“Whether, When, What”: Detection, Localization, and Diarization of Partially Spoofed Audio

Lin Zhang

Supervisor: Prof. Junichi Yamagishi

Mentor: Dr. Xin Wang, Dr. Erica Cooper

National Institute of Informatics, SOKENDAI

Slides by Lin Zhang
National Institute of Informatics

Outline

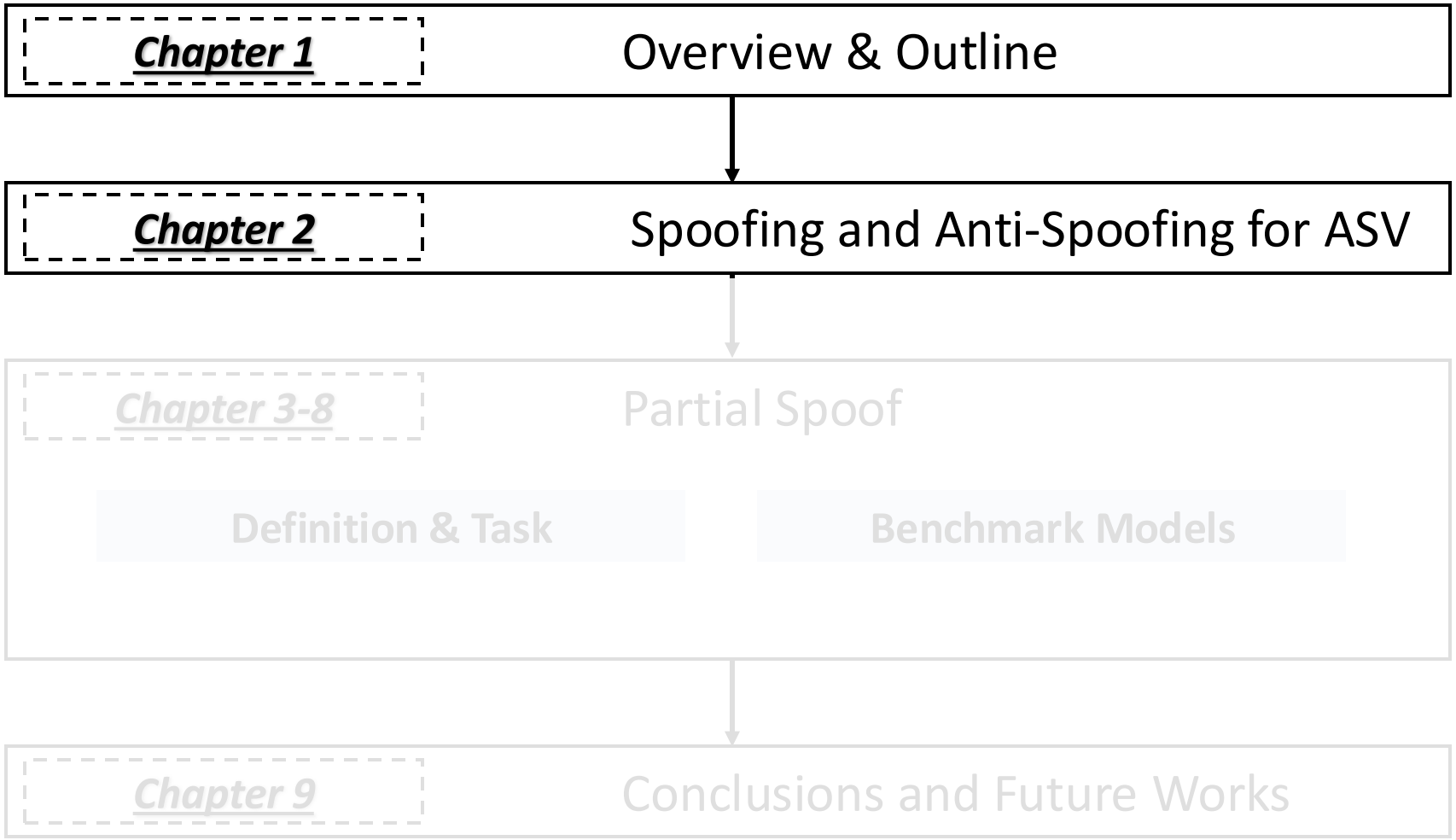


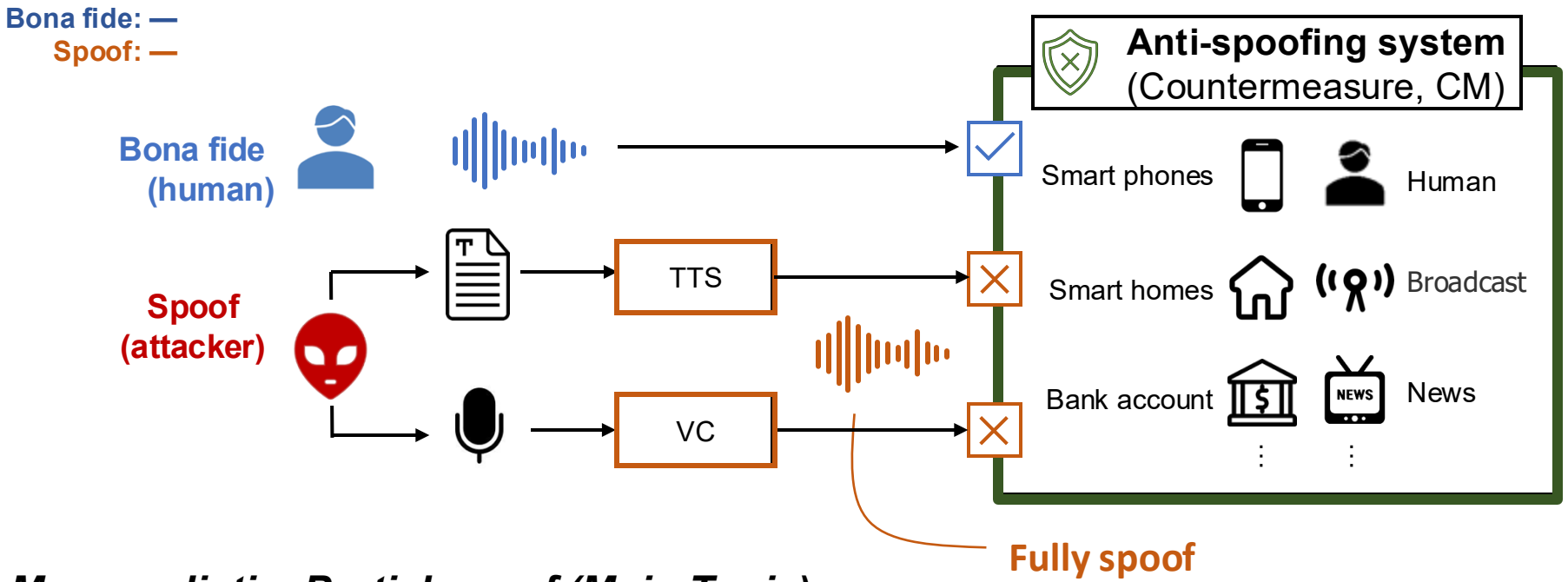
Figure 1: Thesis Outline

1&2 Introduction

Commonly studied scenario: Fully spoof

TTS: Text-to-speech

VC: Voice conversion

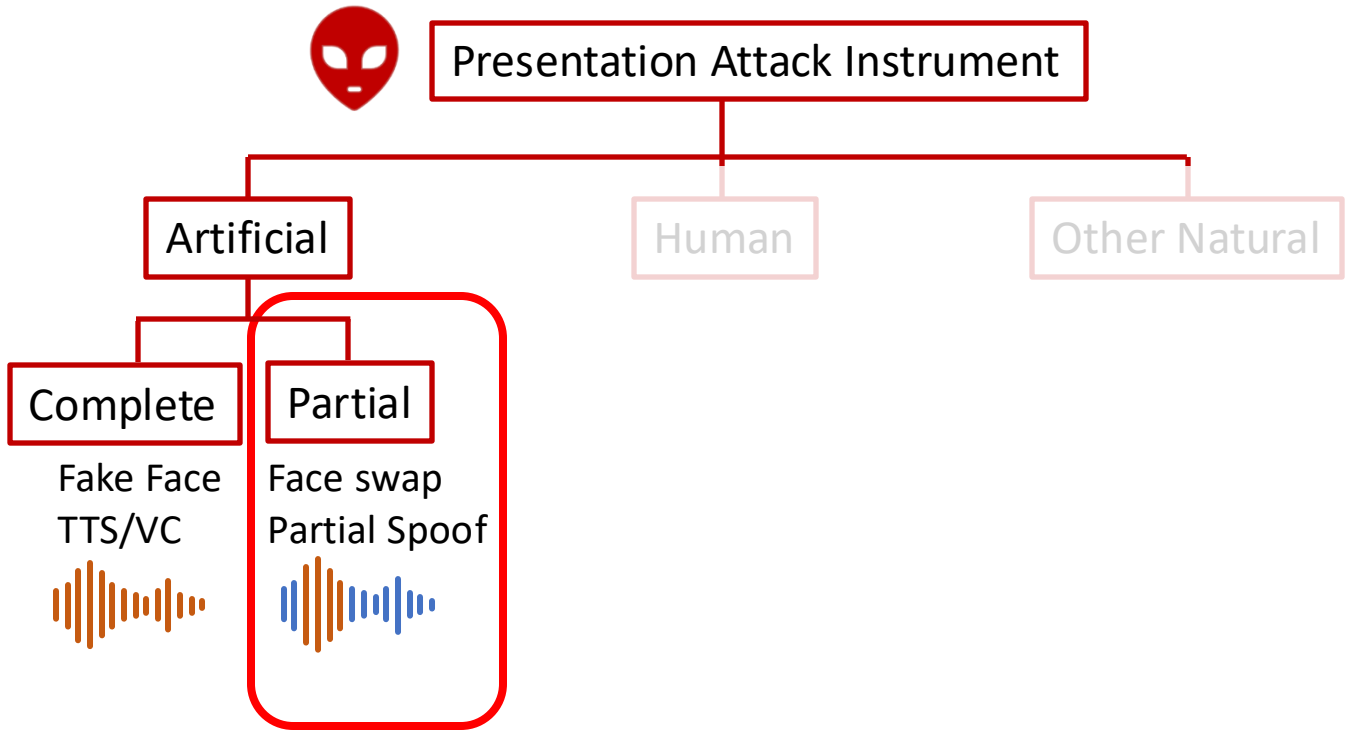


More realistic: Partial spoof (Main Topic)

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.

➢ Definition on ISO/IEC 30107

- **Presentation Attack Instrument (PAI):**
  Biometric characteristic or object used in a presentation attack.
- **Artefact**
  Artificial object or representation presenting a copy of biometric characteristics or synthetic biometric patterns.



ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 30107: Information technology — Biometric presentation attack detection (2023)

Outline

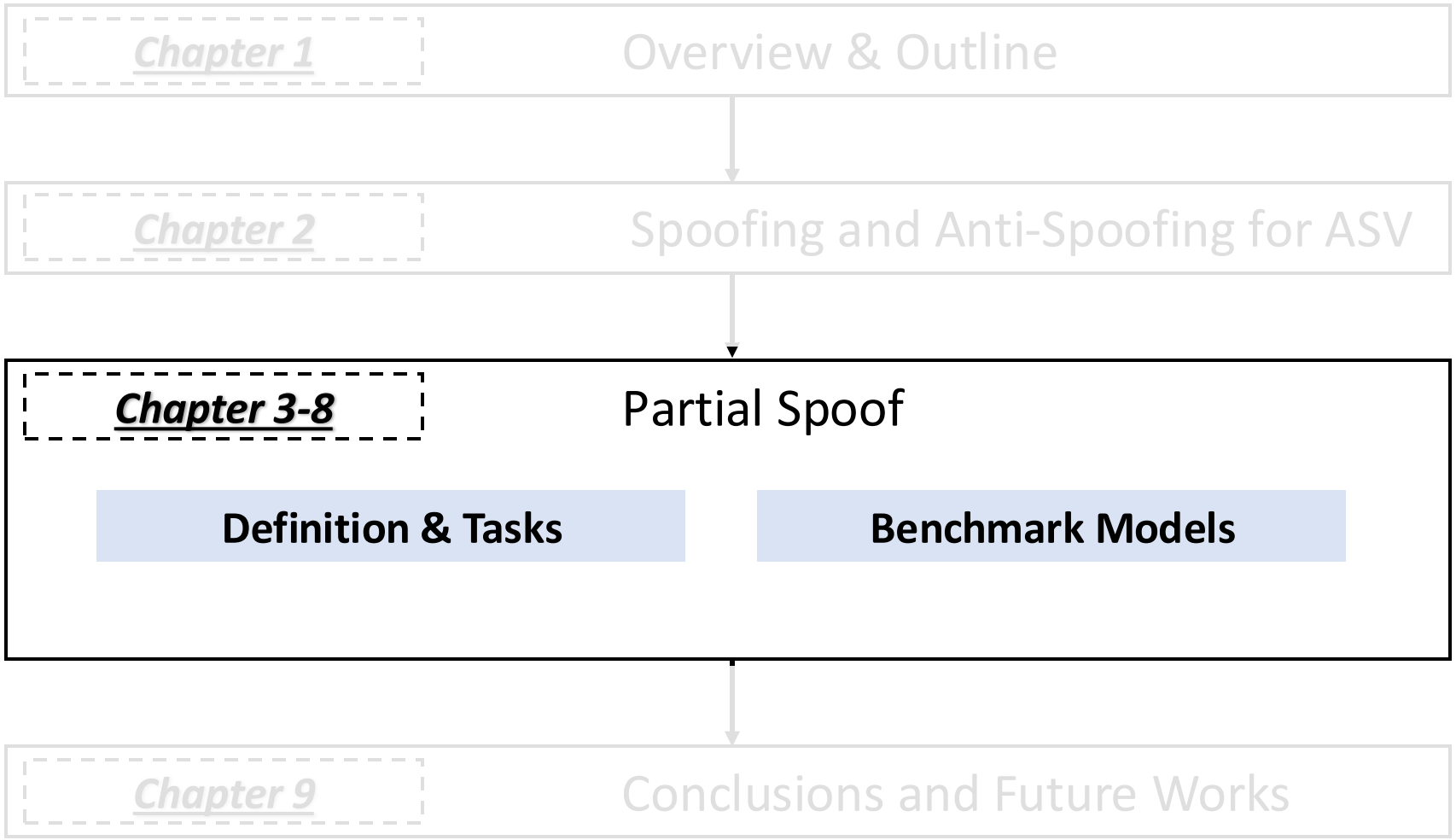


Figure: Thesis Outline

Outline

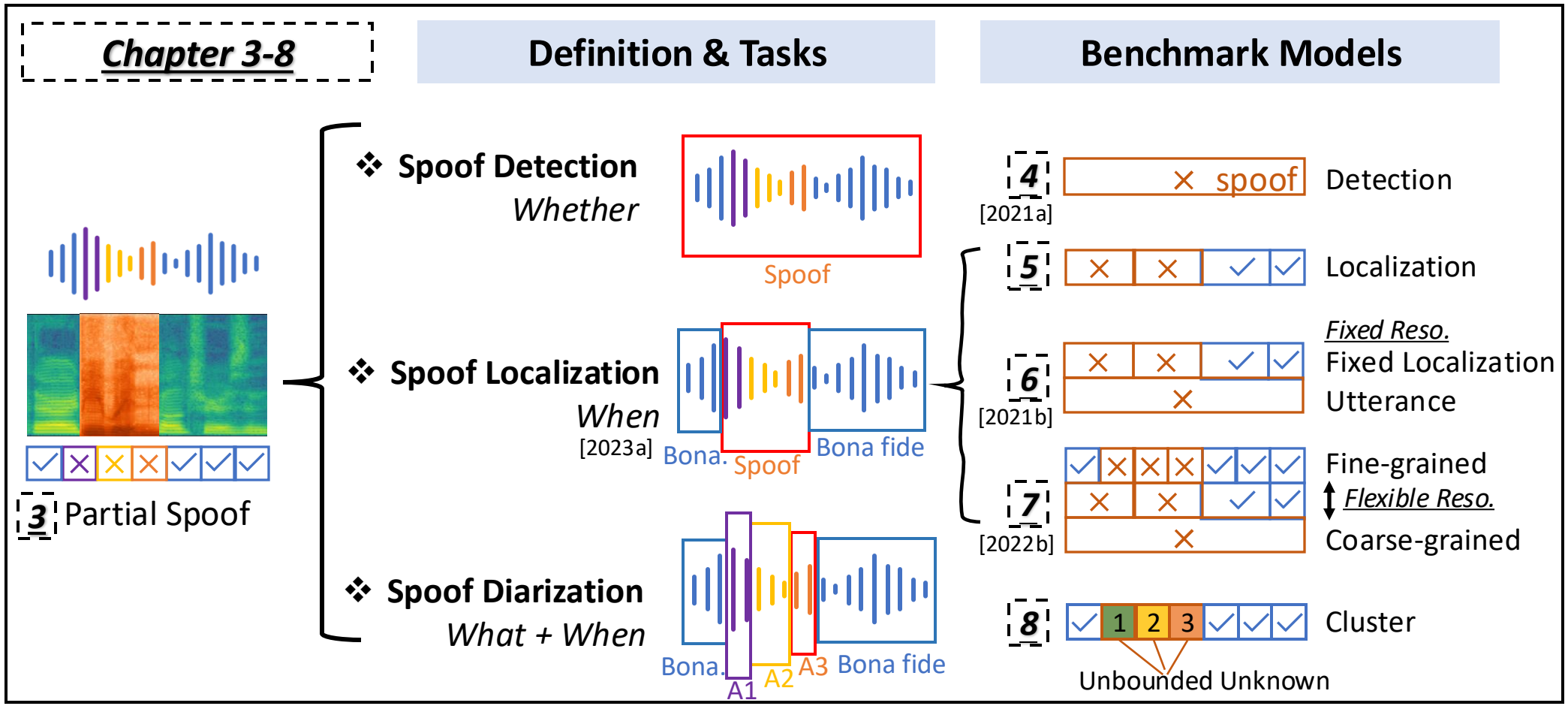


Figure: Thesis Outline & Contribution

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

3 Partial Spoof - Definition

Bona Fide

Tampering Forgery (copy-move)

Tampering Forgery (splicing)

Unit-Selection

Confined to the manipulation of existing bona fide data

Fully Spoof

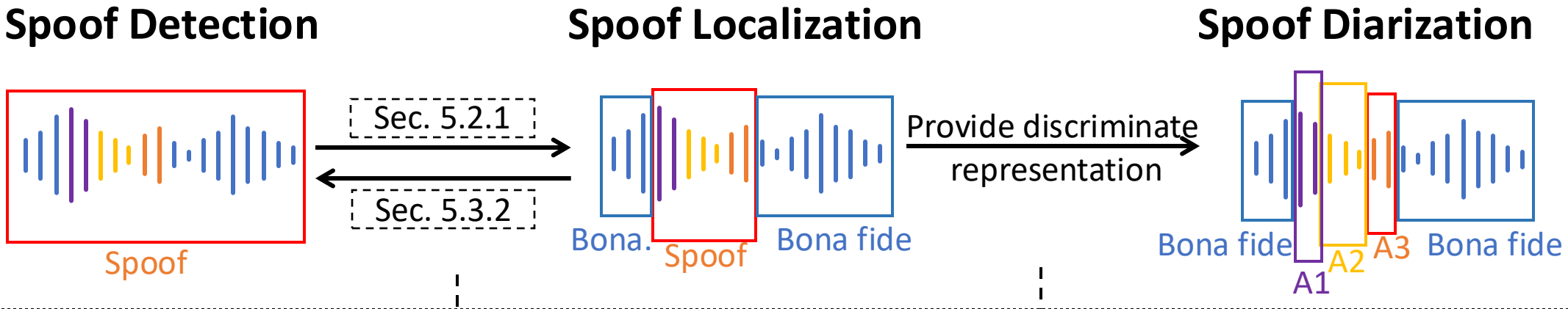
→ May not need to construct an entire utterance

Partial Spoof

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

3 Partial Spoof - Tasks

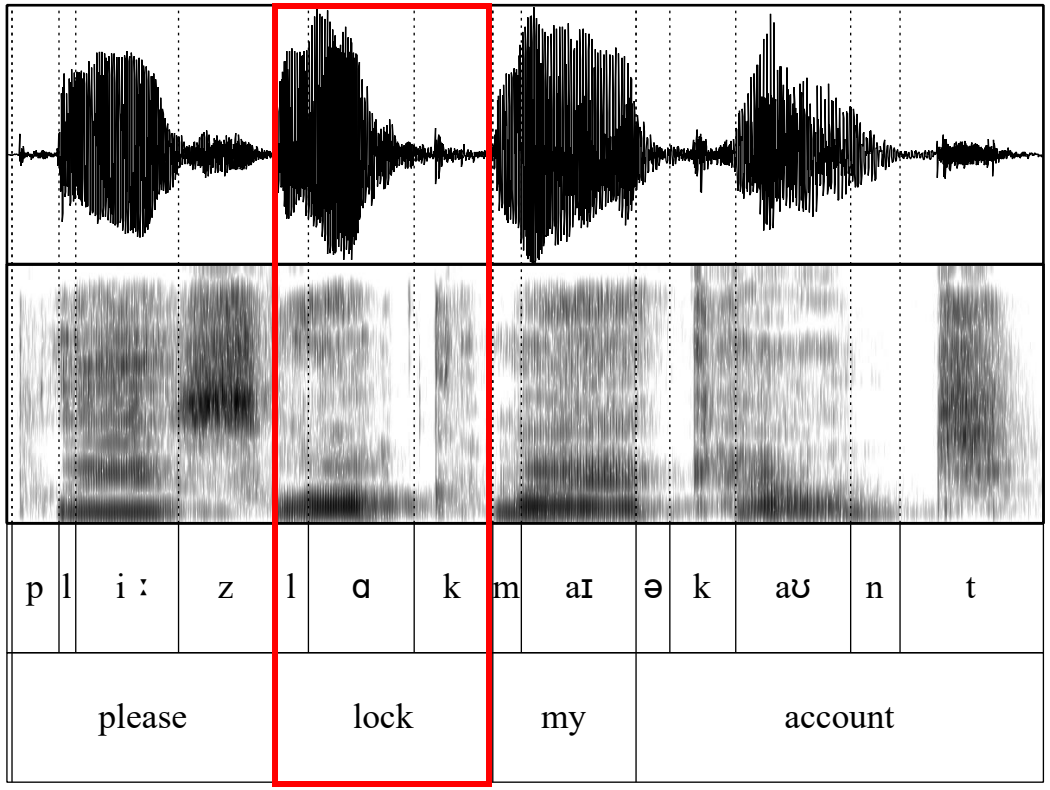
Tasks

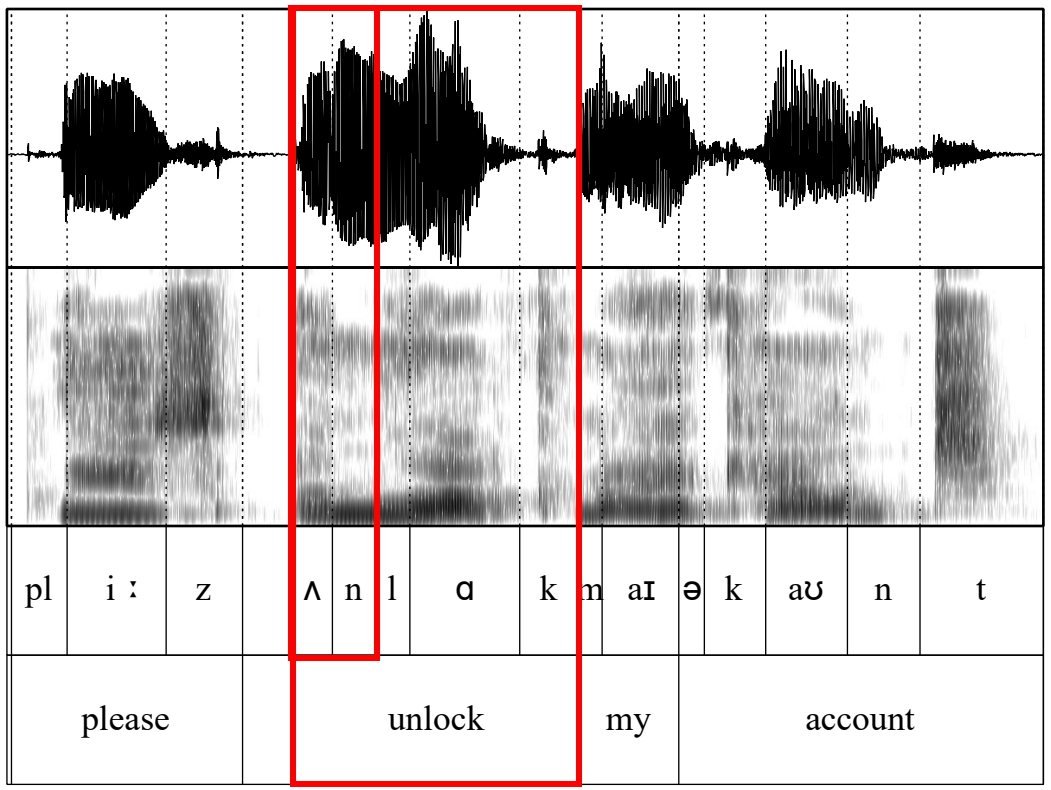


	Spoof Detection	Spoof Localization	Spoof Diarization
Equation	$\boldsymbol{x}_{1:T} \rightarrow c$ $c \in \{bonafide, spoof\}$	$\boldsymbol{x}_{1:T} \rightarrow c_{1:M}$ $c_m \in \{bonafide, spoof\}.$	$\boldsymbol{x}_{1:T} \rightarrow c^{\star}_{1:M}$ $c^{\star}_m \in \{bonafide, A_1, A_2, \cdots\}.$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

3 Partial Spoof – Benchmark Database

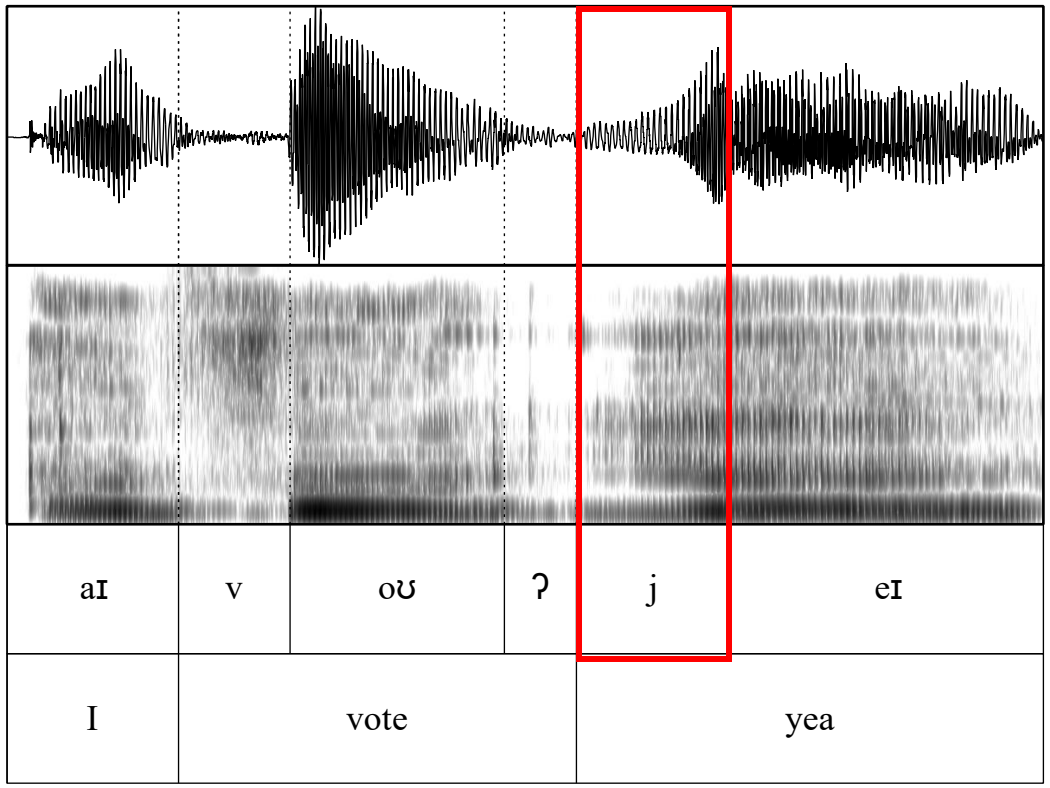
➢ How will attacker manipulate the audio?



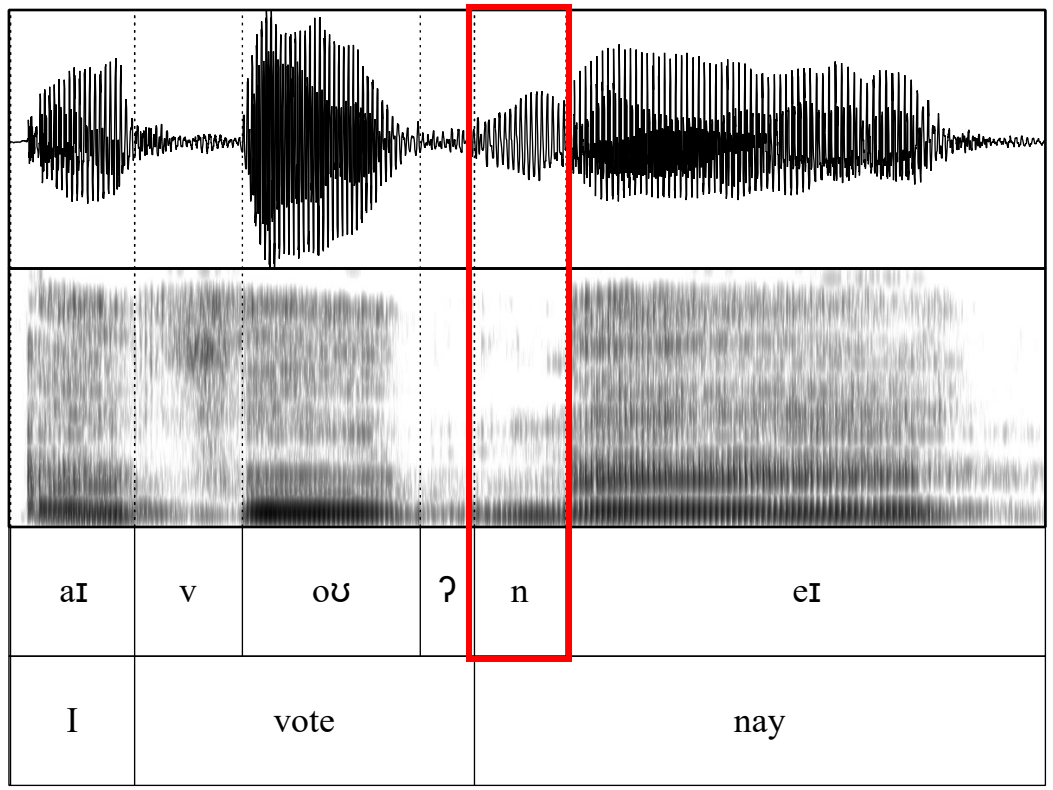


3 Partial Spoof – Benchmark Database

➢ **How will attacker manipulate the audio?**



(yea means “yes”)



(nay means “no”)

Attacker can manipulate segments with **variable duration** to modify utterance.

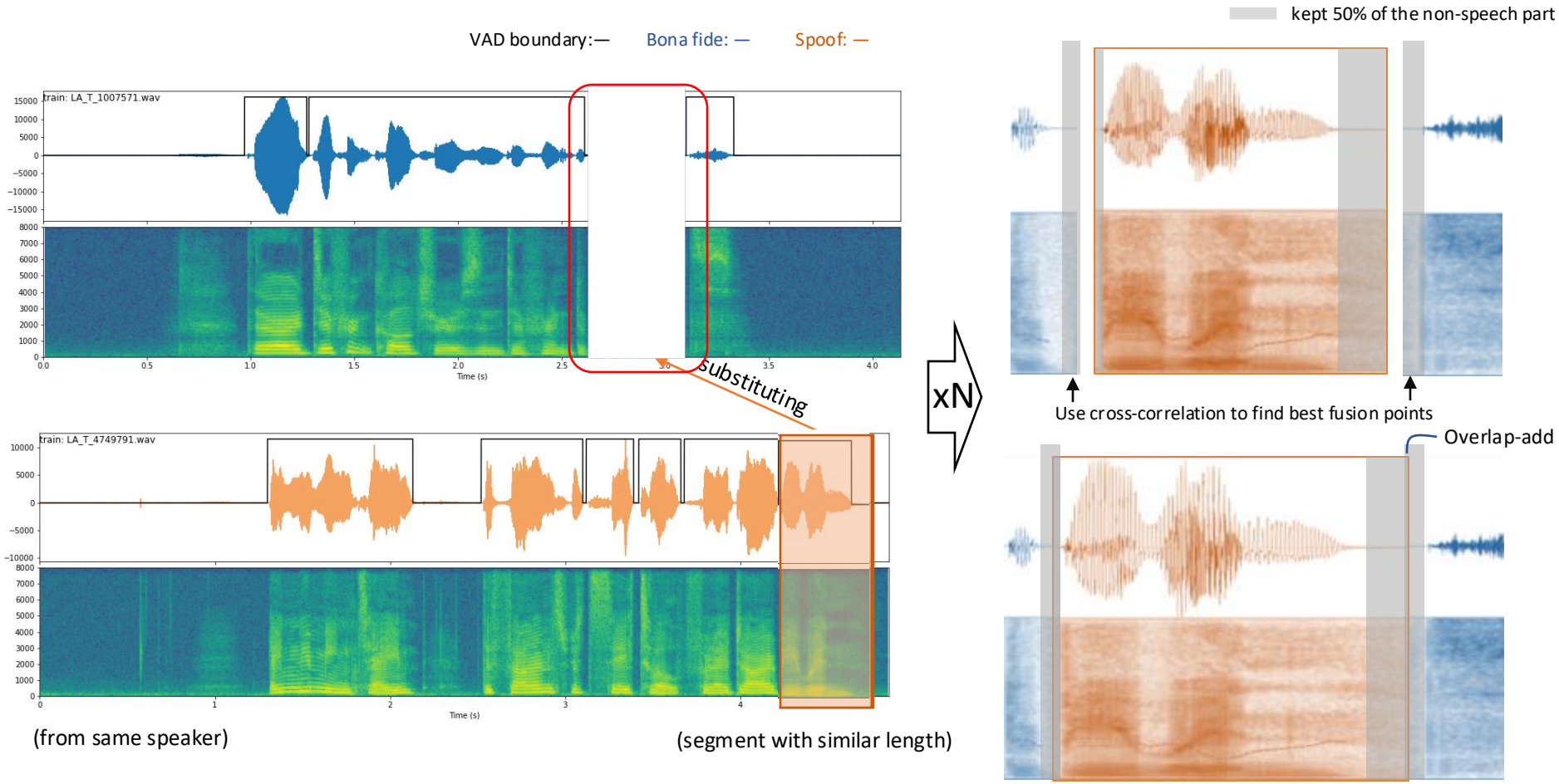
➡ Duration of manipulation should not be fixed. ➡ VAD

VAD: Voice Activity Detection

3 Partial Spoof – Benchmark Database

Processing

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



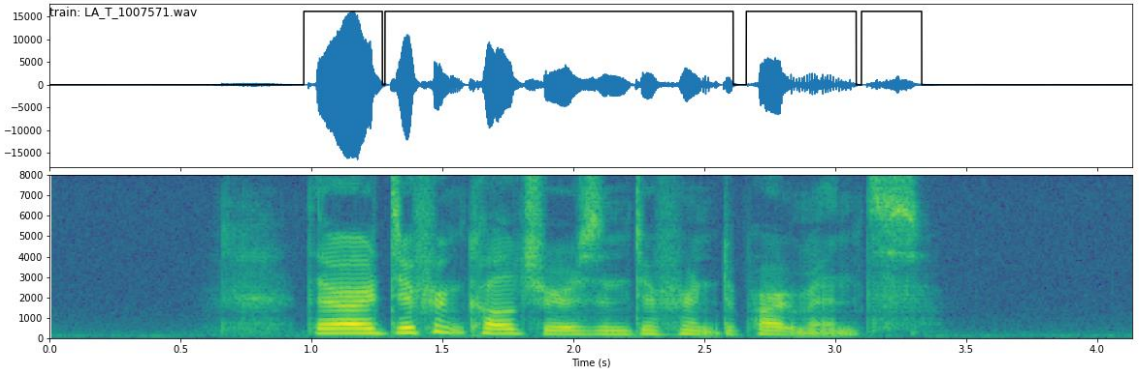
Processing to construct PartialSpoof

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

3 Partial Spoof – Benchmark Database

Source database: ASVspoof 2019 LA database (Wang 2020) (Normalized)

Original bona fide audio:



Partially spoofed audio:

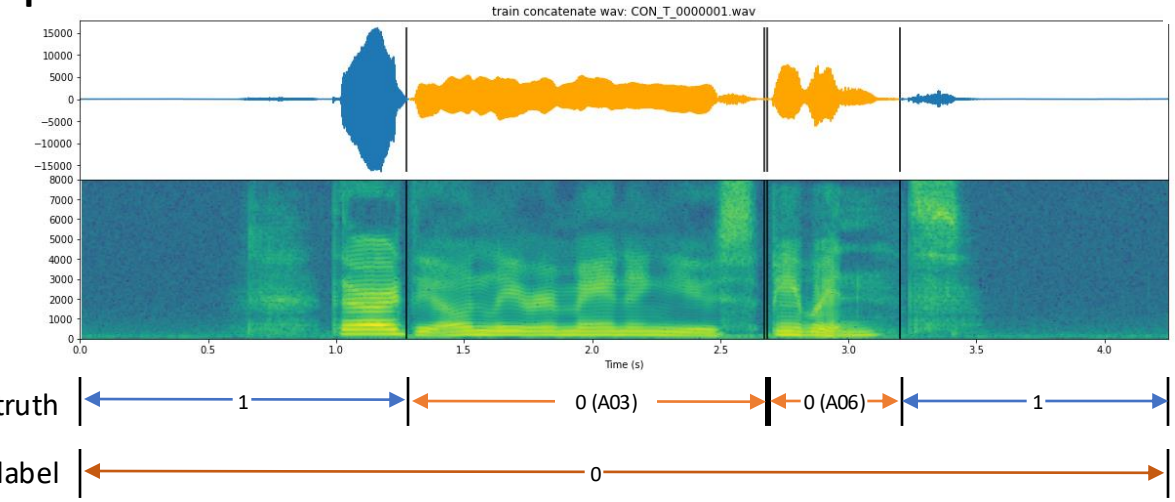


Table 1. *Details of trials in PartialSpoof Database.*

	Subset	#Audio files	Duration (h)	Max #TTS/VC	Audio length (s) min	mean	max
bona fide	Train	2,580	2.43	-	1.36	3.39	11.13
	Dev.	2,548	2.48	-	1.28	3.51	11.39
	Eval.	7,355	6.94	-	0.90	3.40	13.01
spoof	Train	22,800	21.82	6	0.60	3.45	21.02
	Dev.	22,296	21.86	6	0.62	3.53	15.34
	Eval.	63,882	60.74	9	0.48	3.42	18.20

Database

Samples

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

3 Partial Spoof – Related Databases

Table 3.4: Existing databases for the PS scenario in the speech field. (Numbers separated by '/' are shown for train/dev./eval. sets separately. 'Duration-bona' and 'Duration-ps' present the total duration of all bona fide samples and partially spoofed audio samples, separately. '-' indicates unavailable information.)

Database	PartialSpoof [Zhang 2021]	Half-Truth [Yi 2021]	Psynd [Zhang2022]
Year	2021	2021	2022
# Utt. - bona	2,580/2,548/7,355	26,554/8,914/18,144	-
# Utt. - spf	22,800/22,296/63,882	26,554/8,914/18,144	1,963/94/79
Duration-bona	2.43/2.48/6.94	-	-
Duration-ps	21.82/21.86/60.74	-	13
# TTS/VC	6/6/13	1	1
public	yes	yes (from Dec. 2023)	yes
Annotation	Utterance + timestamps	Utterance + timestamps	-
Concatenation	Cross-corr. + overlap-add	pydub	pydub
Language	EN	CN	EN
Source Data	ASVspoof 2019	AISHELL-3	LibriTTS

An approximation to modify the linguistic meaning.

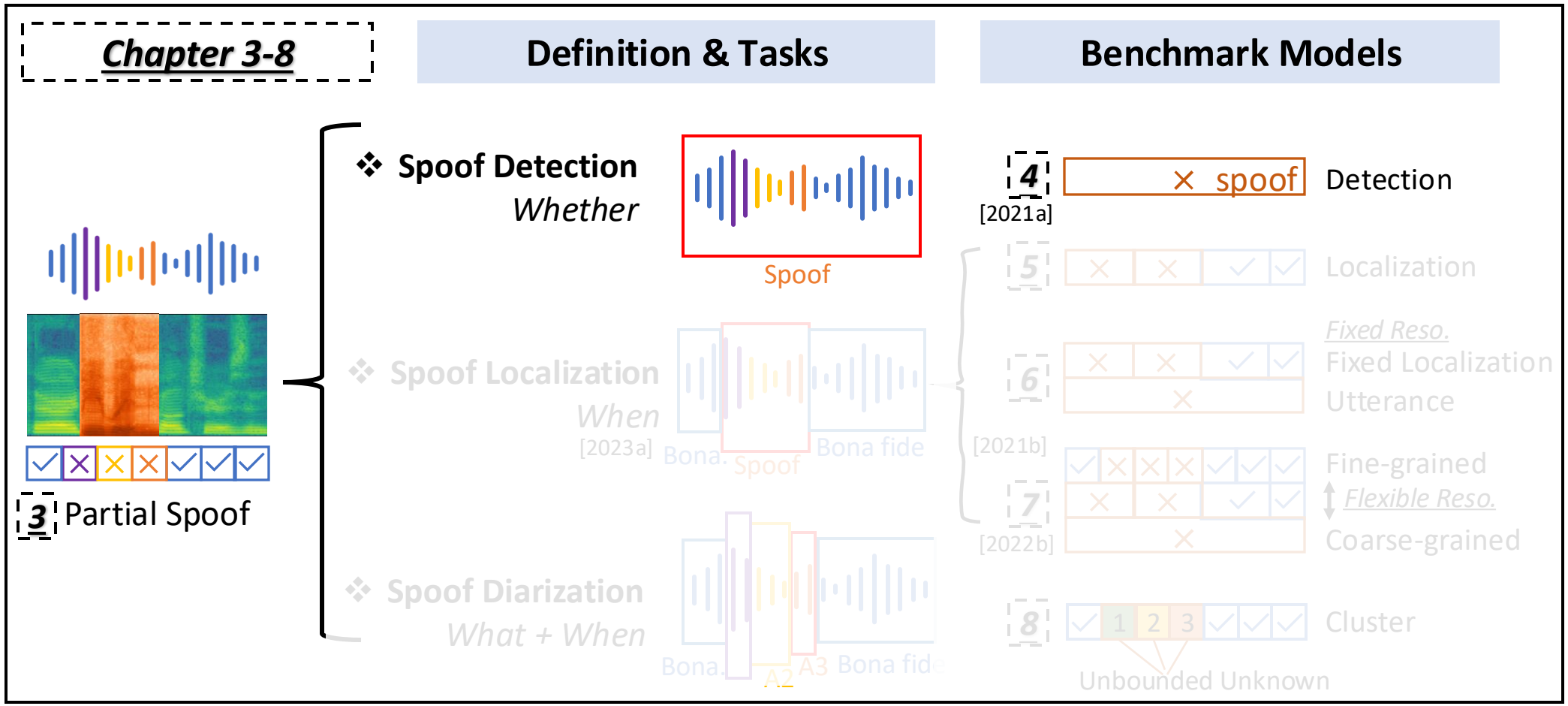
(1) Fixed unit

(2) Limited TTS/VC

PartialSpoof Database:
(1) offers the greatest variety of TTS/VC systems; the ONLY database that considers an open-set scenario.
(2) the ONLY one considers variable-length spoofed regions.

Zhang, L., Wang, X., Cooper, E., Yamagishi, J., Patino, J., & Evans, N. (2021, May 27). PartialSpoof Database - Partially Spoofed Audio Dataset for Anti-spoofing (1.1). Zenodo. https://doi.org/10.5281/zenodo.5112031
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.

Outline



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, A Practical Guide, 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

4 Spoof Detection - Metric

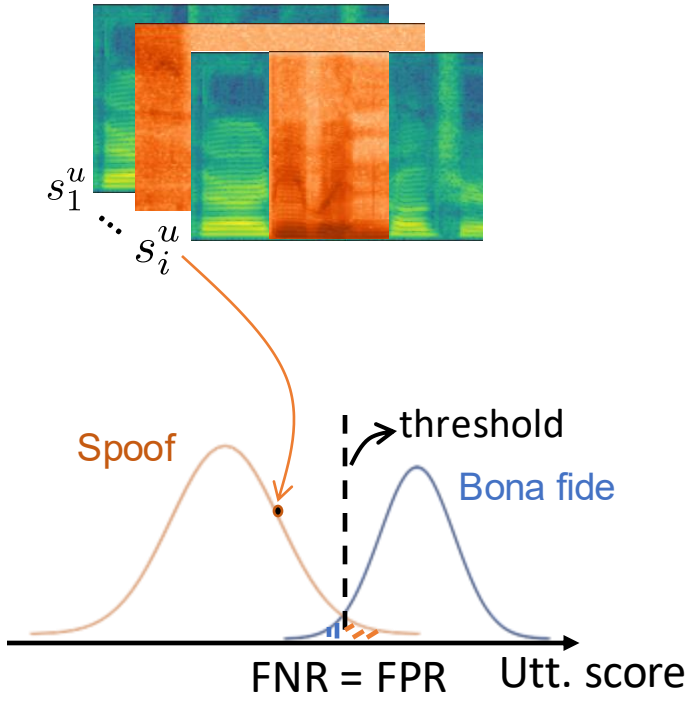
Equal Error Rate (**EER**)

(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

Spoof Detection

- Utterance EER



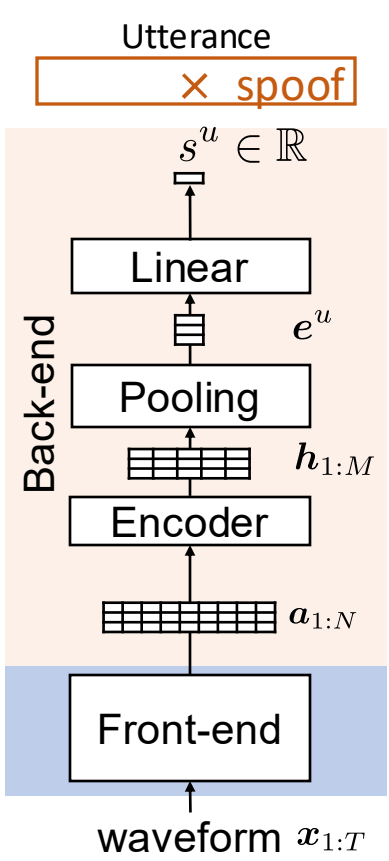
FNR: False Negative Rate; FPR: False Postive Rate

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): ☒

4 Spoof Detection - Model

➢ Whether the input utterance is spoofed?



(a) *Detection*

(Zhang 2021a, Yi 2021, ADD 2022)

$s^u = CM^{utt}(\boldsymbol{x}_{1:T})$ Given an input waveform, CM can derive **an utterance-level score**.

$\boldsymbol{x}_{1:T}$ A waveform with T sampling points.

s^u An utterance-level CM score, indicates how likely the utterance is bona fide.

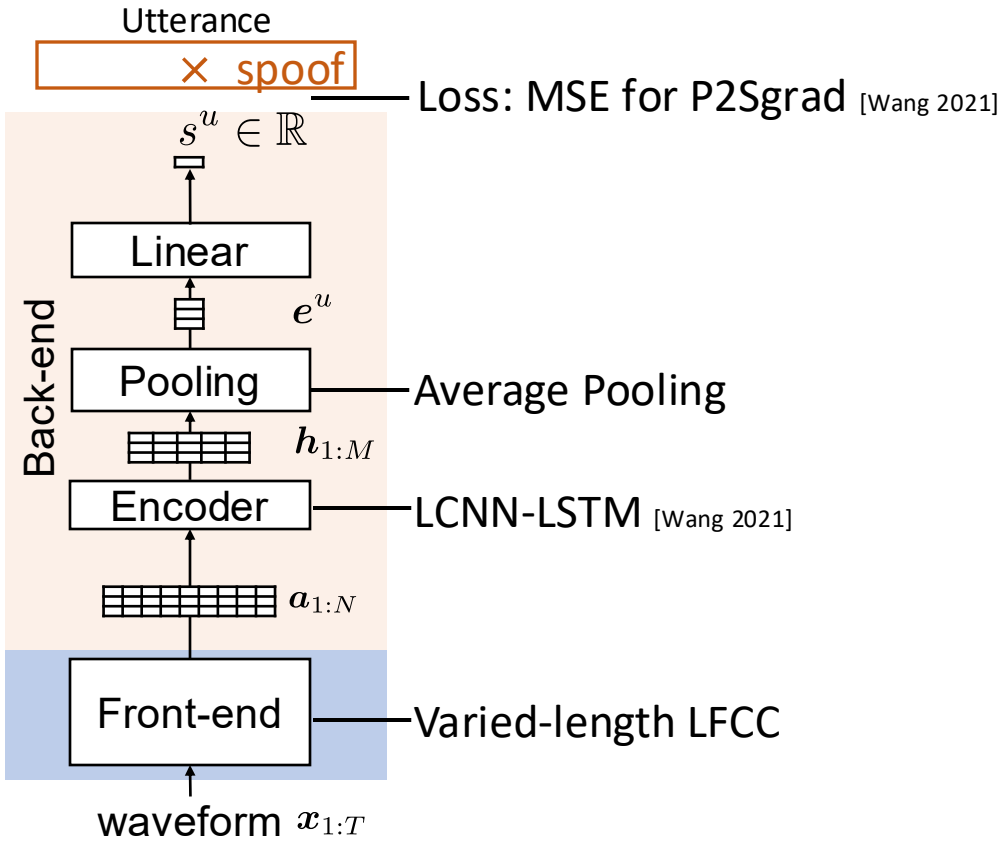
Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans. (2021) An Initial Investigation for Detecting Partially Spoofed Audio. Proc. Interspeech 2021, 4264-4268, doi: 10.21437/Interspeech.2021-738

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

Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., ... & Li, H. (2022, May). Add 2022: the first audio deep synthesis detection challenge. In proc. ICASSP2022, pp. 9216-9220. IEEE.

4 Spoof Detection - Model

➢ Whether the input utterance is spoofed?



(a) *Detection*

(Zhang 2021a, Yi 2021, ADD 2022)

LCNN: Light convolutional neural networks
LSTM: Long Short-Term Memory
LFCC: Linear frequency cepstral coefficient

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
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
Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., ... & Li, H. (2022, May). Add 2022: the first audio deep synthesis detection challenge. In proc. ICASSP2022, pp. 9216-9220. IEEE.

4 Spoof Detection - Results

➢ Whether the input utterance is spoofed? ×

1.1 Spoof detection on the PS scenario is **more difficult** than on the fully spoof scenario.
1.2 **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

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

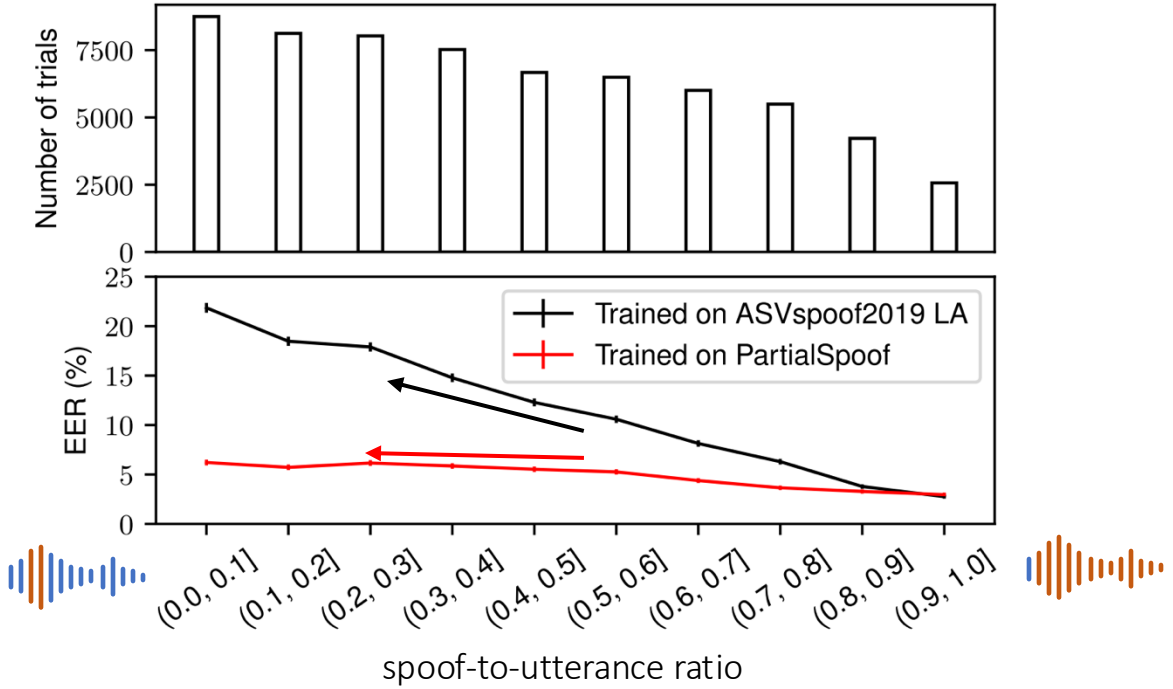


Figure 3. *Break-down results of the cross-scenario study. Top: Histogram of number of trials having different spoof segment ratio. Bottom: EERs for each of the quantized spoof segment ratio.*

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

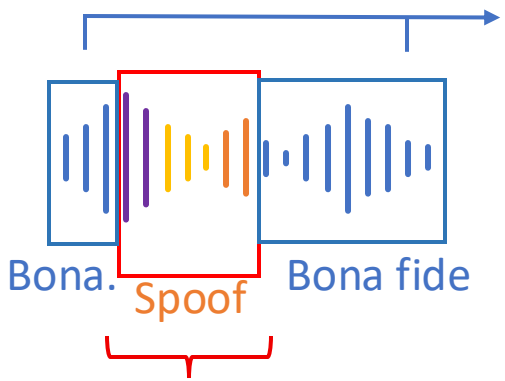
4 Spoof Detection

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

  (Utterance-level)

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

  (segment-level)



Outline

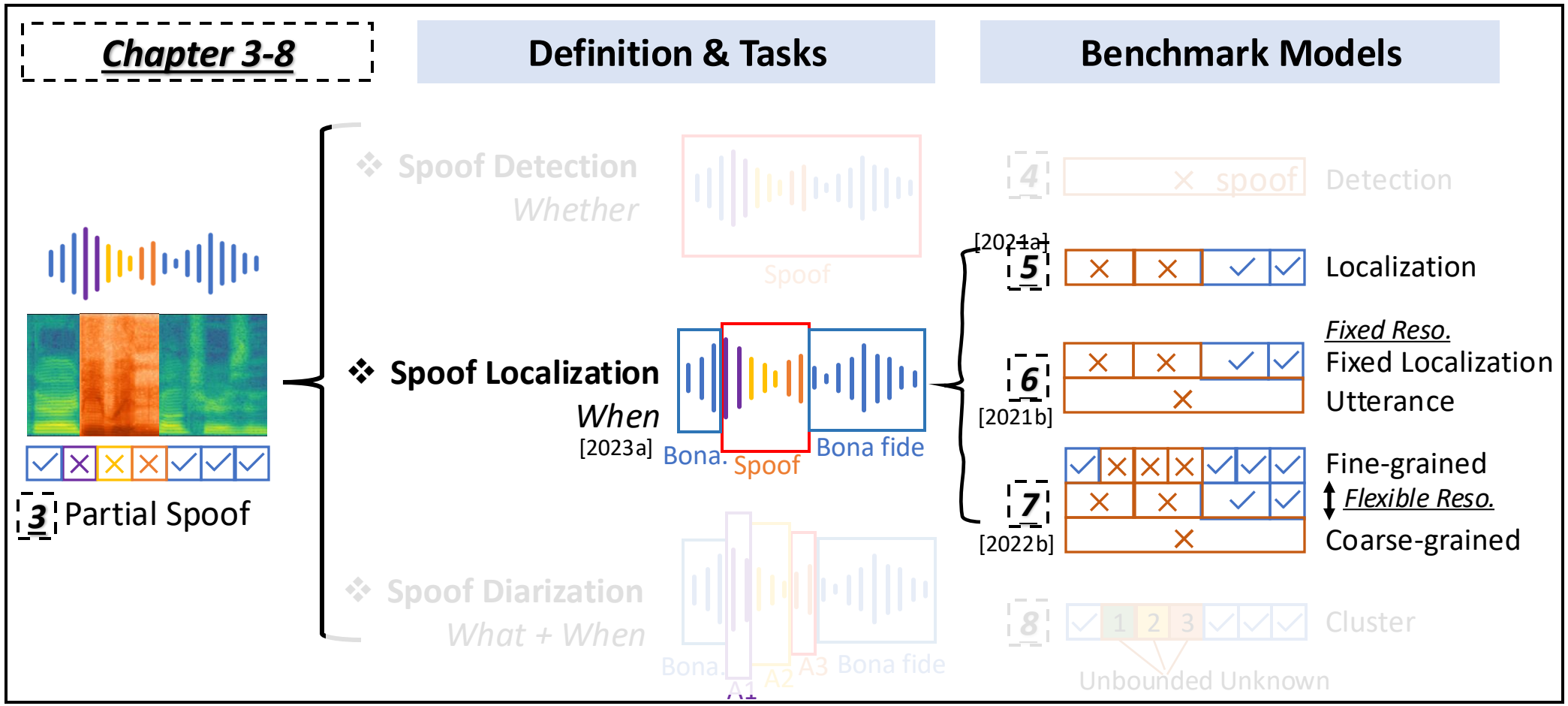


Figure: Thesis Outline & Contribution

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

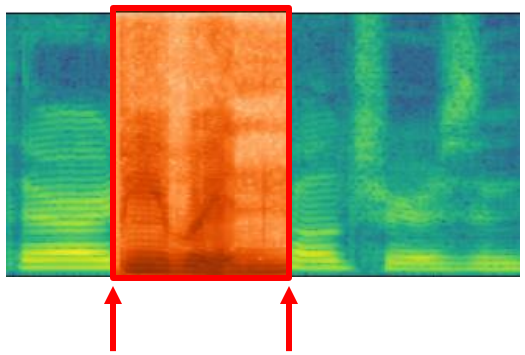
Lin Zhang,.Xin Wang, Erica Cooper, Junichi Yamagishi. (2021) Multi-task Learning in Utterance-level and Segmental-level Spoof Detection. Proc. ASVspoof workshop 2021, 9-15, doi: 10.21437/ASVSPOOF.2021-2

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

5 Spoof 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.

5 Spoof Localization - Metric

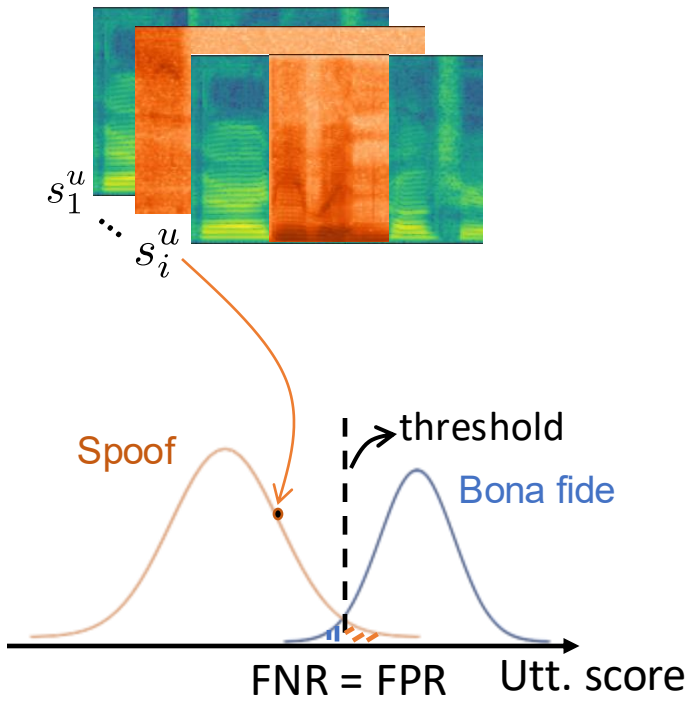
Equal Error Rate (**EER**)

(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

Spoof Detection

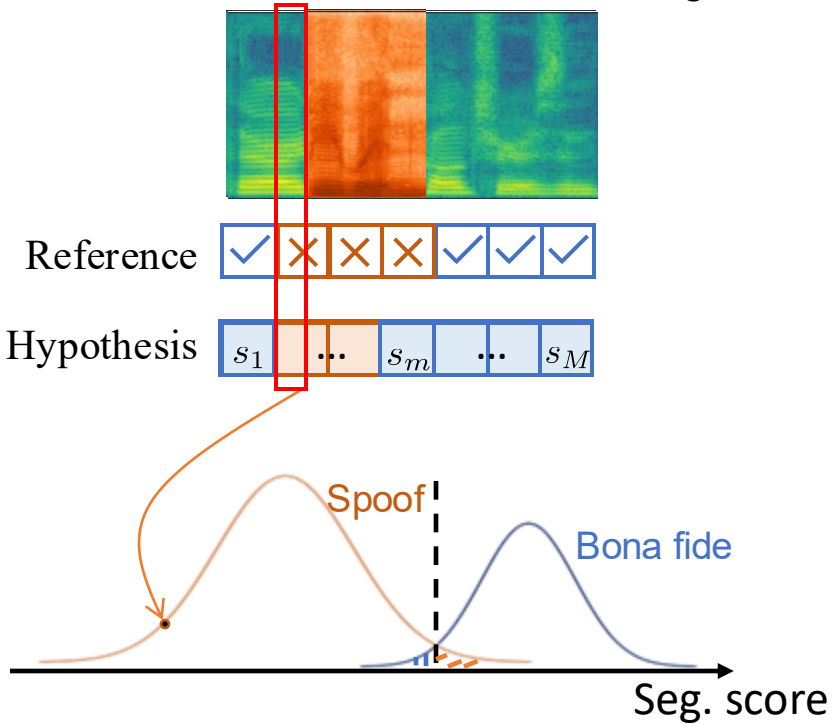
- Utterance EER



Spoof Localization

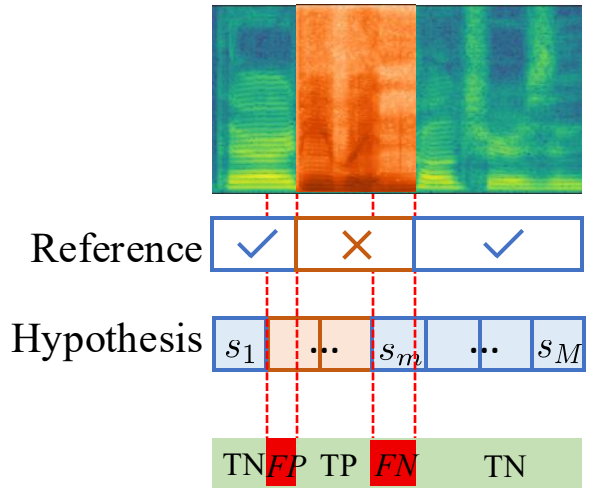
- Segment EER (Point-based)

  Number *of misclassified segments*



- Range-based EER [Zhang 2023]

  Duration *of misclassified regions*



FNR: False Negative Rate; FPR: False Postive Rate

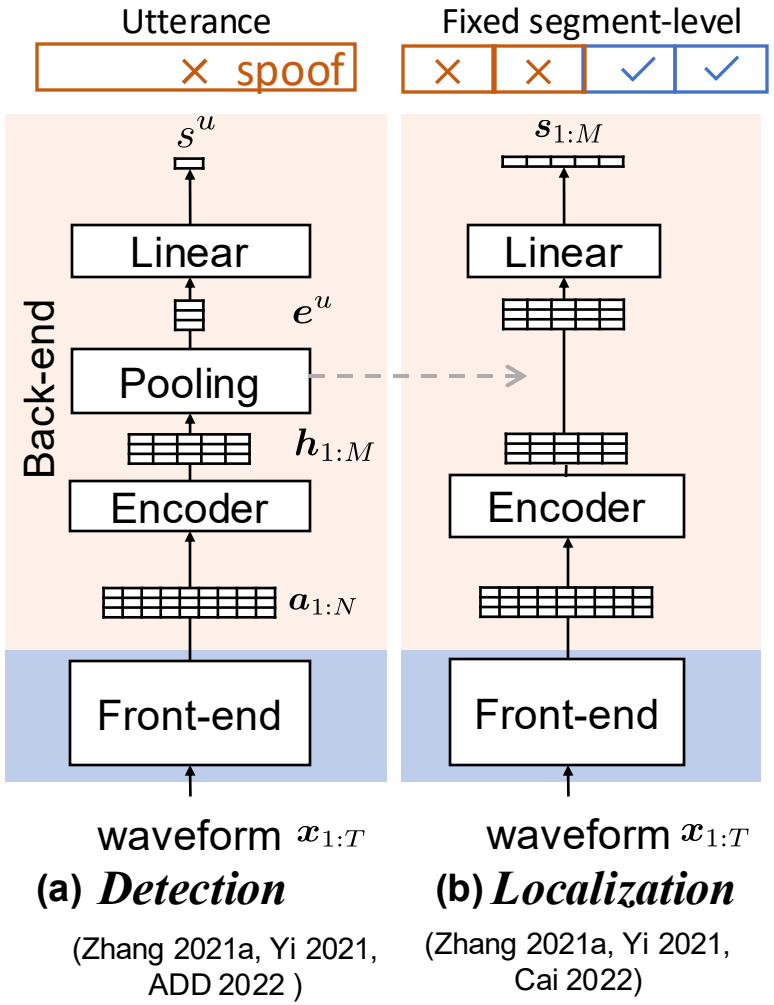
Bona fide (1): ☑ Spoof (0): ☒

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

5 Spoof Localization (single-task)

➢ When do spoofs happen?



$$s^u = CM^{utt}(\boldsymbol{x}_{1:T})$$ derive **an utterance-level score.**

$$\boldsymbol{s}_{1:M} = CM^{seg}(\boldsymbol{x}_{1:T})$$ derive **a sequence of segment-level scores.**

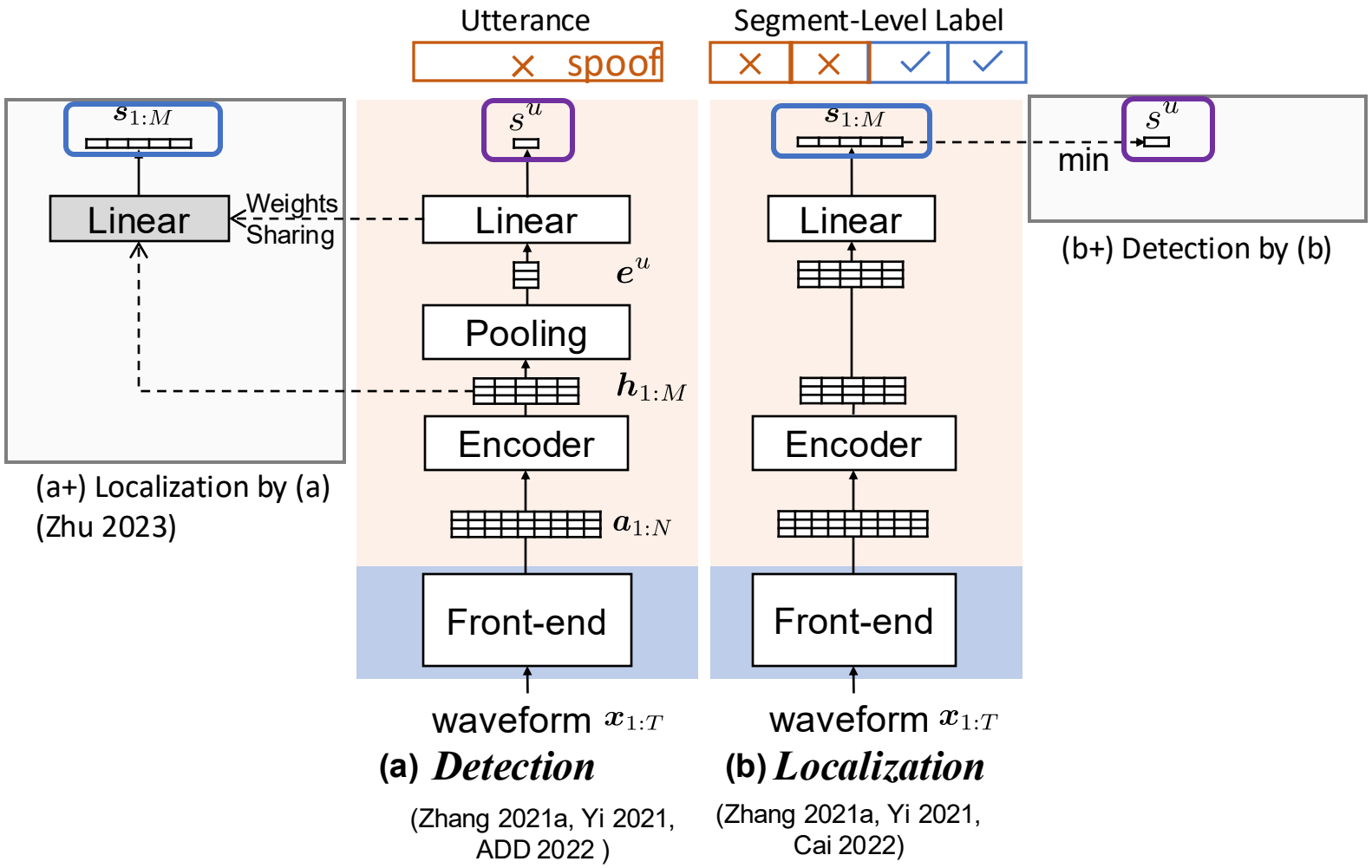
$\boldsymbol{x}_{1:T}$ A waveform with T sampling points.

s^u An utterance-level CM score

$\boldsymbol{s}_{1:M}$ Segment-level CM scores with M segments

Cai, Z., Wang, W., & Li, M. Waveform boundary detection for partially spoofed audio. In *ICASSP 2023*

5 Spoof Localization (single-task)



RangeEER (%) for spoof localization.

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

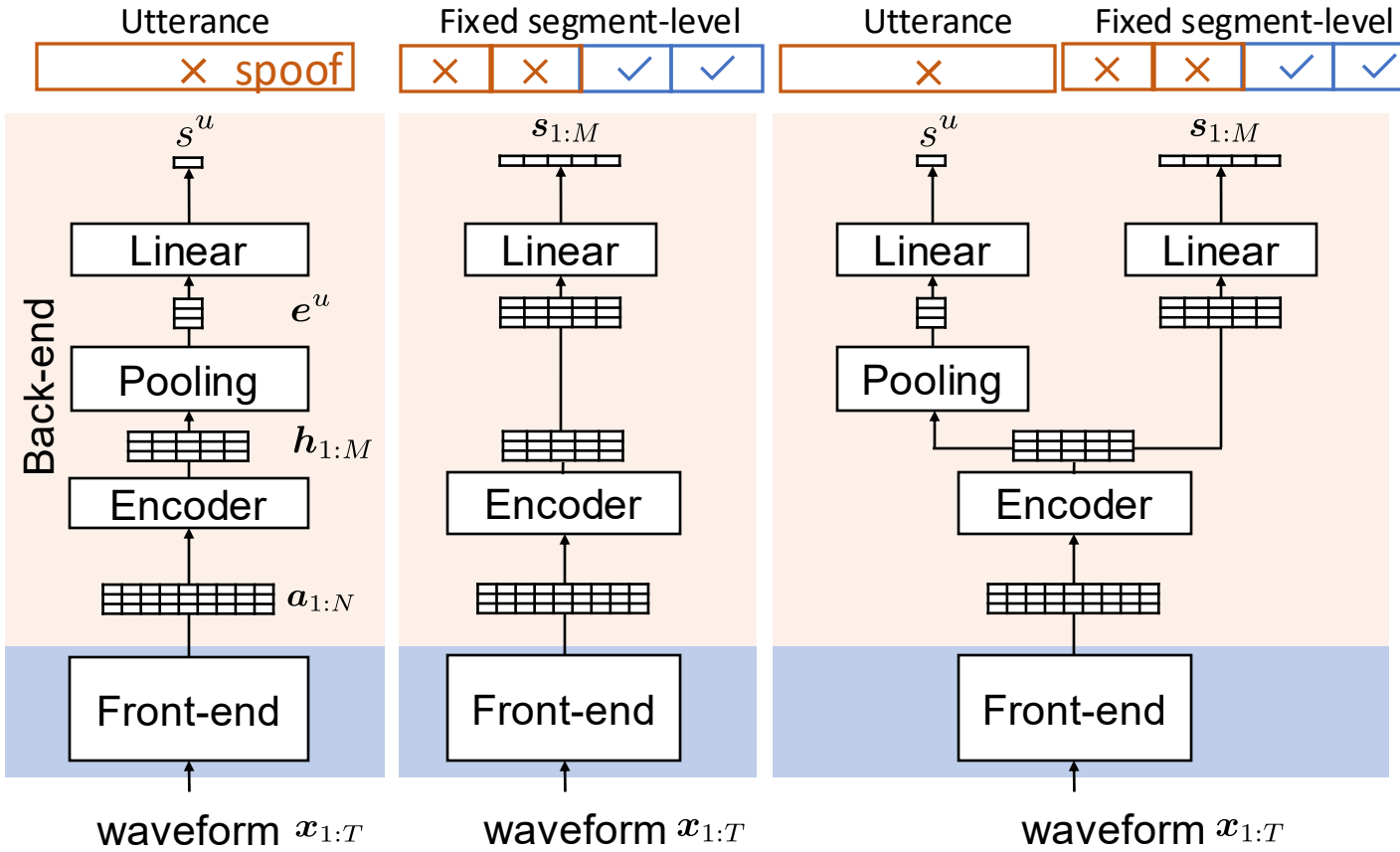
EER (%) for spoof detection.

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

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

6&7 Multi-Task (detection + fixed-resolution localization)

➢ When do spoofs happen?



(Zhang 2021a, Yi 2021, ADD 2022)

(Zhang 2021a, Yi 2021, Cai 2022)

(Zhang 2021b, Wu 2022)

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

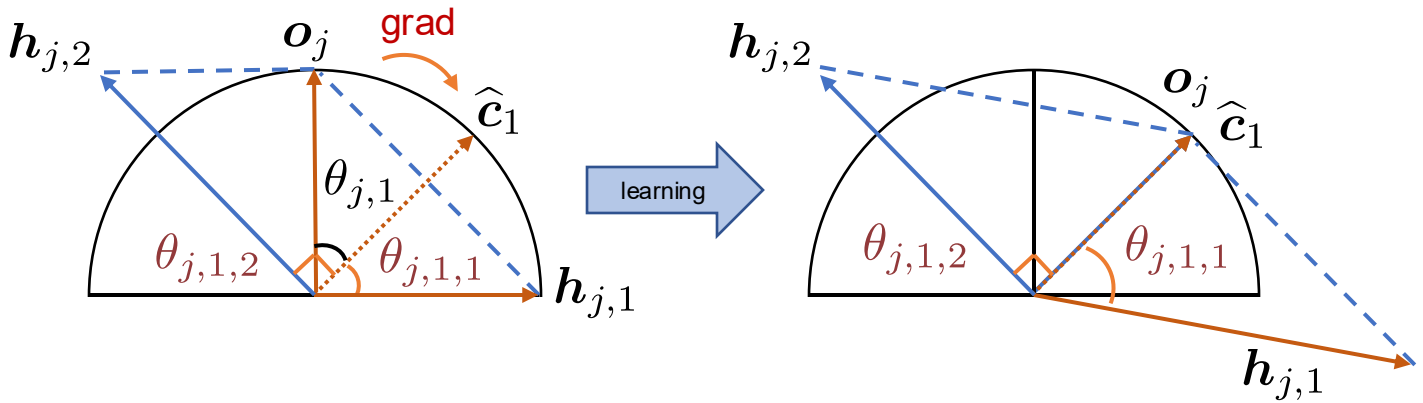
H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.M. Wang, and H. Meng, "Partially fake audio detection by self-attention-based fake span discovery," in Proc. ICASSP, 2022, pp.9236–9240.

6&7 Multi-Task (detection + fixed-resolution localization)

Encoder: SELCNN (Squeeze-and-Excitation LCNN)

Comparison of single-task and multi-task models

	Model Types	Detection EER↓(%) Dev.	Detection EER↓(%) Eval.	RangeEER↓(%) Dev.	RangeEER↓(%) Eval.
Single-task	*Detection*	3.96	6.33	38.91	42.28
	Localization	4.01	7.69	27.20	33.56
	MulBS	2.98	5.90	27.34	33.81



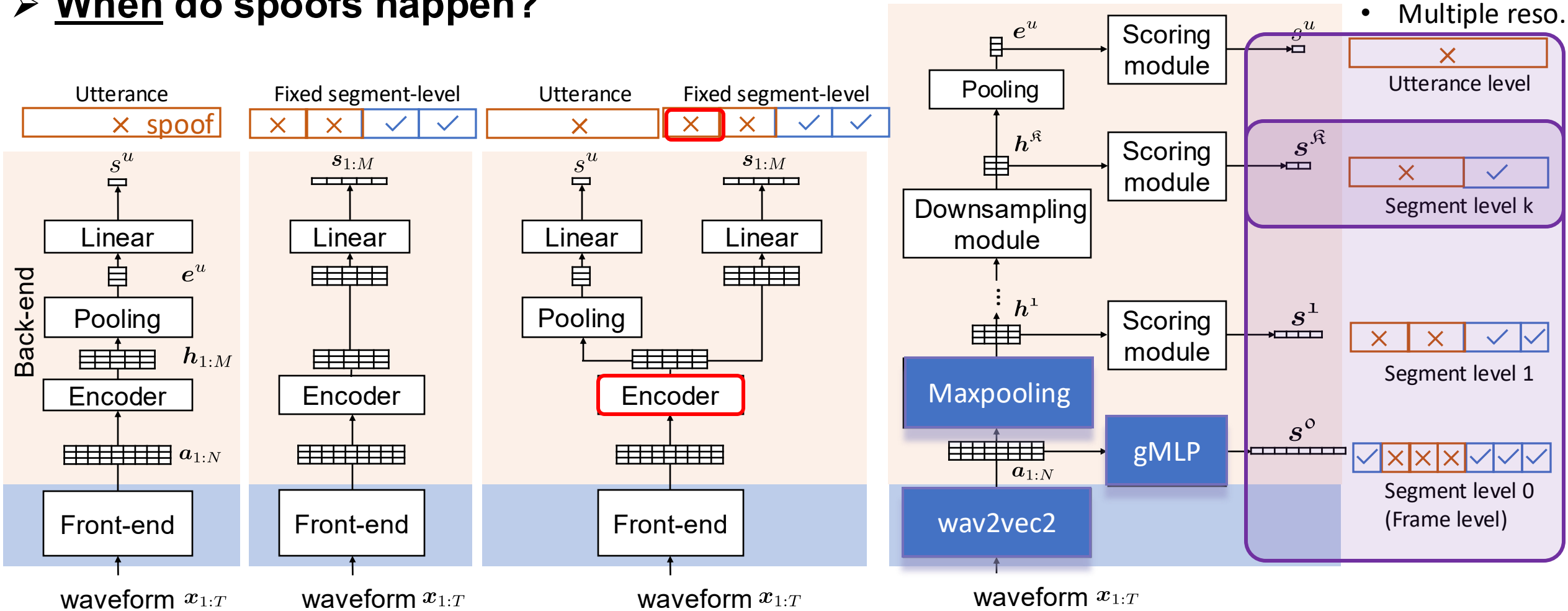
Single-task learning ignores information from another level, and it can force the vector to a wrong direction.

Single-task models cannot handle detection and localization at the same time. A single resolution may not be enough to extract meaningful information in the PS scenario.

6&7 Multi-Task (detection + flexible-resolution localization)

➢ When do spoofs happen?

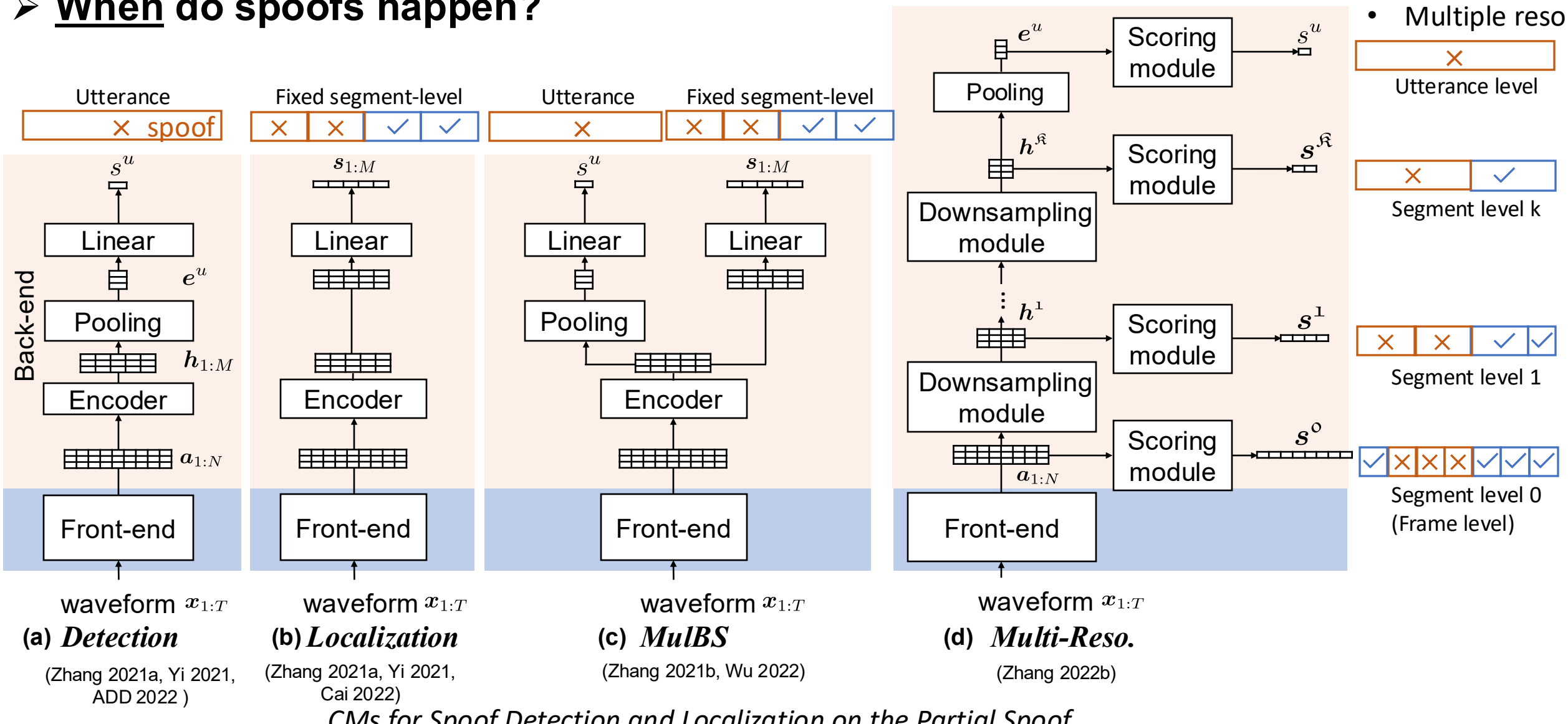
- Single reso.
- Multiple reso.



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.

6&7 Multi-Task (detection + flexible-resolution localization)

➢ When do spoofs happen?

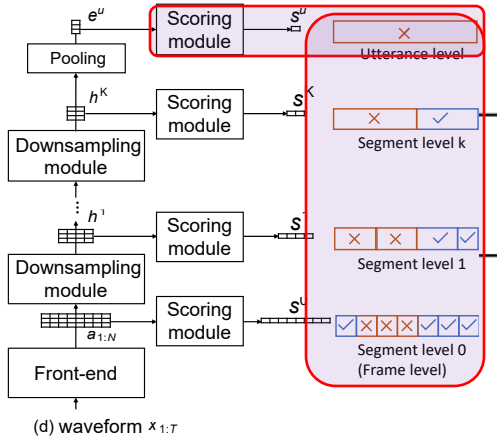


CMs for Spoof Detection and Localization on the Partial Spoof

6&7 Latest Results - Comparison of Different CMs

- **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.	LFCC	LCNN-BLSTM	42.71	6.19
5.2			160 ms			33.76	8.61
6.2			utt.	LFCC	SELCNN-BLSTM	42.28	6.33
			160 ms			33.56	7.69
			160 ms, utt.			33.81	5.90
7.2	*Single reso.*	d	20 ms or utt.	w2v2-large (Baevski 2020)	5gmlp (Liu 2021)	**29.27**	0.64
	Multi reso.	d	20~640, utt.			30.40	**0.49**

1. **Multi-resolution** CM can do detection and localization at different resolutions and **SSL-based** front-end is helpful.

2. For spoof detection, training on the localization task with more fine-grained information can help improve the performance.

3. For spoof localization, training at the finer temporal resolution performs better.

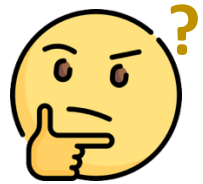
A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS 2020, pp. 12449–12460
H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in Proc. NeurIPS 2021, pp. 9204–9215.

All experiments were repeated three times with different random seeds.

6&7 Latest Results - Cross-Scenario

Can the CMs trained on partially (fully) spoof detect fully (partially) spoofed utterance?

PartialSpoof — ASVspoof2019 LA



Utterance EERs (%) of the cross-scenario study.

Train \ Test	Fully Spoof	Partial Spoof
Fully Spoof	0.83	14.19 ↓
Partial Spoof	0.77	0.64

4. Using **Partial Spoof training data is beneficial.**

6&7 Latest Results - Comparison of Different CMs

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

Performance of different CMs on the PartialSpoof evaluation set.

Sec.	Model ID	Types (Fig. 3.9)	Training Resolutions	Front-end	Back-end	Localization RangeEER(%)	Detection EER(%)
4.2	*Detection*	a	utt.	LFCC	LCNN-BLSTM	42.71	6.19
5.2	*Localization*	b	160 ms			33.76	8.61
	Detection	a	utt.	LFCC	SELCNN-BLSTM	42.28	6.33
6.2	*Localization*	b	160 ms			33.56	7.69
	MulBS	c	160 ms, utt.			33.81	5.90
7.2	*Single reso.*	d	20 ms or utt.	w2v2-large	5gmlp	**29.27**	0.64
	Multi reso.	d	20~640, utt.			30.40	**0.49**

5. **Spoof localization** is a **more complex** task than the spoof detection.

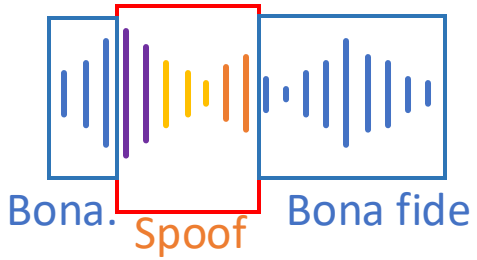
A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS 2020, pp. 12449–12460
H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in Proc. NeurIPS 2021, pp. 9204–9215.

All experiments were repeated three times with different random seeds.

8 Spoof Diarization

- **Spoof Detection: Whether the input utterance is spoofed?**

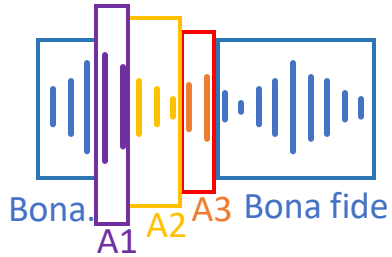
- **Spoof Localization: When do spoofs happen?**



Trace back to the original algorithm, then do further analysis.

VC: Trace back to the original speaker. [Cai 2022]

- **Spoof Diarization: What attacks when?**



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

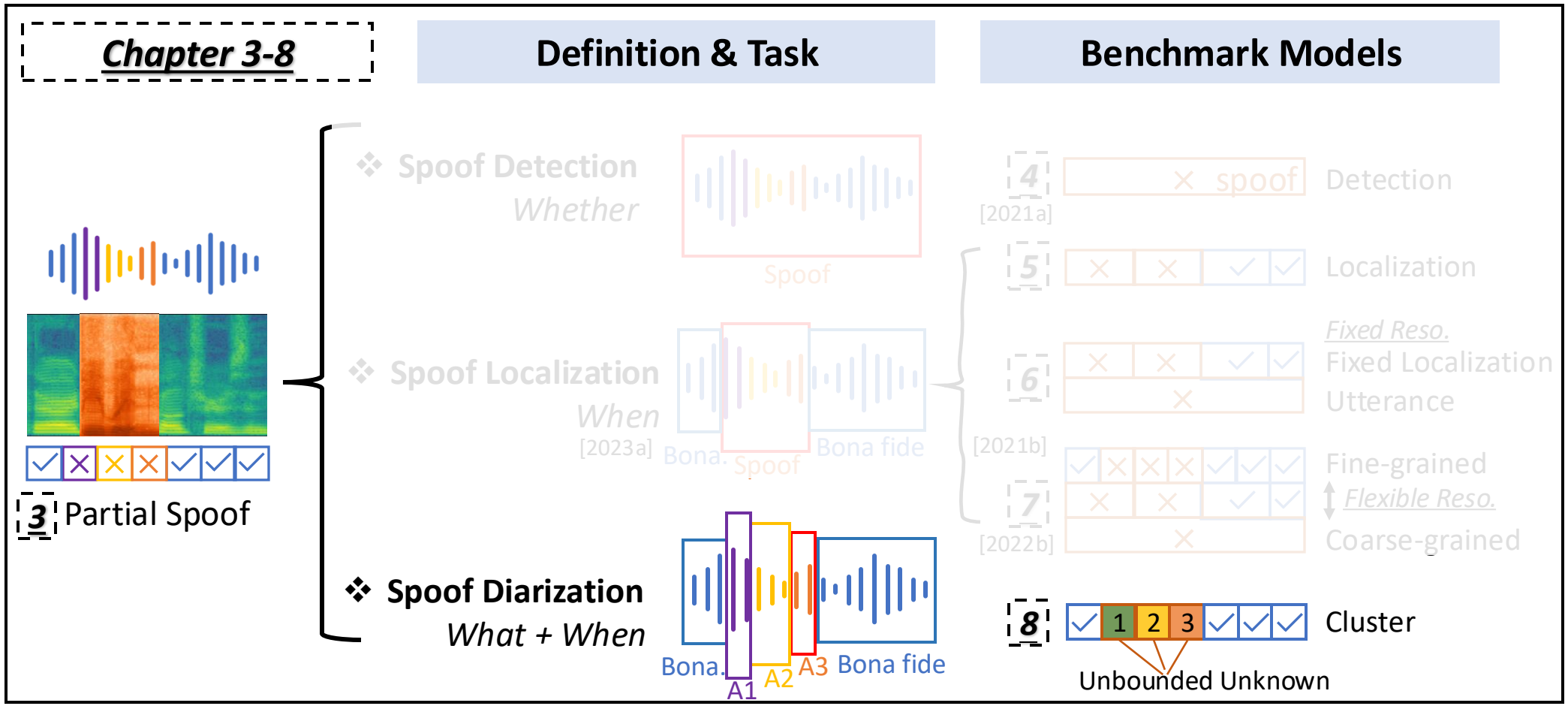


Figure: Thesis Outline & Contribution

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

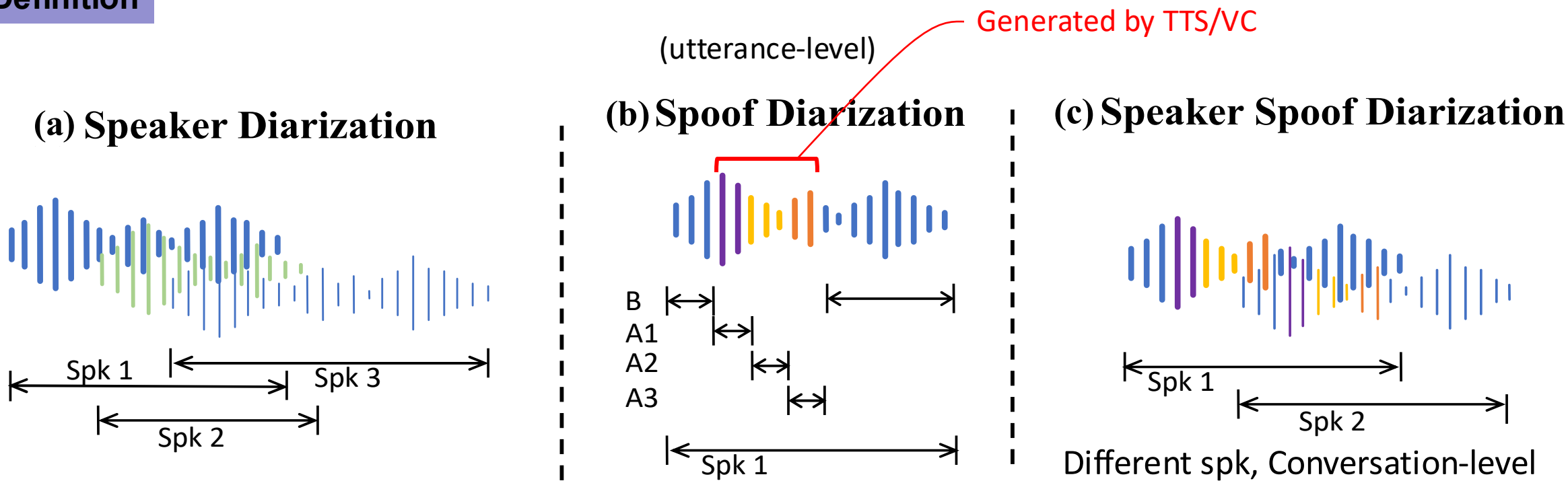
Lin Zhang,.Xin Wang, Erica Cooper, Junichi Yamagishi. (2021) Multi-task Learning in Utterance-level and Segmental-level Spoof Detection. Proc. ASVspoof workshop 2021, 9-15, doi: 10.21437/ASVSPOOF.2021-2

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

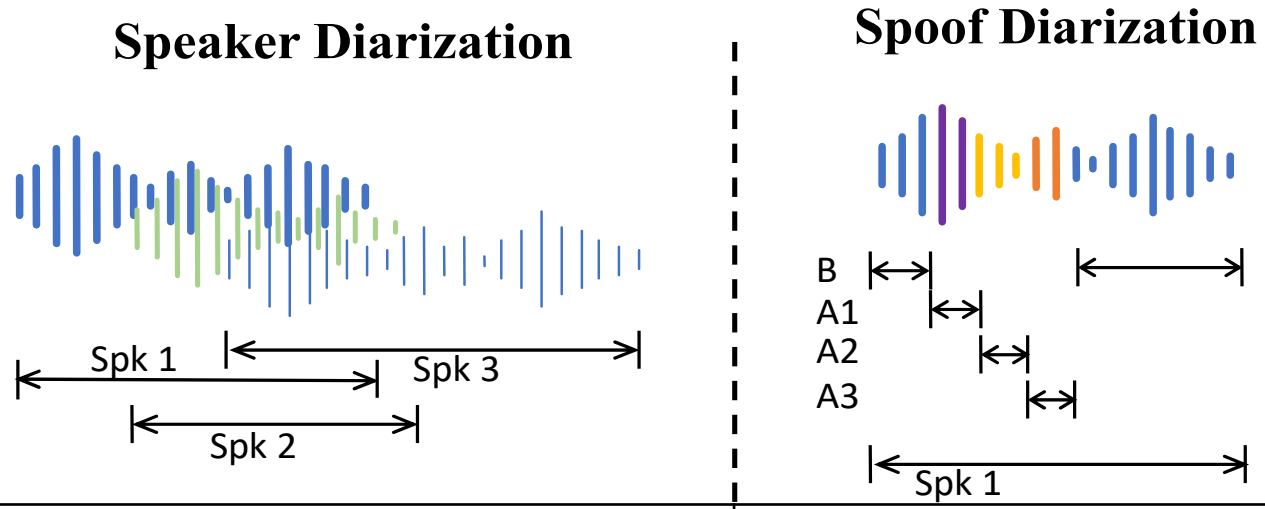
8 Spoof Diarization

Definition



(b.1) One spk, Short duration, Utterance-level
(b.2) One spk, Long duration, Presentation-level

8 Spoof Diarization



		Speaker diarization	Spoof diarization
Similarities		1. Unbounded unknown clusters.	
		2. Different cluster in the audio can be with variable speech duration.	
Differences	3. Duration	**Long conversation** that contains multiple utterances from several speakers.	**Short-duration** diarization, a single utterance consists of different segments for the same speaker.
	4. Strict cluster	Permutation-invariant, no need to identify the speakers by name or definite ID.	Two **primary cluster groups** exist: bona fide, spoof.
	5. Object	(Train on simulated data.) Diarize **real** recording (overlap).	(Train on simulated data.) Diarize **manipulated** recording (no-overlap).

8 Spoof Diarization

ASVspoof 2019

Different processing

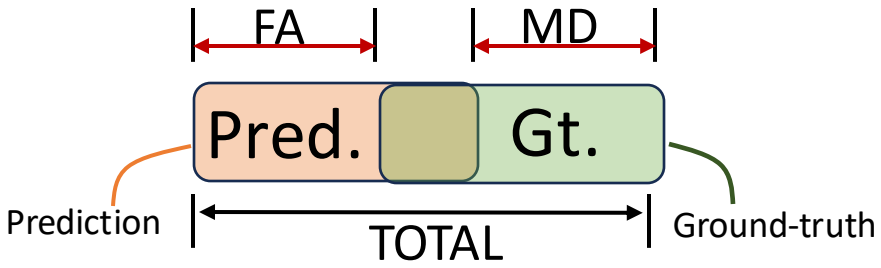
Table 1: Summary of LA spoofing systems. * indicates neural networks. For abbreviations in this table, please refer to Section 3. Note that A04 and A16 use same waveform concatenation TTS algorithm, and A06 and A19 use same VC algorithm.

		Input	Input processor	Duration	Conversion	Speaker represent.	Outputs	Waveform generator	Post process
Train / Dev.	A01	Text	NLP	HMM	AR RNN*	VAE*	MCC, F0	WaveNet*	
	A02	Text	NLP	HMM	AR RNN*	VAE*	MCC, F0, BAP	WORLD	
	A03	Text	NLP	FF*	FF*	One hot embed.	MCC, F0, BAP	WORLD	
	A04	Text	NLP	-	CART	-	MFCC, F0	Waveform concat.	
	A05	Speech (human)	WORLD	-	VAE*	One hot embed.	MCC, F0, AP	WORLD	
	A06	Speech (human)	LPCC/MFCC	-	GMM-UBM	-	LPC	Spectral filtering + OLA	
Eval.	A07	Text	NLP	RNN*	RNN*	One hot embed.	MCC, F0, BA	WORLD	GAN*
	A08	Text	NLP	HMM	AR RNN*	One hot embed.	MCC, F0	Neural source-filter*	
	A09	Text	NLP	RNN*	RNN*	One hot embed.	MCC, F0	Vocaine	
	A10	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*	Mel-spectrograms	WaveRNN*	
	A11	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*	Mel-spectrograms	Griffin-Lim [13]	
	A12	Text	NLP	RNN*	RNN*	One hot embed.	F0+linguistic features	WaveNet*	
	A13	Speech (TTS)	WORLD	DTW	Moment matching*	-	MCC	Waveform filtering	
	A14	Speech (TTS)	ASR*	-	RNN*	-	MCC, F0, BAP	STRAIGHT	
	A15	Speech (TTS)	ASR*	-	RNN*	-	MCC, F0	WaveNet*	
	A16	Text	NLP	-	CART	-	MFCC, F0	Waveform concat.	
	A17	Speech (human)	WORLD	-	VAE*	One hot embed.	MCC, F0	Waveform filtering	
	A18	Speech (human)	MFCC/i-vector	-	Linear	PLDA	MFCC	MFCC vocoder	
	A19	Speech (human)	LPCC/MFCC	-	GMM-UBM	-	LPC	Spectral filtering + OLA	

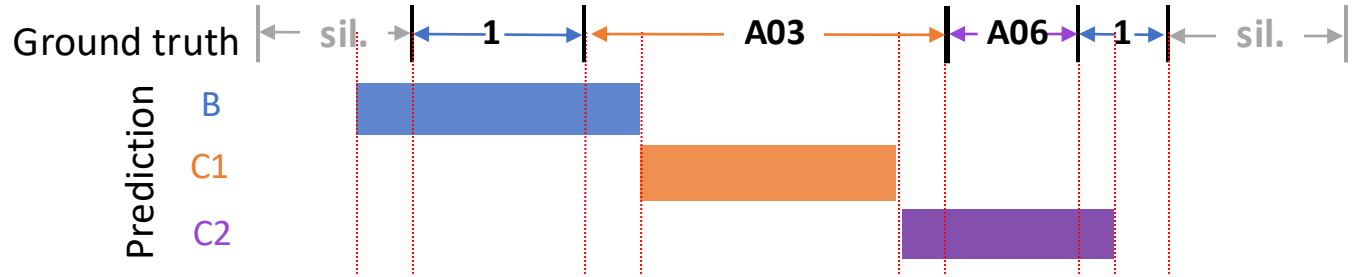
Although our PartialSpoof database is based on the ASVspoof2019, partially spoofed audios could be generated by any unknown spoofing methods.

8 Spoof Diarization

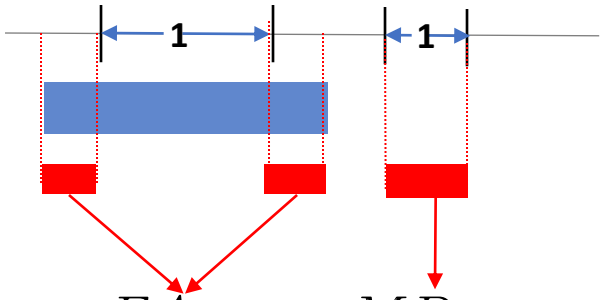
Metric Spoof Jaccard Error Rate



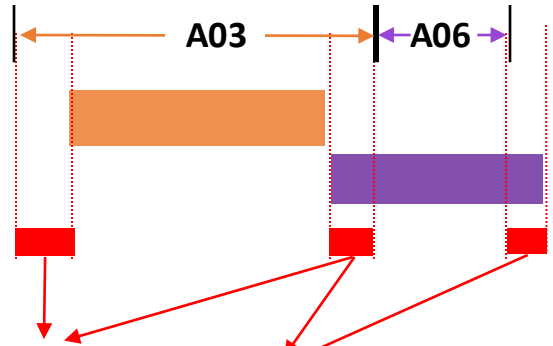
(a) Jaccard Error Rate (JER)



(b) A prediction sample in the PS scenario.



$$JER_{bona} = \frac{FA_{bona} + MD_{bona}}{TOTAL_{bona}}$$



$$\frac{1}{2}\left(\frac{FA_{A03} + MD_{A03}}{TOTAL_{A03}} + \frac{FA_{A06} + MD_{A06}}{TOTAL_{A06}}\right)$$

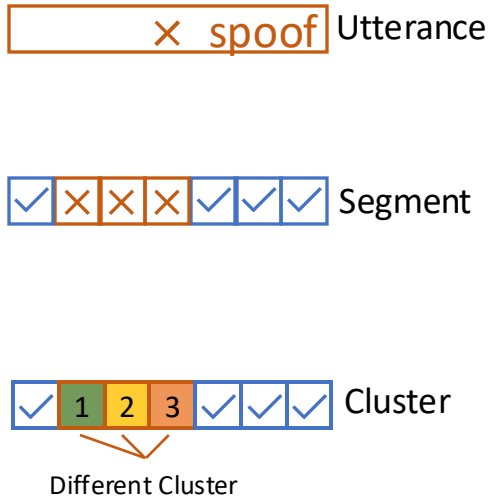
$$JER_{spoof} = \frac{1}{|\mathcal{A}|}\sum_{A_i \in \mathcal{A}} JER_{A_i} = \frac{1}{|\mathcal{A}|}\sum_{A_i \in \mathcal{A}} \frac{FA_{A_i} + MD_{A_i}}{TOTAL_{A_i}}$$

(c) Calculating of JER_{bona} and JER_{spoof}

8 Spoof Diarization

Model

- **1. Whether**
- **2. When**
- **3. What + When**



Supervised Binary Classification

↓ All different spoof sharing the same class ‘spoof’

Supervised Multi Classification

↓ New attacks unknown to the model may occur.

+ Unsupervised Clustering

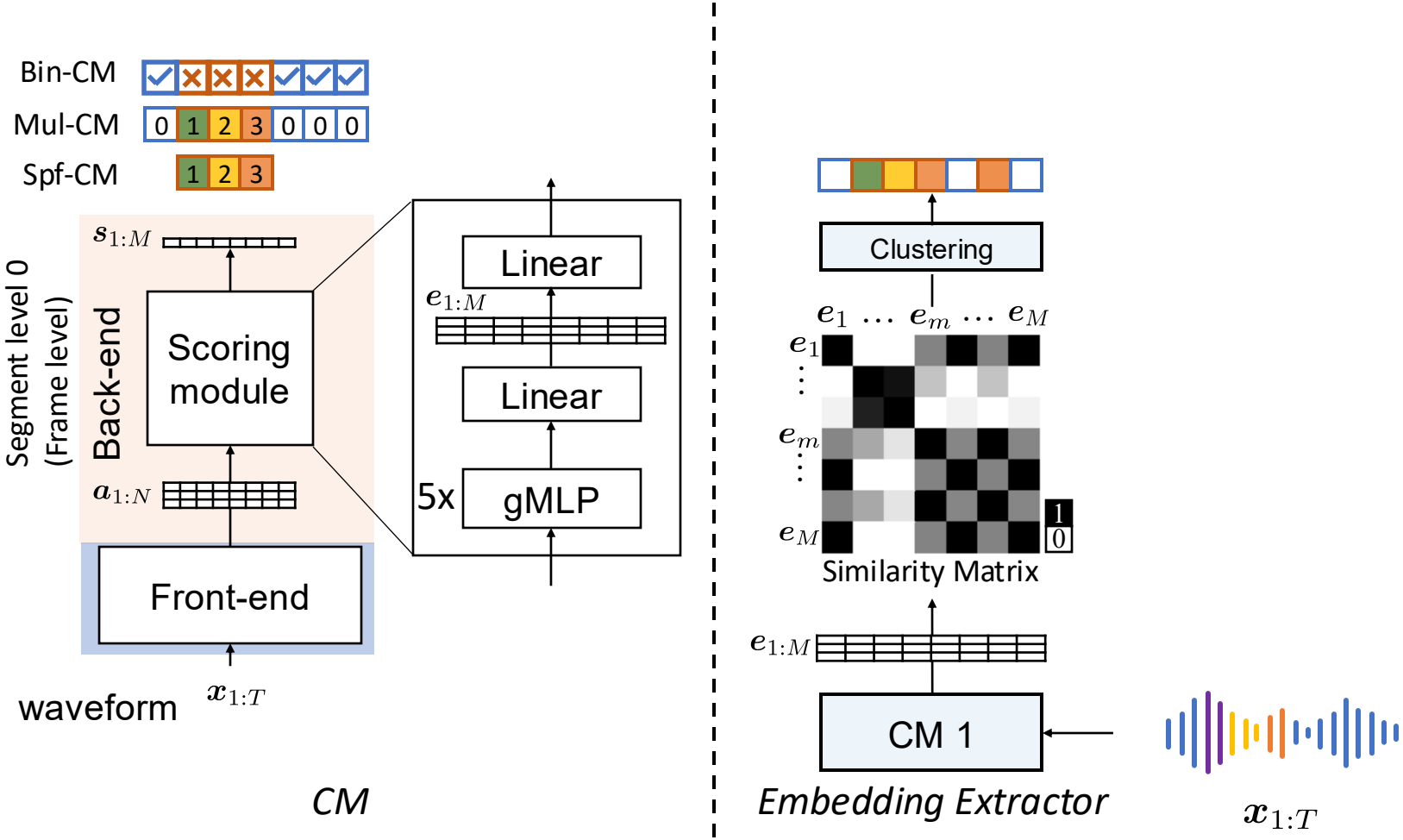
└ Increase the generalizability

When an unseen spoof comes, we can group it with exist cluster or arrange it as a new one.

[10] 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.

Spoof Diarization

How do bona fide and spoof labeling schemes affect the ability of CMs?



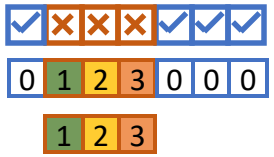
(e) CM-Conditional Clustering (3C) model for spoof diarization

LCM: Label-based CM-constrain
DCM: Distance-based CM-constrain

8 Spoof Diarization

Table 8.1: Results on the different combinations of CMs (clustering was done with the oracle cluster number).

Model			Development set		Evaluation set	
CM-1	CM-2	Condition	JER_{bona}	JER_{spoof}	JER_{bona}	JER_{spoof}
Bin-CM	/	/	6.49	34.63	**17.01**	39.89
Mul-CM	/	/	**4.01**	**4.89**	19.76	**30.18**
Spf-CM	/	/	21.93	20.83	29.91	38.73



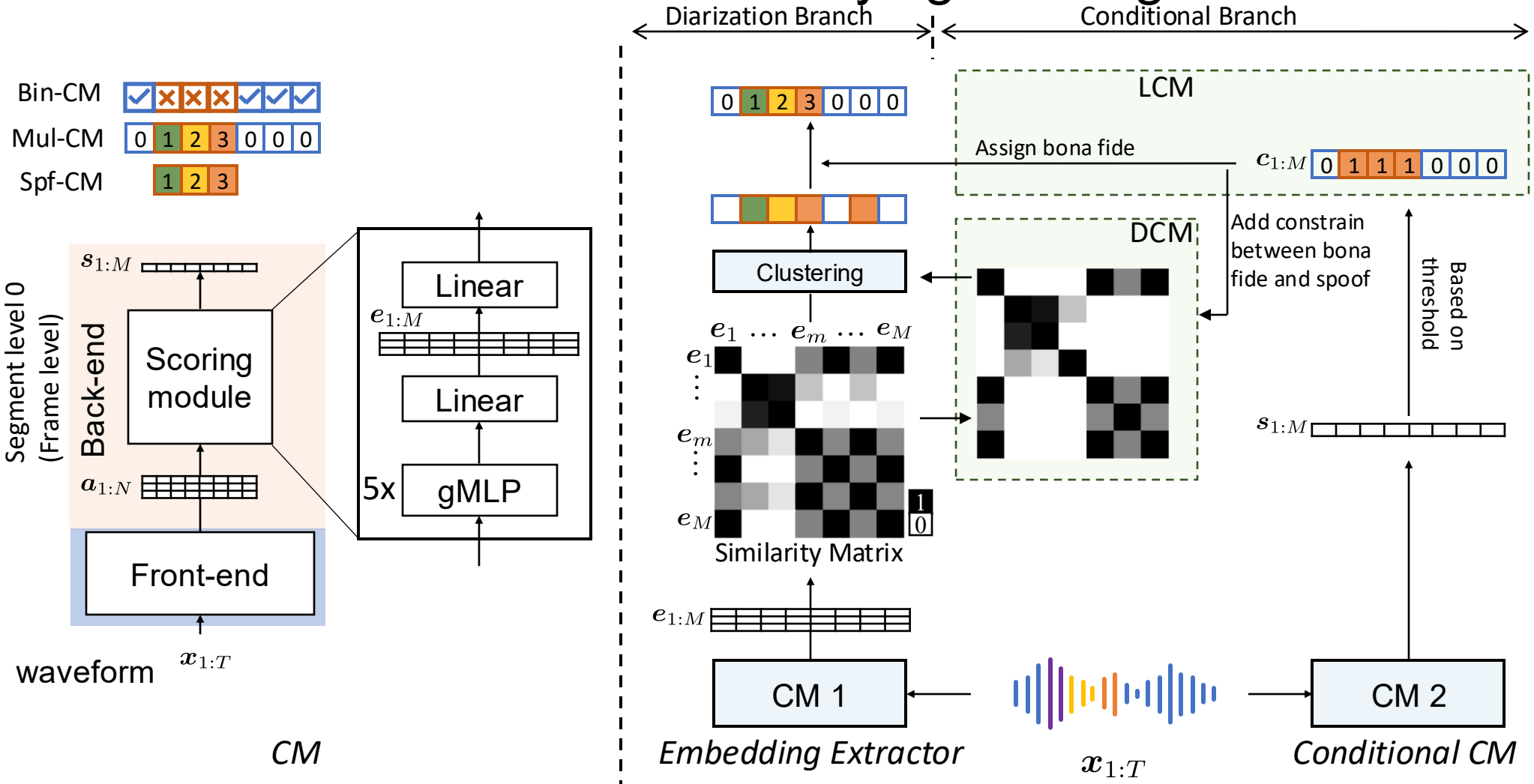
The binary class trained model (Bin-CM) shows superior performance in accurately locating bona fide on the evaluation set
Including bona fide, as in Mul-CM, generally aids in more effectively differentiating between various spoofing methods.

CMs trained with different bona fide and spoof labeling schemes capture distinct information.

→ How do we utilize CMs trained under varying labeling schemes?

8 Spoof Diarization

How do we utilize CMs trained under varying labeling schemes?



(e) CM-Conditional Clustering (3C) model for spoof diarization

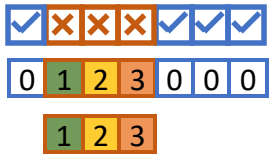
LCM: Label-based CM-constrain

DCM: Distance-based CM-constrain

8 Spoof Diarization

Table 8.1: Results on the different combinations of CMs (clustering was done with the oracle cluster number).

Model			Development set		Evaluation set	
CM-1	CM-2	Condition	JER_{bona}	JER_{spoof}	JER_{bona}	JER_{spoof}
Bin-CM	/	/	6.49	34.63	**17.01**	39.89
Mul-CM	/	/	**4.01**	**4.89**	19.76	**30.18**
Spf-CM	/	/	21.93	20.83	29.91	38.73



8 Spoof Diarization

Table 8.1: Results on the different combinations of CMs (clustering was done with the oracle cluster number).

Model			Development set		Evaluation set	
CM-1	CM-2	Condition	JER_{bona}	JER_{spoof}	JER_{bona}	JER_{spoof}
Bin-CM	/	/	6.49	34.63	**17.01**	39.89
Mul-CM	/	/	**4.01**	**4.89**	19.76	**30.18**
Spf-CM	/	/	21.93	20.83	29.91	38.73
Mul-CM	*Bin-CM*	LCM	4.83	37.65	**14.76**	41.28
	Bin-CM	DCM	6.51	20.93	16.93	31.48
	Mul-CM	LCM	**4.01**	**4.89**	16.16	**27.73**
	Mul-CM	DCM	4.02	**4.89**	17.38	28.68

The conditional CM contributes to enhanced generalizability and mitigate overfitting.

The selection of a conditional CM-2 should take into account the overall performance.

Outline

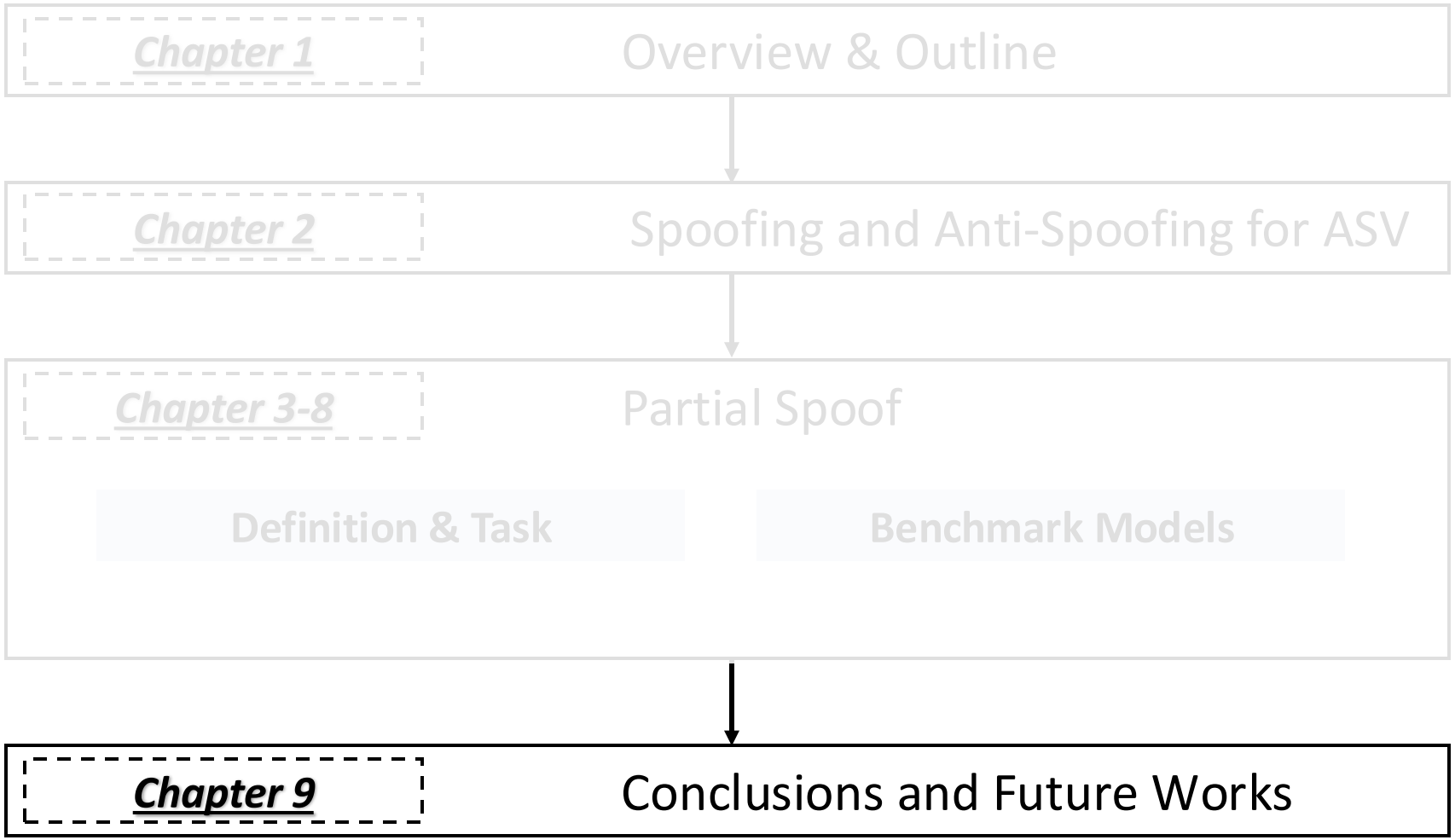


Figure 1: Thesis Outline

9 Summary

1. Spoof detection on the **PS scenario is more challenging** than on the fully spoof scenario.

1. Spoof localization on the PS scenario is a more complex task than detection but is feasible.
2. Single-task trained spoof detection and localization models can predict each other's tasks **without** additional

1. A **single resolution may not be enough** to extract meaningful information in the PS scenario. Trying **more fine-grained resolutions** could be worthwhile.
2. Although the multi-task model can do detection and localization simultaneously, localization did not show obvious

1
2
3

1. Multi-resolution CM can do detection and localization at different resolutions and can improve detection performance significantly with the **SSL-based front-end**,

1. Despite the fact that **spoof diarization is an emerging and complex task**, we managed to achieve promising results under simplified conditions. Solving remaining technical issues awaits future solution.
2. CMs trained with different bona fide and spoof labeling schemes capture distinct information.
3. Properly **integrating different CMs** can improve performance.

Bona fide: ☑
Spoof: ☒

❖ Spoof Diarization
What + When

Bona A1 A2 A3 Bona fide

8

1 2 3 Cluster

Unbounded Unknown

Figure: Thesis Outline & Contribution

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

9 Summary - Conclusion

3 *Partial Spoof Scenario*

- The **PS scenario is a more realistic** spoofing scenario and can pose a threat.
- Three tasks (**Detection, Localization, and Diarization**) are essential for exploring the PS scenario.
- Manipulation on the PS scenario should consider **variable duration**.

4 *Spoof Detection*

- Spoof detection on the PS scenario is **more challenging** than on the fully spoof scenario.
- Spoof detection CMs trained on **fully spoofed data lack generalization ability** when tested with partially spoofed data, while training on **partially spoofed data led to stable** performance when evaluating both scenarios.

5 *Spoof Localization*

- Spoof localization on the PS scenario is **a more complex** task than detection **but is feasible**.
- Single-task trained spoof detection and localization models can predict each other's tasks **without additional labeling or training.**
- The **precise annotations** from the PS scenario are useful for **localization**.

9 Summary - Conclusion

6 ***Fixed-Resolution Model for Spoof Detection and Localization***

- A single resolution may not be enough to extract meaningful information in the PS scenario. Trying **more fine-grained resolutions** could be worthwhile.
- Although the multi-task model can do detection and localization simultaneously, localization did not show obvious improvement.

7 ***Flexible-Resolution Model for Spoof Detection and Localization***

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

8 ***Spoof Diarization***

- Despite the fact that spoof diarization is **an emerging and complex** task, we managed to achieve promising results under simplified conditions. Solving remaining technical issues awaits future resolution.
- CMs trained with **different** bona fide and spoof **labeling schemes** capture distinct information.
- Properly **integrating different CMs** can improve performance.

9 Future Works

- **Intervention analysis on the partial spoof scenario**
  - Impact of spoof-to-utterance ratio
  - Effect of manipulated position and duration

- **More realistic PS scenario**
  - PS scenario in the wild (more complicated and mismatched environment)
  - PS scenario with flexible semantic manipulation
  - Multi-modal partial spoof

- **Advanced spoof diarization scenario and technologies**
  - More advanced diarization models
  - Speaker spoof diarization
  - Watermarking based spoof diarization
  - Multi-level spoof diarization

A Outcomes

- **First-authored papers:**

Journal Paper:

[1] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, Junichi Yamagishi, (2023) 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.

International conference paper:

[2] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, Junichi Yamagishi. (2023) Range-Based Equal Error Rate for Spoof Localization. Proc. INTERSPEECH 2023, 3212-3216, doi: 10.21437/Interspeech.2023-1214 (CORE rank A)

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

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

A Outcomes

- **Co-authored papers (related to anti-spoofing):**

[5] Chang Zeng, **Lin Zhang**, Meng Liu, Junichi Yamagishi. (2022) Spoofing-Aware Attention based ASV Back-end with Multiple Enrollment Utterances and a Sampling Strategy for the SASV Challenge 2022. Proc. Interspeech 2022, 2883-2887, doi: 10.21437/Interspeech.2022-10495 (CORE rank A)

[6] Linjuan Zhang, Kong Aik Lee, **Lin Zhang**, Longbiao Wang, Baoning Niu, Cpaug: Refining Copy-Paste Augmentation for Speech Anti-Spoofing (submitted to ICASSP 2024)

[7] Xiaohui Liu, Meng Liu, **Lin Zhang**, et al.,Deep Spectro-temporal Artifacts for Detecting Synthesized Speech. in Proc. DDAM 2022, 69–75. doi:10.1145/3552466.3556527

A Outcomes

- **Others (related to ASV/Diarization):**

[8] Ruiteng Zhang, Jianguo Wei, Xugang Lu, Wenhuan Lu, Di Jin, **Lin Zhang**, Junhai Xu, and Jianwu Dang. "TMS: Temporal multi-scale in time-delay neural network for speaker verification." *Applied Intelligence* 53, no. 22 (2023): 26497-26517.

[9] Ruiteng Zhang, Jianguo Wei, Xugang Lu, Wenhuan Lu, Di Jin, **Lin Zhang**, Yantao Ji, and Junhai Xu. "Self-supervised learning based domain regularization for mask-wearing speaker verification." *Speech Communication* (2023): 102953.

[10] Ruiteng Zhang, Jianguo Wei, Xugang Lu, Wenhuan Lu, Di Jin, **Lin Zhang**, and Junhai Xu. "Optimal Transport with a Diversified Memory Bank for Cross-Domain Speaker Verification." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023.

[11] Kai Li, Sheng, Li, Xugang Lu, Masato Akagi, Meng Liu, **Lin Zhang**, Chang Zeng, Longbiao Wang, Jianwu Dang, and Masashi Unoki. “Data Augmentation Using McAdams Coefficient-Based Speaker Anonymization for Fake Audio Detection,” Proc. Interspeech 2022.

[12] Ruiteng Zhang, Jianguo Wei, Wenhuan Lu, **Lin Zhang**, Yantao Ji, Junhai Xu, and Xugang Lu. "Cs-rep: Making speaker verification networks embracing re-parameterization." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7082-7086. IEEE, 2022.

References

[0] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 30107: Information technology — Biometric presentation attack detection (2016)

[1] 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

[2] 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] 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.

[4] Lin Zhang, Xin Wang, Erica Cooper, Nicolas Evans and Junichi Yamagishi, “Range-Based Equal Error Rate for Spoof Localization.” (submitted to)

[5] 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

[6] 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

[7] Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., ... & Li, H. (2022, May). Add 2022: the first audio deep synthesis detection challenge. In proc. ICASSP2022, pp. 9216-9220. IEEE.

[8] D. Cai, Z. Cai, and M. Li, “Identifying source speakers for voiceconversion based spoofing attacks on speaker verification systems,” arXiv preprint arXiv:2206.09103, 2022.

[9] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.M. Wang, and H. Meng, “Partially fake audio detection by self-attention-based fake span discovery,” in Proc. ICASSP, 2022, pp.9236–9240.

[10] 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

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

Q&A

- Please find more details (paper, database, github repo) from this QR code.

- PartialSpoof provides fine-grained timestamp labels! Please contact me (zhanglin@nii.ac.jp, partialspoof@gmail.com) if you need other information!