2021-09-03
Interspeech 2021

Comparison of (some) Recent Neural Network Spoofing Countermeasures for Logical Access Scenario

Xin Wang, Junichi Yamagishi

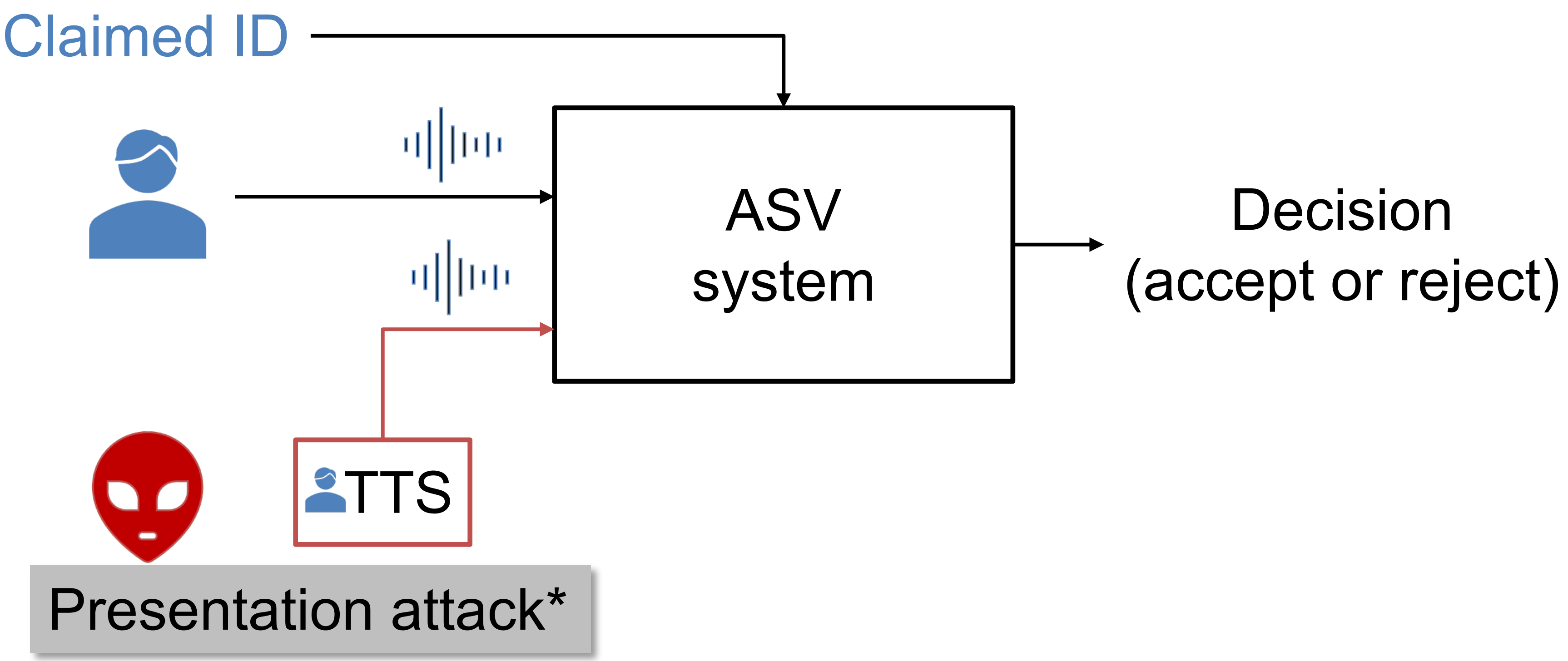
Fri-M-V-7-1

wangxin@nii.ac.jp

Background

Background

ASV & spoofing attack

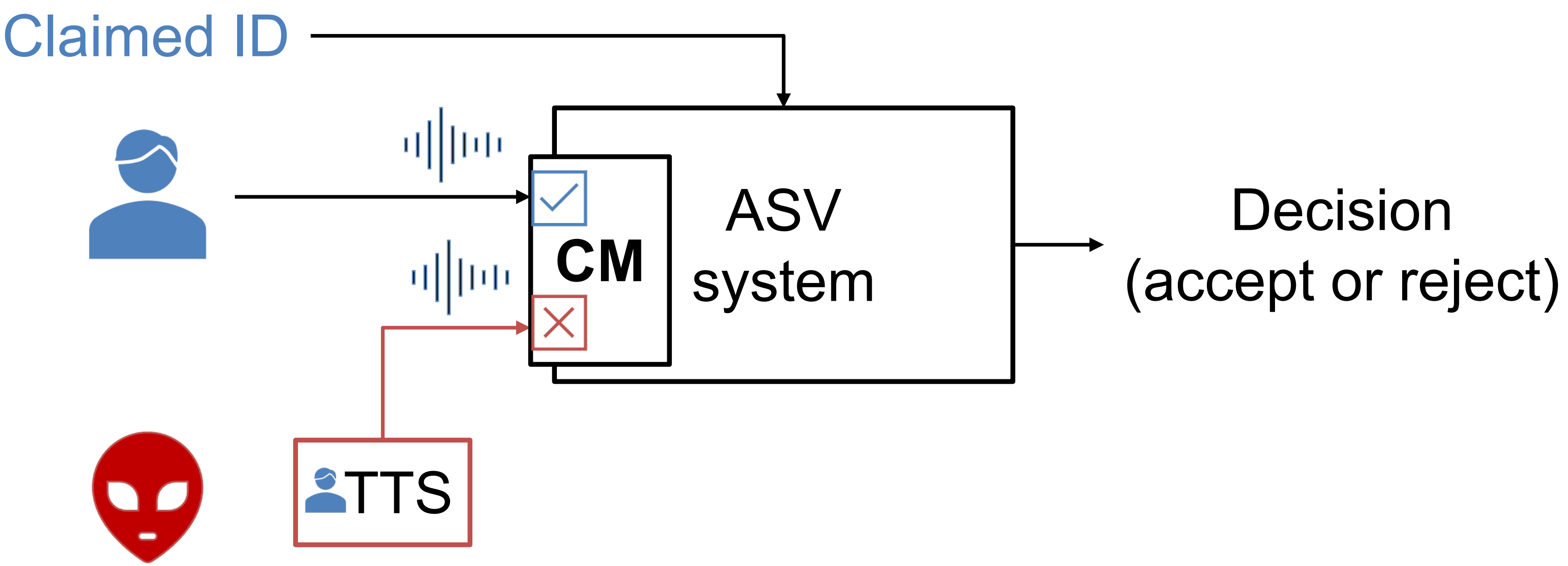


- ASV: Automatic Speaker Verification
- TTS: Text-to-Speech

*from: ISO/IEC CD 30107-1, Information Technology — Biometrics -- Presentation Attack Detection

Background

❑ Spoofing countermeasures (CMs)



- Task: CM(input_speech) → *Bona fide* / *Spoofed*
- Also known as Presentation Attack Detection (PAD)
- Logical access (Wu 2015): TTS, Voice conversion, …

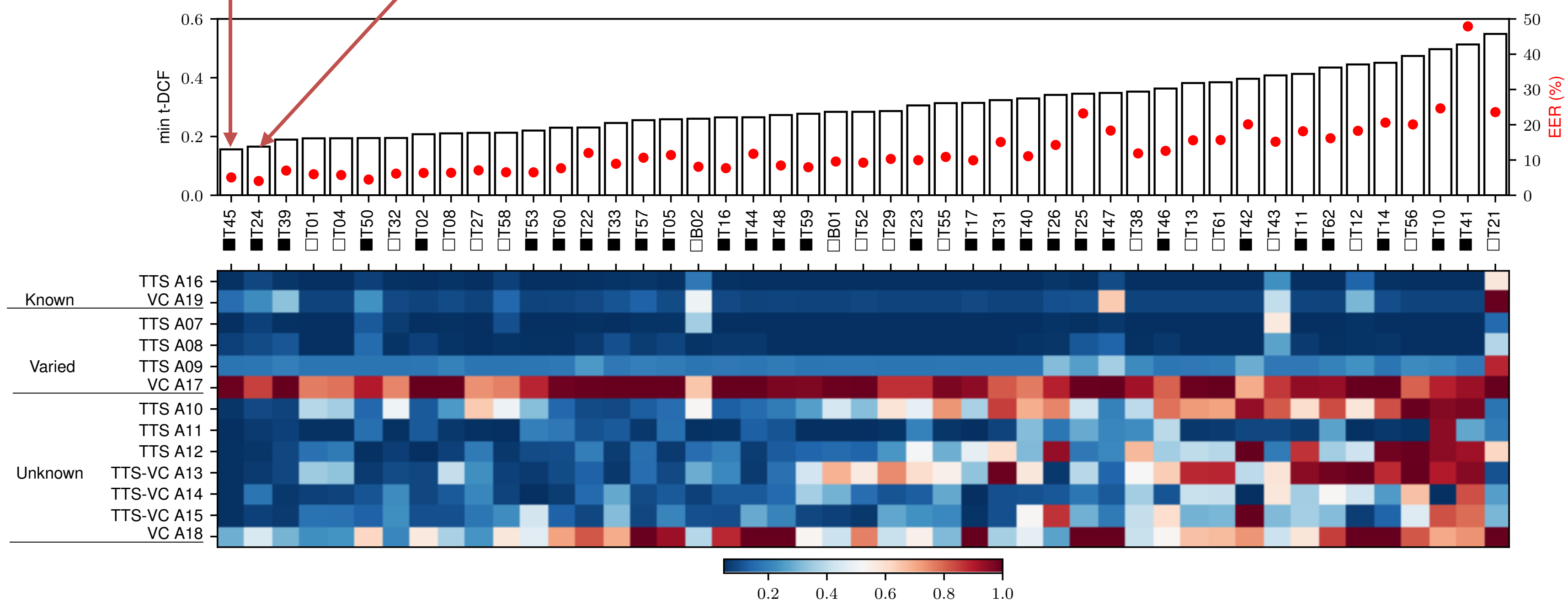
Wu, Z. *et al.* Spoofing and countermeasures for speaker verification: A survey. *Speech Commun.* **66**, 130–153 (2015)

Background

❑ ASVspoof challenge 2019 on LA

- Ranking of single system (not model ensemble)

EER: 5.06%   4.04%



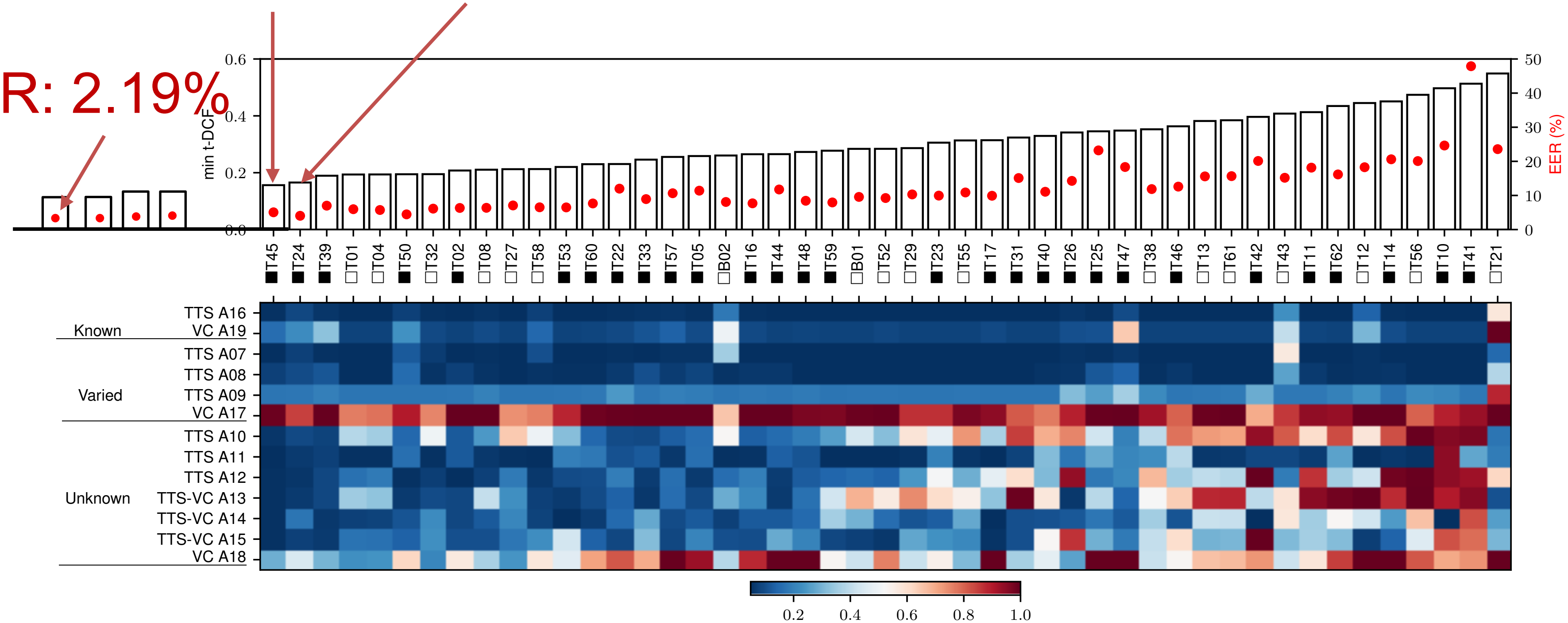
Background

❑ ASVspoof challenge 2019 on LA

- Ranking of single system (not model ensemble)

EER: 5.06%   4.04%

EER: 2.19%

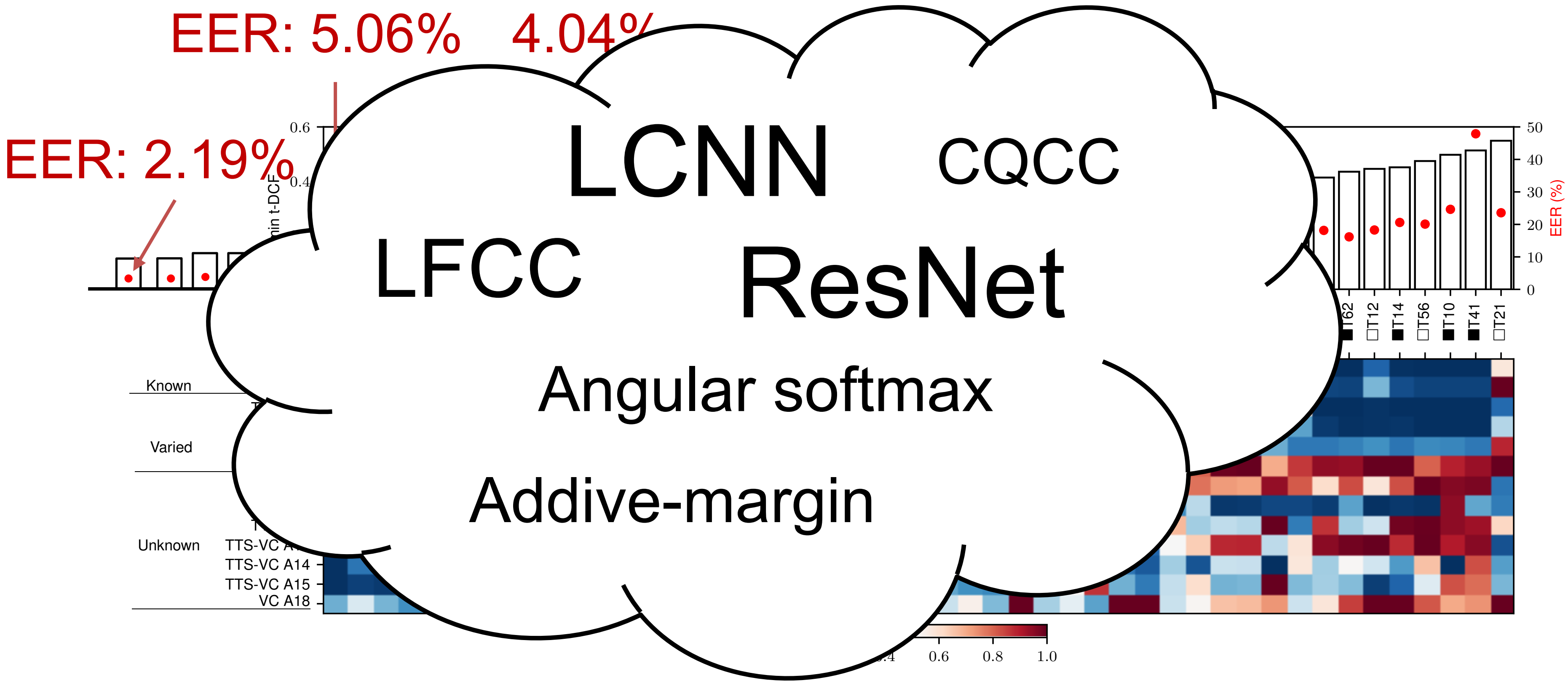


Zhang, Y., Jiang, F. & Duan, Z. One-class learning towards generalized voice spoofing detection. *arXiv Prepr. arXiv2010.13995* (2020)
Wu, Z., Das, R. K., Yang, J. & Li, H. Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks. in *Proc. Interspeech 2020* 1101–1105 (2020). doi:10.21437/Interspeech.2020-1810
Tak, H. *et al.* End-to-end anti-spoofing with RawNet2. *arXiv Prepr. arXiv2011.01108* (2020)
Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G. & Khoury, E. Generalization of Audio Deepfake Detection. in *Proc. Odyssey* 132–137 (2020). doi:10.21437/Odyssey.2020-19

Background

❑ ASVspoof challenge 2019 on LA

- Ranking of single system (not model ensemble)



Zhang, Y., Jiang, F. & Duan, Z. One-class learning towards generalized voice spoofing detection. *arXiv Prepr. arXiv2010.13995* (2020)
Wu, Z., Das, R. K., Yang, J. & Li, H. Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks. in *Proc. Interspeech 2020* 1101–1105 (2020). doi:10.21437/Interspeech.2020-1810
Tak, H. *et al.* End-to-end anti-spoofing with RawNet2. *arXiv Prepr. arXiv2011.01108* (2020)
Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G. & Khoury, E. Generalization of Audio Deepfake Detection. in *Proc. Odyssey* 132–137 (2020). doi:10.21437/Odyssey.2020-19

This study

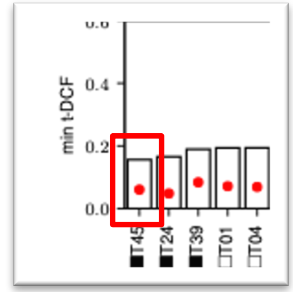
Motivation

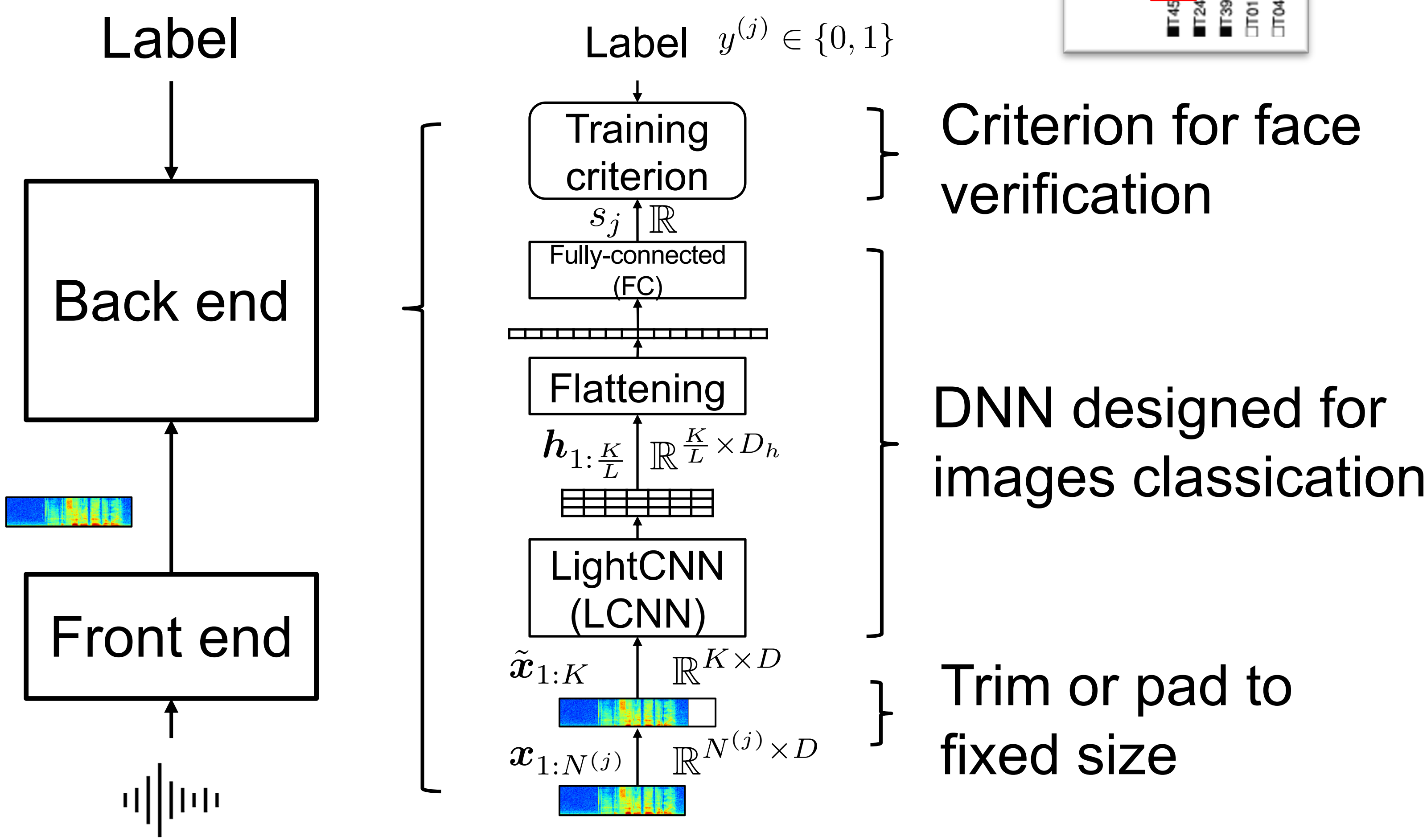
- **Many new CMs using DNNs**
  - New techniques borrowed from face verification, image classification, …
  - Better results on spoofing CMs

- **What have we learned?**
  - Suitable DNN architectures?
  - Good training criterion?
  - Do new techniques outperform others significantly?

Methods

❑ Recent neural CMs – network

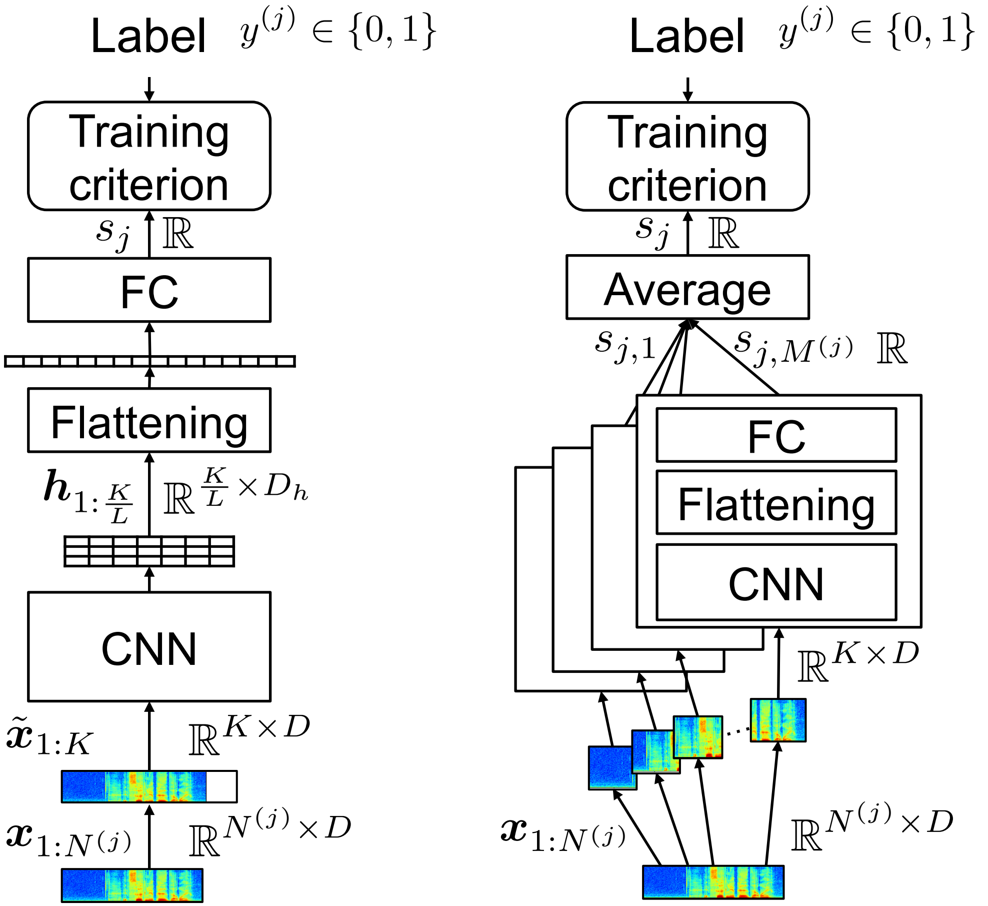




Lavrentyeva, G. *et al.* STC Antispoofing Systems for the ASVspoof2019 Challenge. in *Proc. Interspeech* 1033–1037 (2019). doi:10.21437/Interspeech.2019-1768

Methods

❑ Recent neural CMs – network



Criterion for face verification

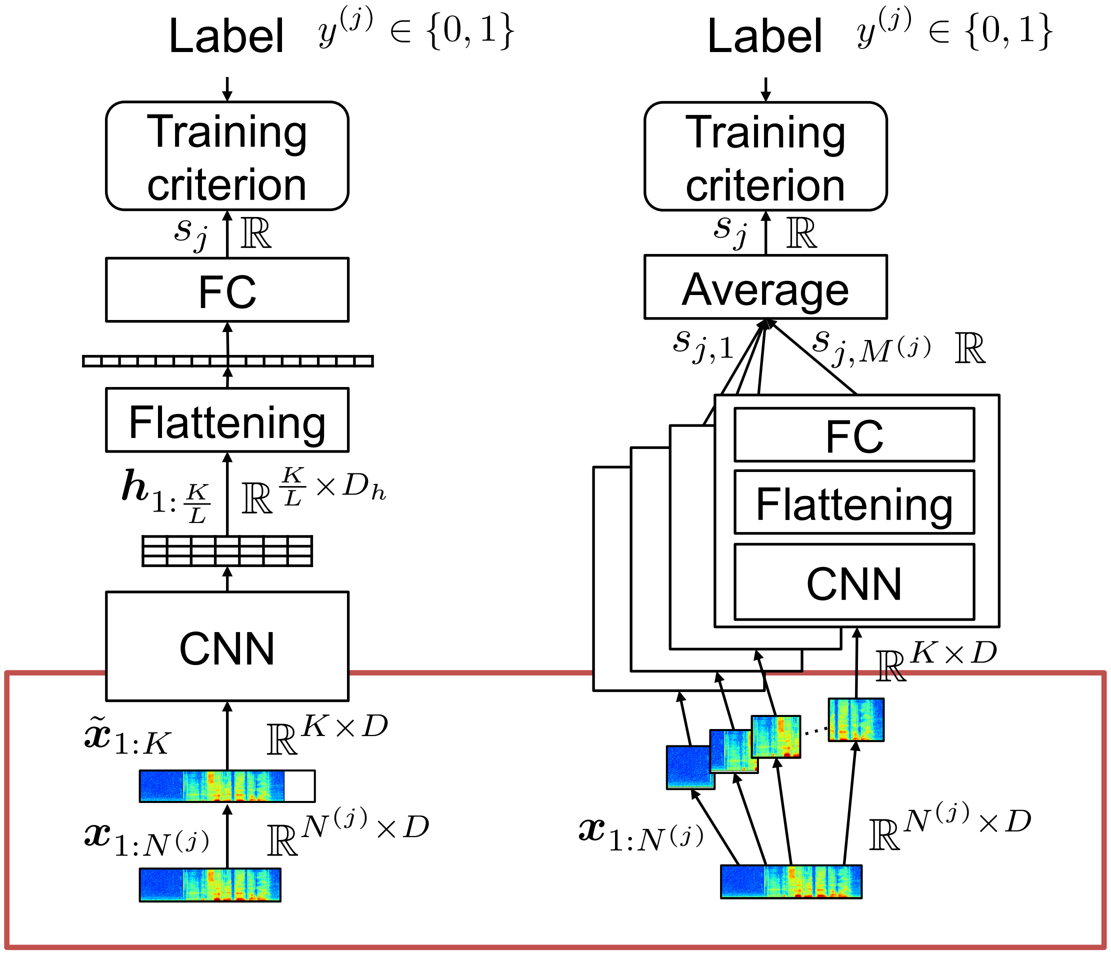
DNN designed for images

Framing (again!)

Lai, C.-I. *et al.* Attentive filtering networks for audio replay attack detection. in *Proc. ICASSP* 6316–6320 (2019)
Wang, Q., Lee, K. A. & Koshinaka, T. Using Multi-Resolution Feature Maps with Convolutional Neural Networks for Anti-Spoofing in ASV. in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop* 138–142 (2020). doi:10.21437/Odyssey.2020-20

Methods

❑ Recent neural CMs – network



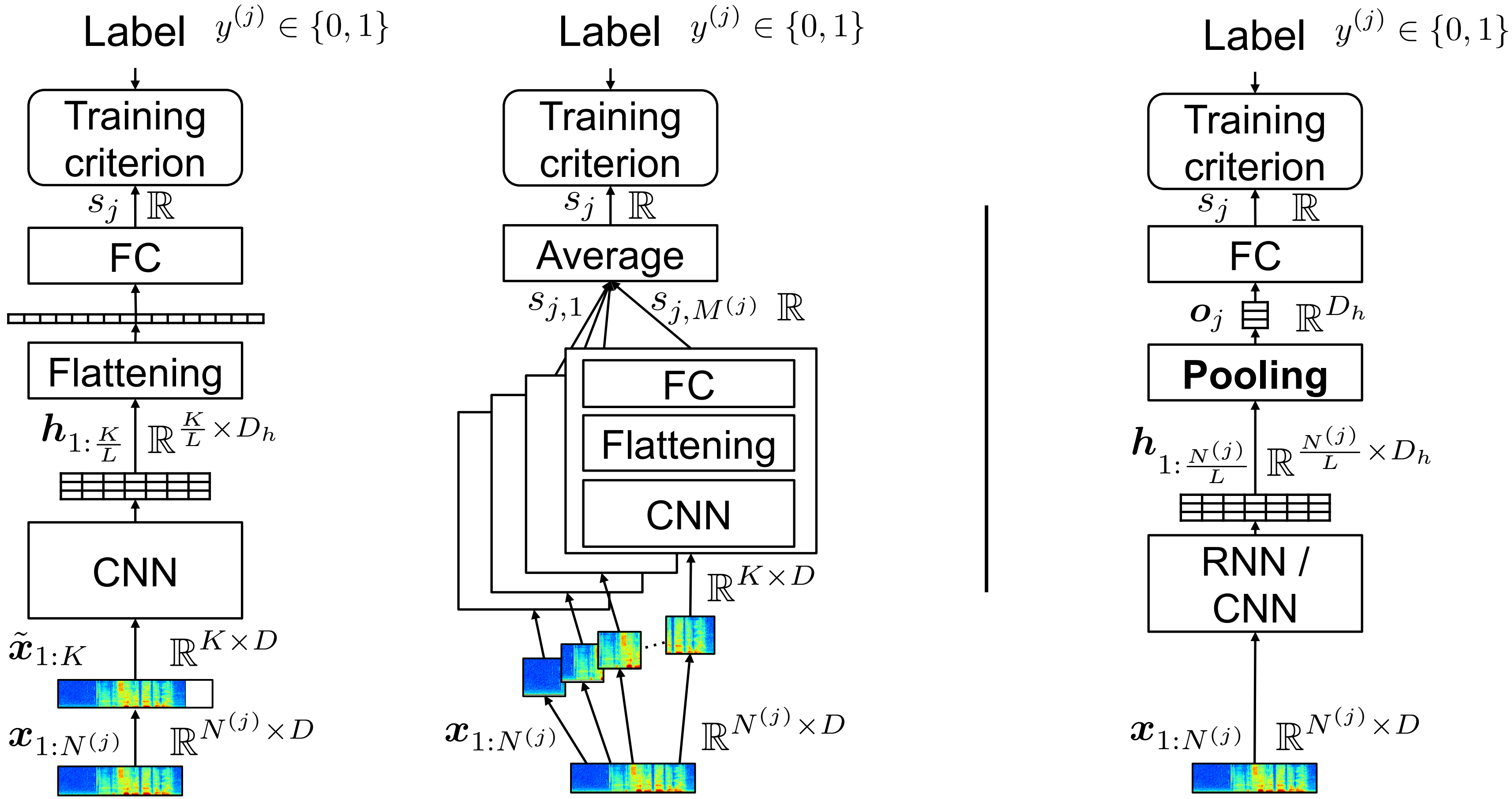
- Trim-pad or framing introduce new artefacts! (Chettri 2020)
- Additional hyper-parameters to tune
- Unnecessary randomness

Chettri, B., Benetos, E. & Sturm, B. L. T. Dataset artefacts in anti-spoofing systems: a case study on the ASVspoof 2017 benchmark. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **28**, 3018–3028 (2020)

Methods

❑ Recent neural CMs – network

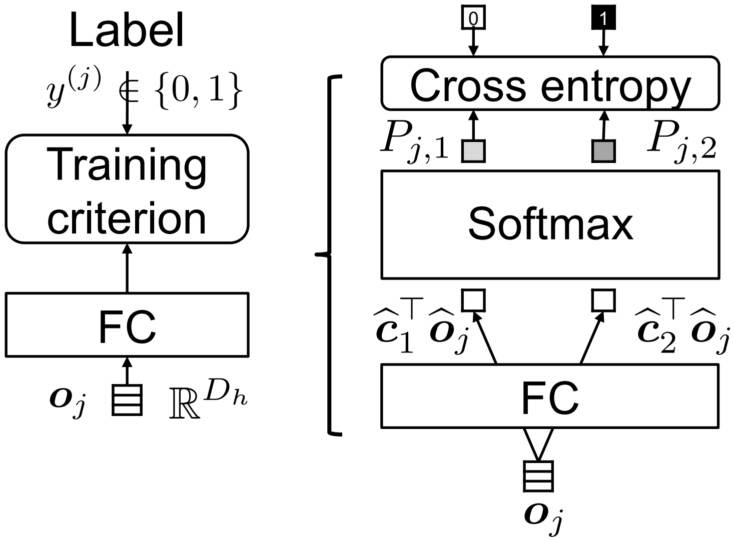
Why not?



- More CMs use trim/pad rather than pooling

Methods

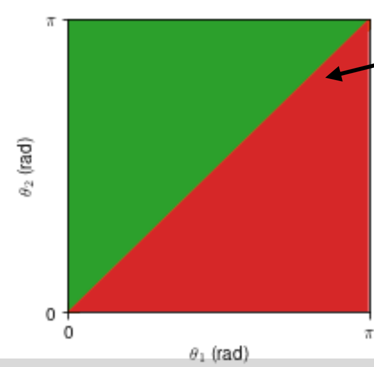
Recent neural CMs – training criterion



$$P_{j,k} = \frac{e^{\alpha \cos \theta_{j,k}}}{e^{\alpha \cos \theta_{j,k}} + \sum_{i=1, i\neq k}^{C} e^{\alpha \cos \theta_{j,i}}},$$

$$\cos \theta_{j,k} = \widehat{\boldsymbol{c}}_k^\top \widehat{\boldsymbol{o}}_j = \frac{\boldsymbol{c}_k^\top \boldsymbol{o}_j}{||\boldsymbol{c}_k|| \cdot ||\boldsymbol{o}_j||}$$

- Plain softmax (with length-normalization on input vector)

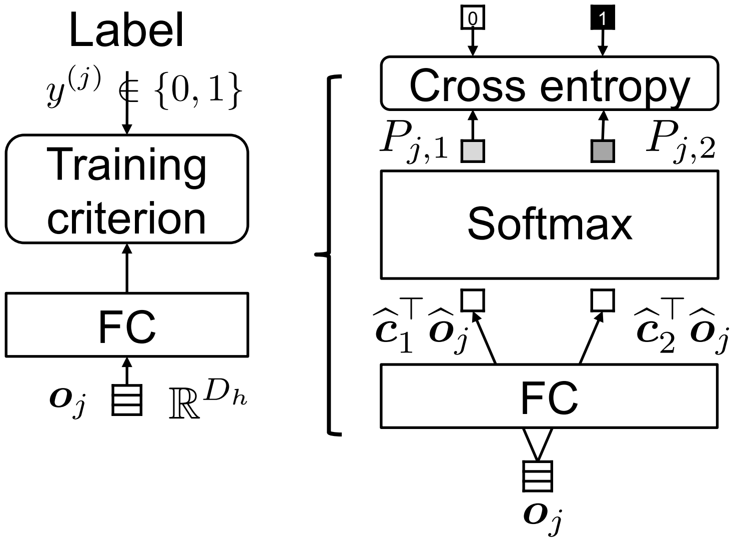


$$\theta_{j,0} = \theta_{j,1}$$

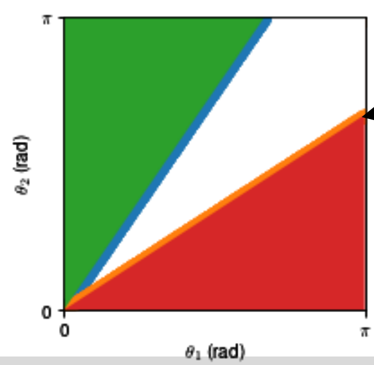
Liu, W. *et al.* Sphereface: Deep hypersphere embedding for face recognition. in *Proc. CVPR* 212–220 (2017).

Methods

❑ Recent neural CMs – training criterion



- More discriminative? --- Introduce margin

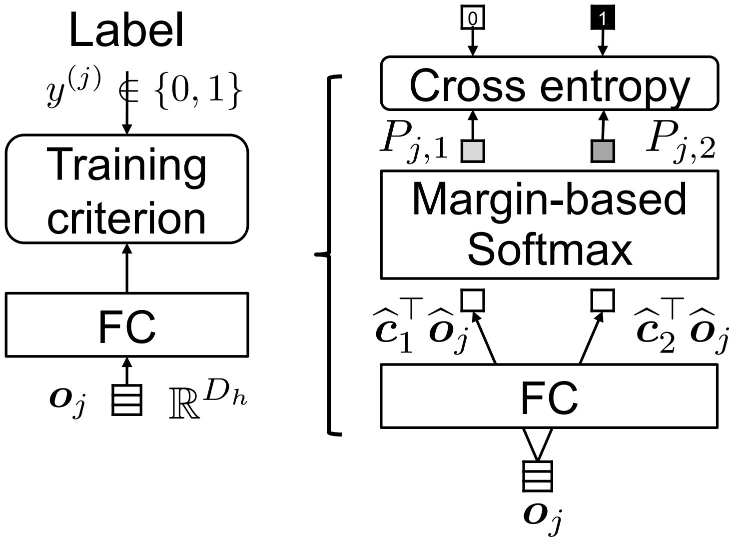


$$\theta_{j,0} = \frac{\theta_{j,1}}{m}$$

Liu, W. *et al.* Sphereface: Deep hypersphere embedding for face recognition. in *Proc. CVPR* 212–220 (2017).

Methods

❑ Recent neural CMs – training criterion

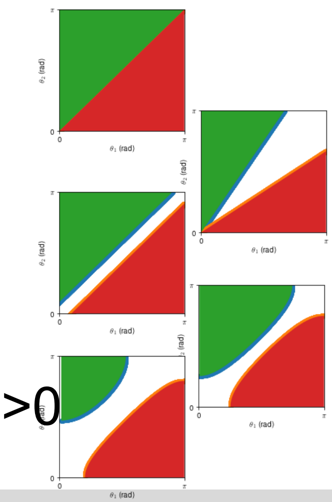


$$P_{j,k} = \frac{e^{\alpha[\cos(m_1\theta_{j,k}+m_2)-m_3]}}{e^{\alpha[\cos(m_1\theta_{j,k}+m_2)-m_3]} + \sum_{i=1,i\neq k}^{C} e^{\alpha\cos\theta_{j,i}}},$$

(Deng 2019, Eq(4))

$$\cos\theta_{j,k} = \widehat{\boldsymbol{c}}_k^\top \widehat{\boldsymbol{o}}_j$$

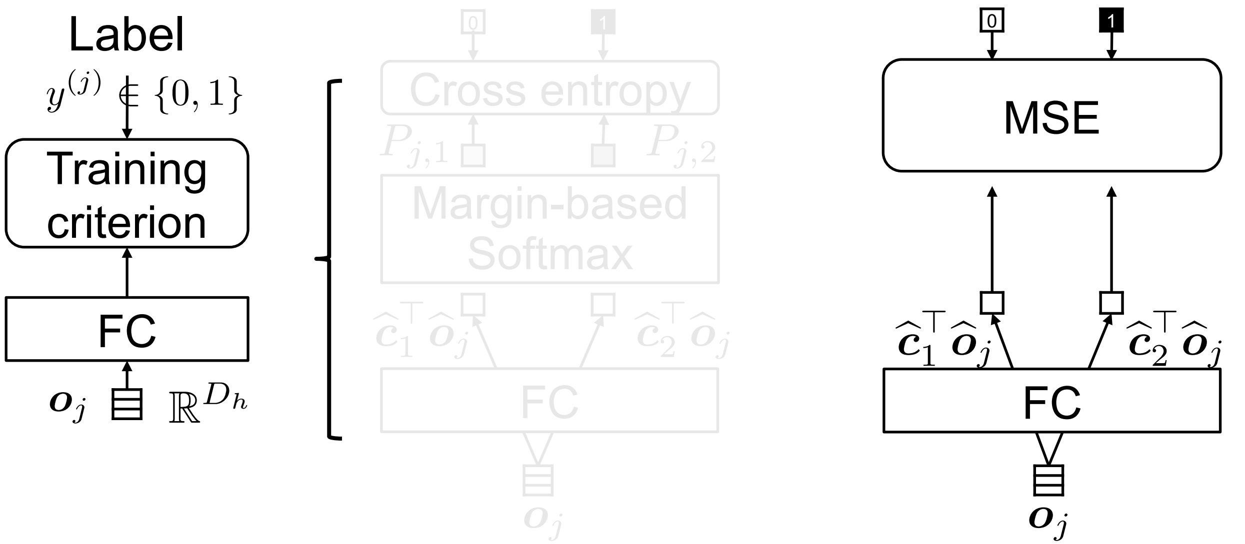
- Plain softmax: $m_1=1, m_2=0, m_3=0$
- Angular softmax(Liu 2017) $m_1>1, m_2=0, m_3=0$
- Additive angular(Deng 2019) $m_1=1, m_2>0, m_3=0$
- Additive margin (AM)(Wang 2018) $m_1=1, m_2=0, m_3>0$
- One-class (OC) (zhang 2020) $m_1=1, m_2=0, m_{3,1}>0, m_{3,2}>0$



Deng, J., Guo, J., Xue, N. & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. in *Proc. CVPR* 4690–4699 (2019).
Liu, W. *et al.* Sphereface: Deep hypersphere embedding for face recognition. in *Proc. CVPR* 212–220 (2017).
Wang, H. *et al.* Cosface: Large margin cosine loss for deep face recognition. in *Proc. CVPR* 5265–5274 (2018).
Zhang, Y., Jiang, F. & Duan, Z. One-class learning towards generalized voice spoofing detection. *arXiv Prepr. arXiv2010.13995* (2020)

Methods

A new training criterion



- Hyper-parameter-free mean square error

$$\mathcal{L}^{(\mathrm{p2s})} = \sum_{k=1}^{C} (\widehat{\boldsymbol{c}}_k^\top \widehat{\boldsymbol{o}}_j - \mathbb{1}(y_j = k))^2$$

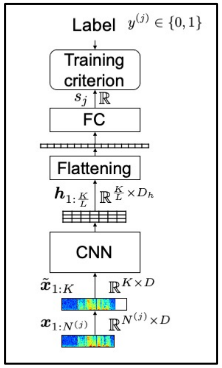
- Gradients will be equal to the so-called P2SGrad (zhang 2019)

Zhang, X. *et al.* P2sgrad: Refined gradients for optimizing deep face models. in *Proc. CVPR* 9906–9914 (2019).

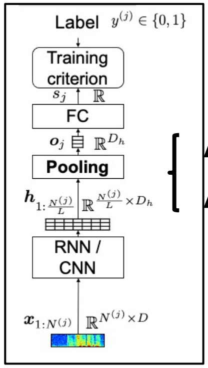
Experiments

❑ On ASVspoof 2019 LA, official protocol

- Front ends
  - Spectra, Linear filter-bank cepstrum coef (LFCC), or Linear filter bank coef (LFB)
- Back ends
  - Base on light-CNN (Lavrentyeva 2019)
  - Three types



Trim-pad



Attentive pooling
Average pooling

- Training criteria: plain softmax (sigmoid), AM, OC, p2s

Lavrentyeva, G. *et al.* STC Antispoofing Systems for the ASVspoof2019 Challenge. in *Proc. Interspeech* 1033–1037 (2019). doi:10.21437/Interspeech.2019-1768

Experiments

On ASVspoof 2019 LA , official protocol

- Multiple training & evaluation rounds (6 rounds)
  - Same CPU card & software dependency

```
For i in N rounds:
      random_seed ← 10^i
      model ← Training(train_dev_set)
      EER & min-tDCF ← model(eval_set)
```

- Significance test between models *A* & *B* (Bengio 2004):

$$z = \frac{2|\mathrm{EER}_A - \mathrm{EER}_B|}{\sqrt{\left[\mathrm{EER}_A(1-\mathrm{EER}_A) + \mathrm{EER}_B(1-\mathrm{EER}_B)\right] \frac{N_{\mathrm{bona}}+N_{\mathrm{spoof}}}{N_{\mathrm{bona}}N_{\mathrm{spoof}}}}}$$

  - Test: $z \geq Z_{\alpha/2}$?
  - $Z_{\alpha/2}$ at α = 0.05 with Holm-Bonferroni correction

Bengio, S. & Mariéthoz, J. A statistical significance test for person authentication. in *Proc. Odyssey* (2004).

Results

❑ EERs

Four types of training criterion

		EERs																							
		AM-softmax						OC-softmax						Sigmoid						MSE for P2SGrad					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	5.59	6.39	7.12	7.79	9.33	10.72	5.98	6.76	7.02	7.14	9.56	11.34	7.00	7.40	8.69	8.89	9.86	10.13	6.81	7.25	8.10	8.28	8.41	10.21
	Attention pooling	4.26	4.32	4.69	5.23	8.55	14.86	4.01	4.54	4.57	5.27	5.33	5.76	3.34	5.07	5.70	5.91	5.93	5.93	3.99	4.87	5.75	5.85	6.20	7.36
	Average pooling	4.24	5.27	5.71	6.23	7.03	7.10	5.81	6.51	6.89	7.64	9.15	10.24	7.04	7.93	8.52	9.16	9.53	9.84	5.06	5.17	5.69	5.75	6.17	6.50
SPEC	Trim-pad	4.84	5.02	5.23	5.87	5.90	6.25	4.41	4.60	4.84	5.04	5.37	6.35	3.09	3.58	4.72	4.82	5.22	5.56	2.94	3.22	3.59	3.73	4.49	4.74
	Attention pooling	4.02	4.08	4.99	5.22	5.57	8.20	4.05	4.55	4.94	5.67	6.31	8.47	3.92	4.04	4.42	4.91	4.95	6.62	4.70	5.03	5.60	5.88	6.35	6.50
	Average pooling	3.96	4.04	4.38	4.52	5.13	5.97	2.81	3.41	4.49	4.50	4.65	4.91	3.29	3.56	3.82	4.45	4.61	5.44	2.37	2.91	3.00	3.94	4.26	4.37
LFCC	Trim-pad	3.04	3.52	4.73	4.89	5.85	7.06	2.93	2.93	2.95	2.99	3.84	4.00	2.54	2.73	2.77	2.91	3.08	3.47	2.31	2.46	2.64	2.65	3.09	3.11
	Attention pooling	2.99	3.48	3.60	3.61	3.62	3.92	2.91	2.96	3.36	3.40	3.61	3.71	3.18	3.19	3.24	3.29	3.39	4.12	2.72	2.81	2.91	3.03	3.22	3.30
	Average pooling	2.46	3.02	3.23	3.34	3.41	3.86	2.23	2.42	2.96	2.96	2.96	4.64	2.67	2.94	3.20	3.40	3.45	3.53	1.92	2.09	2.43	2.50	2.62	3.10

Three types of back ends

6 rounds, sorted based on EER

Three types of front ends

☺ Low EER

☹ High EER

Results

❑ EERs

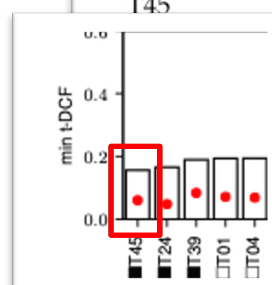
System EER in this study

		AM-softmax						OC-softmax						Sigmoid						MSE for P2SGrad					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	5.59	6.39	7.12	7.79	9.33	10.72	5.98	6.76	7.02	7.14	9.56	11.34	7.00	7.40	8.69	8.89	9.86	10.13	6.81	7.25	8.10	8.28	8.41	10.21
	Attention pooling	4.26	4.32	4.69	5.23	8.55	14.86	4.01	4.54	4.57	5.27	5.33	5.76	3.34	5.07	5.70	5.91	5.93	5.93	3.99	4.87	5.75	5.85	6.20	7.36
	Average pooling	4.24	5.27	5.71	6.23	7.03	7.10	5.81	6.51	6.89	7.64	9.15	10.24	7.04	7.93	8.52	9.16	9.53	9.84	5.06	5.17	5.69	5.75	6.17	6.50
SPEC	Trim-pad	4.84	5.02	5.23	5.87	5.90	6.25	4.41	4.60	4.84	5.04	5.37	6.35	3.09	3.58	4.72	4.82	5.22	5.56	2.94	3.22	3.59	3.73	4.49	4.74
	Attention pooling	4.02	4.08	4.99	5.22	5.57	8.20	4.05	4.55	4.94	5.67	6.31	8.47	3.92	4.04	4.42	4.91	4.95	6.62	4.70	5.03	5.60	5.88	6.35	6.50
	Average pooling	3.96	4.04	4.38	4.52	5.13	5.97	2.81	3.41	4.49	4.50	4.65	4.91	3.29	3.56	3.82	4.45	4.61	5.44	2.37	2.91	3.00	3.94	4.26	4.37
LFCC	Trim-pad	3.04	3.52	4.73	4.89	5.85	7.06	2.93	2.93	2.95	2.99	3.84	4.00	2.54	2.73	2.77	2.91	3.08	3.47	2.31	2.46	2.64	2.65	3.09	3.11
	Attention pooling	2.99	3.48	3.60	3.61	3.62	3.92	2.91	2.96	3.36	3.40	3.61	3.71	3.18	3.19	3.24	3.29	3.39	4.12	2.72	2.81	2.91	3.03	3.22	3.30
	Average pooling	2.46	3.02	3.23	3.34	3.41	3.86	2.23	2.42	2.96	2.96	2.96	4.64	2.67	2.94	3.20	3.40	3.45	3.53	1.92	2.09	2.43	2.50	2.62	3.10

- Reproduce the original CM (T45 in ASVspoof2019 LA)
- Varied results across multiple runs

Top single CMs in ASVspoof2019 LA

TID	min t-DCF	EER [%]
T45	0.1562	5.06
[illegible]	[illegible].1655	4.04
[illegible]	[illegible].1894	7.01
[illegible]	[illegible].1937	5.97
[illegible]	[illegible].1939	5.74
[illegible]	[illegible].2839	9.57
[illegible]	[illegible].2605	8.09
[illegible]	[illegible].0627	0.0



(Nautsch 2019, Tab.2)

Nautsch, A. *et al.* ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biometrics, Behav. Identity Sci.* (2021) doi:10.1109/TBIOM.2021.3059479

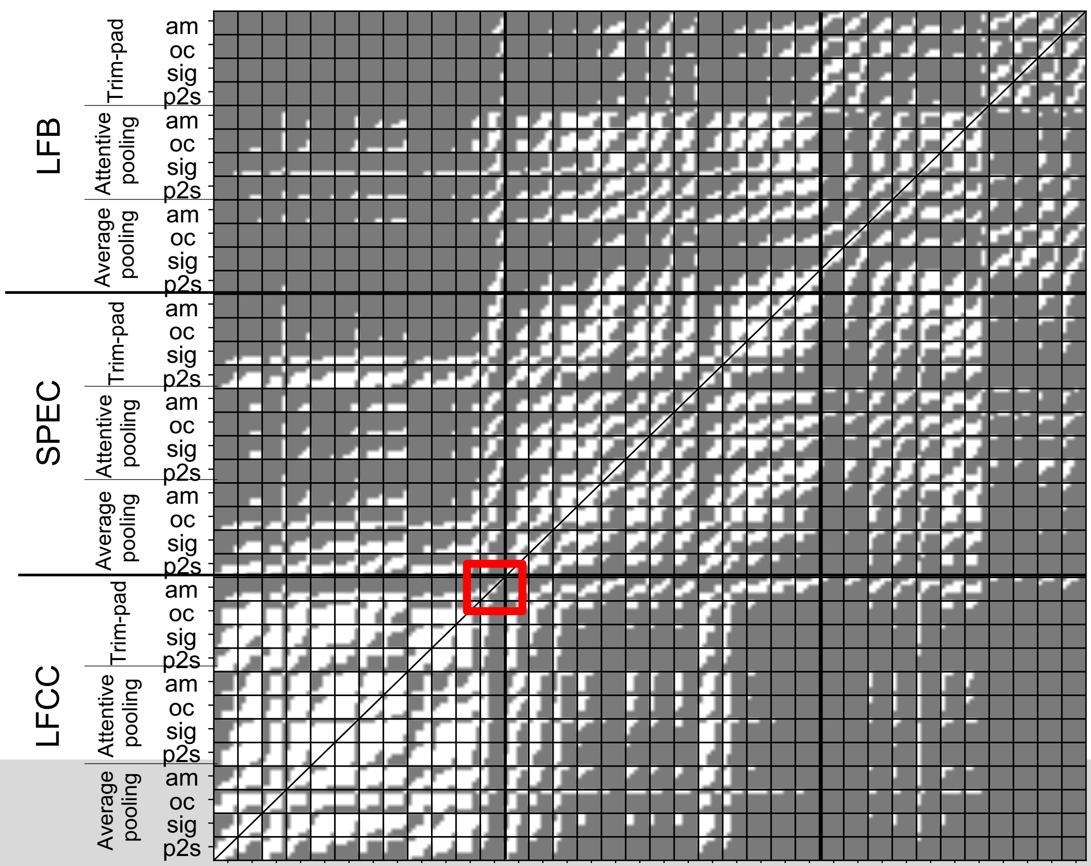
Results

❑ Pairwise significance test

EERs

		AM-softmax						OC-softmax					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	5.59	6.39	7.12	7.79	9.33	10.72	5.98	6.76	7.02	7.14	9.56	11.34
	Attention pooling	4.26	4.32	4.69	5.23	8.55	14.86	4.01	4.54	4.57	5.27	5.33	5.76
	Average pooling	4.24	5.27	5.71	6.23	7.03	7.10	5.81	6.51	6.89	7.64	9.15	10.24
SPEC	Trim-pad	4.84	5.02	5.23	5.87	5.90	6.25	4.41	4.60	4.84	5.04	5.37	6.35
	Attention pooling	4.02	4.08	4.99	5.22	5.57	8.20	4.05	4.55	4.94	5.67	6.31	8.47
	Average pooling	3.96	4.04	4.38	4.52	5.13	5.97	2.81	3.41	4.49	4.50	4.65	4.91
LFCC	Trim-pad	3.04	3.52	4.73	4.89	5.85	7.06	[illegible]	2.93	2.95	2.99	3.84	4.00
	Attention pooling	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.96	3.36	3.40	3.61	3.71
	Average pooling	2.46	3.02	3.23	3.34	3.41	3.86	2.23	2.42	2.96	2.96	2.96	4.64

		Sigmoid						MSE for P2SGrad					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	7.00	7.40	8.69	8.89	9.86	10.13	6.81	7.25	8.10	8.28	8.41	10.21
	Attention pooling	3.34	5.07	5.70	5.91	5.93	5.93	3.99	4.87	5.75	5.85	6.20	7.36
	Average pooling	7.04	7.93	8.52	9.16	9.53	9.84	5.06	5.17	5.69	5.75	6.17	6.50
SPEC	Trim-pad	3.09	3.58	4.72	4.82	5.22	5.56	2.94	3.22	3.59	3.73	4.49	4.74
	Attention pooling	3.92	4.04	4.42	4.91	4.95	6.62	4.70	5.03	5.60	5.88	6.35	6.50
	Average pooling	3.29	3.56	3.82	4.45	4.61	5.44	2.37	2.91	3.00	3.94	4.26	4.37
LFCC	Trim-pad	2.54	2.73	2.77	2.91	3.08	3.47	2.31	2.46	2.64	2.65	3.09	3.11
	Attention pooling	3.18	3.19	3.24	3.29	3.39	4.12	2.72	2.81	2.91	3.03	3.22	3.30
	Average pooling	2.67	2.94	3.20	3.40	3.45	3.53	1.92	2.09	2.43	2.50	2.62	3.10



} 6x6 pairs

- Grey color indicates significant difference ($\alpha = 0.05$)

Intra-system variance can be significant

Results

❑ EERs

System EER in this study

		AM-softmax						OC-softmax						Sigmoid						MSE for P2SGrad					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	5.59	6.39	7.12	7.79	9.33	10.72	5.98	6.76	7.02	7.14	9.56	11.34	7.00	7.40	8.69	8.89	9.86	10.13	6.81	7.25	8.10	8.28	8.41	10.21
	Attention pooling	4.26	4.32	4.69	5.23	8.55	14.86	4.01	4.54	4.57	5.27	5.33	5.76	3.34	5.07	5.70	5.91	5.93	5.93	3.99	4.87	5.75	5.85	6.20	7.36
	Average pooling	4.24	5.27	5.71	6.23	7.03	7.10	5.81	6.51	6.89	7.64	9.15	10.24	7.04	7.93	8.52	9.16	9.53	9.84	5.06	5.17	5.69	5.75	6.17	6.50
SPEC	Trim-pad	4.84	5.02	5.23	5.87	5.90	6.25	4.41	4.60	4.84	5.04	5.37	6.35	3.09	3.58	4.72	4.82	5.22	5.56	2.94	3.22	3.59	3.73	4.49	4.74
	Attention pooling	4.02	4.08	4.99	5.22	5.57	8.20	4.05	4.55	4.94	5.67	6.31	8.47	3.92	4.04	4.42	4.91	4.95	6.62	4.70	5.03	5.60	5.88	6.35	6.50
	Average pooling	3.96	4.04	4.38	4.52	5.13	5.97	2.81	3.41	4.49	4.50	4.65	4.91	3.29	3.56	3.82	4.45	4.61	5.44	2.37	2.91	3.00	3.94	4.26	4.37
LFCC	Trim-pad	3.04	3.52	4.73	4.89	5.85	7.06	2.93	2.93	2.95	2.99	3.84	4.00	2.54	2.73	2.77	2.91	3.08	3.47	2.31	2.46	2.64	2.65	3.09	3.11
	Attention pooling	2.99	3.48	3.60	3.61	3.62	3.92	2.91	2.96	3.36	3.40	3.61	3.71	3.18	3.19	3.24	3.29	3.39	4.12	2.72	2.81	2.91	3.03	3.22	3.30
	Average pooling	2.46	3.02	3.23	3.34	3.41	3.86	2.23	2.42	2.96	2.96	2.96	4.64	2.67	2.94	3.20	3.40	3.45	3.53	1.92	2.09	2.43	2.50	2.62	3.10

- Less variance when
  - Using hyper-para-free training criterion
  - Avoid trim-pad and use pooling

Top single CMs in ASVspoof2019 LA

TID	min t-DCF	EER [%]
T45	0.1562	5.06
T24	0.1655	4.04
T39	0.1894	7.01
T01	0.1937	5.97
T04	0.1939	5.74
B01	0.2839	9.57
B02	0.2605	8.09
Perfect	0.0627	0.0

(Nautsch 2019, Tab.2)

Nautsch, A. *et al.* ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biometrics, Behav. Identity Sci.* (2021) doi:10.1109/TBIOM.2021.3059479

Results

EERs

System EER in this study

		AM-softmax						OC-softmax						Sigmoid						MSE for P2SGrad					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	5.59	6.39	7.12	7.79	9.33	10.72	5.98	6.76	7.02	7.14	9.56	11.34	7.00	7.40	8.69	8.89	9.86	10.13	6.81	7.25	8.10	8.28	8.41	10.21
	Attention pooling	4.26	4.32	4.69	5.23	8.55	14.86	4.01	4.54	4.57	5.27	5.33	5.76	3.34	5.07	5.70	5.91	5.93	5.93	3.99	4.87	5.75	5.85	6.20	7.36
	Average pooling	4.24	5.27	5.71	6.23	7.03	7.10	5.81	6.51	6.89	7.64	9.15	10.24	7.04	7.93	8.52	9.16	9.53	9.84	5.06	5.17	5.69	5.75	6.17	6.50
SPEC	Trim-pad	4.84	5.02	5.23	5.87	5.90	6.25	4.41	4.60	4.84	5.04	5.37	6.35	3.09	3.58	4.72	4.82	5.22	5.56	2.94	3.22	3.59	3.73	4.49	4.74
	Attention pooling	4.02	4.08	4.99	5.22	5.57	8.20	4.05	4.55	4.94	5.67	6.31	8.47	3.92	4.04	4.42	4.91	4.95	6.62	4.70	5.03	5.60	5.88	6.35	6.50
	Average pooling	3.96	4.04	4.38	4.52	5.13	5.97	2.81	3.41	4.49	4.50	4.65	4.91	3.29	3.56	3.82	4.45	4.61	5.44	2.37	2.91	3.00	3.94	4.26	4.37
LFCC	Trim-pad	3.04	3.52	4.73	4.89	5.85	7.06	2.93	2.93	2.95	2.99	3.84	4.00	2.54	2.73	2.77	2.91	3.08	3.47	2.31	2.46	2.64	2.65	3.09	3.11
	Attention pooling	2.99	3.48	3.60	3.61	3.62	3.92	2.91	2.96	3.36	3.40	3.61	3.71	3.18	3.19	3.24	3.29	3.39	4.12	2.72	2.81	2.91	3.03	3.22	3.30
	Average pooling	2.46	3.02	3.23	3.34	3.41	3.86	2.23	2.42	2.96	2.96	2.96	4.64	2.67	2.94	3.20	3.40	3.45	3.53	1.92	2.09	2.43	2.50	2.62	3.10

- Potential techniques:
  - Front-end:
    - LFCC
  - Network:
    - average pooling
  - Training criterion:
    - simple softmax (sigmoid)
    - MSE with P2SGrad

Top single CMs in ASVspoof2019 LA

TID	min t-DCF	EER [%]
T45	0.1562	5.06
T24	0.1655	4.04
T39	0.1894	7.01
T01	0.1937	5.97
T04	0.1939	5.74
B01	0.2839	9.57
B02	0.2605	8.09
Perfect	0.0627	0.0

(Nautsch 2019, Tab.2)

Nautsch, A. *et al.* ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biometrics, Behav. Identity Sci.* (2021) doi:10.1109/TBIOM.2021.3059479

Results

❑ EERs

System EER in this study

		AM-softmax						OC-softmax						Sigmoid						MSE for P2SGrad					
		I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI	I	II	III	IV	V	VI
LFB	Trim-pad	5.59	6.39	7.12	7.79	9.33	10.72	5.98	6.76	7.02	7.14	9.56	11.34	7.00	7.40	8.69	8.89	9.86	10.13	6.81	7.25	8.10	8.28	8.41	10.21
	Attention pooling	4.26	4.32	4.69	5.23	8.55	14.86	4.01	4.54	4.57	5.27	5.33	5.76	3.34	5.07	5.70	5.91	5.93	5.93	3.99	4.87	5.75	5.85	6.20	7.36
	Average pooling	4.24	5.27	5.71	6.23	7.03	7.10	5.81	6.51	6.89	7.64	9.15	10.24	7.04	7.93	8.52	9.16	9.53	9.84	5.06	5.17	5.69	5.75	6.17	6.50
SPEC	Trim-pad	4.84	5.02	5.23	5.87	5.90	6.25	4.41	4.60	4.84	5.04	5.37	6.35	3.09	3.58	4.72	4.82	5.22	5.56	2.94	3.22	3.59	3.73	4.49	4.74
	Attention pooling	4.02	4.08	4.99	5.22	5.57	8.20	4.05	4.55	4.94	5.67	6.31	8.47	3.92	4.04	4.42	4.91	4.95	6.62	4.70	5.03	5.60	5.88	6.35	6.50
	Average pooling	3.96	4.04	4.38	4.52	5.13	5.97	2.81	3.41	4.49	4.50	4.65	4.91	3.29	3.56	3.82	4.45	4.61	5.44	2.37	2.91	3.00	3.94	4.26	4.37
LFCC	Trim-pad	3.04	3.52	4.73	4.89	5.85	7.06	2.93	2.93	2.95	2.99	3.84	4.00	2.54	2.73	2.77	2.91	3.08	3.47	2.31	2.46	2.64	2.65	3.09	3.11
	Attention pooling	2.99	3.48	3.60	3.61	3.62	3.92	2.91	2.96	3.36	3.40	3.61	3.71	3.18	3.19	3.24	3.29	3.39	4.12	2.72	2.81	2.91	3.03	3.22	3.30
	Average pooling	2.46	3.02	3.23	3.34	3.41	3.86	2.23	2.42	2.96	2.96	2.96	4.64	2.67	2.94	3.20	3.40	3.45	3.53	1.92	2.09	2.43	2.50	2.62	3.10

- **1.92** is the lowest EER so far
- But not significantly different from 2.23, 2.31
- SOTA can be achived, but should be reported with caution

Other recent CMs after ASVspoof2019 challenge

System	EER (%)	min t-DCF
CQCC + GMM [3]	9.57	0.237
LFCC + GMM [3]	8.09	0.212
Chettri et al. [22]	7.66	0.179
Monterio et al. [14]	6.38	0.142
Gomez-Alanis et al. [16]	6.28	-
Aravind et al. [18]	5.32	0.151
Lavrentyeva et al. [21]	4.53	0.103
ResNet + OC-SVM	4.44	0.115
Wu et al. [17]	4.07	0.102
Tak et al. [19]	3.50	0.090
Chen et al. [15]	3.49	0.092
Proposed	**2.19**	**0.059**

(Zhang 2020, Tab.3)

Zhang, Y., Jiang, F. & Duan, Z. One-class learning towards generalized voice spoofing detection. *arXiv Prepr. arXiv2010.13995* (2020)

Summary

Messages

- **Suitable DNN architectures?**
  - Trim-pad can work
  - But pooling is more suitable
    - No trim, no information loss
    - No pad, no extra computation

- **Good training criterion?**
  - Plain softmax works
  - MSE for p2sgrad is simple and effective

Messages

Are new CMs good?

- Many CMs outperform ASVspoof2019 LA single systems
- But not necessarily outperform each other

Statistical analysis and multiple-runs are recommended for ASVspoof2019 LA!

Discussion

Variance in training

- Variance is there no matter how we tune learning rate, batch size … (see appendix 1-2)
- How to reduce it?

Fusion of model? (see appendix)

- Same model but different training rounds
  - They are similar
  - No significant gains by fusion
- Models with different front-ends are good to fuse (EER: 1%)