Speaker Verification: Pre-trained model, Attention Augmented, and Contrastive Learning

LI Zhe

Department of Electronic and Information Engineering

Hong Kong Polytechnic University

lizhe.li@connect.polyu.hk

27/06/2022

Biography

LI Zhe

- A first-year PhD student, The Hong Kong Polytechnic University
- PhD Supervisor Prof. MAK Man-Wai
- Research Interests： Roubst Speaker Verification & Diarization, Multimodal Speaker Recognition
- Homepage： http://lizhe.link
- Contact： **lizhe.li@connect.polyu.hk**

X-vector

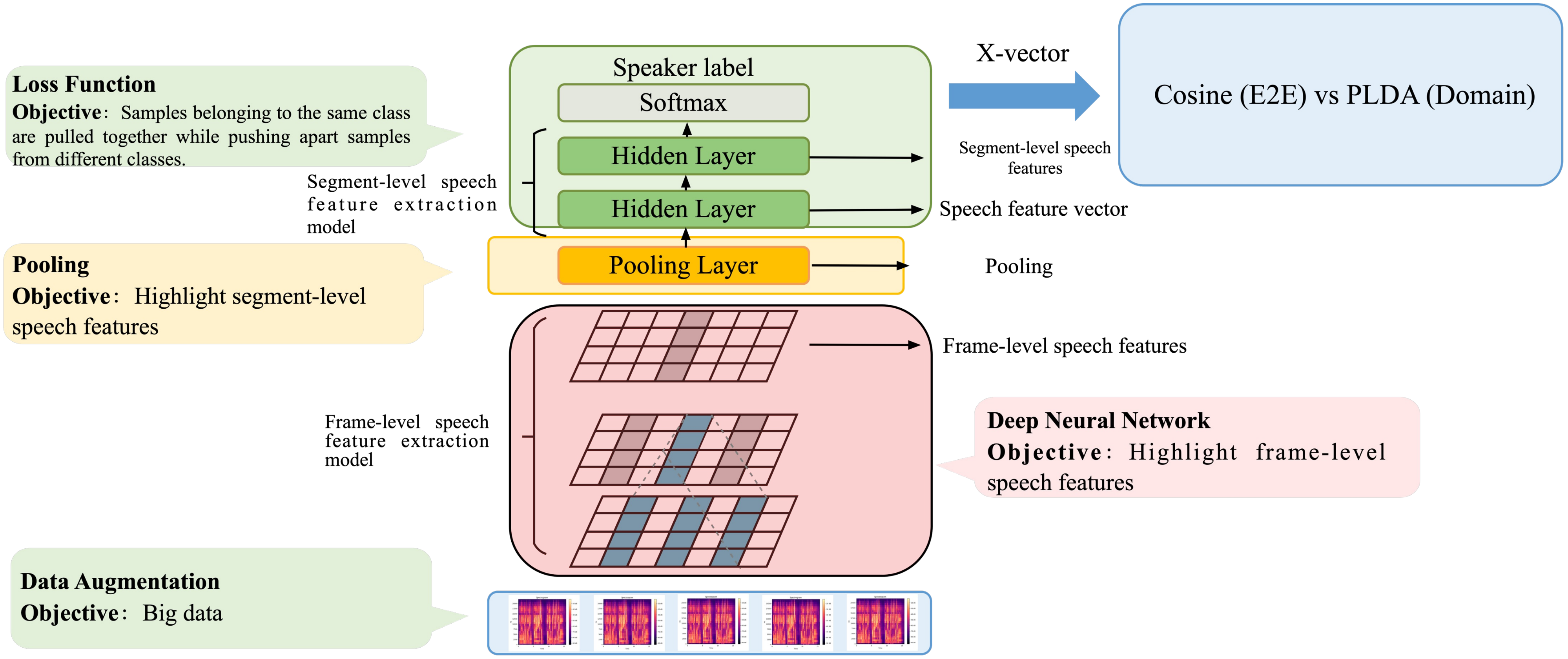

Figure 1. X-vector architecture.

The material in this slide is extracted from the presentation of Prof. HE Liang in The Symposium on Speaker Recognition Reasearch and Application 2021

Pre-Trained model

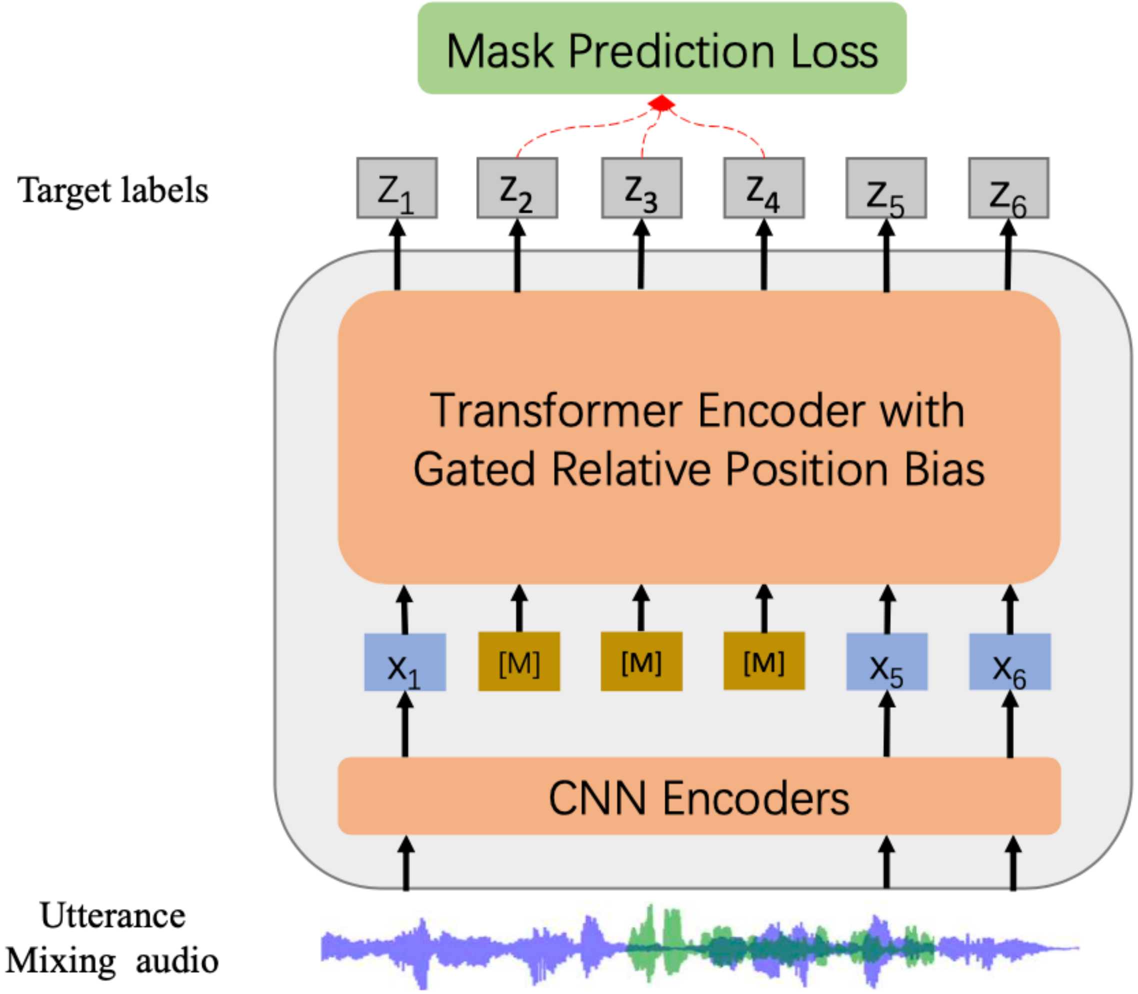

Figure 1. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing [1].

- Highlights:
  - Self-supervised learning
  - Denoising Masked Speech Modeling
  - Large-scale training data

- Fine-tune:
  - Fine-tuning Strategies
  - Further Pre-training

Experimental Results

Preliminary Results on Voxceleb1

Feature	EER (%)		
	Vox1-O	Vox1-E	Vox1-H
ECAPA-TDNN [2]	1.010	1.240	2.320
HuBERT Base	0.989	0.822	1.678
WavLM Baese+	0.840	0.928	1.758
HuBERT Large *	0.585	0.654	1.342
WavLM Large*	**0.383**	**0.480**	**0.986**

Preliminary Results on CN-Celeb

Feature	EER (%)	**minDCF(%)**
Fbank + ECAPA-TDNN	8.7920	0.4976
WavLM Large + ECAPA-TDNN	**8.3980**	**0.4762**

Attention Augmented

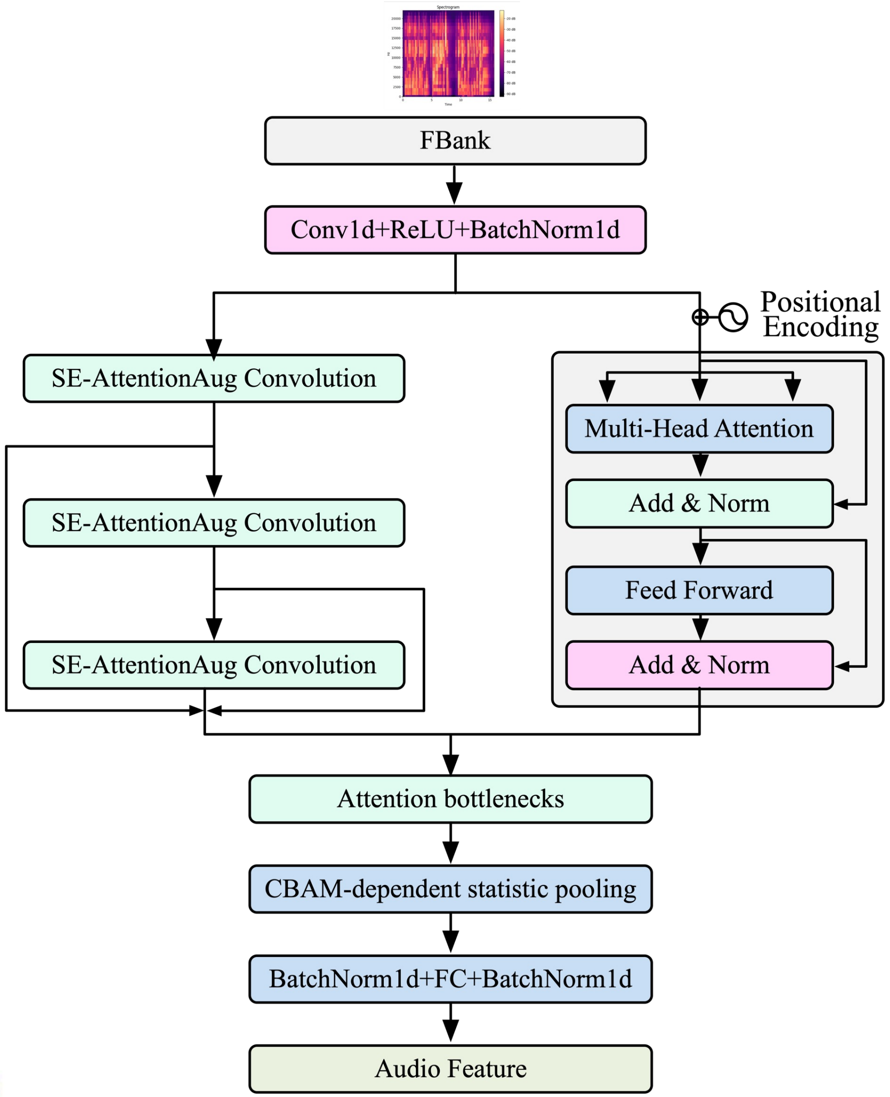

Figure 2. High-level illustration of Our proposed model.

- Contributions:
  - Multi-task and multi-scale feature extraction
  - Multi-layer feature aggregation and summation
  - CBAM-dependent statistics pooling

Attention Augmented

- CBAM: Convolutional Block Attention Module

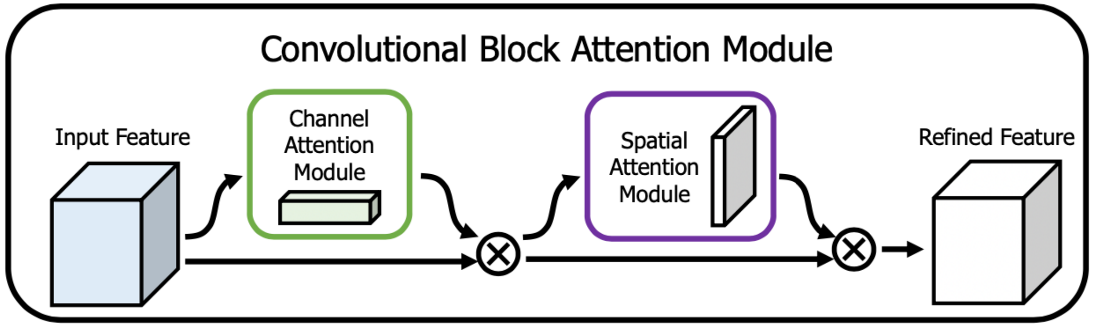

Figure 3. The overview of CBAM. The module has two sequential sub-modules: channel and spatial [3].

- Attention Augmented Convolution

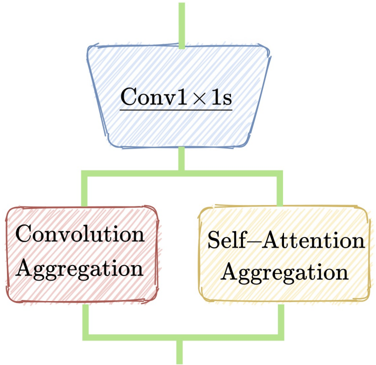

Figure 4. Attention Augmented Convolution [4].

Contrastive learning

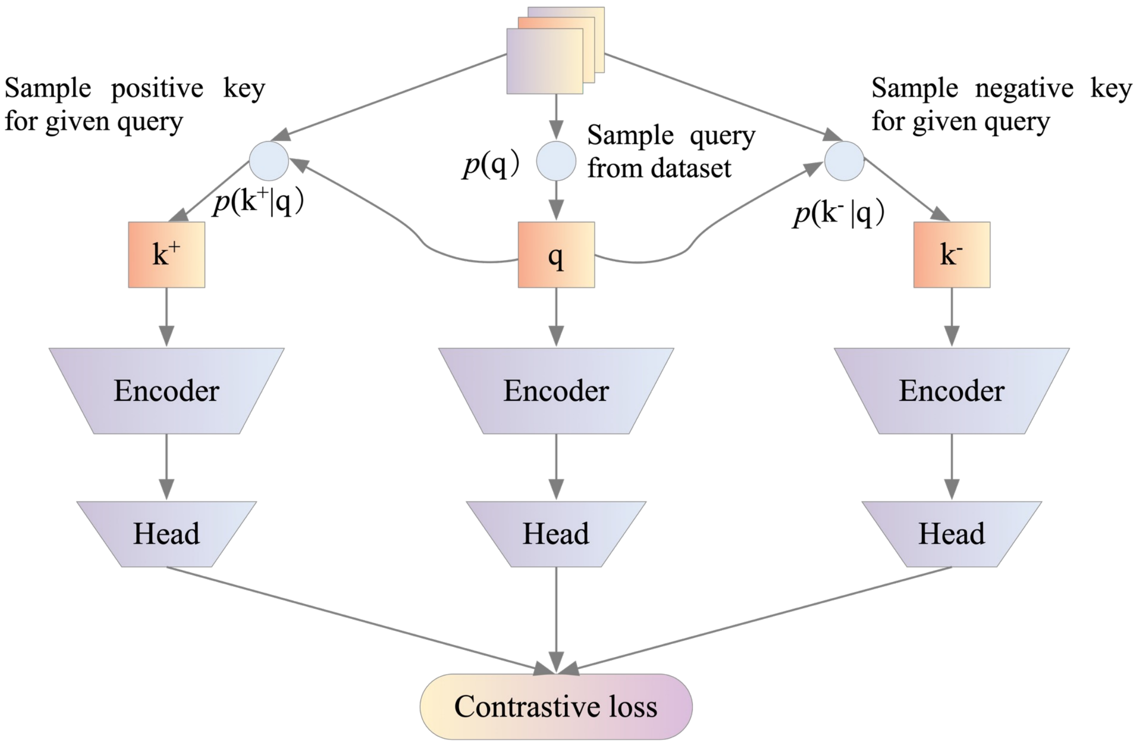

Figure 1. High-level illustration of contrastive learning [5].

- **Objective**: Clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes [6-9].

- Sampled Data q (query)
- Data for comparison k (key)
  - (k^+,q), positive pair
  - (k^-,q), negative pair

Contrastive learning

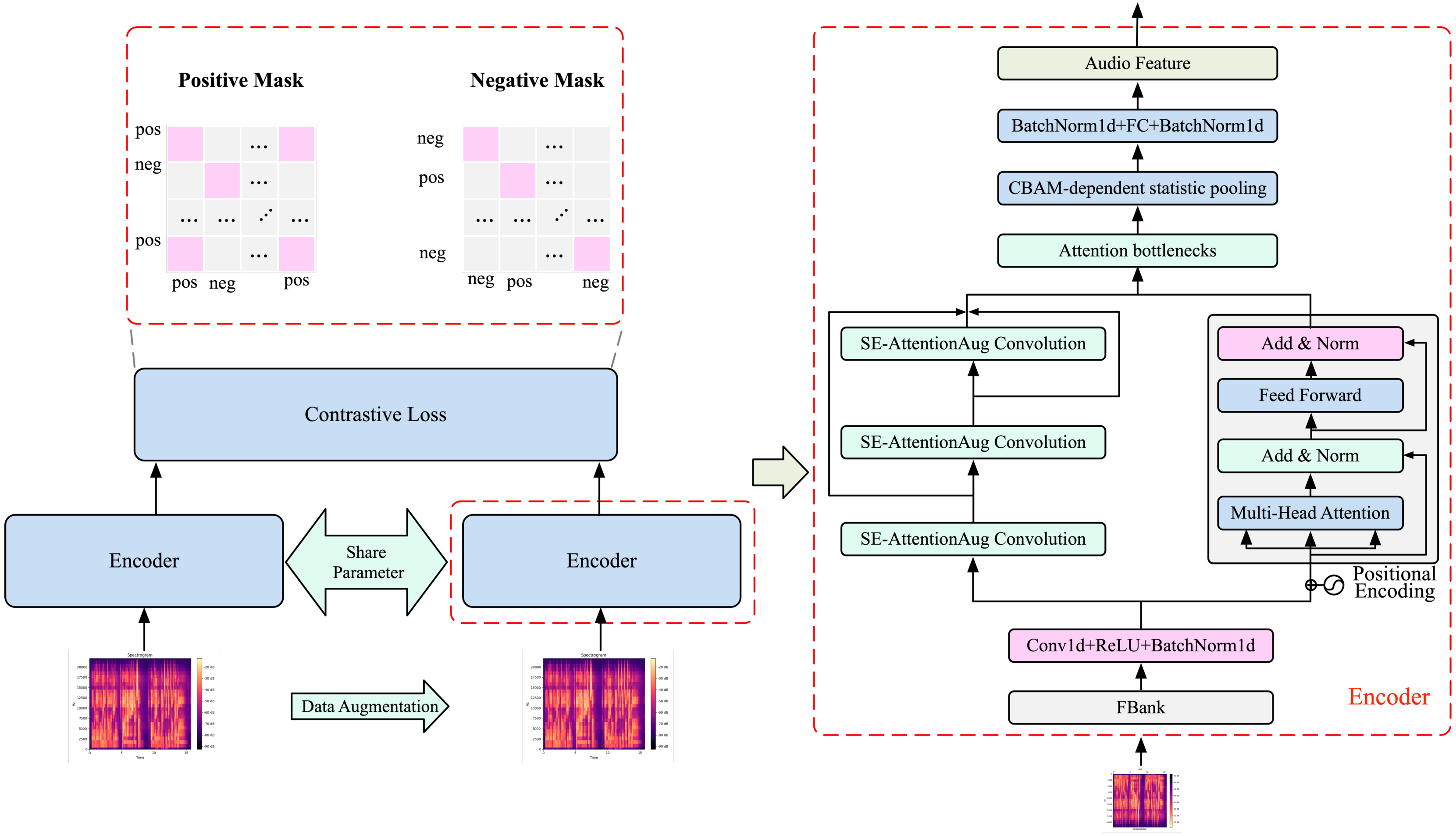

Figure 2. High-level illustration of contrastive learning for speaker verification.

- Motivations:
  - Robust representation
  - Effectively leverage label information
  - Positive samples will be more closely aligned

Contrastive Loss

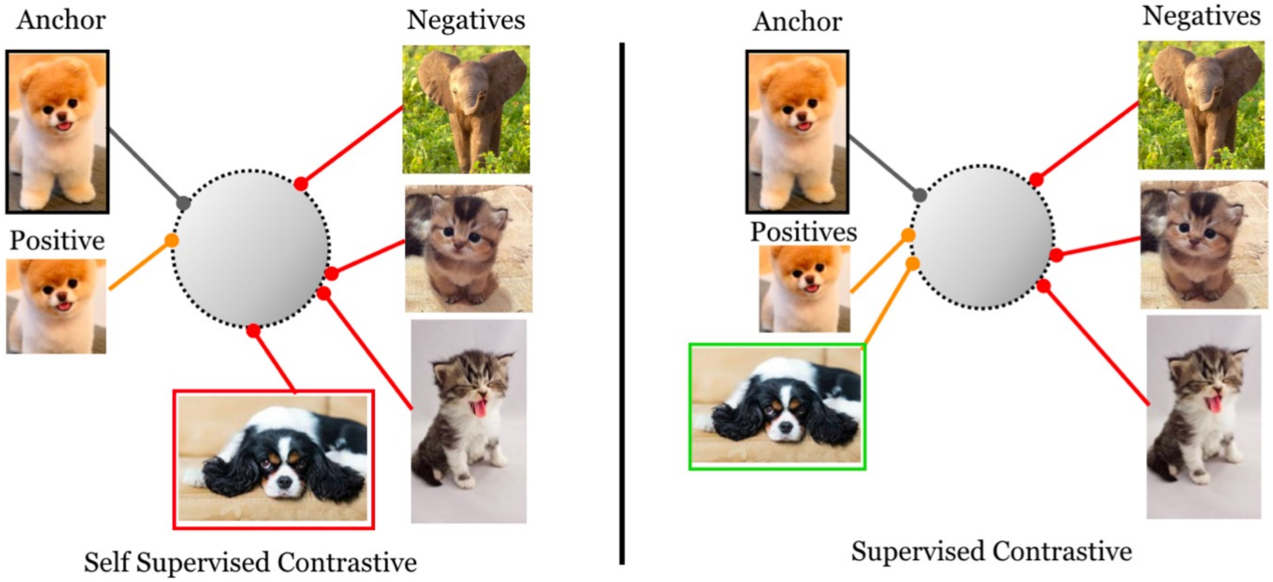

Figure 3. Self-supervised contrastive loss vs supervised contrastive loss Loss [10].

- Self-Supervised Contrastive Loss:

$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_{j(i)}/\tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau\right)}$$

- Supervised Contrastive Loss:

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p/\tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau\right)}$$

- $\boldsymbol{z_i}$ is anchor. $\boldsymbol{z_p}, \boldsymbol{z_{j(i)}}$ augmented data, $\boldsymbol{z_a}$ is negative samples, $\boldsymbol{P(i)}$ is the set of positive data, $\boldsymbol{A(i)}$ is the set of negative data.

Summary

- **Data** Cover as much data as possible and use data augmentation strategies.
- **Model** Robust speech features representations from different aspects.
- **Pooling** Highlight segment-level speech features
- **Loss** Clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes.
- **Verification** Cosine (E2E) vs PLDA (Domain), adversarial domain mismatch, score calibration

References

[1] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2021). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. arXiv preprint arXiv:2110.13900.

[2] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification}}. Proc. Interspeech 2020, 3830-3834.

[3] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[4] Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., & Huang, G. (2022). On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 815-825).

[5] Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. IEEE Access, 8, 193907-193934.

[6] Tao, R., Lee, K. A., Das, R. K., Hautamäki, V., & Li, H. (2022, May). Self-supervised speaker recognition with loss-gated learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6142-6146). IEEE.

[7] Cai, D., Wang, W., & Li, M. (2022). Incorporating Visual Information in Audio Based Self-Supervised Speaker Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1422-1435

[8] Xia, W., Zhang, C., Weng, C., Yu, M., & Yu, D. (2021, June). Self-supervised text-independent speaker verification using prototypical momentum contrastive learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6723-6727). IEEE

[9] Tang, Y., Wang, J., Qu, X., & Xiao, J. (2021, July). Contrastive learning for improving end-to-end speaker verification. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

[10] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 18661-18673.

Try to enjoy your research!

Try to do meaningful research forever!

You are more than what you have become!

PolyU