Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation

Hemlata Tak[1], Massimiliano Todisco[1], Xin Wang[2], Jee-weon Jung[3]
Junichi Yamagishi[2] and Nicholas Evans[1]

[1]EURECOM, France,
[2]National Institute of Informatics, Japan,
[3]Naver Corporation, South Korea

June 29, 2022

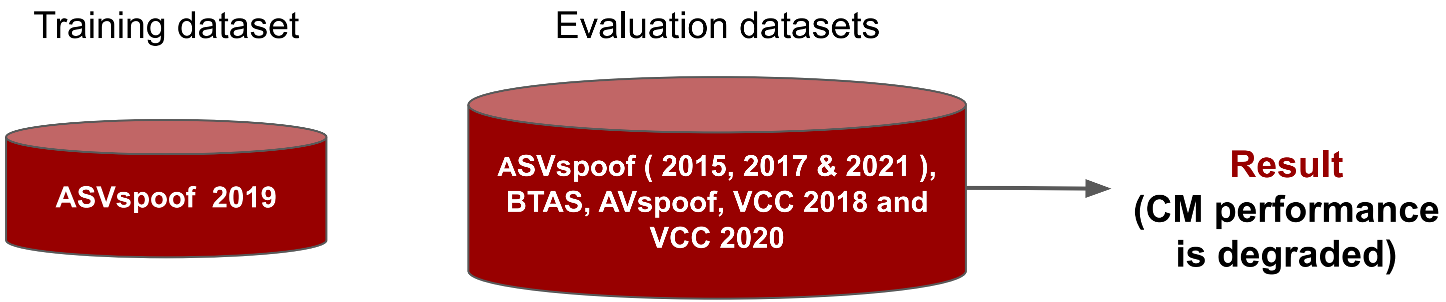
Contents

- Motivation
- Proposed framework
- Experimental setup
- Results
- Conclusions

Motivation

Challenges:

- Lack of generalisation and domain mismatch between training and testing data [1,2].
- Lack of sufficiently representative training data.



[1] D. Paul, M. Sahidullah, et al.," Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora", in Proc. ICASSP 2017.
[2] R. K Das, H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing", in Proc. ICASSP 2020.

Key idea

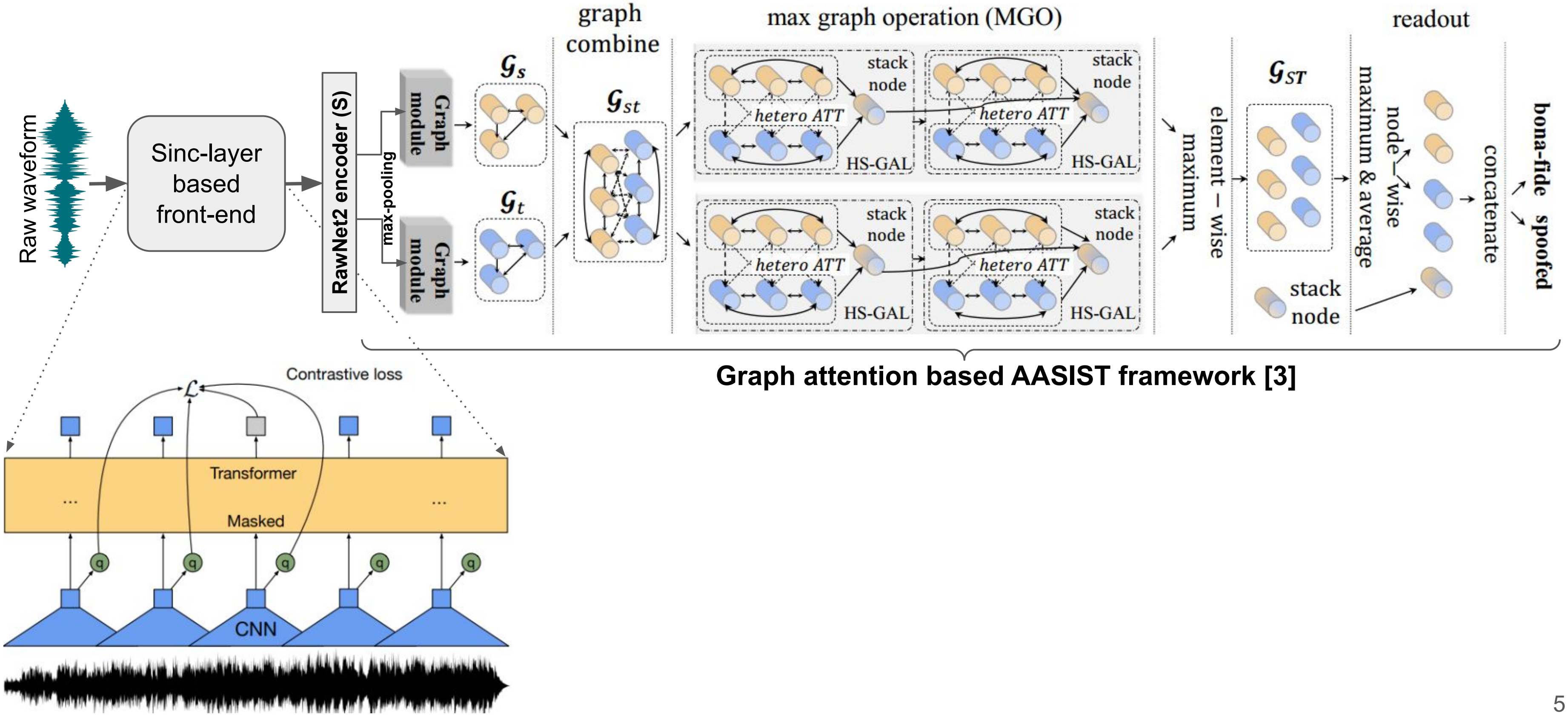
- Use larger and diverse representative training database.
    - Advantage: better generalisation
    - Disadvantage: It’s impractical - never enough

To use **self-supervised models** trained on large diverse datasets to learn more generalised representation.

Data augmentation

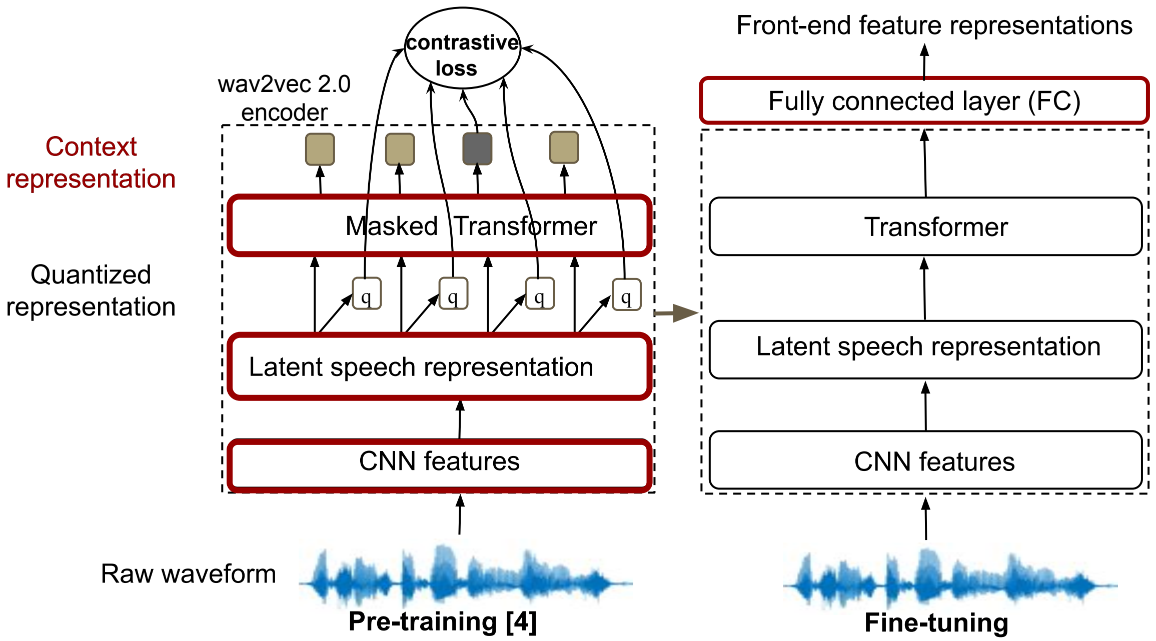
This work combines both techniques to improve generalizability of CMs

Proposed framework



[3] J. Jung, H. Heo, H. Tak et al., “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks,” in Proc. ICASSP, 2022.
[4] A. Babu, C. Wang, et al., “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” arXiv preprint arXiv:2111.09296, 2021.c

Wav2vec 2.0 (XLSR) Model



Fine-tuning: Add a simple linear layer on top of the transformer layer and jointly optimize using weighted cross entropy loss with a lower learning rate.

[4] A. Babu, C. Wang, et al., “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” arXiv preprint arXiv:2111.09296, 2021.

Experimental setup

Datasets:

- ASVspoof 2019 (training set) [6] for fine-tuning
- ASVspoof 2021 (evaluation set) [7]
  - Logical access (LA)
  - Speech Deepfake detection (DF)

Metrics: Min t-DCF [8] & Equal Error Rate

Baseline: An integrated spectro-temporal graph attention network (**AASIST**) [3].

RawBoost Data augmentation [9] applied ***on-the-fly*** to existing training database.

[3] J. Jung, H. Heo, H. Tak et al., "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," in Proc. ICASSP, 2022.
[6] X. Wang, J. Yamagishi et al., "ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech", in CSL, 2020.
[7] J. Yamagishi , X. Wang et al., "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection", in ASVspoof 2021 workshop, 2021.
[8] T. Kinnunen, H. Delgado et al., "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," in IEEE/ACM TASLP, vol. 28, 2020.
[9] H. Tak, M. Kamble, et al., "RawBoost: A Raw Data Boosting and Augmentation Method applied to Automatic Speaker Verification Anti-Spoofing," in Proc. ICASSP, 2022.

RawBoost data augmentation

- To introduce extrinsic variability stemming from, e.g., encoding, transmission effects and compression effects into training data.

- **Rawboost** [9] processes:
    1. Linear and non-linear convolutive noise
    2. Impulsive signal-dependent additive noise
    3. Stationary signal-independent additive noise

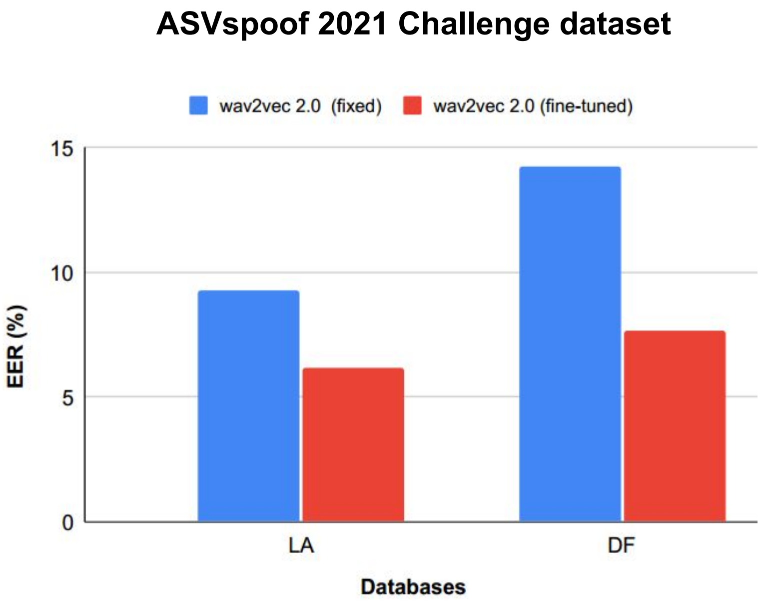
Source code:

RawBoost: https://github.com/TakHemlata/RawBoost-antispoofing

[9] H. Tak, M. Kamble, et al., “RawBoost: A Raw Data Boosting and Augmentation Method applied to Automatic Speaker Verification Anti-Spoofing,” in Proc. ICASSP, 2022.

Benefits of fine-tuning SSL front-end

- Fine-tune SSL front-end on the **ASVspoof 2019 LA** train data to achieved better performance [10].



[10] X. Wang and J. Yamagishi, “Investigating self-supervised front ends for speech spoofing countermeasures,” in Proc. Speaker Odyssey, 2022 (to appear).

Results

ASVspoof 2021 LA evaluation set

front-end	SA	DA	Pooled EER	Pooled min t-DCF
sinc-layer	×	×	11.47 (11.95)	0.5081 (0.5139)
wav2vec 2.0	×	×	6.15 (6.46)	0.3577 (0.3587)
sinc-layer	✓	×	8.73 (11.61)	0.4285 (0.5203)
wav2vec 2.0	✓	×	4.48 (6.15)	0.3094 (0.3482)
sinc-layer	✓	✓	7.65 (7.87)	0.3894 (0.3960)
wav2vec 2.0	✓	✓	**0.82 (1.00)**	**0.2066 (0.2120)**

~90%
relative improvement

ASVspoof 2021 DF evaluation set

front-end	SA	DA	Pooled EER
sinc-layer	×	×	21.06 (22.11)
wav2vec 2.0	×	×	7.69 (9.48)
sinc-layer	✓	×	23.22 (25.08)
wav2vec 2.0	✓	×	4.57 (7.70)
sinc-layer	✓	✓	24.42 (25.38)
wav2vec 2.0	✓	✓	**2.85 (3.69)**

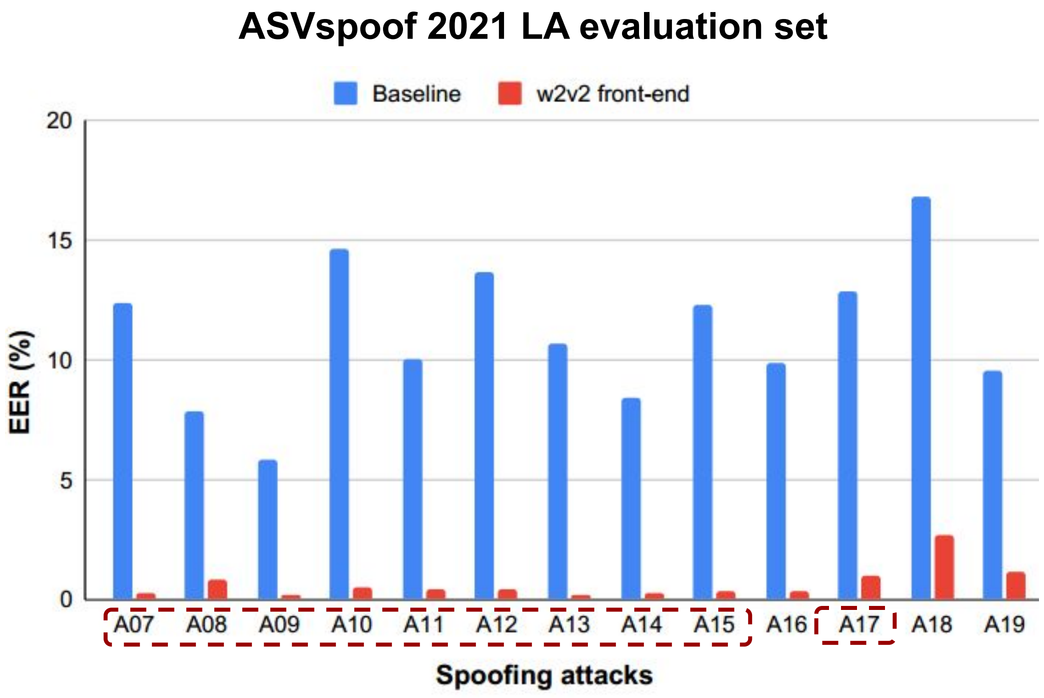
~88%
relative improvement

- Best single system results on ASVspoof 2021 challenge LA and DF task till date.
- Improve robustness in detecting previously unseen spoofing attacks (**100+ spoofing attacks** in DF database).

[11] C. Veaux, J. Yamagishi, et al., "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017, http://dx.doi.org/10.7488/ds/1994.
[12] Y. Zhao, W.-C. Huang,et al., "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion", Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge, pp. 80--98, 2020.

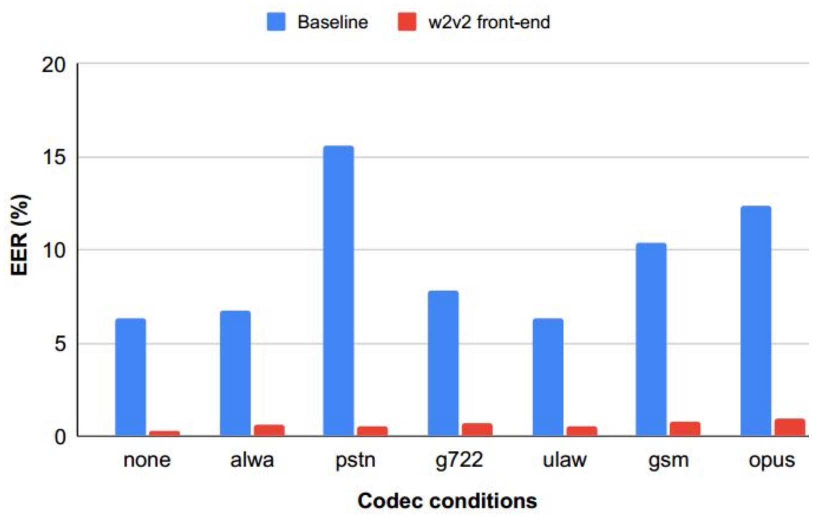
Results for each spoofing attack

- Improve generalisation towards unseen spoofing attacks.

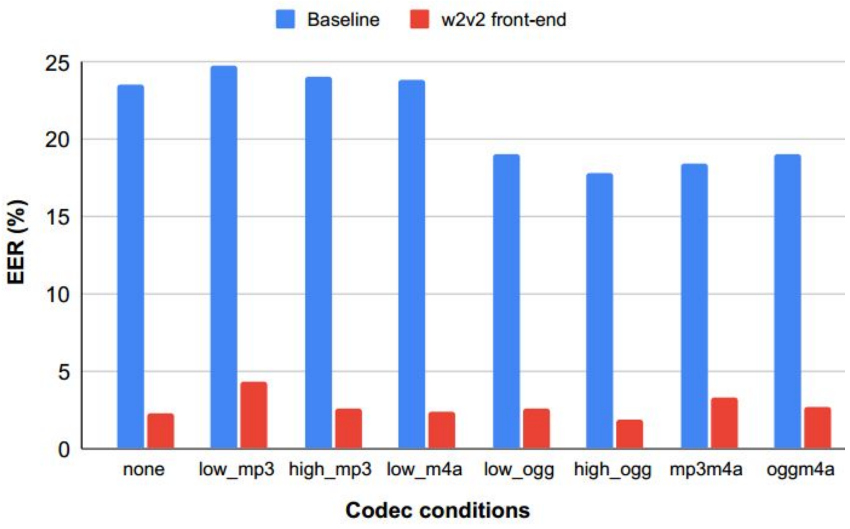


Results for each codec condition

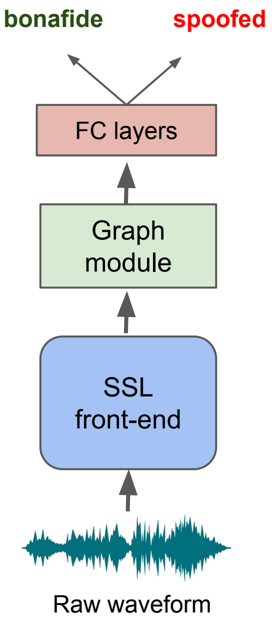
ASVspoof 2021 LA evaluation set



ASVspoof 2021 DF evaluation set



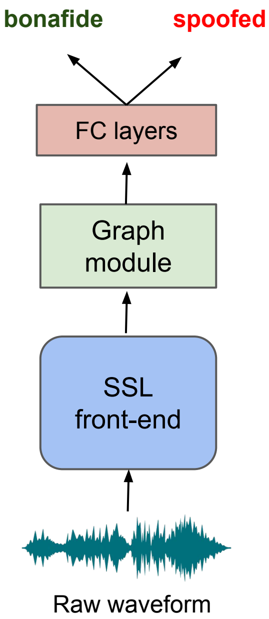
Results using simpler CM



Train on ASVspoof 2019 LA training dataset

Front-end	DA	Test Database	Pooled EER (%)	Pooled min t-DCF
Wav2vec 2.0 **(fixed)**	Yes	ASVspoof 2021 LA	2.26	0.2407
Wav2vec 2.0 **(fixed)**	Yes	ASVspoof 2021 DF	7.28	-
Wav2vec 2.0 **(fine-tuned)**	Yes	ASVspoof 2021 LA	1.19	0.2175
Wav2vec 2.0 **(fine-tuned)**	Yes	ASVspoof 2021 DF	4.38	-

Results using simpler CM



Train on ASVspoof 2019 LA training dataset

Front-end	DA	Test Database	Pooled EER (%)	Pooled min t-DCF
Wav2vec 2.0 **(fixed)**	Yes	ASVspoof 2021 LA	2.26	0.2407
Wav2vec 2.0 **(fixed)**	Yes	ASVspoof 2021 DF	7.28	-
Wav2vec 2.0 **(fine-tuned)**	Yes	ASVspoof 2021 LA	1.19	0.2175
Wav2vec 2.0 **(fine-tuned)**	Yes	ASVspoof 2021 DF	4.38	-

Conclusions

- Our results are **best single system results** reported so far on **ASVspoof 2021 LA and DF** database.
- SSL front-end improves **domain robustness** in detecting previously **unseen spoofing attacks.**
- Our system achieved **~90% and ~88% relative improvement** over **baseline system** for LA and DF database.
- **RawBoost (DA)** further improves robustness in more realistic challenging environments:
    - telephony and audio codec
    - compression effect in DF dataset

Source code:
https://github.com/TakHemlata/SSL_Anti-spoofing


ODYSSEY
June 28th-July 1st 2022, Beijing, China
THE SPEAKER AND LANGUAGE
RECOGNITION WORKSHOP

Thank you

Motivation

Possible directions:

- Use larger and diverse representative training database.
    - Advantage: better generalisation
    - Disadvantage: To collect sufficient amount of spoof data requires more efforts and technically demanding.
- Data augmentation
    - To further enhance the performance in challenging environments such as telephonic, audio codec and compression.

Potential solution: Can we use **self-supervised models** trained on large diverse datasets to learn more generalised representation?

This work combines both techniques to improve generalizability of CMs

Self-attentive aggregation (SA) layer

- Explore self-attentive aggregation layer [5] to extract the discriminative feature representations.

$$\mathbf{S} \in \mathbb{R}^{C\times F\times T}$$

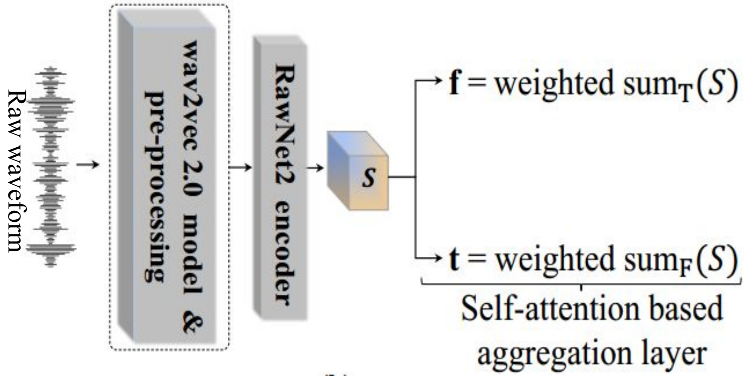
$$W_{F\times T} = \mathrm{Softmax}(\mathrm{conv2d}(\mathrm{BN}(\mathrm{SeLU}(\mathrm{conv2d}(\mathbf{S})))))$$

spectral feature $\mathbf{f}_{C\times F} = \sum_{T} \mathbf{S}W$

weighted sum along the time

temporal feature $\mathbf{t}_{C\times T} = \sum_{F} \mathbf{S}W$

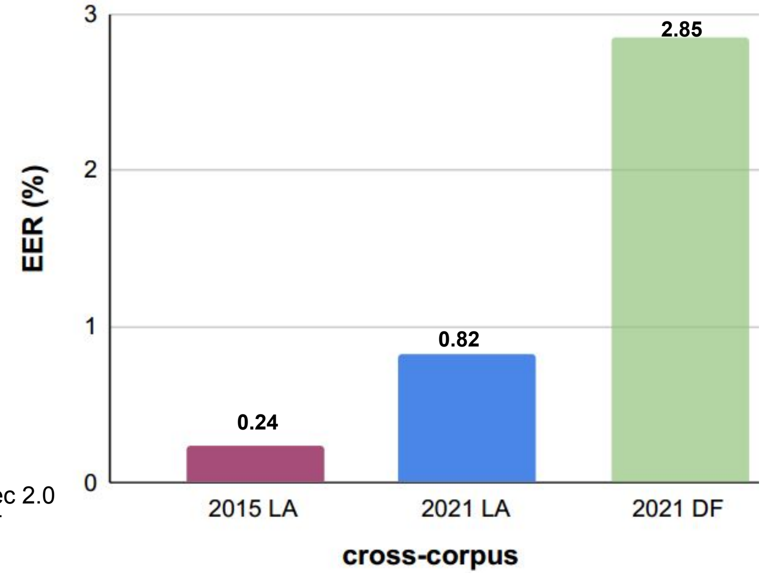
weighted sum along the Freq.



[5] K. Okabe, T. Koshinaka, et al., “Attentive statistics pooling for deep speaker embedding,” in Proc. INTERSPEECH, 2018.

Cross database evaluations

- CM system is trained on **ASVspoof 2019 LA training dataset** and tested cross databases.



Front-end: wav2vec 2.0
Back-end: AASIST

Motivation

Possible directions:

- Use larger and diverse representative training database.
  - Advantage: better generalisation
  - Disadvantage: It’s impractical- never enough

- Data augmentation
  - Further enhance the performance in challenging environments.

Potential solution: Can we use **self-supervised models** trained on large diverse datasets to learn more generalised representation?

This work combines both techniques to improve generalizability of CMs

Self-attentive aggregation (SA) layer

- Explore self-attentive aggregation layer [5] to extract the discriminative feature representations.
- Aggregate the information using learnable weight along the spectral and temporal domains.

$$\mathbf{S} \in \mathbb{R}^{C \times F \times T}$$

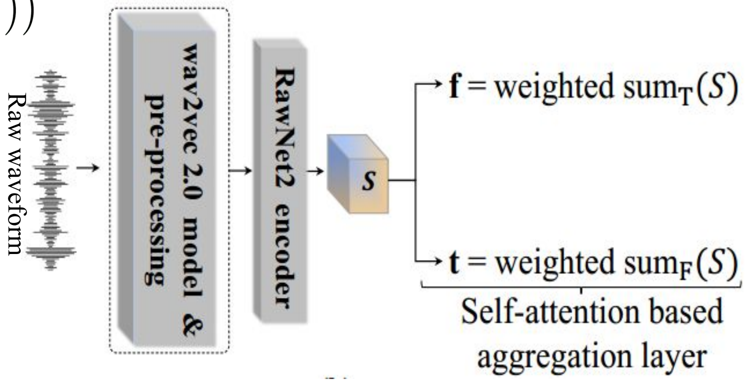
$$W_{F \times T} = \mathrm{Softmax}(\mathrm{conv2d}(\mathrm{BN}(\mathrm{SeLU}(\mathrm{conv2d}(\mathbf{S})))))$$

spectral feature $\mathbf{f}_{C \times F} = \sum_{T} \mathbf{S}W$

weighted sum along the time

temporal feature $\mathbf{t}_{C \times T} = \sum_{F} \mathbf{S}W$

weighted sum along the Freq.



[5] K. Okabe, T. Koshinaka, et al., “Attentive statistics pooling for deep speaker embedding,” in Proc. INTERSPEECH, 2018.

RawBoost data augmentation

- To introduce extrinsic variability stemming from, e.g., encoding, transmission effects and compression effects into training data.

- **Rawboost** [9] processes:
    1. Linear and non-linear convolutive noise
    2. Impulsive signal-dependent additive noise
    3. Stationary signal-independent additive noise

1–2 → **ASVspoof 2021 LA database**
Suitable for telephony application
(encoding & transmission)

3 → **ASVspoof 2021 DF database**
(mp3 compression method)

[9] H. Tak, M. Kamble, et al., “RawBoost: A Raw Data Boosting and Augmentation Method applied to Automatic Speaker Verification Anti-Spoofing,” in Proc. ICASSP, 2022.