Self supervised learning for speaker verification

Victor Miara

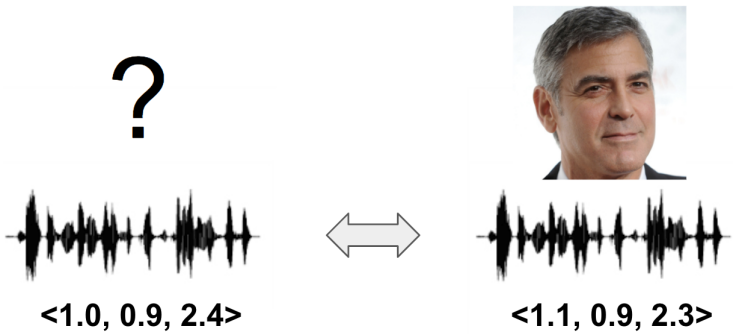
EPITA Research Laboratory
Supervised by Reda Dehak
Following the work of Théo Lepage

Seminaire – January 2024

Speaker verification

Goal:

Verify that an audio **utterance** corresponds to the **identity claimed** by the speaker.

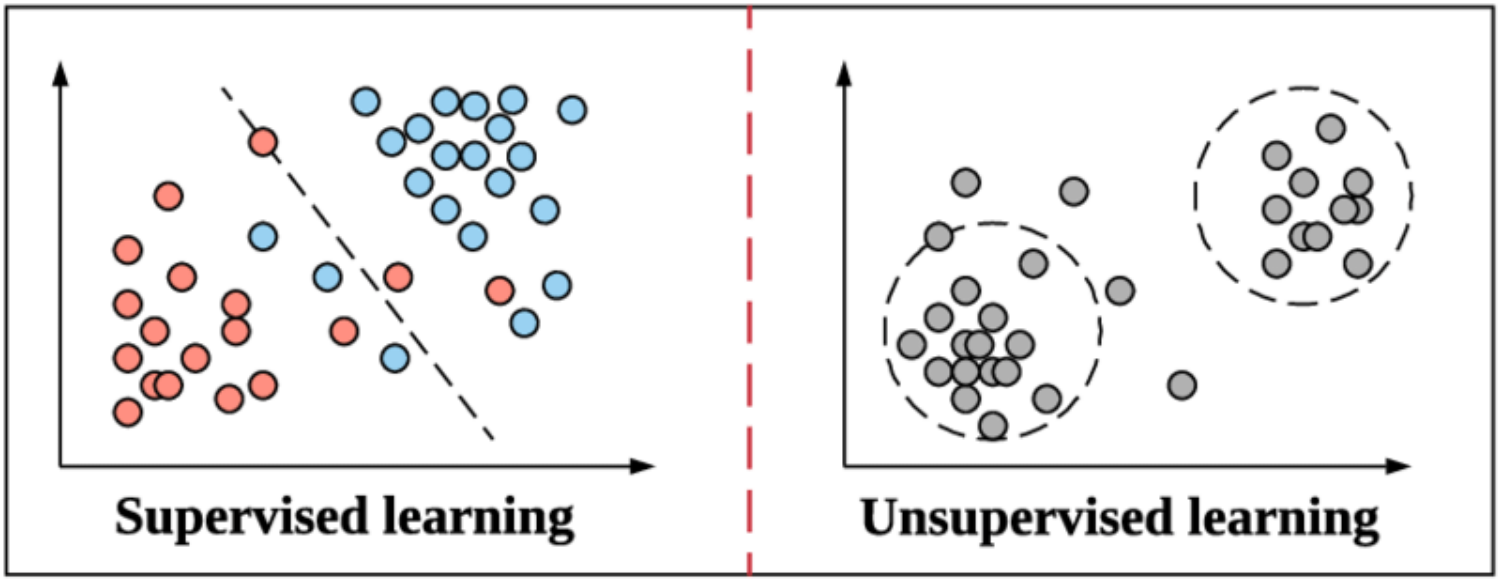


Self supervised learning

Being label-dependant is very constraining

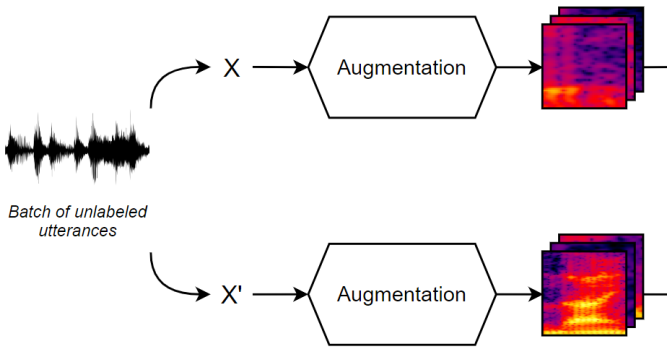
$\Longrightarrow$ What is **self supervised learning**?

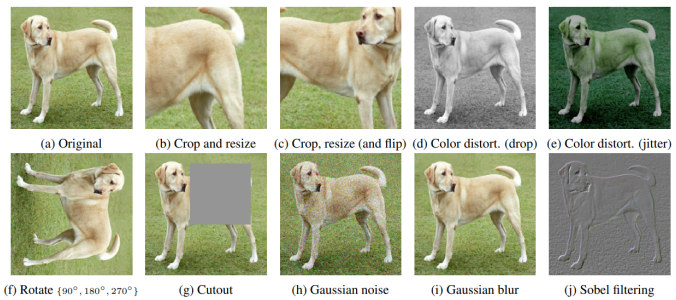
$\Longrightarrow$ How to use **SSL** for **speaker verification**?

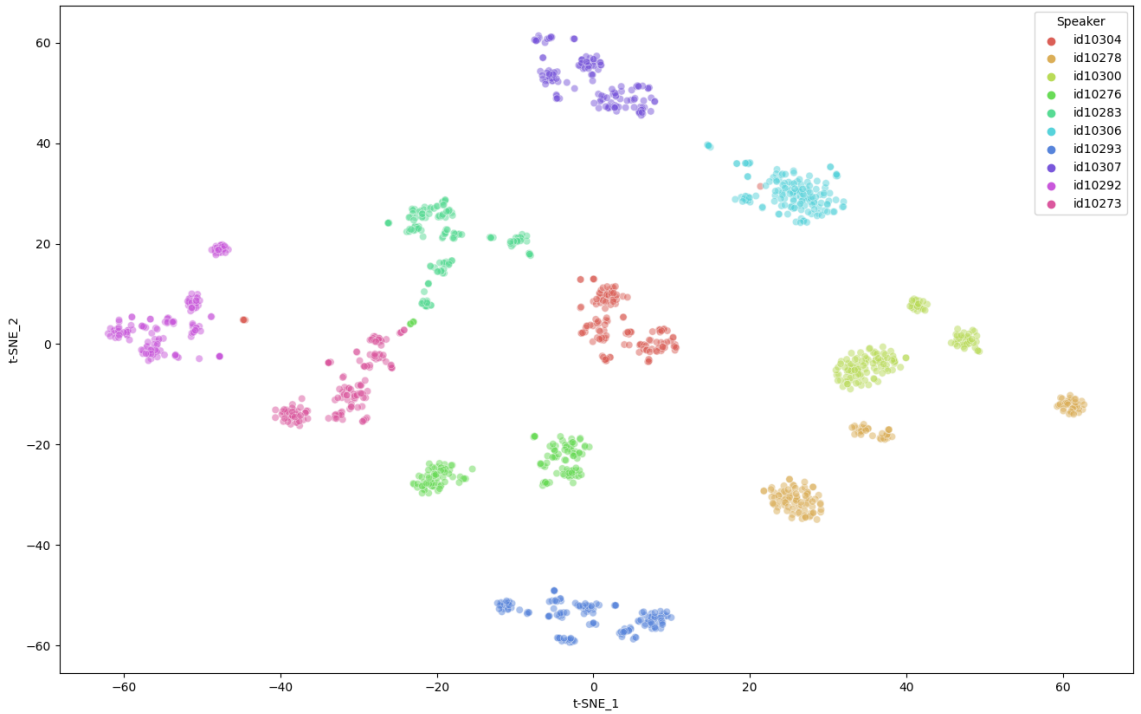


Self supervised method

- Generate two versions of the same audio
- Minimize the distance between their **latent representations**







Learn representations with contrastive learning

Assumption: Each utterance in the mini-batch belong to a unique speaker.

Objective: Learn embeddings that have small intra-speaker and large inter-speaker distances.

$$\mathcal{L}_{InfoNCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\cos\left(\mathbf{e}_{i,1,aug},\mathbf{e}_{i,2,aug}\right)}}{\sum_{j=1}^{N}e^{\cos\left(\mathbf{e}_{i,1,aug},\mathbf{e}_{j,2,aug}\right)}}$$

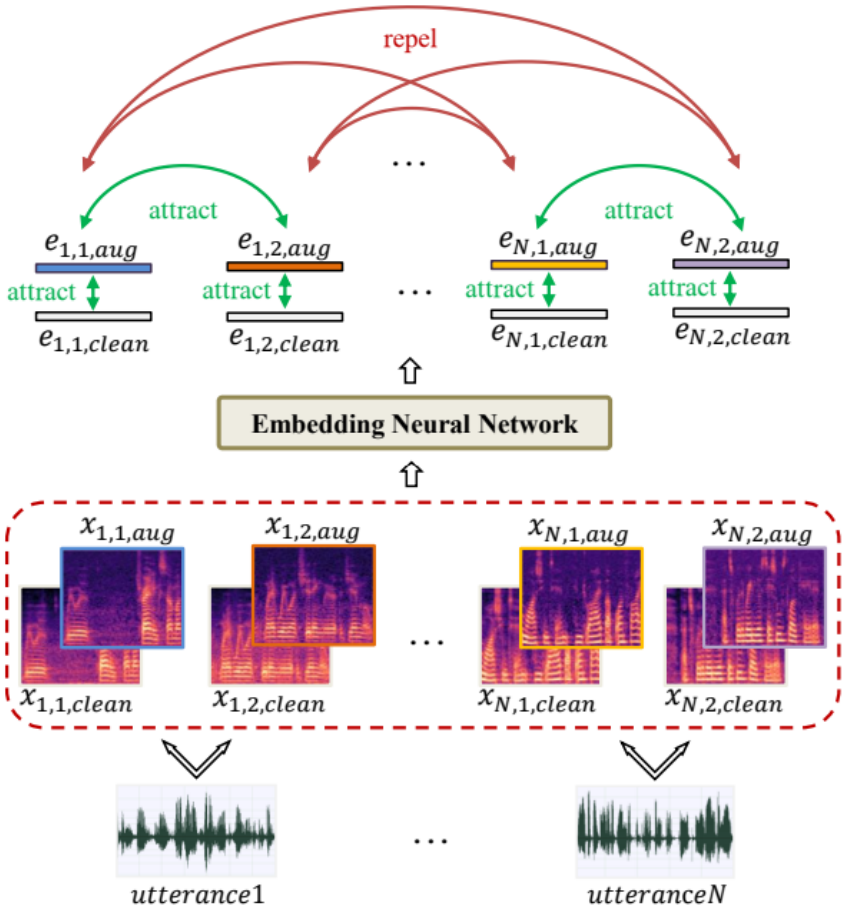
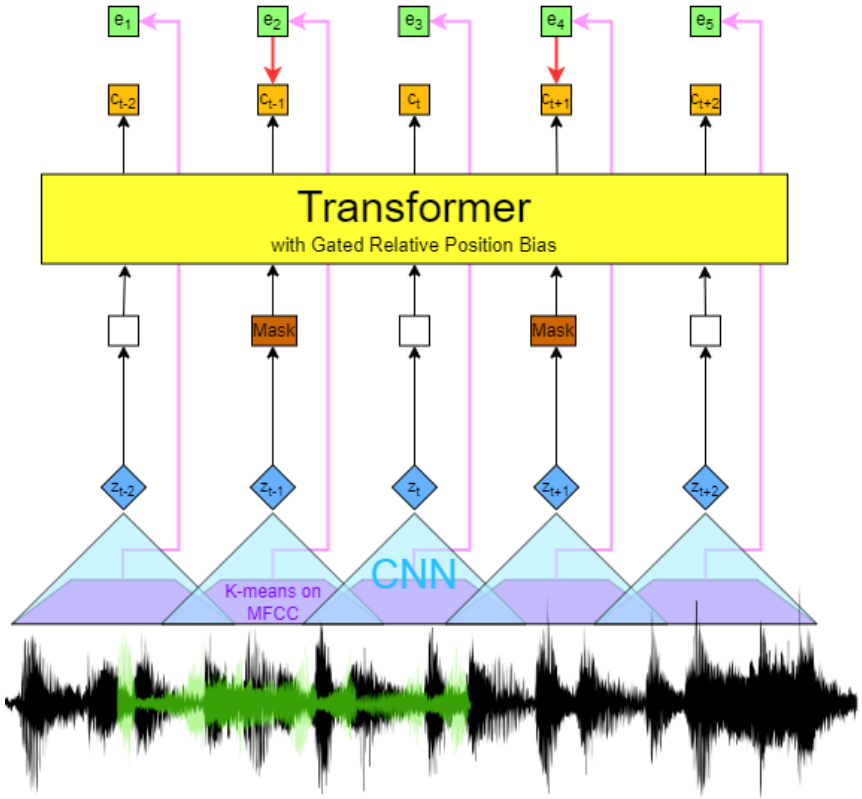


Figure: SimCLR [**?**] model architecture.

WavLM [Sanyuan Chen, 2021]

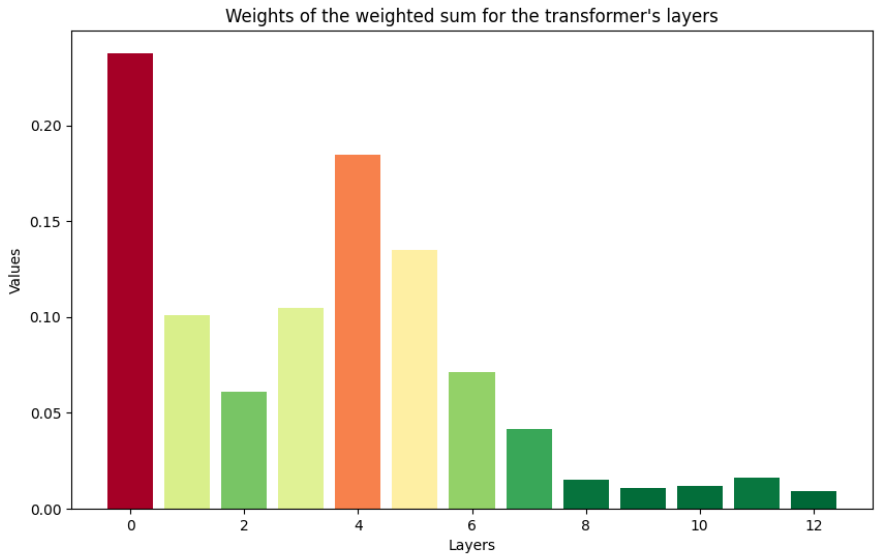
- **Same architecture** than Hubert [Wei-Ning Hsu, 2021] and Wav2Vec2
- Improved **data augmentation**
  - Mixing audio with different utterances

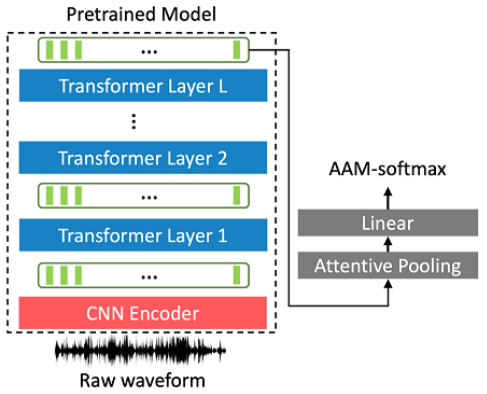


WavLM

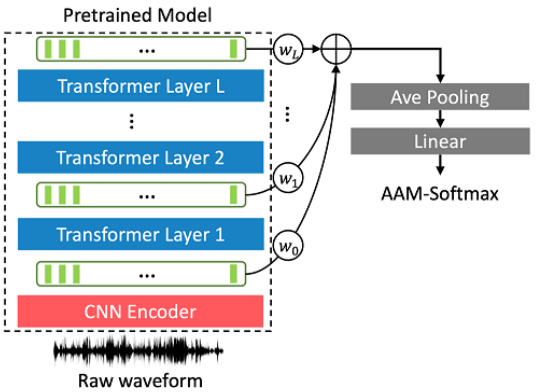
Weighted sum

- **Weighted sum** on the transformer layers
  - Learn the weights during 1st training phase
- **Last half** layers are used more for **speech information**
  - **First half** layers are used more for **speaker information**





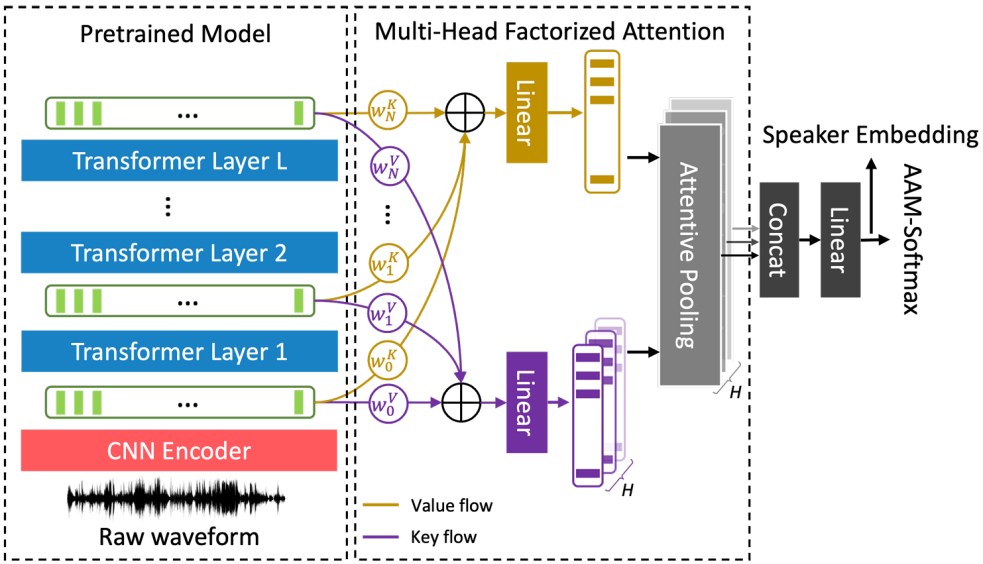
(a) Top Layer Attentive Pooling

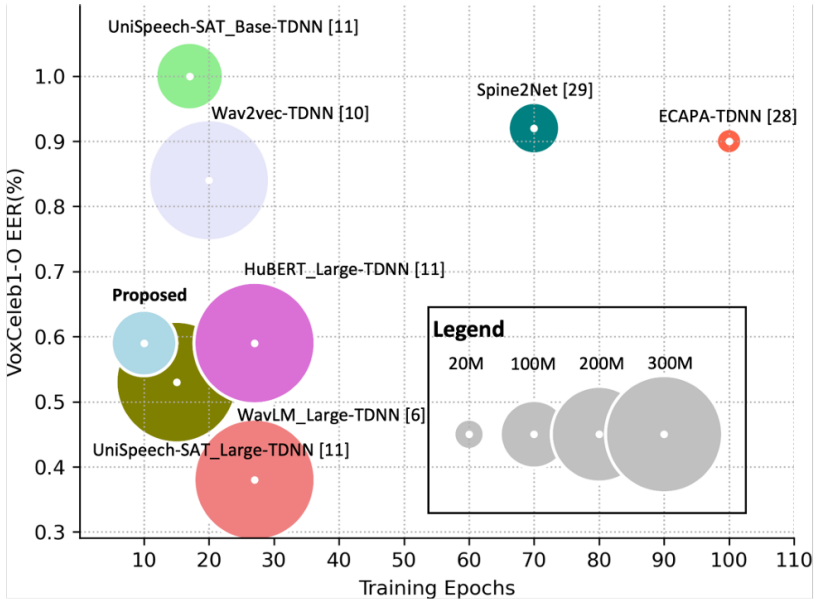


(b) Layer-wise Weighted Average Pooling

Multi Head Factorized Attentive Pooling

- Use **attention** to weight the layer outputs
- Very lightweight backend





Multi Head Factorized Attentive Pooling [Peng et al., 2023]

- Very powerful model with attention
- Focuses on easier task $\Longrightarrow$ learn the channel characteristics
- Model doesn't converge

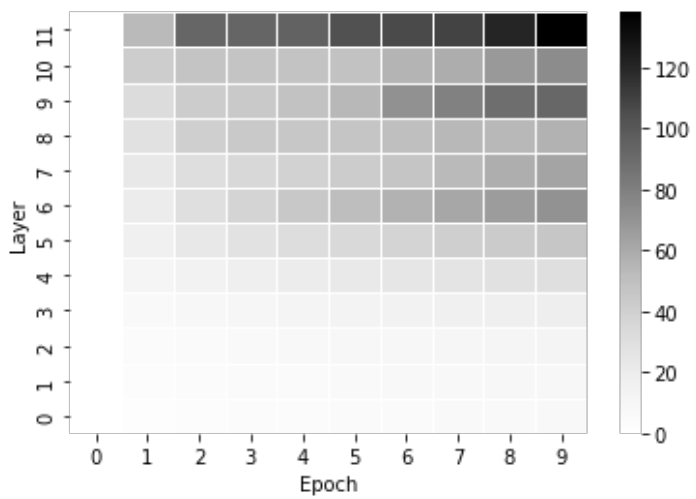


Figure: Supervised Layer Modification

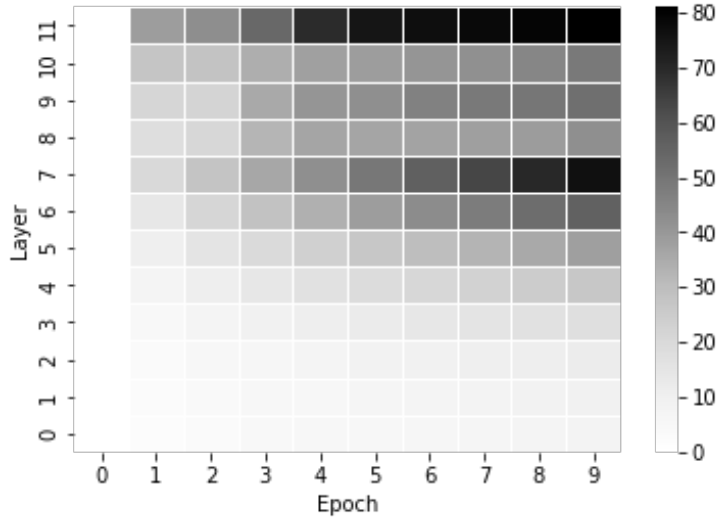
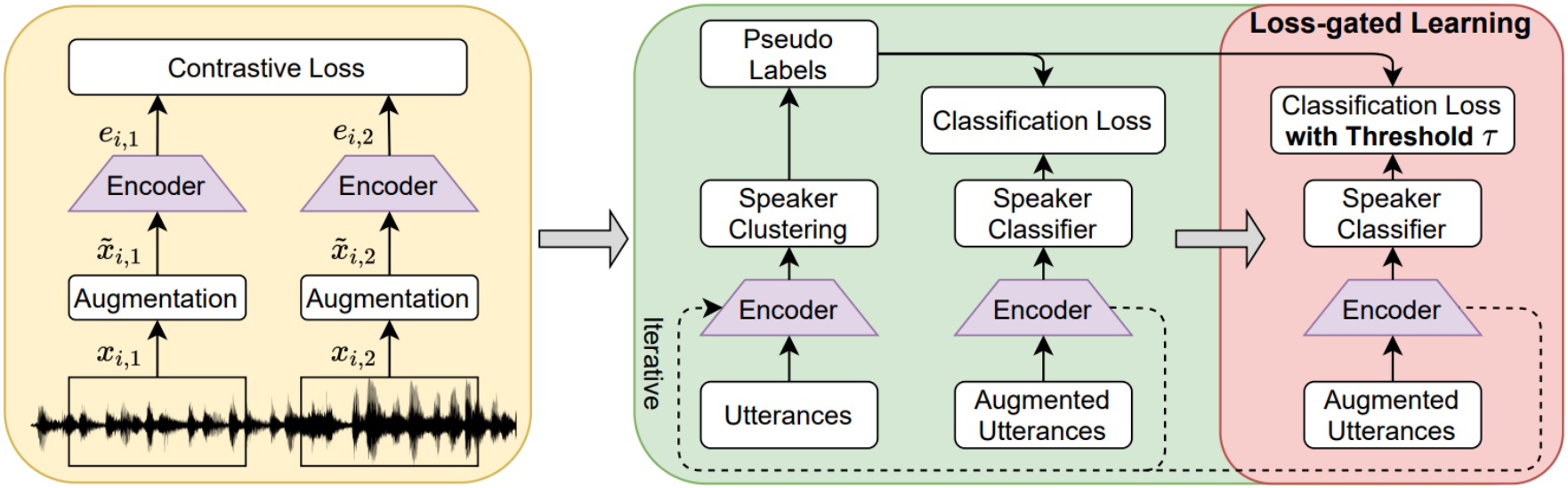


Figure: Self Supervised Layer Modification

Pseudo labels

- Extract embeddings from dataset using baseline ssl model
- Clusterize these embeddings
- Consider each cluster as one speaker id and labelise the dataset

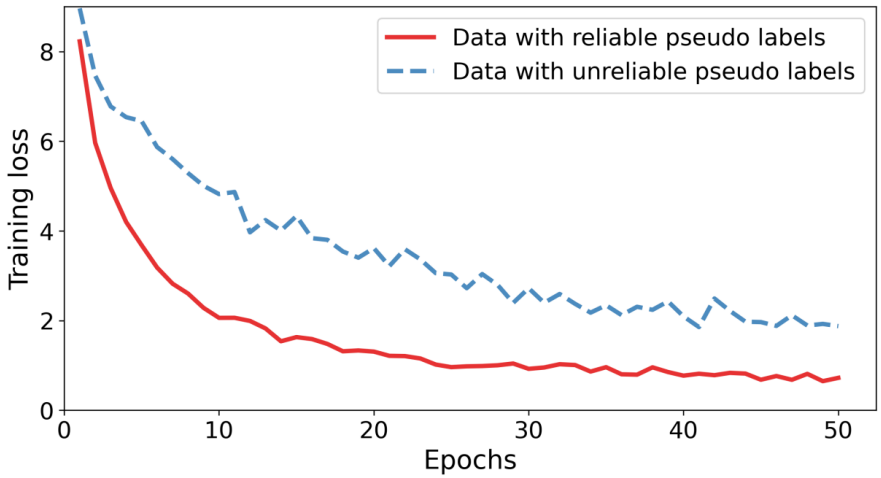


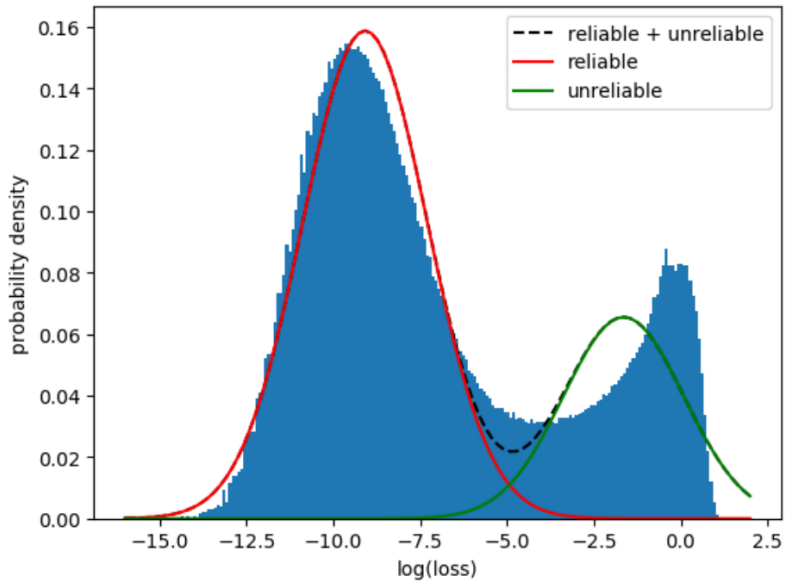
Stage I: Contrastive learning

Stage II: Iterative self-supervised learning

Loss Gate

- Higher **training loss** for unreliable pseudo labels
- Consider only samples with loss under threshold
- Threshold chosen by hand [Ruijie Tao, 2021] or dynamically [Bing Han, 2022] with a GMM





Loss Gate

- Set loss to 0 for sample with losses below threshold
- Otherwise AAM softmax

$$L_{DLG} = \sum_{i=1}^{N} \mathbb{1}_{l_i < \tau} \log \frac{e^{s(\cos(\theta_{y_i,i} + m))}}{Z}$$

Figure: Dynamic Loss Gate

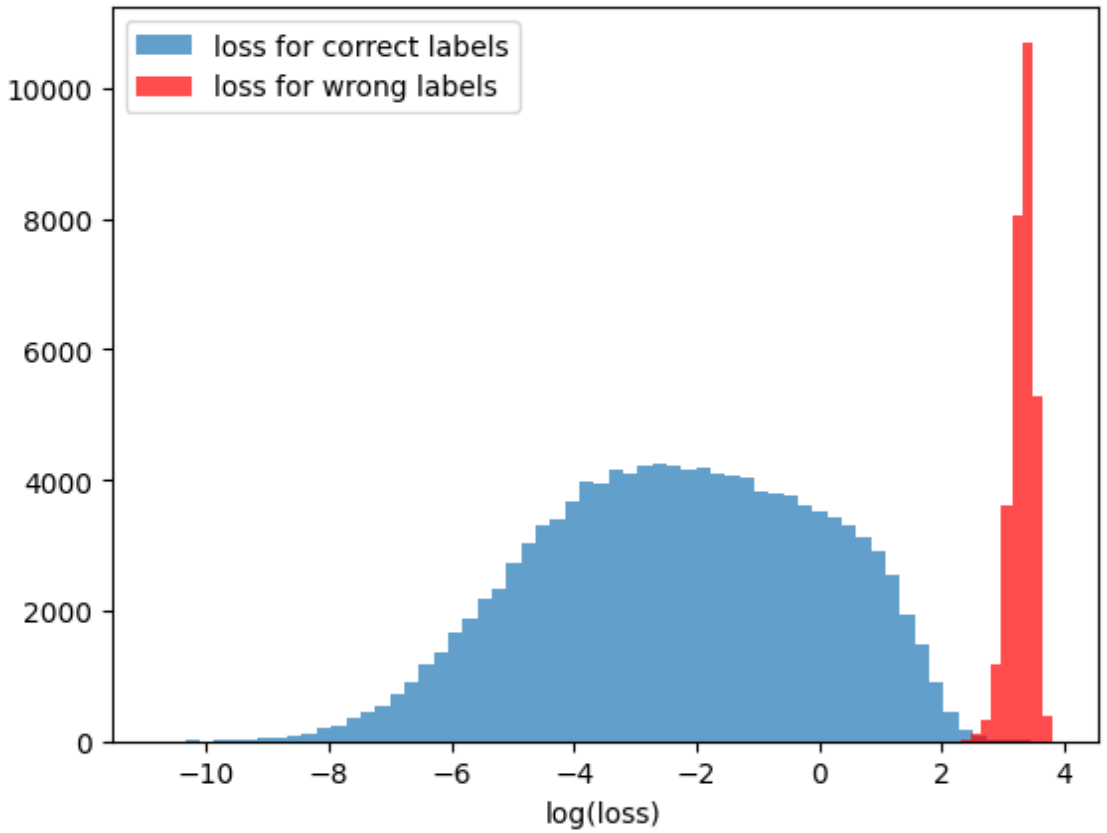
Dynamic label correction

- Make use of the non reliable labels
- Assume that the model's prediction is the true label
- Maximize the similarity of prediction between output of **clean** and **augmented** sample

$$L_{LC} = \sum_{i=1}^{N} \mathbb{1}_{l_i > \tau, \max(\hat{p_i}) > \tau_2} H(\hat{p}_i \mid p_i)$$

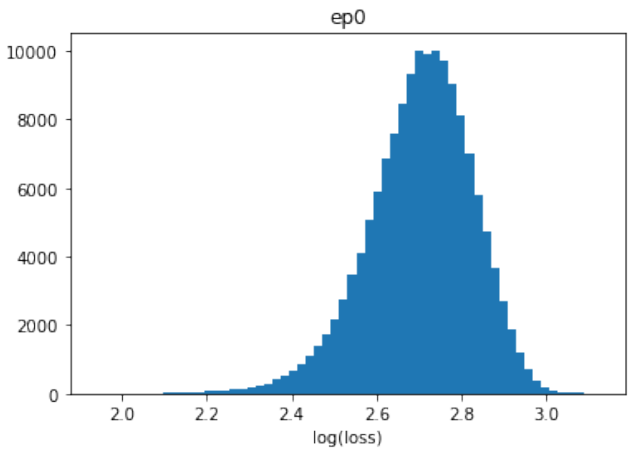
Loss distribution on simulated dataset

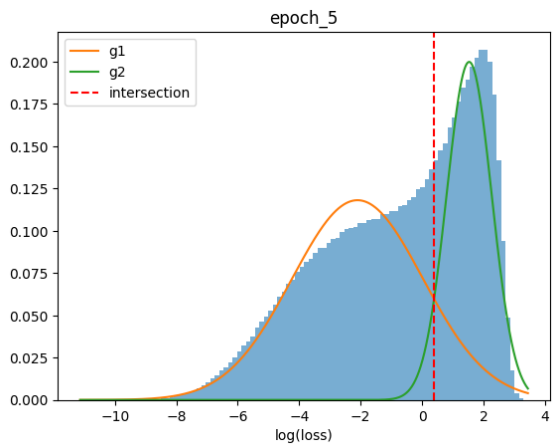
- Simulate errors in dataset
  - Assign a random label to 20% of the dataset (voxceleb1)
- Loss is very separable by a threshold

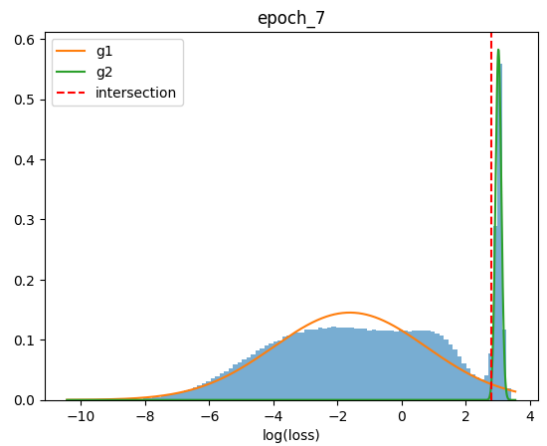


Loss distribution on real dataset

- Evolution of loss distribution with model training
- Separation by threshold is a lot less trivial
- The id for the non reliable samples is the id of a speaker that is close in the latent space







Results

Stage	Loss	Model					EER (%)	Min DCF
Supervised training from scratch	AAM Softmax	WavLM MHFA					0.76	0.05
Self supervised learning from scratch	DLG-LC	DINO					3.16	0.23
Iterative clustering	AAM Softmax (margin = 0.2)	WavLM MHFA	iter1	-	-	0.1	1.56	0.10
						0.2	1.54	0.10
			iter2	-	0.2	0.1	**1.37**	**0.09**
						0.2	1.41	0.09
					0.1	0.1	1.44	0.10
						0.2	1.51	0.10
			iter3	0.2	0.2	0.1	1.50	0.09
						0.2	1.43	0.09
					0.1	0.1	1.46	0.10
						0.5	1.42	0.11

Results

Method	Margin	Threshold	Iteration	EER (%)	Min DCF
AAMSoftmax	0.2	-	1	1.50	0.09
	0.1	-	1	1.44	0.10
AAM + LGL	0.2	1.5	1	1.35	0.09
		dynamic	1	1.27	0.08
			2	1.16	0.08
	0.1	1.5	1	1.32	0.09
		0	1	1.33	0.10
		dynamic	1	1.30	0.08
AAM + LGL + LC	0.2	dynamic	1	1.17	0.08
			2	1.01	0.076
			3	1.04	0.078
LM-FT	0.5 (5 sec audio)		2	**0.99**	**0.063**

References


Bing Han, Zhengyang Chen, Y. Q. (2022).
Self-supervised speaker verification using dynamic loss-gate and label correction.


Peng, J., Plchot, O., Stafylakis, T., Mošner, L., Burget, L., and Černocký, J. (2023).
An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification.


Ruijie Tao, Kong Aik Lee, R. K. D. V. H. H. L. (2021).
Self-supervised speaker recognition with loss-gated learning.


Sanyuan Chen, Chengyi Wang, Z. C. Y. W. S. L. Z. C. J. L. N. K. T. Y. X. X. J. W. L. Z. S. R. Y. Q. Y. Q. J. W. M. Z. X. Y. F. W. (2021).
Wavlm: Large-scale self-supervised pre-training for full stack speech processing.


Wei-Ning Hsu, Benjamin Bolte, Y.-H. H. T. K. L. R. S. A. M. (2021).
Hubert: Self-supervised speech representation learning by masked prediction of hidden units.