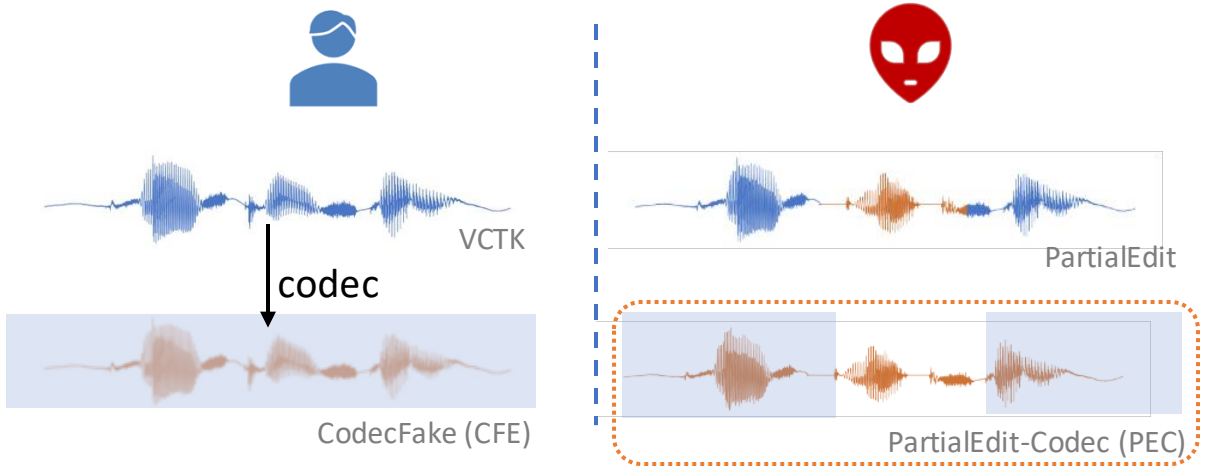


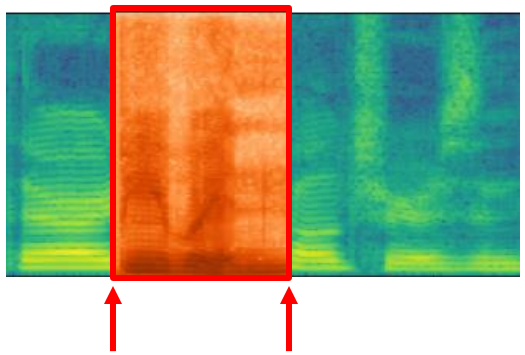
Table 4: *Comparison of localization EER (%) on PartialEdit-Codec with different training settings. △ indicates datasets used as bona fide, while ◯ represents datasets used as deep-fake. (CFE: CodecFake-Encodec; PEC: PartialEdit-Codec)*

	Train on				Test on	
	VCTK	CFE	PEC	PartialEdit	I	II
I	△			◯	3.57	47.14
II		△	◯		10.73	5.30

Artifacts introduced by codec processing may mislead the model when they are not recognized as bona fide during training.

3 Fake Localization

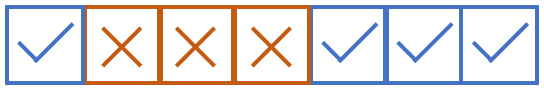
➢ When do spoofs happen?



(1) Change points / boundaries + classification

Quantify

(2) Uniform segmentation



- **Uniform Segmentation**
  - LCNN [Zhang 2021, Yi 2021], Multi-reso. [Zhang 2022],
  - Vigo [Vieites 2024], MFMS [Zhang 2024]
- **Boundary-aware**
  - BDR [Cai 2023]
- **Uniform segmentation + Boundary-aware**
  - BAM [Zhong 2024] (KLASSify [2025]), CFPRF [Wu 2024],
  - BFC-Net [Zhou 2025], Pindrop [2025]
- **Segment-wise comparison (inconsistency):**
  - TDL [Xie 2024], PET [He 2025]
- **Multi-modal**
  - Temporal action localization
  - BA-TFD [Cai 2023], UMMAFormer [Zhang 2023], Vigo [Vieites 2024],
  - MFMS [Zhang 2024], Pindrop [2025], etc.

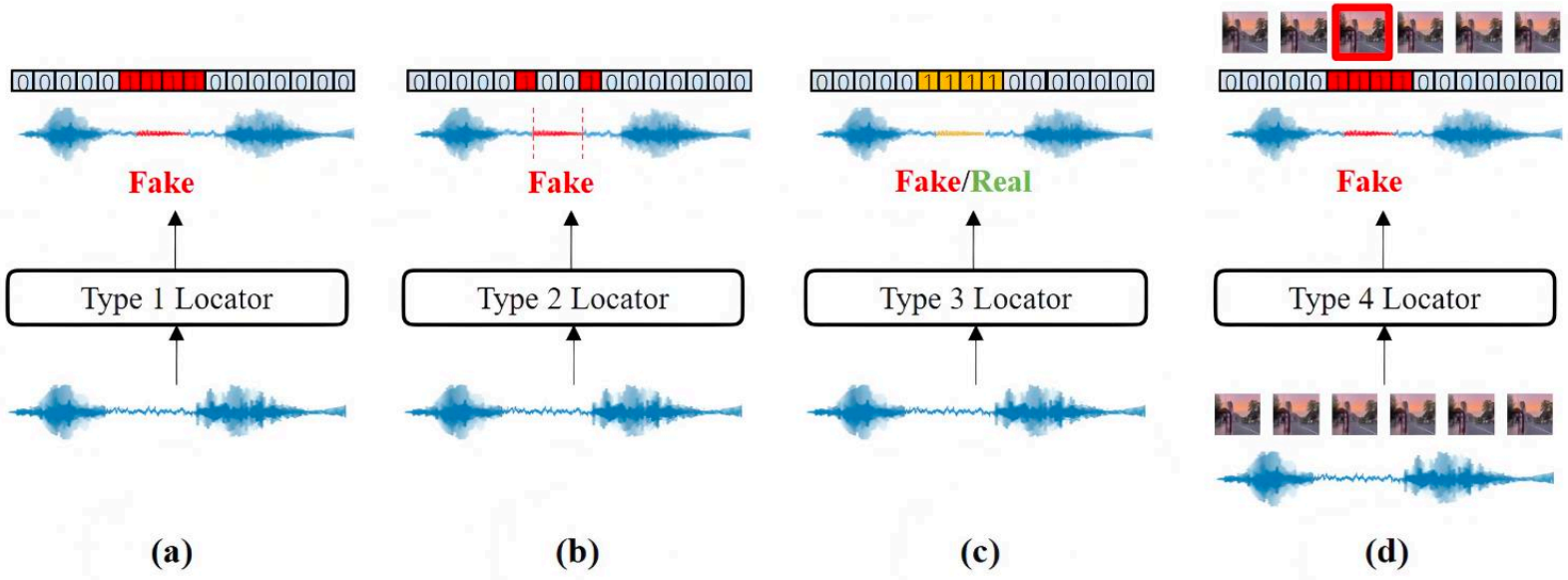
BAM: J. Zhong, B. Li, and J. Yi, “Enhancing partially spoofed audio localization with boundary-aware attention mechanism,” in Proc. Interspeech, 2025.
BFC-Net: Zhou, Y., Xue, Z., Senhadji, L., Shu, H. and Wu, J., 2025. BFC-Net: Boundary-Frame cross graph attention network for partially spoofed audio localization. *Neurocomputing*, p.130867.
PET: He, J., Yi, J., Tao, J. and Zeng, S., 2025, April. PET: High-Frequency Temporal Self-Consistency Learning for Partially Deepfake Audio Localization. In *ICASSP 2025*
Pindrop: Klein, Nicholas, Hemlata Tak, et al. "Pindrop it! Audio and Visual Deepfake Countermeasures for Robust Detection and Fine-Grained Localization." In Proc. ACM Multimedia 2025, pp. 13700-13706
KLASSify: Ivan Kukanov, and Jun Wah Ng. "KLASSify to Verify: Audio-Visual Deepfake Detection Using SSL-based Audio and Handcrafted Visual Features." In Proc. ACM Multimedia 2025, pp. 13707-13713.

3 Fake Localization

TABLE III
THE STRENGTHS AND WEAKNESS FOR EACH TYPE OF METHOD.

Type	Properties	Strengths	Weakness	Related methods
1	Frame-level Authenticity	It is straightforward and constitutes the majority in existing research.	It may fail to locate the manipulation regions when the splicing clips are bona fide	SPF[26], TDL[19]
2	Boundary Perception	It focuses on detecting stitching traces to avoid relying entirely on frame-level authenticity.	It may fail when the splicing boundaries are hidden intentionally	CFPRF[57], BAM[35]
3	Frame-level Inconsistency	It focuses on the inconsistency between frames instead of the authenticity, overcoming the weakness of the former two types.	For long lasting audio, the information in the utterances may change, the effectiveness needs further validation	PET[33], AGO[34], GNCL[58]
4	Multi-Modality Fusion	It integrates multimodal forgery information and represents a new trend in recent research.	It focuses more on the visual modality, and further explorations are needed in audio modality.	UMMAFormer[59], W-TDL[60]

AV-Deepfake1M
AV-Deepfake1M++



Manipulated Regions Localization For Partially Deepfake Audio: A Survey

Jiayi He, Jiangyan Yi *Member, IEEE,*, Jianhua Tao* *Senior Member, IEEE*, Siding Zeng, and Hao Gu

https://arxiv.org/pdf/2506.14396

Localizing Audio-Visual Deepfakes via Hierarchical Boundary Modeling

Xuanjun Chen[1], Shih-Peng Cheng[2*], Jiawei Du[2], Lin Zhang[3], Xiaoxiao Miao[4]
Chung-Che Wang[2], Haibin Wu[5†], Hung-yi Lee[1], Jyh-Shing Roger Jang[2]

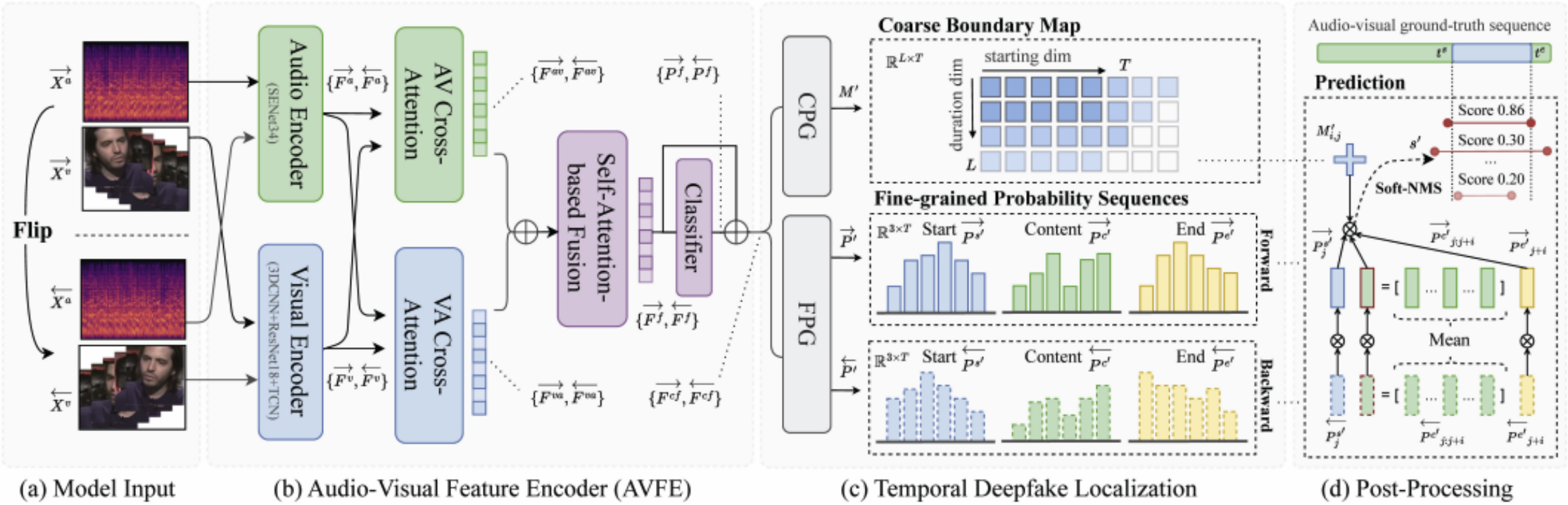


Figure 1: The main stem of Hierarchical Boundary Modeling Network (HBMNet) used in training and inference. Auxiliary branch attachments share the main stem architecture but have independent weights, indicated with structures omitted for clarity.

Chen, Xuanjun, Shih-Peng Cheng, Jiawei Du, Lin Zhang, Xiaoxiao Miao, Chung-Che Wang, Haibin Wu, Hung-yi Lee, and Jyh-Shing Roger Jang. "Localizing Audio-Visual Deepfakes via Hierarchical Boundary Modeling." *arXiv preprint arXiv:2508.02000* (2025).

3 Fake Localization

(top-1 on AV-Deepfake1M++)

Pindrop it! Audio and Visual Deepfake Countermeasures for Robust Detection and Fine Grained-Localization

Nicholas Klein* nklein@pindrop.com
Hemlata Tak* hemlata.tak@pindrop.com
James Fullwood* james.fullwood.i@pindrop.com
Krishna Regmi* krishna.regmi@pindrop.com
Leonidas Spinoulas* leonidas.spinoulas@pindrop.com
Ganesh Sivaraman* gsivaraman@pindrop.com
Tianxiang Chen* tchen@pindrop.com
Elie Khoury* ekhoury@pindrop.com

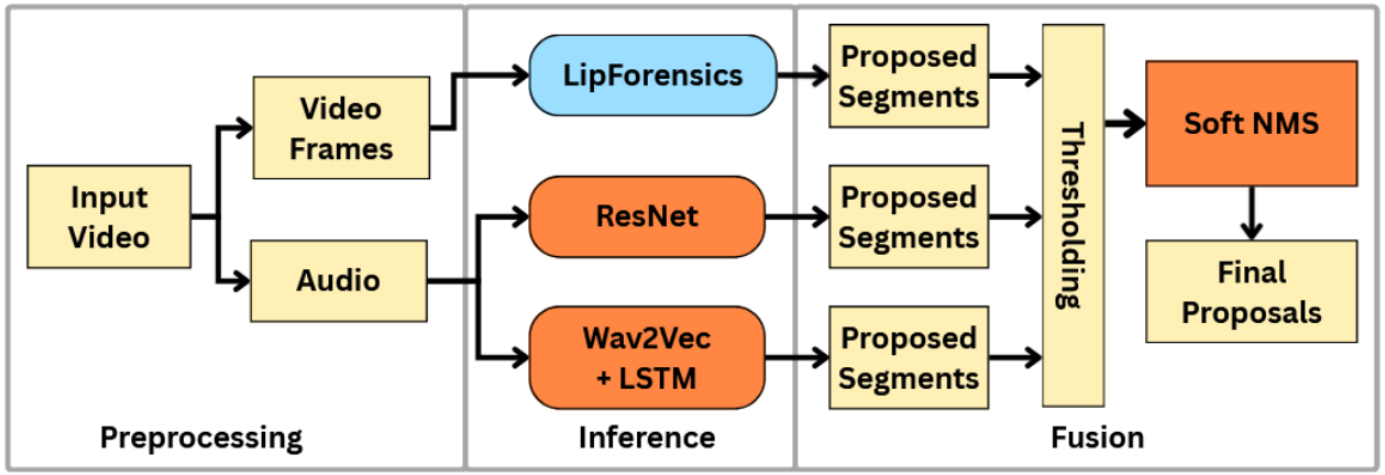


Figure 4: Task 2 overview.

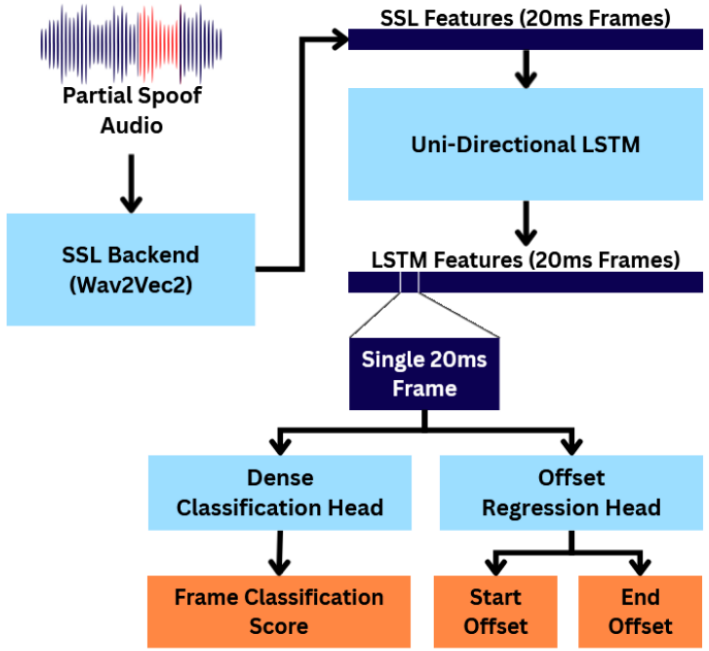


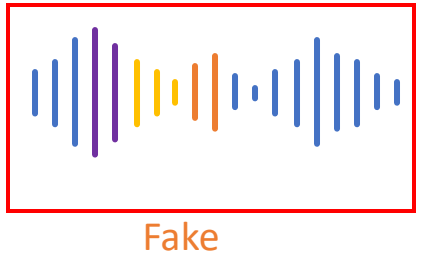
Figure 6: Wav2Vec-based SSL pipeline with LSTM for fine-grained frame-level detection and localization task.

Boundary-aware + uniform segmentation

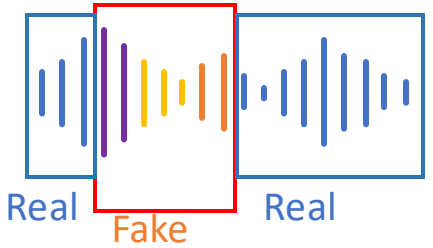
Klein, N., Tak, H., Fullwood, J., Regmi, K., Spinoulas, L., Sivaraman, G., Chen, T. and Khoury, E., 2025, October. Pindrop it! Audio and Visual Deepfake Countermeasures for Robust Detection and Fine-Grained Localization. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 13700-13706).

3 Fake Diarization

➢ **Fake Detection: Whether the input utterance is spoofed?**



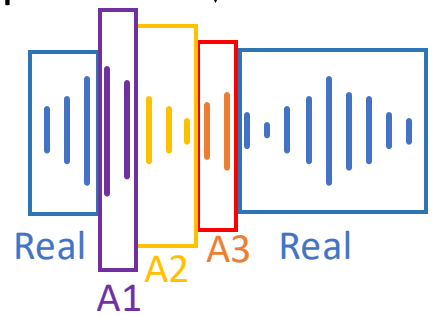
➢ **Fake Localization: When do spoofs happen?**



Trace back to the original algorithm, then do further analysis.
VC: Trace back to the original speaker. [Cai 2022]
Audio contains segments created using multiple generative models or through a recursive generation process

➢ **Fake Diarization: What spoofed when?**

Current studies mainly focus on cluster (without identifying) all segments.



https://sstc-challenge.github.io/

Danwei Cai, Zexin Cai, and Ming Li, “Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems,” arXiv preprint arXiv:2206.09103, 2022.

Outline

0. Introduction

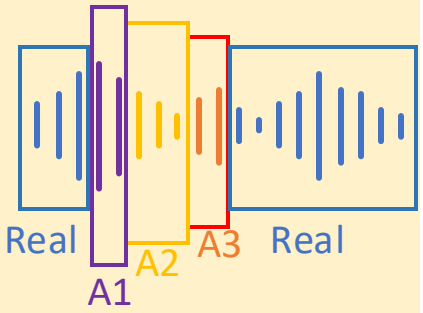
1. Database

Whether

2. Detection

When

3. Localization



What + When

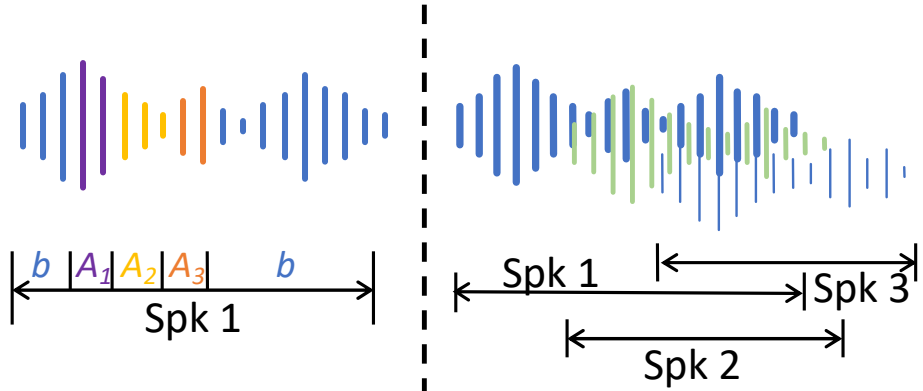
4. Diarization

How

5. Analysis

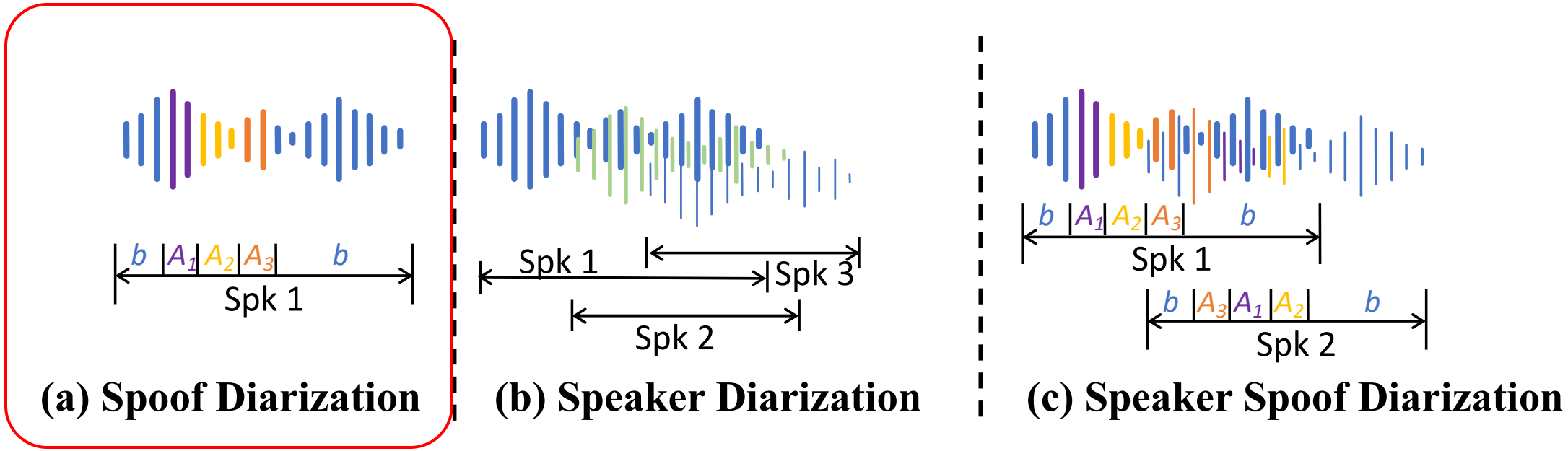
6. Summary & Open Challenges

4 Fake Diarization - Definition



		(a) Spoof Diarization	(b) Speaker Diarization
Similarities		1. Unbounded unknown clusters (spoofing methods or speakers).	
		2. Class-homogeneous regions in both tasks can have variable durations.	
Differences	3. Duration of turns	Detecting **short-turn** spoofed speech is crucial	Detecting short (word level) turns in speaker diarization systems depends on the application.
	4. Two primary groups of clusters	**Two primary groups**: bona fide, spoof.	Speakers may vary for each audio

4 Fake Diarization - Definition

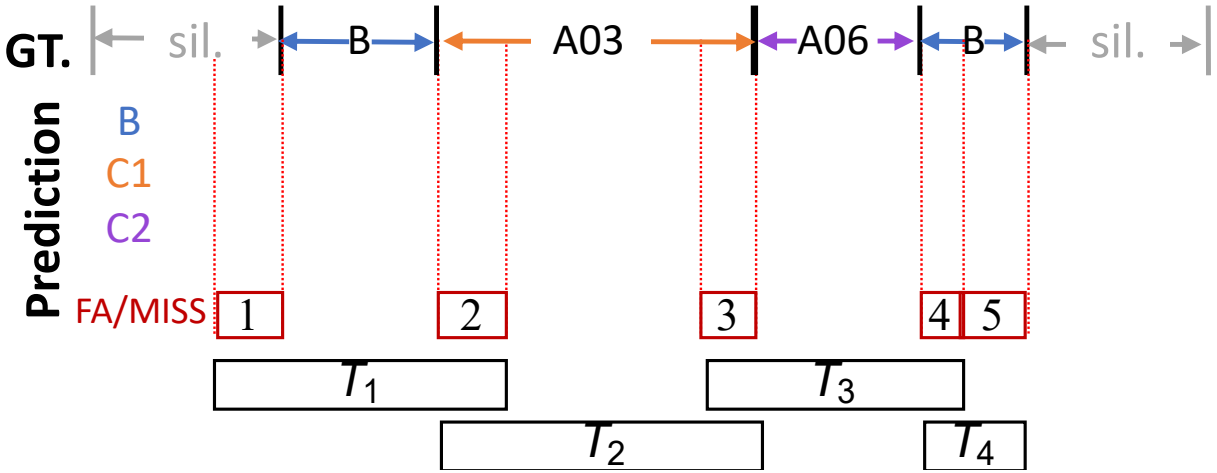


(a) Spoof Diarization (b) Speaker Diarization (c) Speaker Spoof Diarization

4 Fake Diarization - Metric

Metric Spoof Jaccard error rate

Spoof diarization has two important items:
(1) differentiating spoofed from bona fide segments, and
(2) discriminating different spoofing methods



$$\mathrm{JI}_{\mathrm{bona},j} = \frac{\mathrm{FA}_{\mathrm{bona}} + \mathrm{MISS}_{\mathrm{bona}}}{\mathrm{TOTAL}_{\mathrm{bona}}} = \frac{\boxed{1}\boxed{2} + \boxed{4}\boxed{5}}{\boxed{T_1}\boxed{T_4}}$$

$$\mathrm{JER}_{\mathrm{spoof},j} = \frac{1}{2}\left(\frac{\mathrm{FA}_{\mathrm{A03}} + \mathrm{MISS}_{\mathrm{A03}}}{\mathrm{TOTAL}_{\mathrm{A03}}} + \frac{\mathrm{FA}_{\mathrm{A06}} + \mathrm{MISS}_{\mathrm{A06}}}{\mathrm{TOTAL}_{\mathrm{A06}}}\right)$$

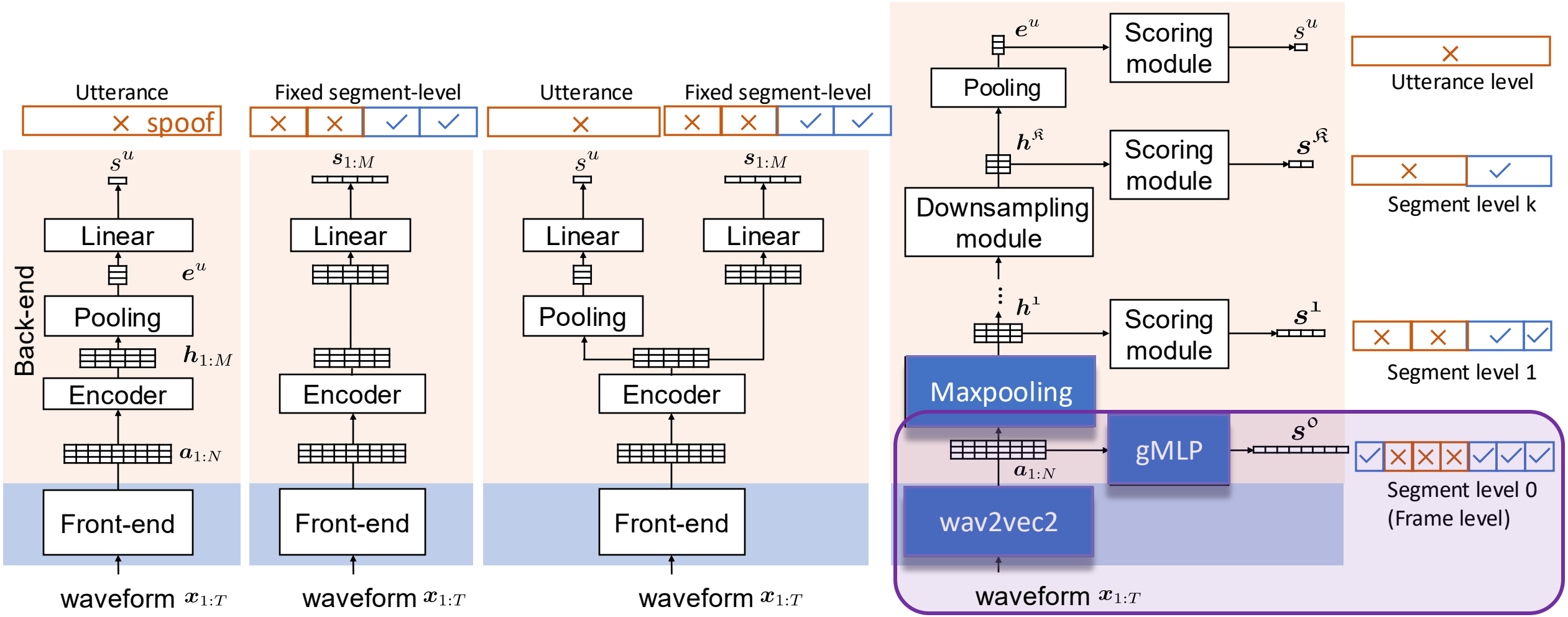
$$= \frac{1}{2}\left(\frac{\mathrm{Null} + \boxed{2}\boxed{3}}{\boxed{T_2}} + \frac{\boxed{3}\boxed{4} + \mathrm{Null}}{\boxed{T_3}}\right)$$

$$\mathrm{JI}_{\mathrm{global\_bona}} = \frac{1}{|\mathcal{D}|}\sum_{j\in\mathcal{D}} \mathrm{JI}_{\mathrm{bona},j}$$

$$\mathrm{JER}_{\mathrm{global\_spoof}} = \frac{1}{\sum_{j\in\mathcal{D}}|\mathcal{A}_j|}\sum_{j\in\mathcal{D}}\sum_{A_i\in\mathcal{A}_j} \mathrm{JER}_{A_i,j}$$

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans, Mireia Diez, Federico Landini and Junichi, 2024. Spoof Diarization:" What Spoofed When" in Partially Spoofed Audio. in Proc. Interspeech 2024

4 Recap Previous Models

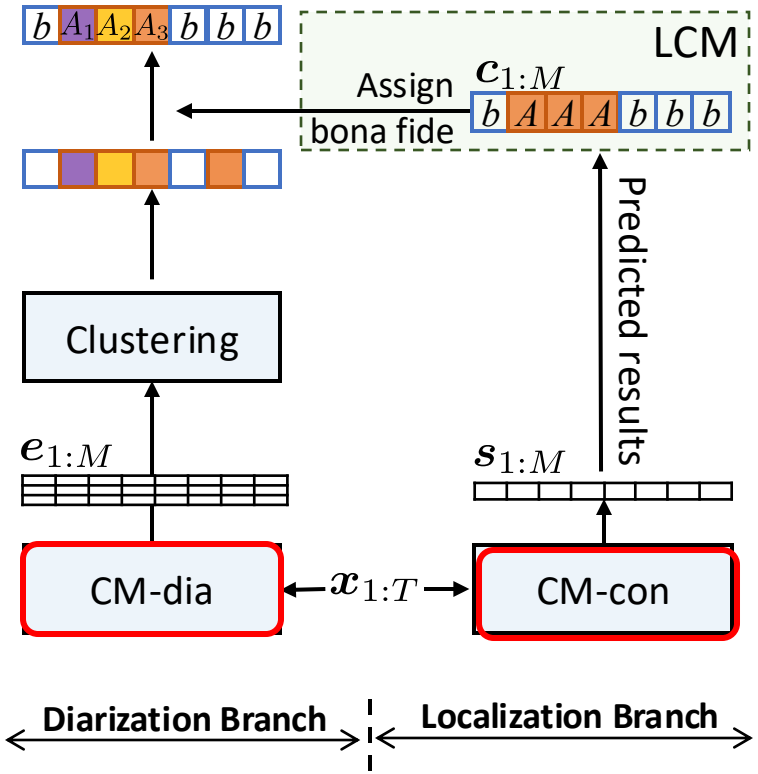


Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans and Junichi Yamagishi, "The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance," in IEEE/ACM TASLP, vol. 31, pp. 813-825, 2023, doi: 10.1109/TASLP.2022.3233236.

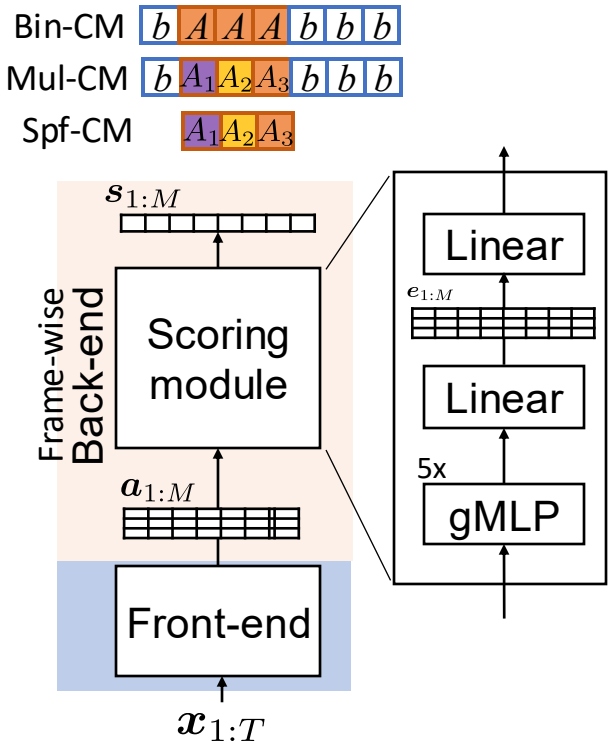
4 Fake Diarization - Model

Model

3C model: CM-condition clustering



(a) 3C model with two branches



(b) Structure of CMs in (a)

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans, Mireia Diez, Federico Landini and Junichi, 2024. Spoof Diarization:" What Spoofed When" in Partially Spoofed Audio. in Proc. Interspeech 2024

4 Fake Diarization - Result

How do labeling schemes affect the ability of CMs?

Table 1: *Results on the PartialSpoof database.*

Model		Development set		Evaluation set	
Dia.	Loc.	JI_{bona} (%)	$\text{JER}_{\text{spoof}}$ (%)	JI_{bona} (%)	$\text{JER}_{\text{spoof}}$ (%)
Bin-CM	/	4.37	20.45	16.85	33.13
Mul-CM	/	4.49	5.21	19.66	28.05
Spf-CM	/	26.17	20.85	32.30	38.51

- Mul-CM obtains overall best results
- Spf-CM underperforms in JER_spoof, revealing that specific treatment of the bona fide class is needed during system training.

4 Fake Diarization - Result

How do labeling schemes affect the ability of CMs?

Table 1: *Results on the PartialSpoof database.*

Model Dia.	Loc.	Development set JI_{bona} (%)	Development set JER_{spoof} (%)	Evaluation set JI_{bona} (%)	Evaluation set JER_{spoof} (%)
Bin-CM	/	4.37	20.45	16.85	33.13
Mul-CM	/	4.49	5.21	19.66	28.05
Spf-CM	/	26.17	20.85	32.30	38.51
Mul-CM	Bin-CM	4.49	5.27	15.15	34.13
	Mul-CM	4.59	5.31	17.08	35.34
Spf-CM	Bin-CM	4.52	5.71	15.18	36.03
	Mul-CM	4.62	5.81	17.10	37.78

- For Mul-CM: localization branch allows for better localization capabilities but it negatively impacts diarizing spoofing methods
  - For Spf-CM: both JI_bona and JER_spoof show an improvement.
- Mul-CM < Spf-CM: Dia-Mul-CM model extracts better embeddings for diarizing spoofing methods.
  - Bin-CM is specialized in distinguishing bona fide from spoof.

4 Fake Diarization

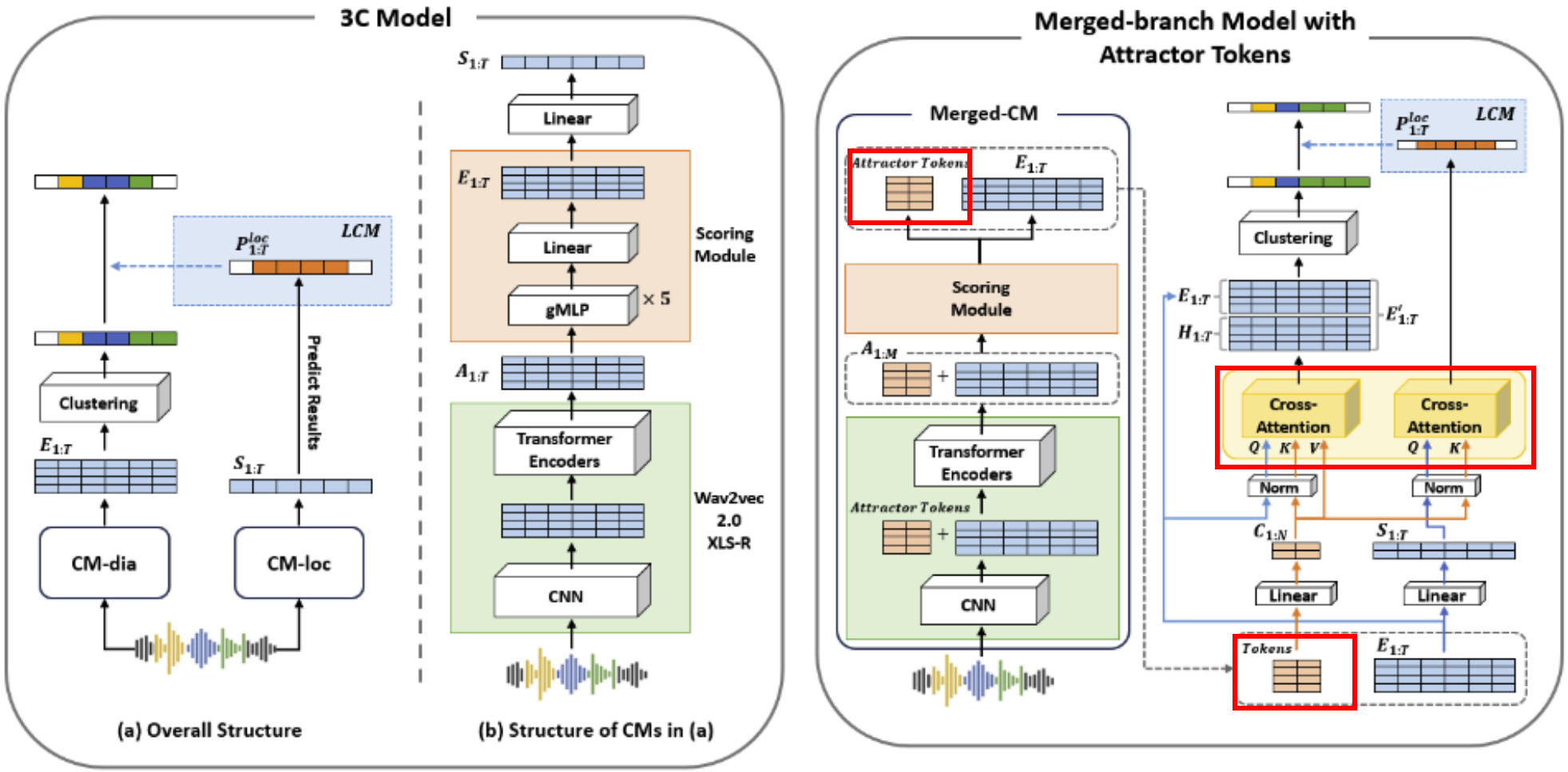


Fig. 2: The overall architecture of the baseline 3C model and our proposed Merged-branch model with attractor tokens. CM-dia and CM-loc represent the models for the diarization and localization branches, respectively. LCM denotes the label-based CM constraint module, and Norm indicates an ℓ_2-normalization step.

Koo, Kyo-Won, Chan-yeong Lim, Jee-weon Jung, Hye-jin Shim, and Ha-Jin Yu. "Token-based Attractors and Cross-attention in Spoof Diarization." *in Proc. ASRU 2025*

Outline

0. Introduction

1. Database

Whether

2. Detection

When

3. Localization

What + When

4. Diarization

How

5. Analysis

6. Summary & Open Challenges

5 Analysis - DSP

Analyzing the Impact of Splicing Artifacts in Partially Fake Speech Signals

Viola Negroni, Davide Salvi, Paolo Bestagini, Stefano Tubaro

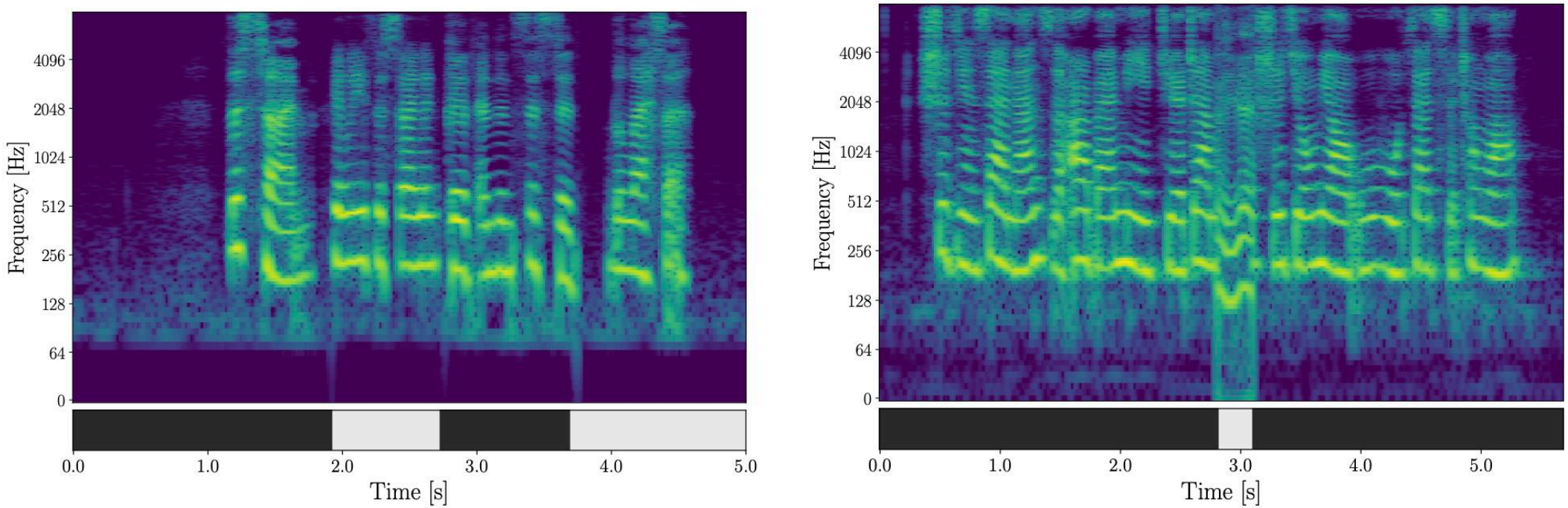


Figure 3: Frequency domain analysis of one example track per dataset: *CON_D_0000001.wav* from PartialSpoof (left) and *ADD2023_T2_D_00000036.wav* from HAD (right). STFT window size: 2048 samples, Hop size: 256 samples, no zero-padding.

Although the induced splicing artifacts (caused by spectral leakage) at the concatenation points are generally inaudible, they can be easily exposed.

Negroni, V., Salvi, D., Bestagini, P., Tubaro, S. (2024) Analyzing the impact of splicing artifacts in partially fake speech signals. Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 79-85, doi: 10.21437/ASVspoof.2024-12

How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio?

Tianchi Liu[1,2], *Lin Zhang*[3], *Rohan Kumar Das*[4], *Yi Ma*[2], *Ruijie Tao*[2], *Haizhou Li*[5,2]

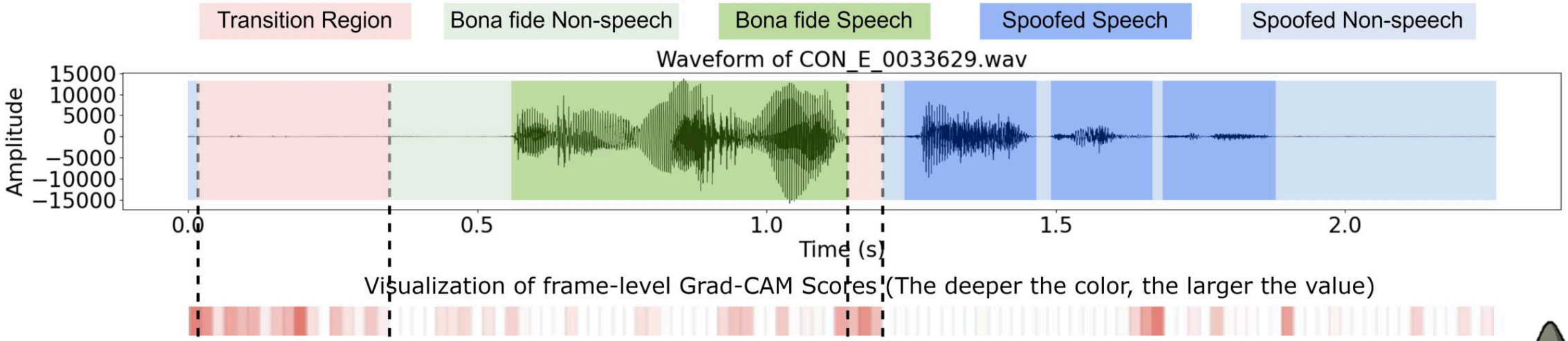


Figure 4: *Visualization of the waveform and frame-level Grad-CAM scores for CON_E_0033629.wav from the evaluation s*

TL;DR: We utilize Grad-CAM to interpret CMs' decisions and find that CMs prioritize the artifacts of transition regions.

Interspeech 2024 https://arxiv.org/abs/2406.02483

Outline

0. Introduction

1. Database

Whether

2. Detection

When

3. Localization

What + When

4. Diarization

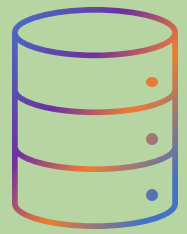
How

5. Analysis

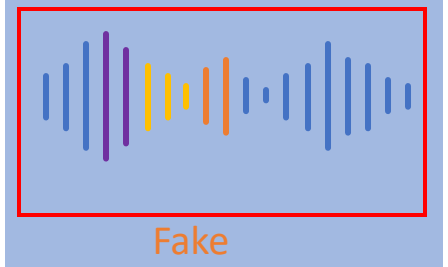
6. Summary & Open Challenges

6 Summary

0. Introduction

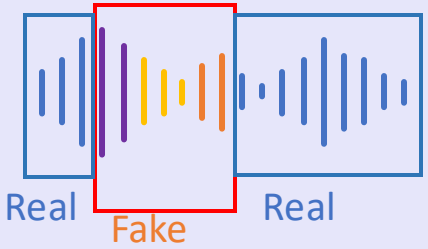


1. Database



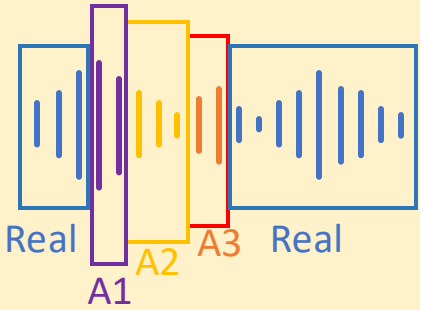
Whether

2. Detection



When

3. Localization



What + When

4. Diarization

How

5. Analysis

6. Summary & Open Challenges

6 Open Challenges

- **Analysis - Intervention analysis on the partial spoof scenario**
  - Impact of spoof-to-utterance ratio
  - Effect of manipulated position and duration
- **More realistic scenario**
  - Noisy, wild scenario (more complicated and mismatched environment)
  - Code-switch editing
- **Advanced spoof diarization scenario and technologies**
  - More advanced diarization models
  - Speaker spoof diarization
  - Watermarking based spoof diarization
  - Multi-level spoof diarization

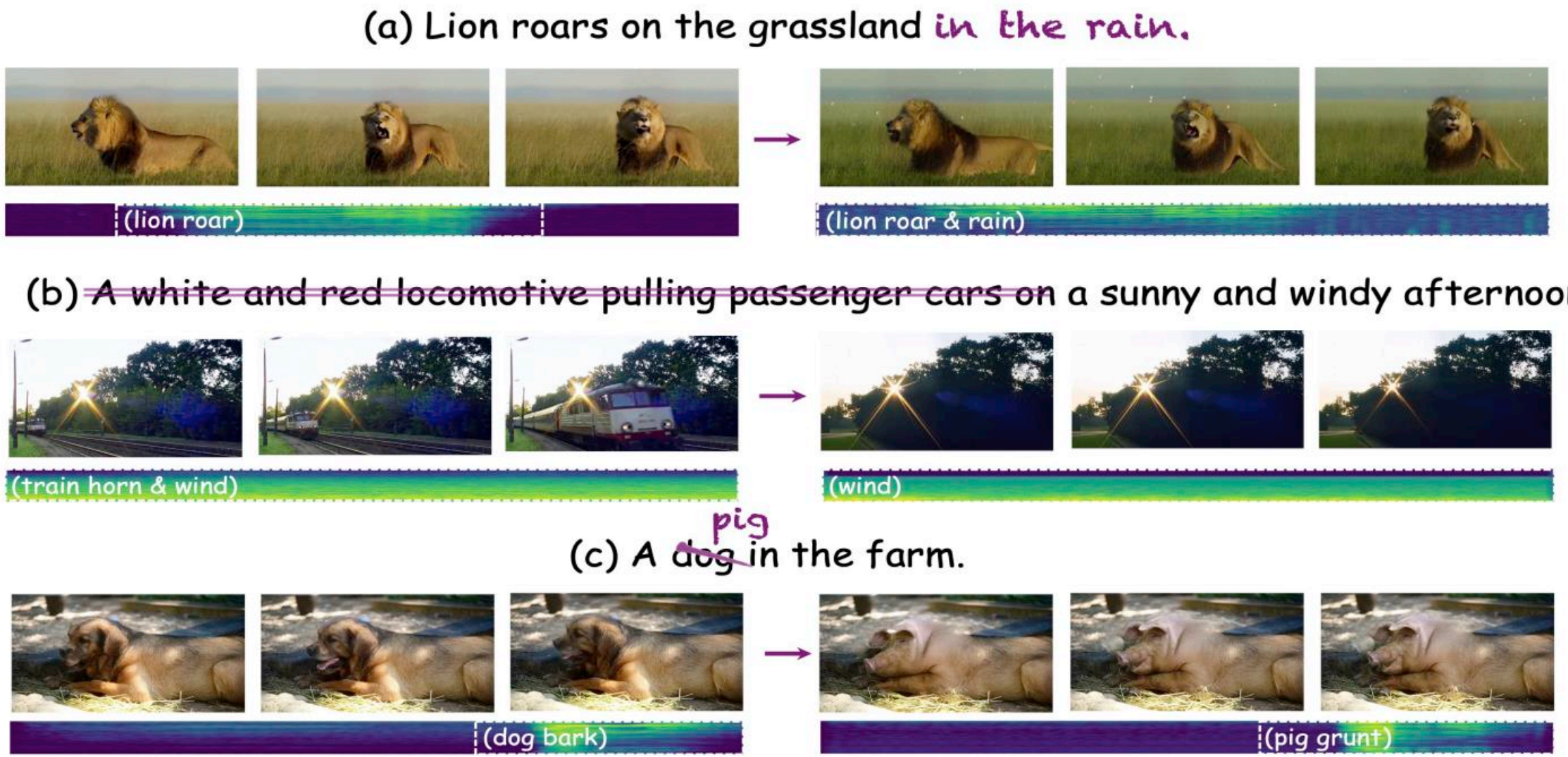
https://deepfakes1m.github.io

6 Open Challenges

- **Editing Beyond Speech Content**
  - **Prosody:** Fine-Grained and Interpretable Neural Speech Editing, https://arxiv.org/abs/2407.05471
  - ***Emotion:*** Towards Emotionally Consistent Text-Based Speech Editing: Introducing EmoCorrector and The ECD-TSE Dataset, *Interspeech 2025, arxiv.org/abs/2505.20341*
  - ***Sound*:** Recomposer: Event-roll-guided generative audio editing. *Google*, arxiv.org/abs/2509.05256 🐶 🐱 🚗 🚪
  - **Background:** SeamlessEdit: Background Noise Aware Zero-Shot Speech Editing with in-Context Enhancement, arxiv.org/abs/2505.14066
  - **Video2audio:** ThinkSound: Chain-of-Thought Reasoning in Multimodal Large Language Models for Audio Generation and Editing. *Alibaba*, NeurIPS 2025, arxiv.org/abs/2506.21448.
- **Editing Beyond Audio**
  - **Image:** ImgEdit: A Unified Image Editing Dataset and Benchmark, arxiv.org/abs/2505.20275
  - **Video:** Movie Gen: A cast of media foundation models. *META*, arxiv.org/abs/2410.13720 (no sound)
  - **Object:** Object-AVEdit: An Object-level Audio-Visual Editing Model, arxiv.org/abs/2510.00050

6 Open Challenges

Fu, Youquan, Ruiyang Si, Hongfa Wang, Dongzhan Zhou, Jiacheng Sun, Ping Luo, Di Hu, Hongyuan Zhang, and Xuelong Li. "Object-AVEdit: An Object-level Audio-Visual Editing Model." *arXiv preprint arXiv:2510.00050* (2025).



Reference 1/3

Introduction:

Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc¸i, M. Sahidullah, and A. Sizov, “ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge,” in Proc. Interspeech 2015, pp. 2037–2041.

T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection,” in Proc. Interspeech, 2017, pp. 2–6.

A. Nautsch, X. Wang, etc, “ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech,” IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021.

J. Yamagishi, X. Wang , etc, “ASVspoof2021: accelerating progress in spoofed and deep fake speech detection,” in Proc. ASVspoof 2021 Workshop, 2021.

Database:

PartialSpoof: Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans. (2021) An Initial Investigation for Detecting Partially Spoofed Audio. Proc. Interspeech 2021, 4264-4268

HAD: Yi, J., Bai, Y., Tao, J., Ma, H., Tian, Z., Wang, C., Wang, T., Fu, R. (2021) Half-Truth: A Partially Fake Audio Detection Dataset. Proc. Interspeech 2021, 1654-1658

ADD 2022: Yi, Jiangyan, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang et al. "Add 2022: the first audio deep synthesis detection challenge." In ICASSP 2022, pp. 9216-9220

ADD 2023: Yi, Jiangyan, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang et al. "Add 2023: the second audio deepfake detection challenge." *arXiv preprint arXiv:2305.13774* (2023).

Psynd: Bowen Zhang and Terence Sim. Localizing fake segments in speech. In Proc. ICPR 2022, pages 3224–3230. IEEE, 2022.

LAV-DF: Cai, Zhixi, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. "Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization." *Computer Vision and Image Understanding* 236 (2023): 103818.

AV-Deepfake1M: Cai, Z., Ghosh, S., Adatia, A.P., Hayat, M., Dhall, A., Gedeon, T. and Stefanov, K., 2024, October. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In Proc. *MM*

LlamaPartialSpoof: Luong, H.T., Li, H., Zhang, L., Lee, K.A. and Chng, E.S., 2025, April. LlamaPartialSpoof: An llm-driven fake speech dataset simulating disinformation generation. In ICASSP 2025

PartialEdit: You Zhang, Baotong Tian, Lin Zhang, and Zhiyao Duan. "PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing." in Interspeech2025

AV-Deepfake1M++: Cai, Z., Kuckreja, K., Ghosh, S., Chuchra, A., Khan, M.H., et. al. 2025. AV-Deepfake1M++: A Large-Scale Audio-Visual Deepfake Benchmark with Real-World Perturbations. *arXiv preprint arXiv:2507.20579*.

LENS-DF: Liu, Xuechen, Wanying Ge, Xin Wang, and Junichi Yamagishi. "LENS-DF: Deepfake Detection and Temporal Localization for Long-Form Noisy Speech." IJCB 2025.

Reference 2/3

X. Wang, J. Yamagishi, M. Todisco, et al. ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech[J]. Computer Speech and Language, vol. 64, pp. 101--114, 2020

Zen, Heiga, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. "LibriTTS: A corpus derived from librispeech for text-to-speech." in Proc. Interspeech 2019, 1526-1530

Le, Matthew, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson et al. "Voicebox: Text-guided multilingual universal speech generation at scale." *Advances in neural information processing systems* 36 (2023): 14005-14034.

P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," in Proc. ACL, 2024, pp. 12 442–12 462.

A.Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan et al., "Audiobox: Unified audio generation with natural language prompts," arXiv:2312.15821, 2023.

Wang, Helin, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. "SSR-Speech: Towards Stable, Safe and Robust Zero-shot Text-based Speech Editing and Synthesis" in Proc. ICASSP, 2025

Detection and localization

N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," in Proc. NeurIPS 2018, 2018, p. 1924–1934.

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans. (2021) An Initial Investigation for Detecting Partially Spoofed Audio. Proc. Interspeech 2021, 4264-4268, doi: 10.21437/Interspeech.2021-738

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi. (2021) Multi-task Learning in Utterance-level and Segmental-level Spoof Detection. Proc. ASVspoof workshop 2021, 9-15, doi: 10.21437/ASVSPOOF.2021-2

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans and Junichi Yamagishi, "The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 813-825, 2023, doi: 10.1109/TASLP.2022.3233236.

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans and Junichi Yamagishi, (2023) Range-Based Equal Error Rate for Spoof Localization. Proc. INTERSPEECH 2023, 3212-3216, doi: 10.21437/Interspeech.2023-1214

Vigo: Pérez-Vieites, Diego, Juan José Moreira-Pérez, Ángel Aragón-Kifute, Raquel Román-Sarmiento, and Rubén Castro-González. "Vigo: Audiovisual Fake Detection and Segment Localization." In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11360-11364. 2024.

BDR: Z. Cai and M. Li, "Integrating frame-level boundary detection and deepfake detection for locating manipulated regions in partially spoofed audio forgery attacks," (2024) Comput. Speech Lang., vol. 85,

BAM: J. Zhong, B. Li, and J. Yi, "Enhancing partially spoofed audio localization with boundary-aware attention mechanism," in Proc. Interspeech, 2025.

CFPRF: 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 Proc. ACM MM, 2024, pp. 7395–7403.

BFC-Net: Zhou, Y., Xue, Z., Senhadji, L., Shu, H. and Wu, J., 2025. BFC-Net: Boundary-Frame cross graph attention network for partially spoofed audio localization. *Neurocomputing*, p.130867.

PET: He, J., Yi, J., Tao, J. and Zeng, S., 2025, April. PET: High-Frequency Temporal Self-Consistency Learning for Partially Deepfake Audio Localization. In *ICASSP 2025*

TDL: Xie, Y., Cheng, H., Wang, Y. and Ye, L., 2024, April. An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection. In *ICASSP 2024*

Reference 3/3

MFMS: Zhang, Yi, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Weibin Yao, Zhe Li et al. "Mfms: Learning modality-fused and modality-specific features for deepfake detection and localization tasks." In *Proc. ACM Multimedia*, pp. 11365-11369. 2024.

Pindrop: Klein, Nicholas, Hemlata Tak, et al. "Pindrop it! Audio and Visual Deepfake Countermeasures for Robust Detection and Fine-Grained Localization." In Proc. ACM Multimedia 2025, pp. 13700-13706

KLASSify: Ivan Kukanov, and Jun Wah Ng. "KLASSify to Verify: Audio-Visual Deepfake Detection Using SSL-based Audio and Handcrafted Visual Features." In Proc. ACM Multimedia 2025, pp. 13707-13713. 2025.

Chen, Xuanjun, Shih-Peng Cheng, Jiawei Du, Lin Zhang, Xiaoxiao Miao, Chung-Che Wang, Haibin Wu, Hung-yi Lee, and Jyh-Shing Roger Jang. "Localizing Audio-Visual Deepfakes via Hierarchical Boundary Modeling." *arXiv preprint arXiv:2508.02000* (2025).

Localization Survey: He, Jiayi, Jiangyan Yi, Jianhua Tao, Siding Zeng, and Hao Gu. "Manipulated Regions Localization For Partially Deepfake Audio: A Survey." *arXiv preprint arXiv:2506.14396* (2025).

Diarization

Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans, Mireia Diez, Federico Landini and Junichi, 2024. Spoof Diarization:" What Spoofed When" in Partially Spoofed Audio. in Proc. Interspeech 2024

Koo, Kyo-Won, Chan-yeong Lim, Jee-weon Jung, Hye-jin Shim, and Ha-Jin Yu. "Token-based Attractors and Cross-attention in Spoof Diarization." *in Proc. ASRU 2025*

Analysis

Negroni, Viola, Davide Salvi, Paolo Bestagini, and Stefano Tubaro. "Analyzing the impact of splicing artifacts in partially fake speech signals." arXiv preprint arXiv:2408.13784 (2024).

Liu, Tianchi, Lin Zhang, Rohan Kumar Das, Yi Ma, Ruijie Tao, and Haizhou Li. "How do neural spoofing countermeasures detect partially spoofed audio?." in Proc. Interspeech 2024, 1105-1109, doi: 10.21437/Interspeech.2024-2009

Thank you

If you have any questions or comments, please feel free to contact with me:

zlin@ieee.org, partialspoof@gmail.com

I will be attending ASA/ASJ and ASRU (12.2-12.11) at Honolulu, HI. Welcome discussion. 😊