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

ISCA Speaker Odyssey 2022

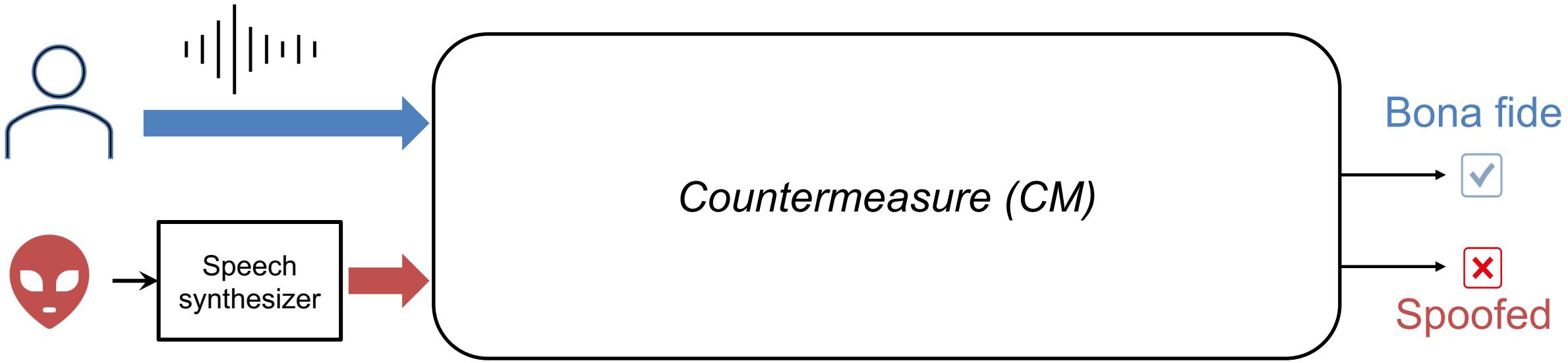
Investigating self-supervised front ends for speech spoofing countermeasures

Xin Wang, Junichi Yamagishi
National Institute of Informatics

wangxin@nii.ac.jp

Introduction

Introduction



- **Anti-spoofing as binary classification**
  - protect automatic speaker verification (ASV)
  - detect fake-voice-based phone scam

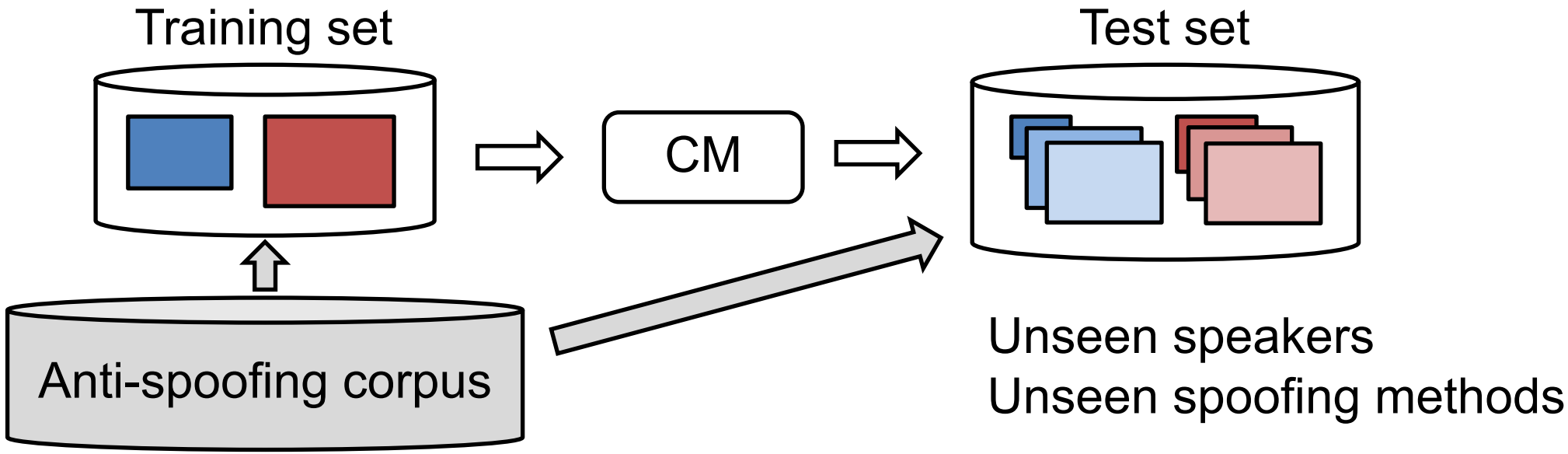

EDITORS' PICK | Oct 14, 2021, 07:01am EDT | 81,432 views

Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=88e208875591

Introduction

❑ CM development flow

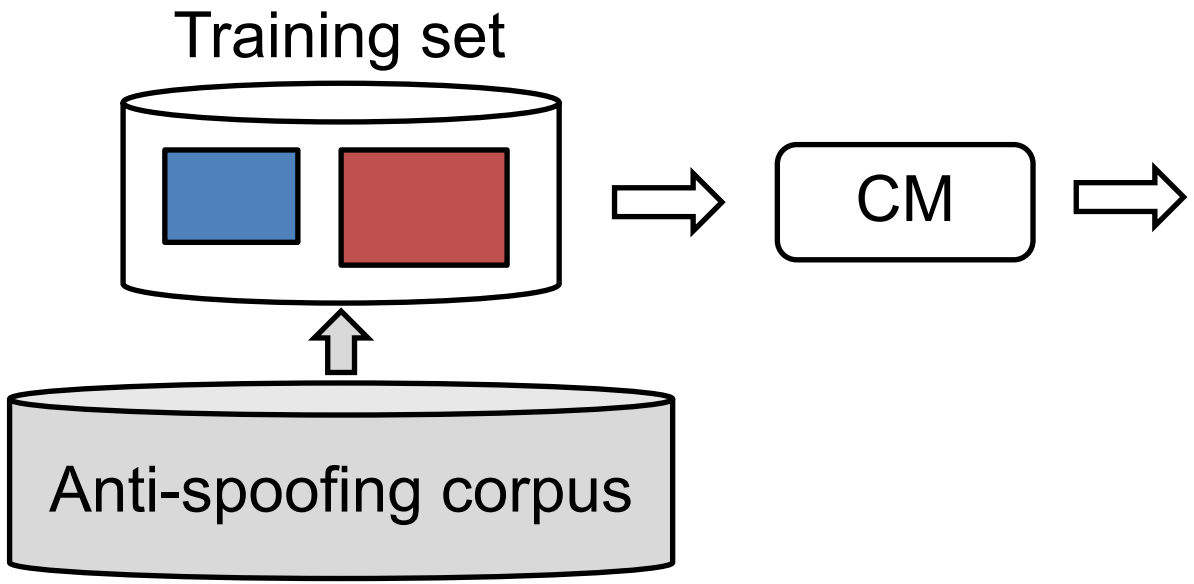


- Corpus w/ data protocol:
  - ASVspoof (Wu 2015, Todisco 2019), BTAS (Korshunov 2016), FMFCC-A (Zhang 2021), ADD (Yi 2022)

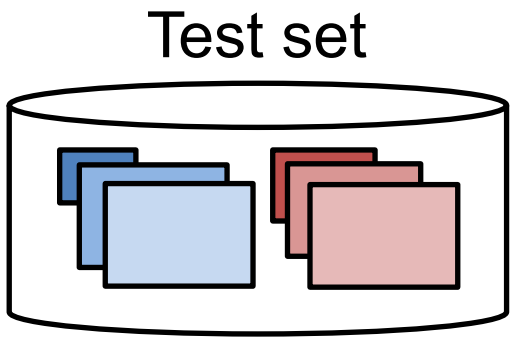
Korshunov, P. *et al.* Overview of BTAS 2016 speaker anti-spoofing competition. in *Proc. BTAS* 1–6 (2016). doi:10.1109/BTAS.2016.7791200
Zhang, Z., Gu, Y., Yi, X. & Zhao, X. FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection. *arXiv Prepr. arXiv2110.09441* (2021)
Wu, Z. *et al.* ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. in *Proc. Interspeech* 2037–2041 (2015). doi:10.21437/Interspeech.2015-462
Todisco, M. *et al.* ASVspoof 2019: future horizons in spoofed and fake audio detection. in *Proc. Interspeech* 1008–1012 (2019). doi:10.21437/Interspeech.2019-2249
Yi, J. *et al.* Add 2022: the first audio deep synthesis detection challenge. in *Proc. ICASSP* 9216–9220 (2022).

Introduction

- **CM development flow**



Space of all possible bona fide and spoofed data



En, Fr, Zh, Jp, …

Speaker A, B, C, …

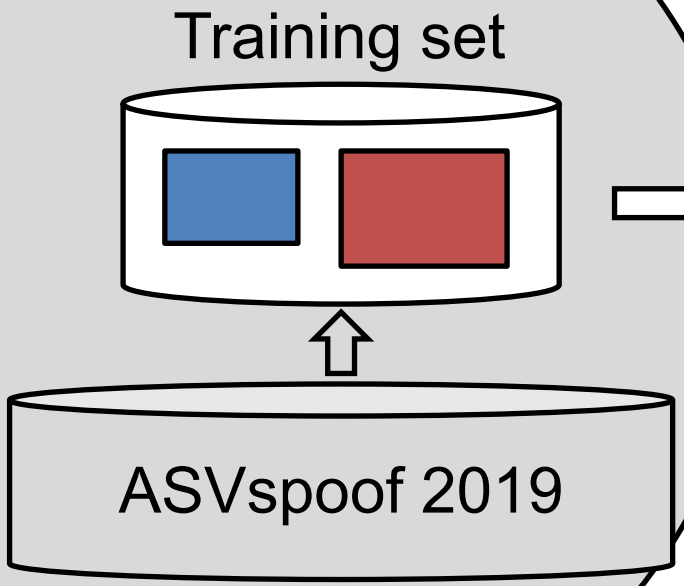
wav, mp3, m4a, ogg …

New spoofing methods

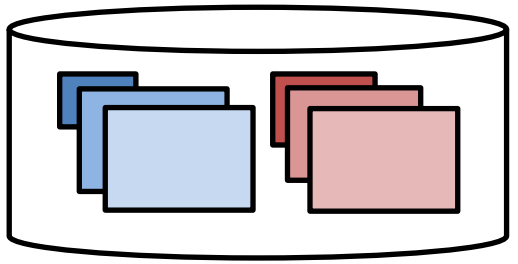
Introduction

- **Questions** (Paul 2017, Das 2020, Müller 2022)

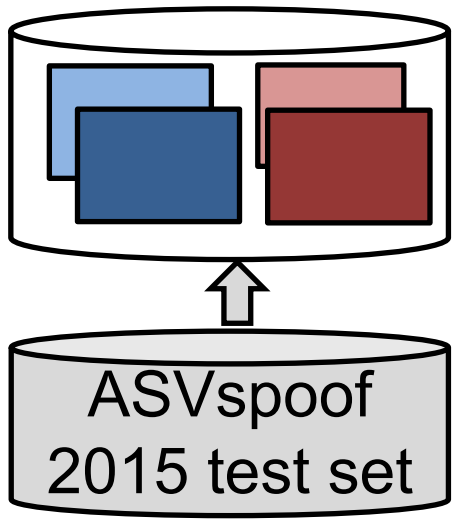
More training data?



EER <8%



EER >20%*



- Does CM generalize well: no
- **How to improve?**

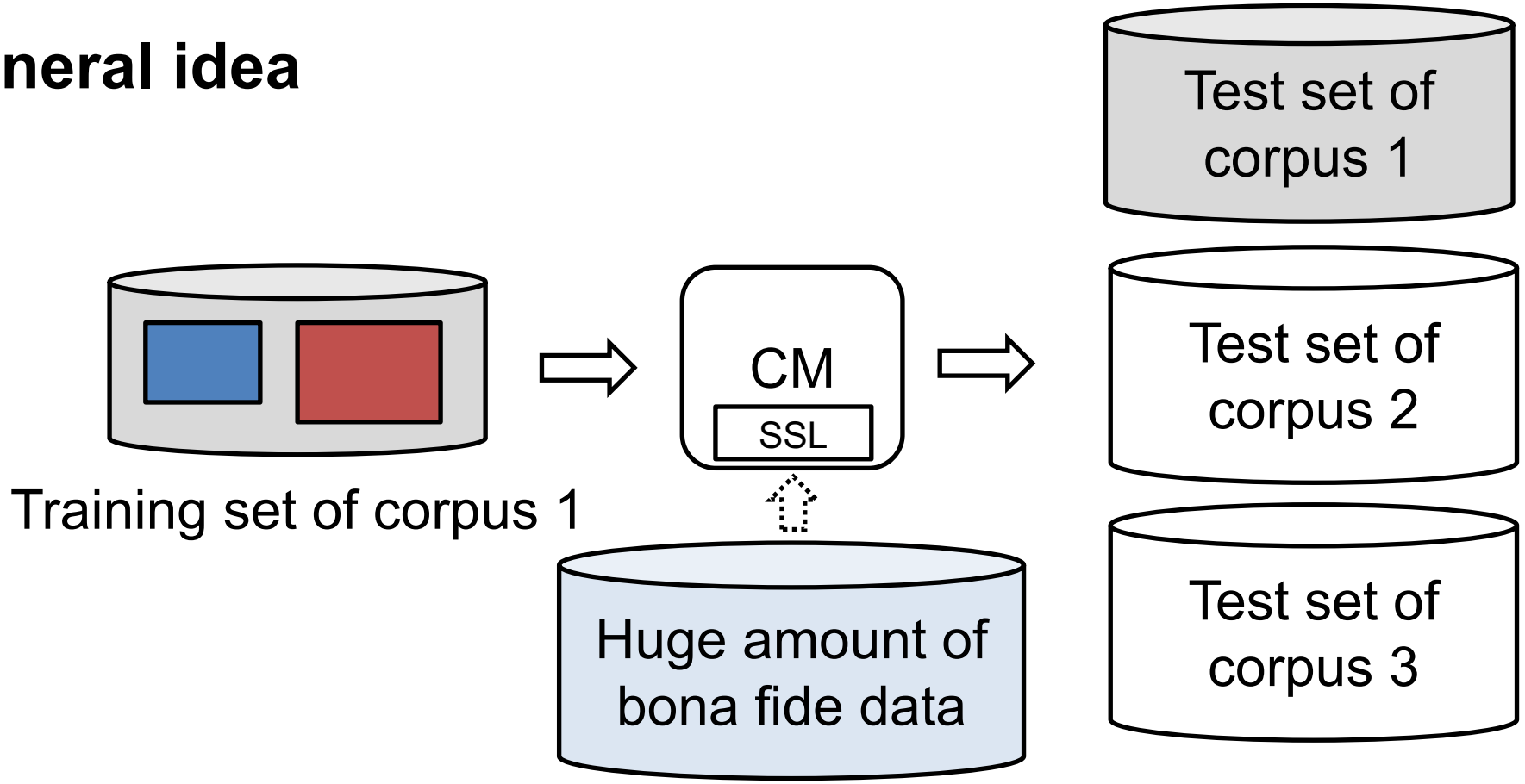
* From CQSPIC system in Table 1&2 of (Das 2020)

Paul, D., Sahidullah, M. & Saha, G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. in *Proc. ICASSP* 2047–2051 (2017). doi:10.1109/ICASSP.2017.7952516
Das, R. K., Yang, J. & Li, H. Assessing the scope of generalized countermeasures for anti-spoofing. in *Proc. ICASSP* 6589–6593 (2020). doi:10.1109/ICASSP40776.2020.9053086
Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does Audio Deepfake Detection Generalize? *ArXiv Preprint ArXiv:2203.16263*. 2022.

Methods

Method

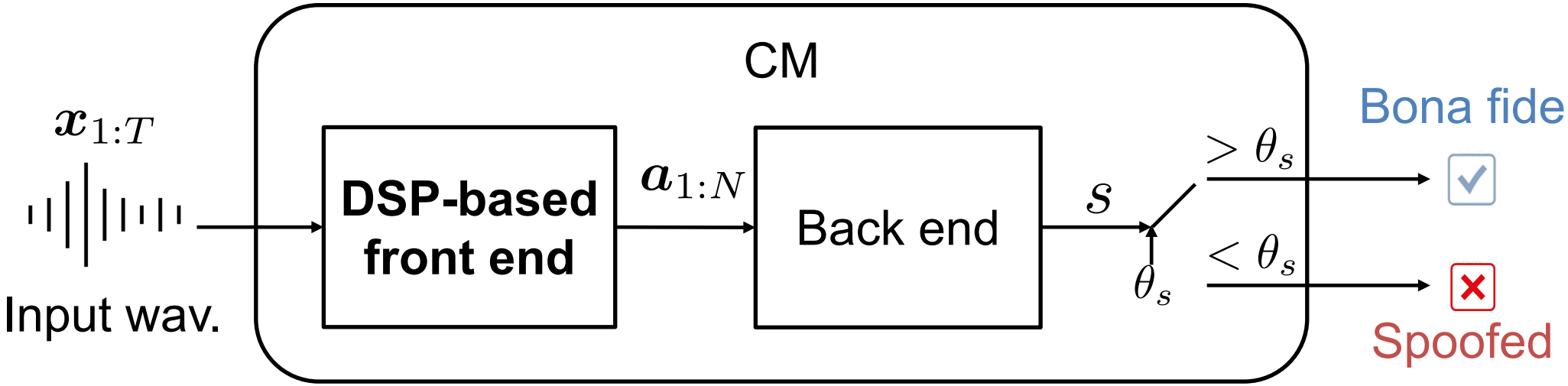
❑ General idea



- Feature extraction using a pre-trained self-supervised learing (SSL) speech model

Method

❑ CM structure - baseline



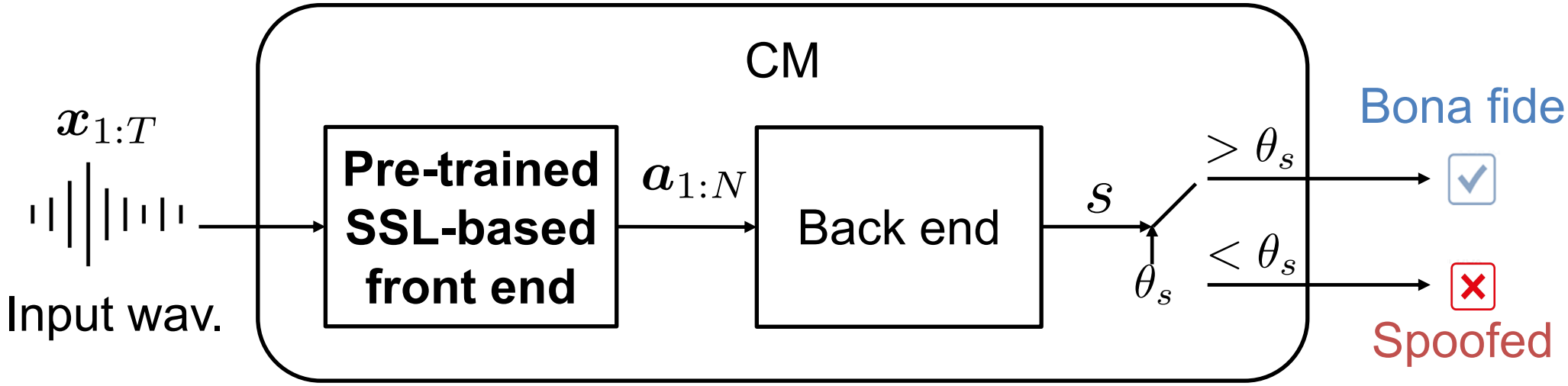
- Front end: linear frequency cepstrum coefficients (LFCCs)(Sahidullah 2015)

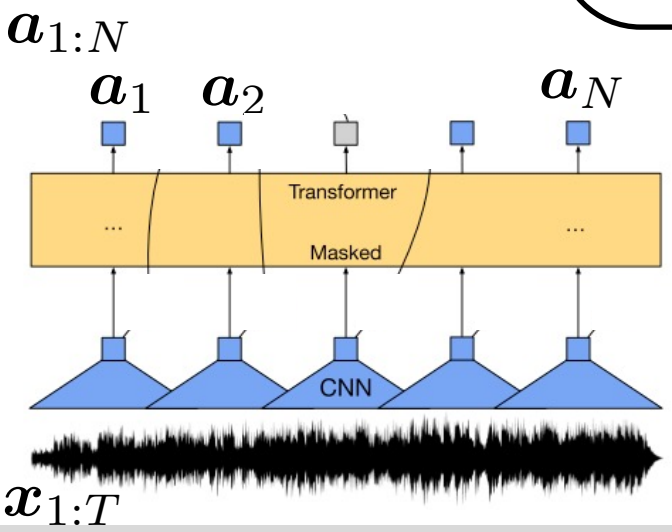
$$\underbrace{\boldsymbol{x}_{1:T}}_{(x_1, x_2, \cdots, x_T) \in \mathbb{R}^{1\times T}} \xrightarrow{\mathrm{DCT}\left(\log(\boldsymbol{W}_{\mathrm{fb}} \cdot |\mathrm{STFT}(\boldsymbol{x}_{1:T})|^2)\right)} \underbrace{\boldsymbol{a}_{1:N}}_{(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N) \in \mathbb{R}^{D\times N}}$$

DSP: digital signal processing
T: number of waveform sampling points, *N*: number of frames, *D*: number of dimensions in each frame
Sahidullah, M., Kinnunen, T. & Hanilçi, C. A comparison of features for synthetic speech detection. in *Proc. Interspeech* 2087–2091 (2015).

Method

❑ CM structure - investigated





For example, wav2vec 2.0 (Baevski 2020)

$$x_{1:T} \xrightarrow{\text{Transformer}\left(\text{CNN}(x_{1:T})\right)} a_{1:N}$$

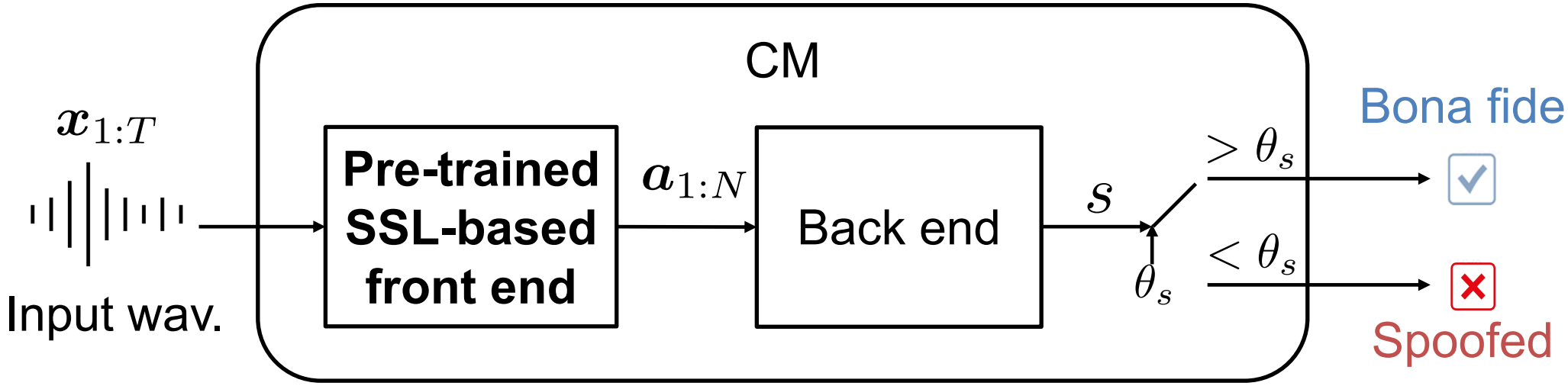
$$\underbrace{(x_1, x_2, \cdots, x_T)}_{x_{1:T}} \in \mathbb{R}^{1\times T} \qquad \underbrace{(a_1, a_2, \cdots, a_N)}_{a_{1:N}} \in \mathbb{R}^{D\times N}{}^{*}$$

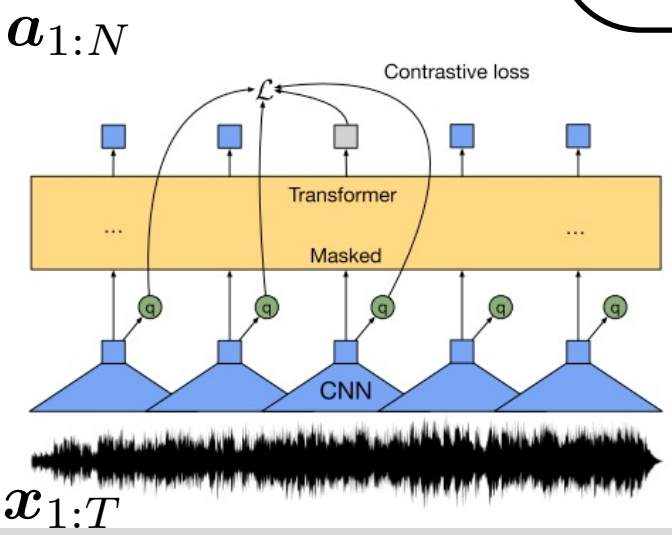
D and *N* are decided by the CNN configuration

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. in *Proc. NIPS* vol. 33 12449–12460 (2020).

Method

❑ CM structure - investigated



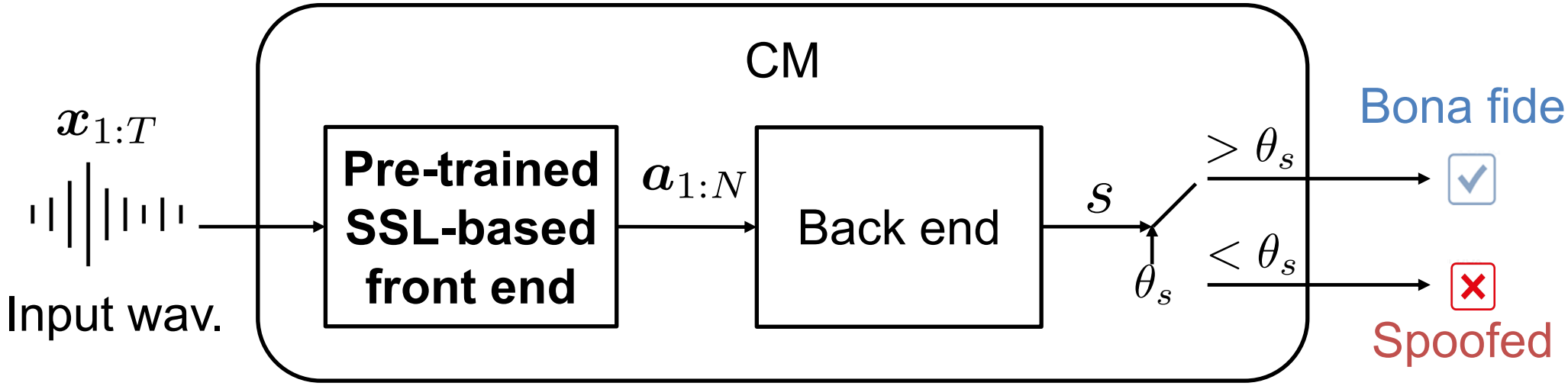


For example, wav2vec 2.0 (Baevski 2020)

- *pre-trained* using a self-contrastive loss and *a huge amount of bona fide waveforms*
  - Languages, speakers, background noises, reverberation, codec …
- many pre-trained SSL models available

Method

❑ CM structure - investigated

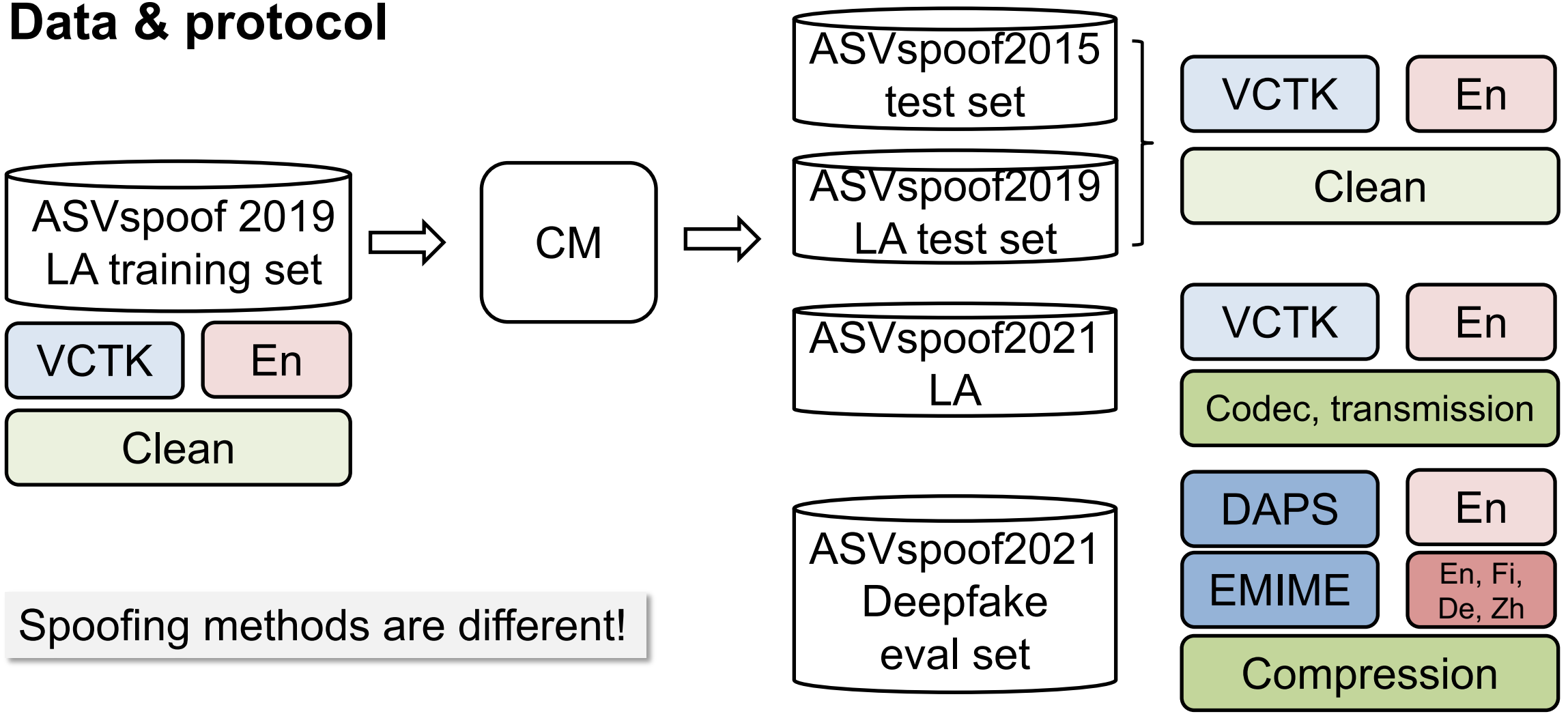


- Should is be fine-tuned?
- How to design back end?
  - DNN, or a single output layer

Experiments

Experiment

❑ Data & protocol



Spoofing methods are different!

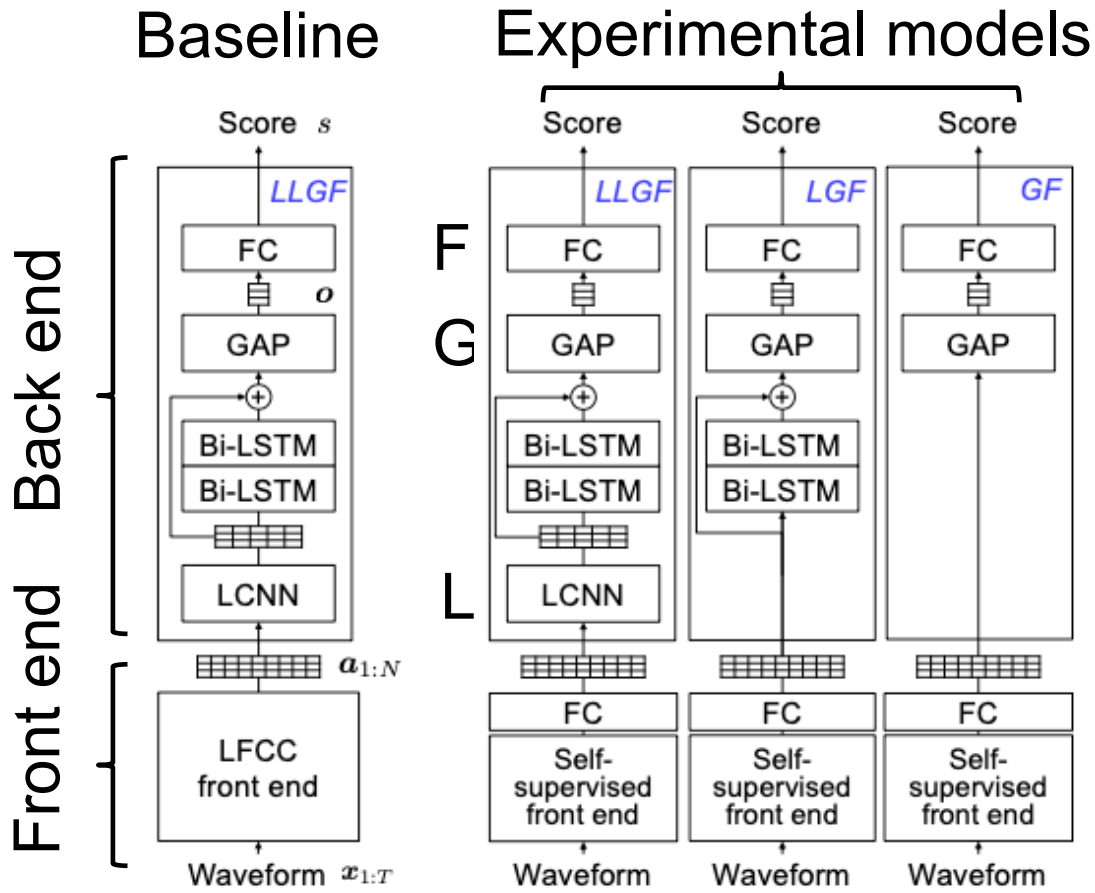
LA: logical access, DF: deepfake

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, and others. VCTK Corpus: English Multi-Speaker Corpus for Cstr Voice Cloning Toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR). 2016.
Gautham J. Mysore, "Can we automatically transform speech recorded on common consumer devices in realworld environments into professional production quality speech? - A dataset, insights, and challenges," IEEE Signal Process. Lett., vol. 22, no. 8, pp. 1006–1010, 2015.
M. Wester, "The EMIME bilingual database," The University of Edinburgh, Tech. Rep., 2010.

Experiment

❑ CM configuration – back end

- Baseline: LCNN-variant (*LLGF*) (Wang 2021)
  - LCNN + LSTM + GAP + FC
- Experimental models
  - Which type of back end?
    - *LLGF*: LCNN + LSTM + GAP + FC
    - *LGF*: LSTM + GAP + FC
    - *GF*: GAP + FC



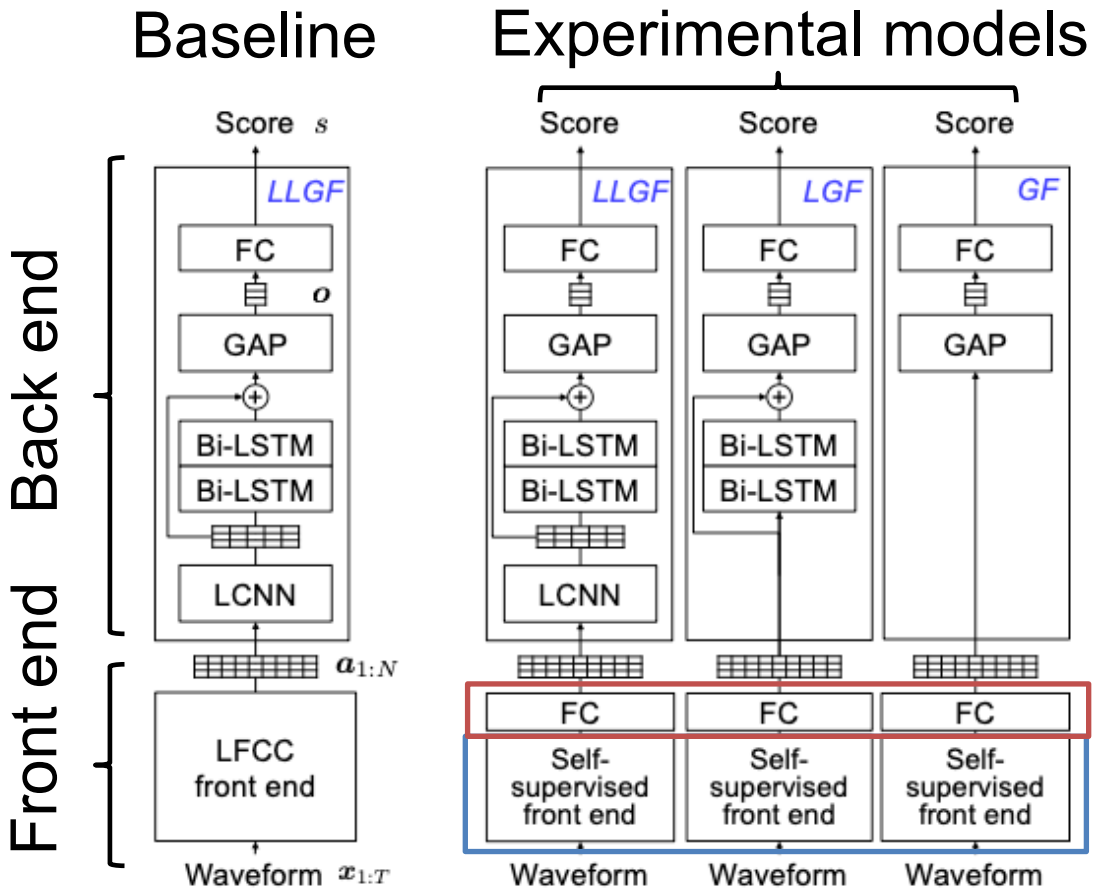
Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A Light CNN for Deep Face Representation with Noisy Labels. *IEEE Transactions on Information Forensics and Security* 13 (11). IEEE: 2884–2896. doi:10.1109/TIFS.2018.2833032. 2018.

Wang, X. & Yamagishi, J. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. in *Proc. Interspeech* 4259–4263 (2021). doi:10.21437/Interspeech.2021-702

Experiment

❑ CM configuration – front end

- Baseline: LFCC
  - Configuration from ASVspoof 2019*
- Experimental models
  - SSL front end: Wav2vec 2.0 XLSR
  - Should we fine-tune the front end on the ASVspoof 2019 training set**?**
    - If yes:
      1. Initialize front end using pre-trained SSL
      2. Train the whole CM w/ a small learning rate
    - If no:
      1. Initialize front end using pre-trained SSL
      2. Fix the front end, train the back end

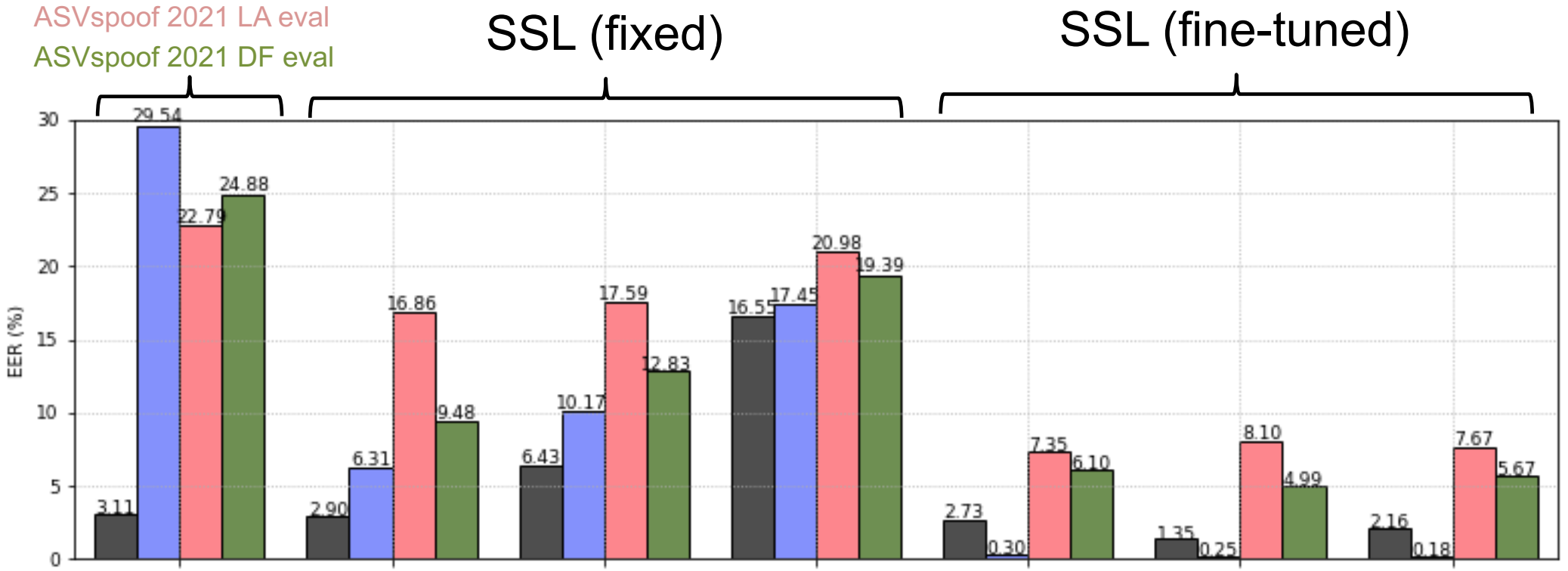


* https://www.asvspoof.org/asvspoof2019/ASVspoof_2019_baseline_CM_v1.zip

Results in EER (%)

Baseline EERs on

ASVspoof 2019 LA
ASVspoof 2015 LA
ASVspoof 2021 LA eval
ASVspoof 2021 DF eval



Front end	LFCC	W2V-XLSR, fixed			W2V-XLSR, fine-tuned		
Back end	LLGF	LLGF	LGF	GF	LLGF	LGF	GF

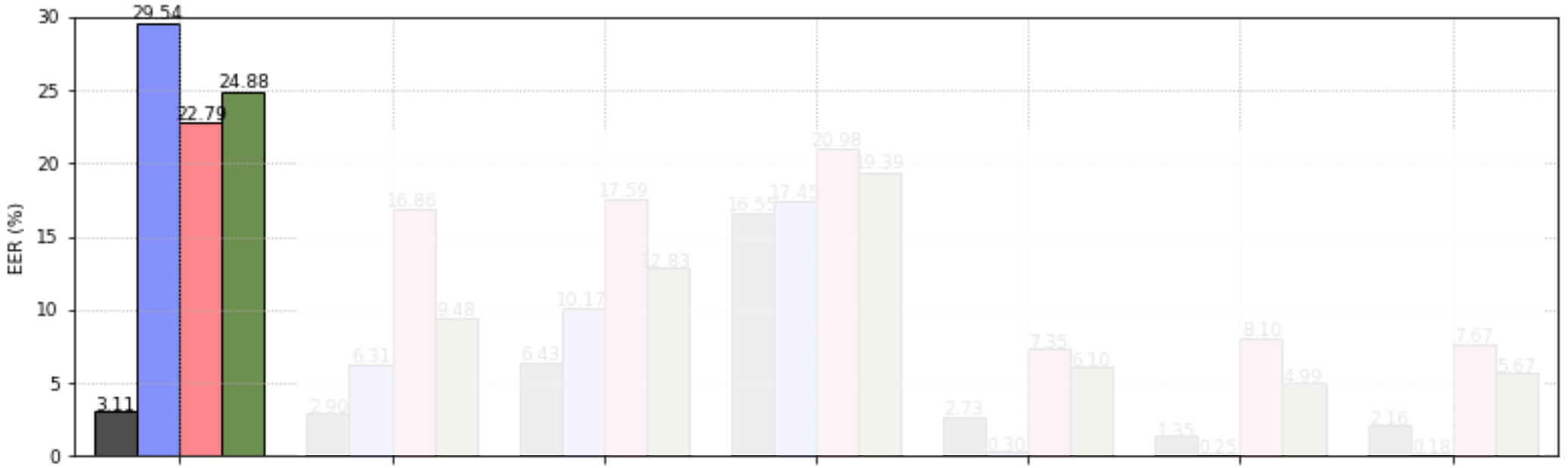
See full results in the paper
See min-tDCF and other results in Appendix: https://arxiv.org/abs/2111.07725

Results in EER (%)

Baseline:

- low EER on ASVspoof 2019 LA test set
- high EER on ASVspoof **2015**, 2021 !

2019 LA
2015 LA
2021 LA eval
2021 DF eval



Front end	LFCC	W2V-XLSR, fixed			W2V-XLSR, fine-tuned		
Back end	LLGF	LLGF	LGF	GF	LLGF	LGF	GF

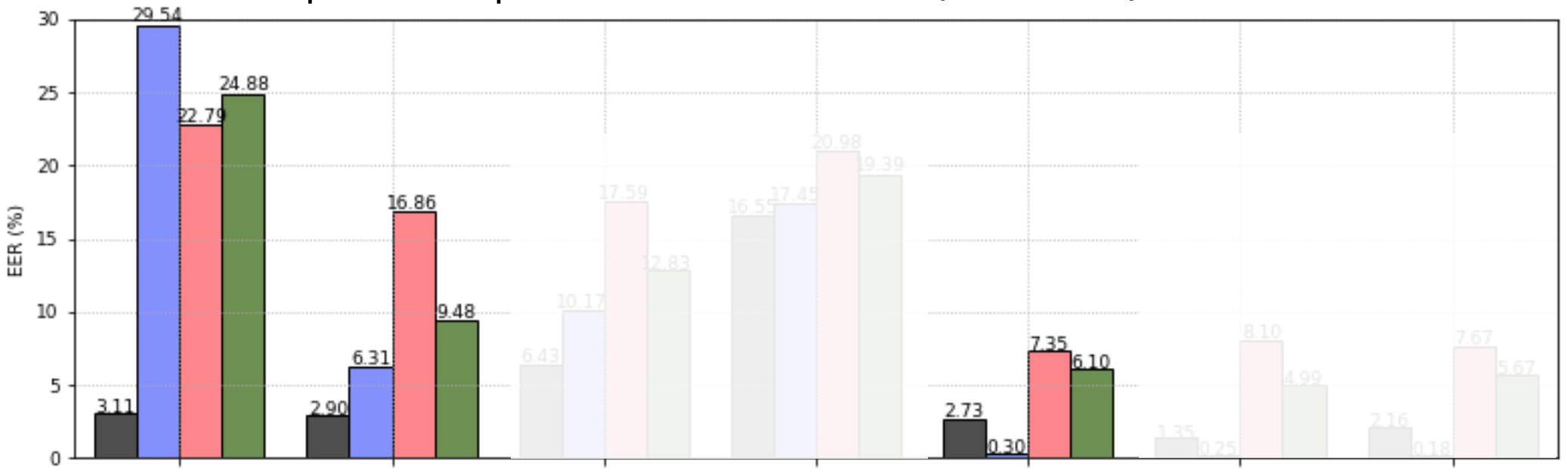
See full results in the paper
See min-tDCF and other results in Appendix: https://arxiv.org/abs/2111.07725

Results in EER (%)

Using a pre-trained SSL front end is good

Fine-tune pre-trained SSL front end **is better**

2019 LA
2015 LA
2021 LA eval
2021 DF eval



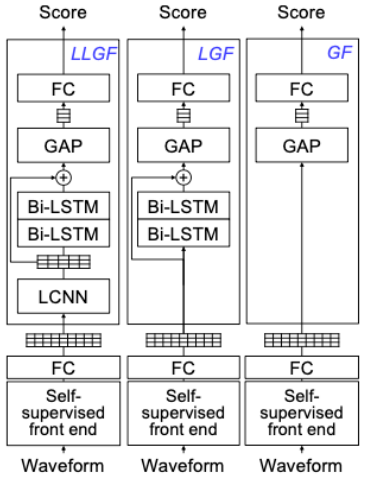
Front end	LFCC	W2V-XLSR, fixed			W2V-XLSR, fine-tuned		
Back end	LLGF	LLGF	LGF	GF	LLGF	LGF	GF

See full results in the paper
See min-tDCF and other results in Appendix: https://arxiv.org/abs/2111.07725

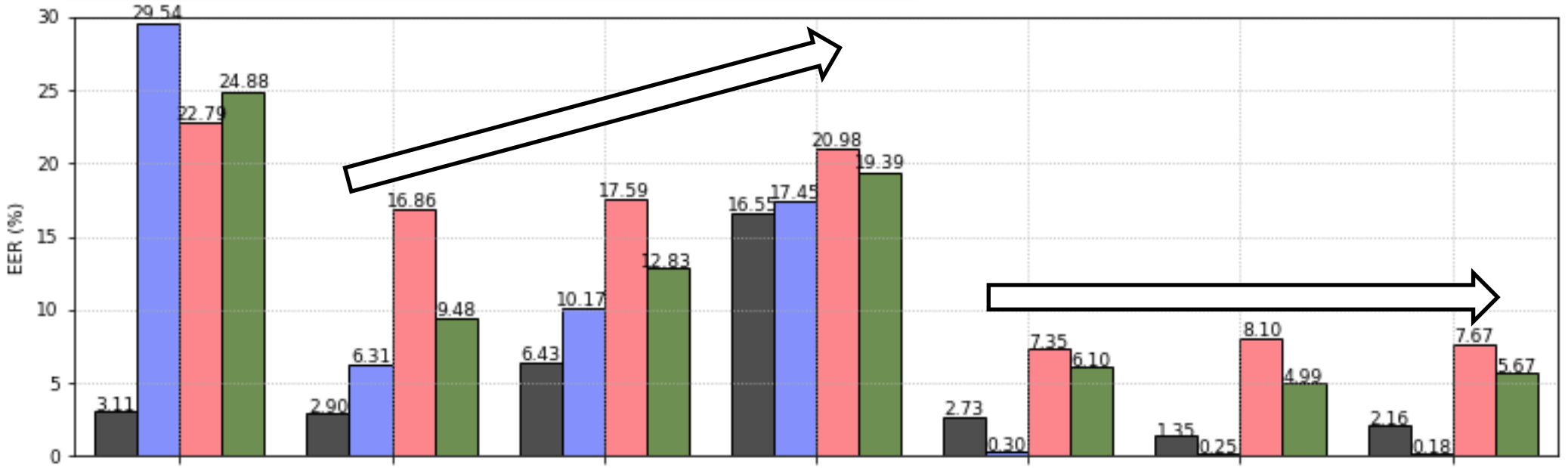
Results in EER (%)

Should we fine tune the SSL front end?
- **Yes**

Which type of back end is better?
- If we fine-tune the front end, difference is small



2019 LA
2015 LA
2021 LA eval
2021 DF eval



Front end	LFCC	W2V-XLSR, fixed			W2V-XLSR, fine-tuned		
Back end	LLGF	LLGF	LGF	GF	LLGF	LGF	GF

See full results in the paper
See min-tDCF and other results in Appendix: https://arxiv.org/abs/2111.07725

Summary

Summary

- **Do SSL front ends generalize better than baseline?**
  - Yes
  - Recommendations from this study
    - ✓Large SSL model pre-trained on diverse data
    - ✓Fine-tuning the SSL front end

- **However, the best model is not sufficiently generaliable**
  - https://arxiv.org/pdf/2203.14553.pdf
  - **EER > 30%** on FMFCC-A Mandarin data (Zhang 2021)
  - **EER > 15%** on WaveFake English/Japanese data(Fank 2021)

Zhang, Z., Gu, Y., Yi, X. & Zhao, X. FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection. *arXiv Prepr. arXiv2110.09441* (2021)
Frank, J. & Schönherr, L. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. in *Proc. NeurIPS Datasets and Benchmarks 2021* accepted (2021).

Summary

❑ Other findinds in the paper & appendix

- **Why SSL outperformed DSP front end?**
  - It ignores spurious “evidence” in high-frequency band of training data
- Which pre-trained SSL model is better?
- Detailed EERs on each codec, spoofing methods …

❑ Findings from related works

- Data augmentation to CM is useful (Martin-Donas 2022, Tak 2022)
- Back-end can be further improved (Tak 2022)

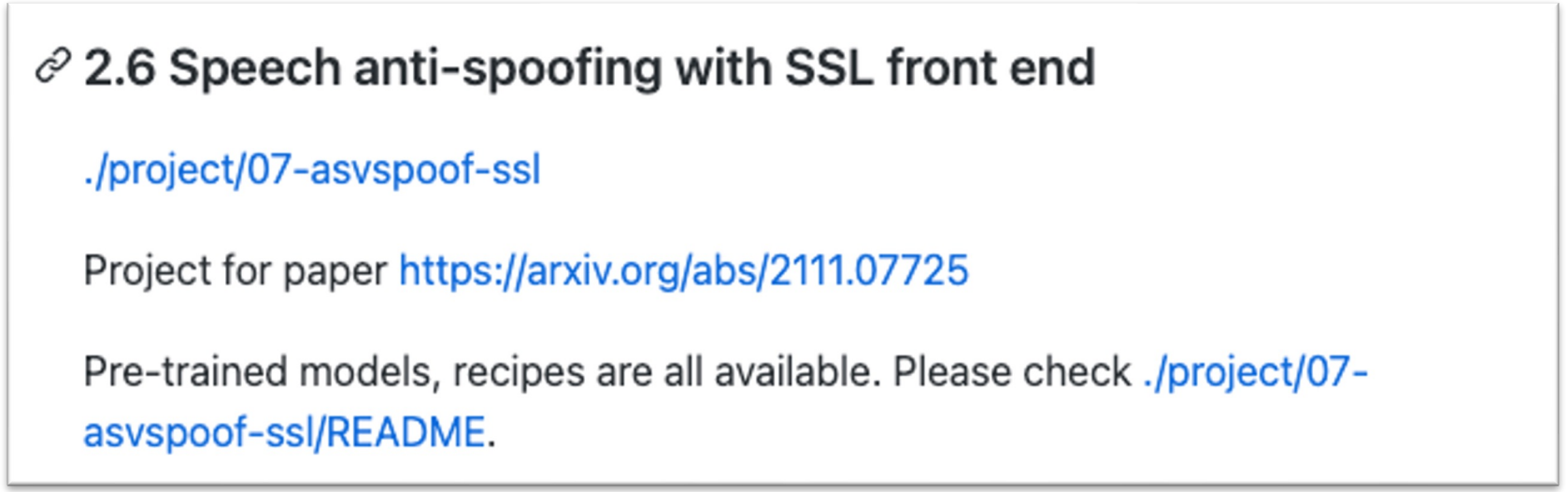
NII

It is comparable to XLS-53 without data augmentation in (Juan M Martin-Donas 2022)

Juan M Martin-Donas, and Aitor Álvarez. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In *Proc. ICASSP*, 9241–9245. 2022.
Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *Proc. Odyssey 2022 The Speaker and Language Recognition Workshop*. Vol. accepted. 2022.

Thank you

https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts#26-speech-anti-spoofing-with-ssl-front-end

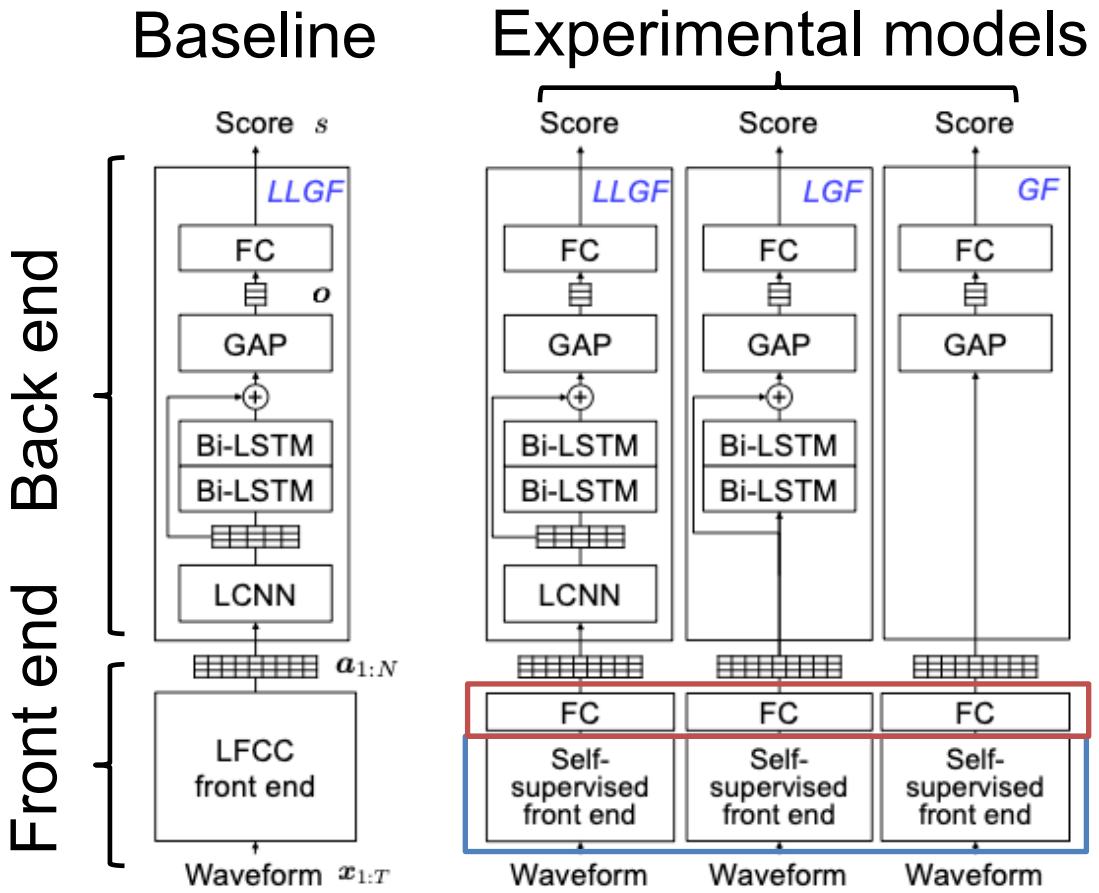


Appendix

Experiment

❑ CM configuration – front end

- Baseline: LFCC
  - Configuration from ASVspoof 2019*
  - 60 dim. per frame (s, Δ, Δ^2)
- Experimental models
  - Which SSL model?
  - Should we fine-tune the front end on the ASVspoof 2019 training set**?**
    - If yes:
      1. Initialize front end using pre-trained SSL
      2. Train the whole CM w/ a small learning rate
    - If no:
      1. Initialize front end using pre-trained SSL
      2. Fix the front end, train the back end only



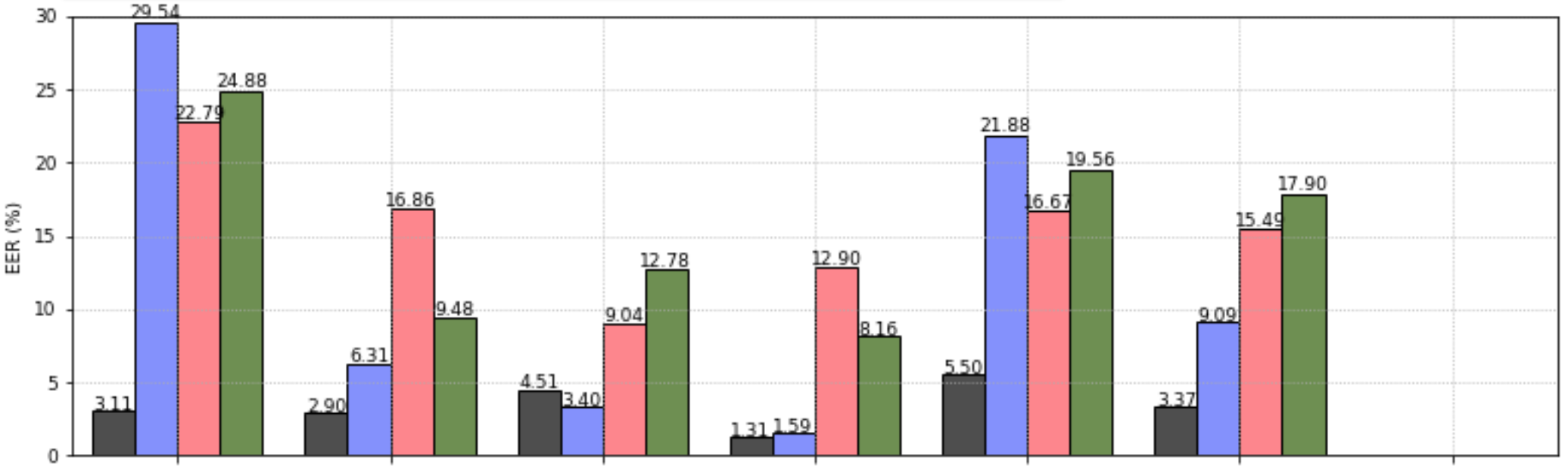
* https://www.asvspoof.org/asvspoof2019/ASVspoof_2019_baseline_CM_v1.zip

Results in EER (%)

Which pre-trained SSL front end?

ID	Model type	Data for pre-training	#.para
W2V-XLSR	Wav2vec (xlsr)	LibriSpeech [24], CommonVoice [25], BABEL [26]	317 m
W2V-Large2	Wav2vec (w2v_large)	CommonVoice, Switchboard [27], Libri-Light [28], Fisher [29]	317 m
W2V-Large1	Wav2vec (w2v_vox_new)	Libri-Light	317 m
W2V-Small	Wav2vec (w2v_small)	Librispeech	95 m
HuBERT-XL	HuBERT (extra_large)	Libri-Light	964 m

2019 LA
2015 LA
2021 LA eval
2021 DF eval



SSL model size	Large	Extra Large	Large	Large	Small
Multi-lingual?	Yes (>50)	En	Yes (>50)	En	En
SSL data (hour)	~55k	>60k	En >61k, else 4k	>60k	~1k

Front end	LFCC	W2V-XLSR, fixed	HuBERT-XL, fixed	W2V-Large2, fixed	W2V-Large1, fixed	W2V-Small, fixed
Back end	LLGF					

See full results in the paper
See min-tDCF and other results in Appendix: https://arxiv.org/abs/2111.07725

Results in EER (%)

☺ Low EER ☹ High EER

Not fine-tuned (pre-trained model was fixed) ↓ fine-tuned ↓

Front end →
Back end →

Front end	LFCC			W2V-XLSR, fixed									W2V-XLSR, fine-tuned								
Back end	LLGF			LLGF			LGF			GF			LLGF			LGF			GF		
	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
2019 LA	2.98	3.03	3.33	1.47	3.45	3.77	6.01	6.32	6.95	15.96	16.72	16.98	2.31	2.80	3.08	1.28	1.28	1.50	1.96	2.25	2.27
2015 LA	29.42	27.98	31.21	3.97	6.78	8.18	10.04	10.95	9.51	16.90	17.55	17.89	0.25	0.41	0.24	0.24	0.19	0.31	0.21	0.17	0.17
2021 LA prog.	15.82	15.81	24.40	9.85	17.29	20.17	16.76	13.77	15.63	20.06	20.88	21.25	7.58	6.38	6.56	10.63	9.19	6.27	7.65	7.16	7.82
2021 LA eval.	20.93	20.38	27.06	10.97	18.91	20.71	20.23	16.02	16.52	20.30	21.16	21.48	7.62	7.26	7.18	9.66	8.11	6.53	7.99	7.42	7.61
2021 DF prog.	28.38	23.60	31.12	2.67	5.09	7.02	6.92	7.91	8.39	19.30	20.26	20.63	4.40	4.33	4.14	3.38	3.75	3.55	3.97	4.23	4.94
2021 DF eval.	24.37	23.05	27.22	7.14	9.94	11.35	13.26	13.23	12.00	18.88	19.48	19.81	5.44	6.68	6.18	4.75	5.23	4.98	5.04	6.10	5.88

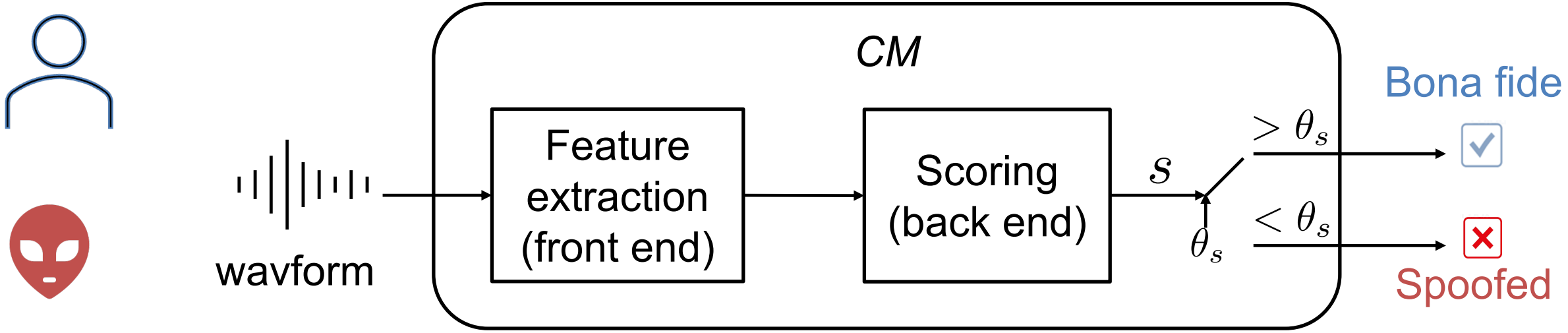
Test set →

(⇓ results are copied)

Front end	HuBERT-XL, fixed			W2V-XLSR, fixed			W2V-Large2, fixed			W2V-Large1, fixed			W2V-Small, fixed		
Back end	LLGF														
	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
2019 LA	3.55	4.04	5.93	1.47	3.45	3.77	0.86	0.99	2.08	4.47	5.67	6.36	2.61	3.48	4.01
2015 LA	3.27	3.25	3.69	3.97	6.78	8.18	1.39	1.39	1.99	19.66	22.33	23.65	10.40	7.58	9.28
2021 LA prog.	7.63	6.61	9.55	9.85	17.29	20.17	11.40	10.50	10.92	18.25	21.00	22.32	20.28	18.91	18.57
2021 LA eval.	9.55	7.03	10.54	10.97	18.91	20.71	13.19	12.57	12.94	13.86	16.77	19.38	16.11	14.79	15.56
2021 DF prog.	4.16	4.32	5.11	2.67	5.09	7.02	1.86	2.12	3.36	8.22	10.32	12.92	5.34	7.80	7.87
2021 DF eval.	13.07	12.87	12.39	7.14	9.94	11.35	7.44	7.77	9.26	19.26	18.68	20.75	17.74	17.00	18.97

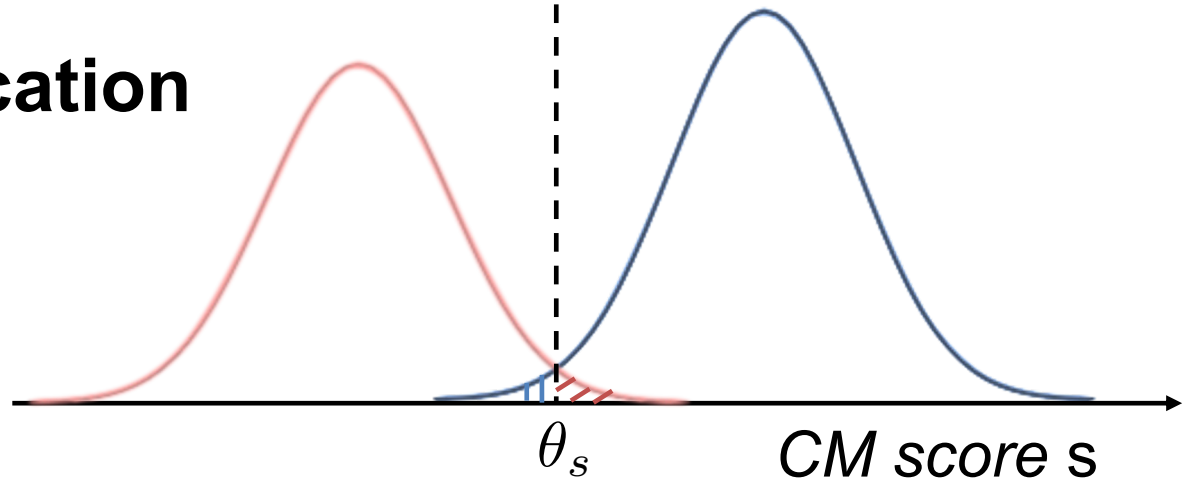
The three train-evaluation rounds (I, II, III)

Introduction



❑ Anti-spoofing as binary classification

- $\text{waveform} \mapsto s \in \mathbb{R}$
- Evaluation metrics
  - equal error rate (EER)
  - min-tDCF(Kinnuen 2020)



ASVspoof 2021 DF

Compression Methods:

Cond.	Compression (Quality)	VBR settings (kbps)
C1	None	-
C2	mp3 (low)	~80-120
C3	mp3 (high)	~220-260
C4	m4a (low)	~20-32
C5	m4a (high)	~96-112
C6	ogg (low)	~80-96
C7	ogg (high)	~256-320
C8	mp3 (low) => m4a (high)	~80-120=>~96-112
C9	ogg (low) => m4a (high)	~80-96=>~96-112

No compression

Single compression

Double compression

Audio data:

	Conditions								
	C1	C2	C3	C4	C5	C6	C7	C8	C9
ASVspoof / VCTK									
VCC 2018									
VCC 2020									

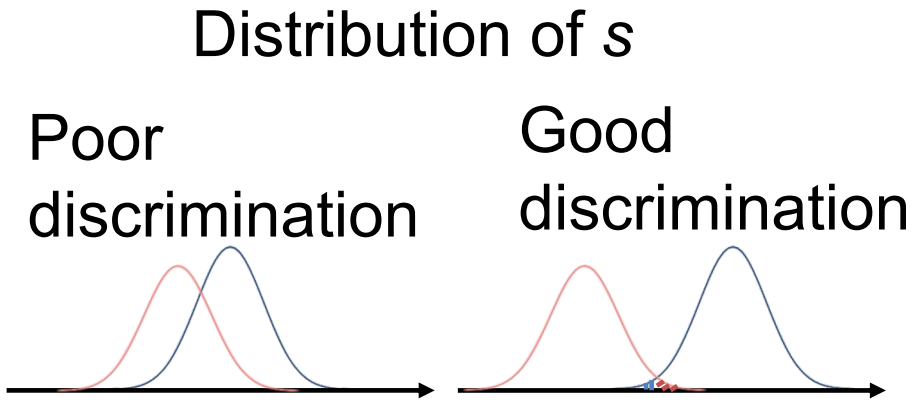
Progress & Eval Set

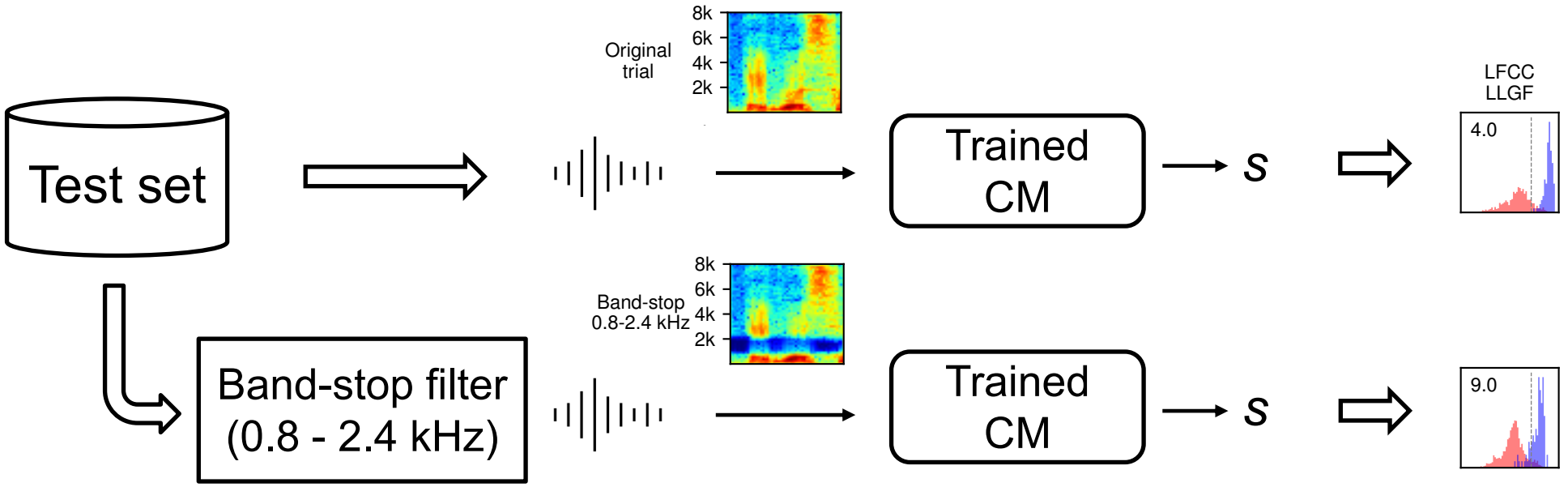
Only Eval Set

Analysis

Analysis

- **Sub-band analysis after training**
  - Insipred by (Tak 2020)
  - Given a trained CM:



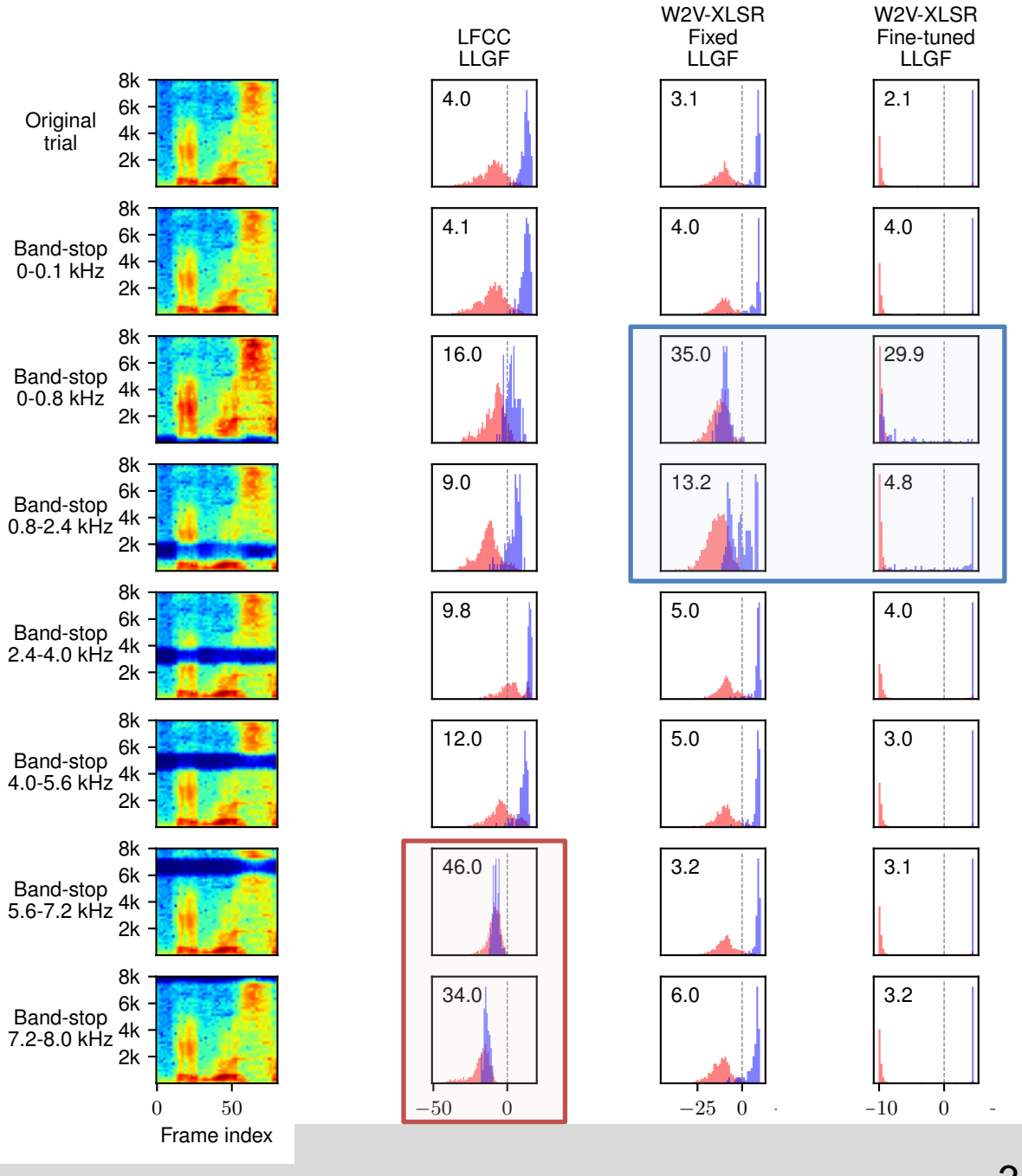


Analysis

- **Results on ASVspoof 2019 LA test subset**

  - Baseline replies more on high-freq. “evidence” ?

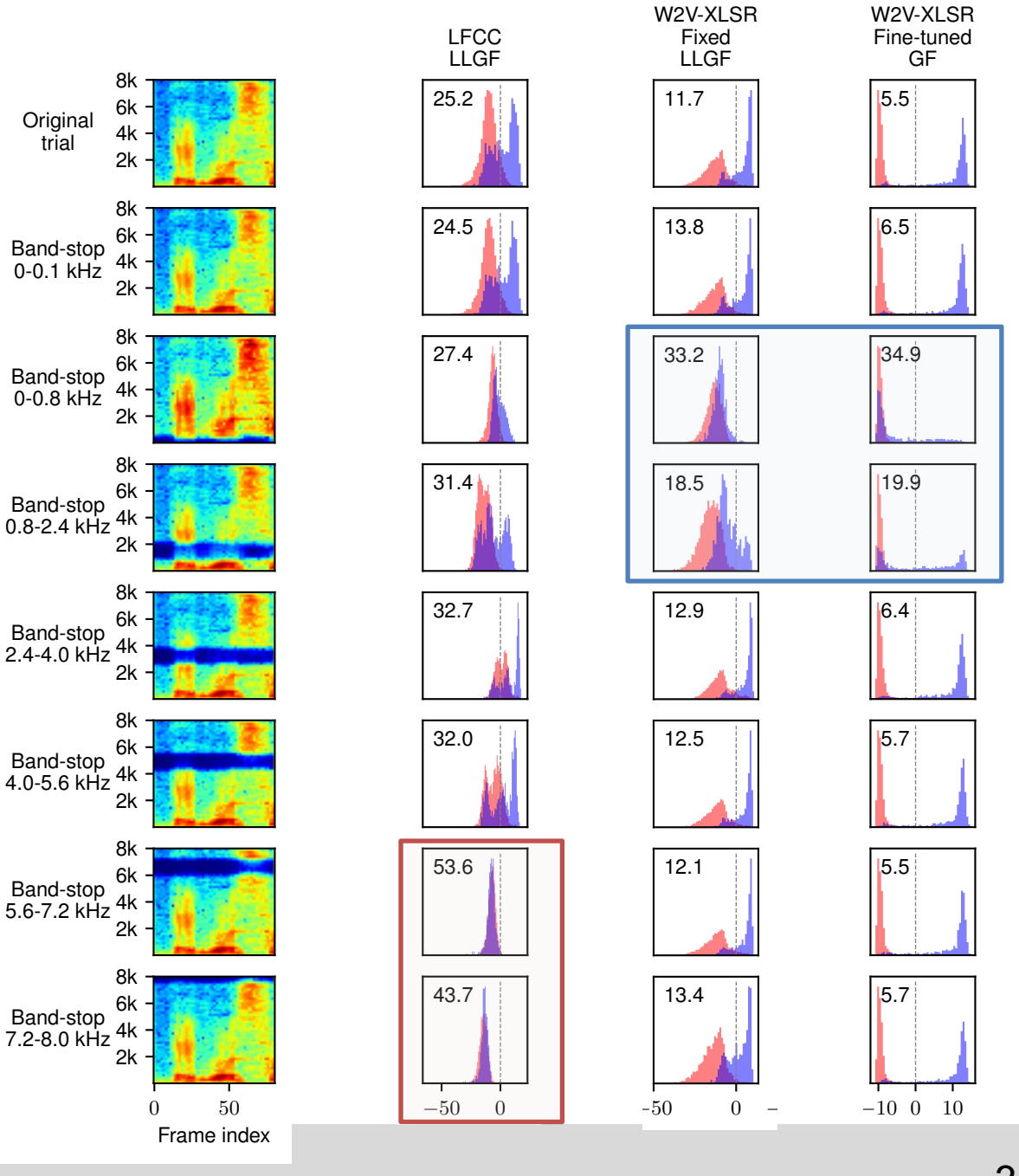
  - SSL front end is more sensitive to low freq. band (F1 in speech formant)



See full results in the paper

Analysis

- **Results on ASVspoof 2021 DF test subset**
  - Baseline replies more on high-freq. “evidence” ?
  - SSL front end is more sensitive to low freq. band (F1 in speech formant)
  - Similar patterns on other test sets and SSL front ends



See full results in the paper