Towards Supervised Performance on Speaker Verification with Self-Supervised Learning by Leveraging Large-Scale ASR Models

Victor Miara, Théo Lepage, Réda Dehak

EPITA Research Laboratory (LRE), France

INTERSPEECH 2024

Code: `https://github.com/theolepage/wavlm_ssl_sv`

Speaker Verification (SV)

<u>**Objective:**</u> Compute **similarity** between two **speaker representations** extracted from pre-trained model on speaker classification [1, 2]

Learn **speaker discriminative** representations that:

- **minimize intra-speaker** distance
- **maximize inter-speaker** distance
- discard non-speaker information (noise, channel, …)

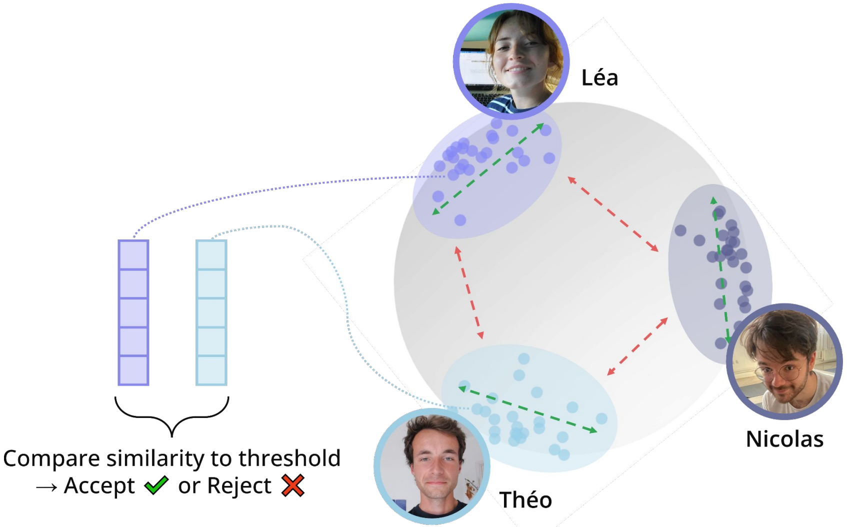


Figure 1. Learning speaker embeddings space for speaker verification systems.

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudan- pur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in ICASSP, 2018.
[2] J. S. Chung, J. Huh, and S. Mun, “Delving into VoxCeleb: Environment Invariant Speaker Recognition,” in Odyssey, 2020.

Self-Supervised Learning for SV

- Why self-supervised learning ?
  - **labeled data** is scarce and expensive
  - leverage abundance of **unlabeled data**
  - learning **meaningful representations** directly from the data

- SSL methods for SV
  - **Contrastive** → SimCLR [4]
  - **Knowledge distillation** → DINO [5]

- Emergence of SSL in **Automatic Speech Recognition** (ASR)
  - wav2vec [1], HuBERT [2]
  - WavLM [3] → Masked **speech denoising** and prediction

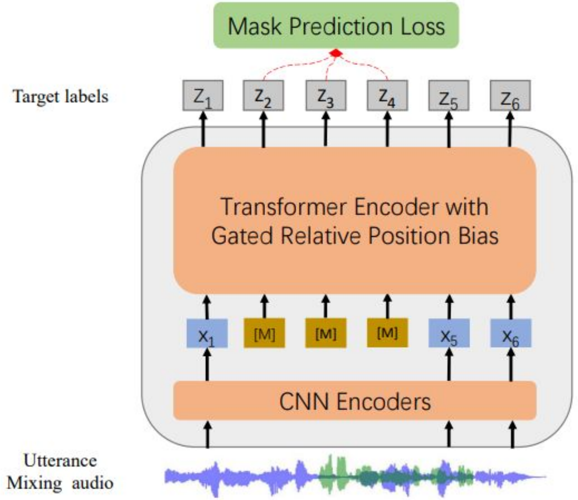


Figure 2. WavLM model architecture [3]

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec:Unsupervised Pre-Training for Speech Recognition,” in Interspeech, 2019
[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE TASLP, 2021
[3] S. Chen, et al., “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”, in IEEE JSTSP 2022
[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in ICML, 2020.
[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, “Emerging Properties in Self-Supervised Vision Transformers”, in ICCV, 2021

Learning speaker representations from ASR models

- Progressive abstraction of **speaker information** across Transformer layers
- Extracting information during training
  - Weighted sum of hidden states with learned weights
  - Multi-Head Factorized Attention (MHFA)

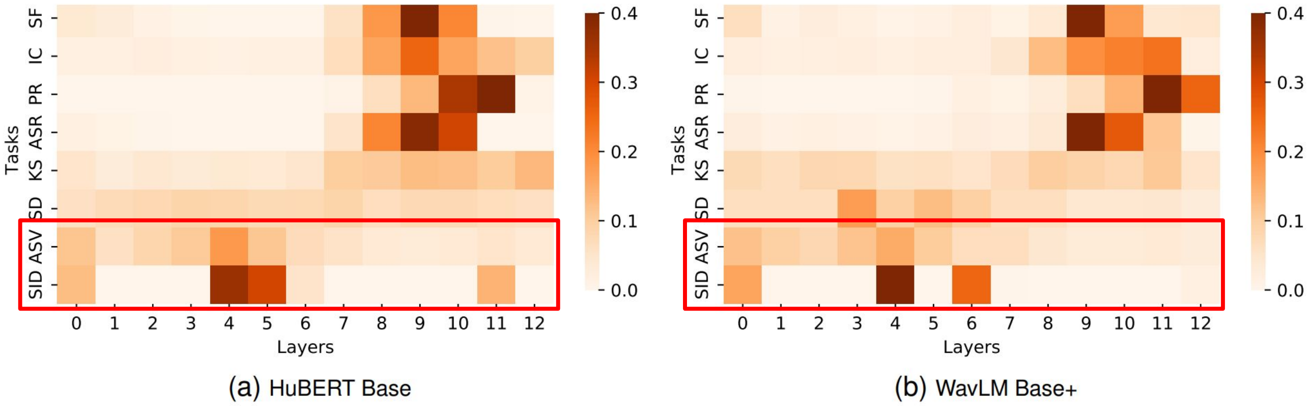
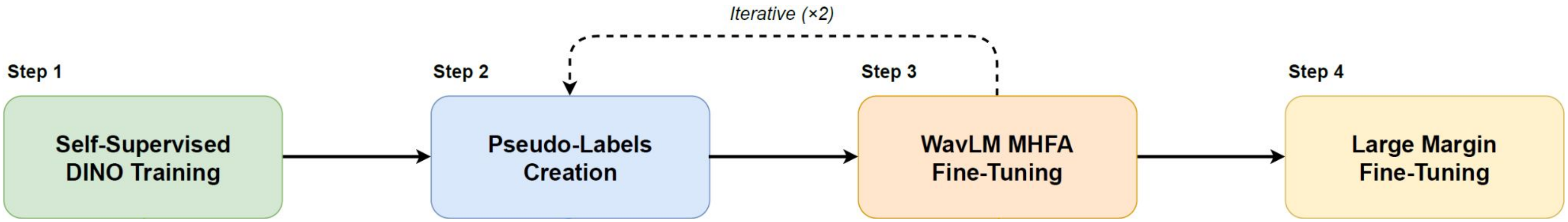


Figure 3. From [1]: Weight analysis per layer when fine-tuning for different tasks of the SUPERB Benchmark

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”, in IEEE JSTSP 2022

Method – Overview

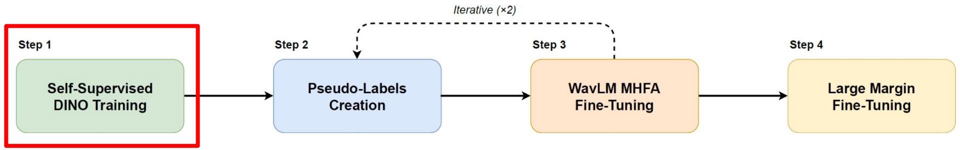


- **Clustering** training samples embeddings

- **Supervised** training on **pseudo-labels**
- 2 iterations

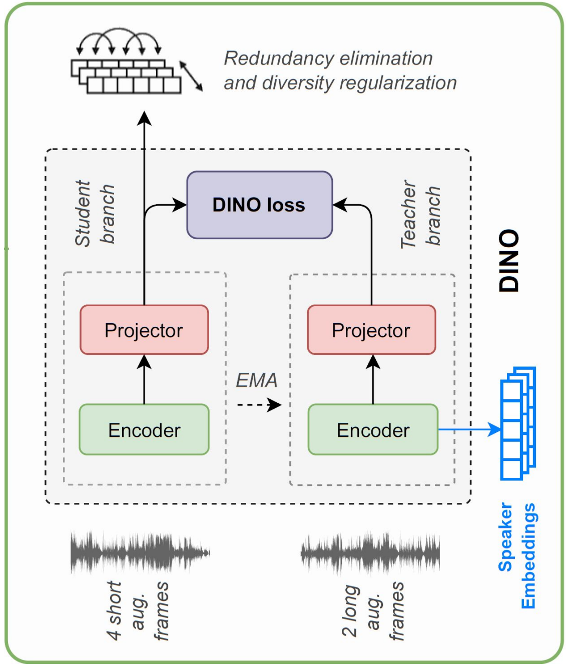
- Audio samples **duration**
  - 3s → 5s
- Margins
  - 0.2 → 0.5
- 2 epochs

Method – DINO-based SSL for SV



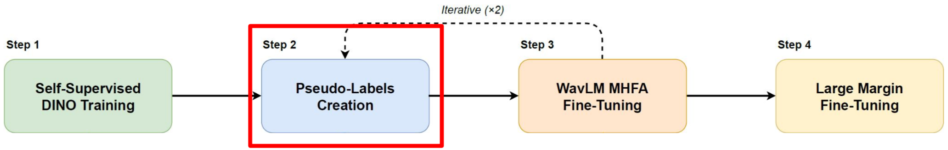
- Self-distillation using DINO [1] framework
- **Minimize** CE between **teacher** and **student** distributions
- Avoid **collapse**
  - sharpening
  - centering

$$\mathcal{L}_{\text{DINO}} = \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in V \\ x' \neq x}} H\left(P_t(x), P_s\left(x'\right)\right)$$



[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, “Emerging Properties in Self-Supervised Vision Transformers”, in ICCV, 2021

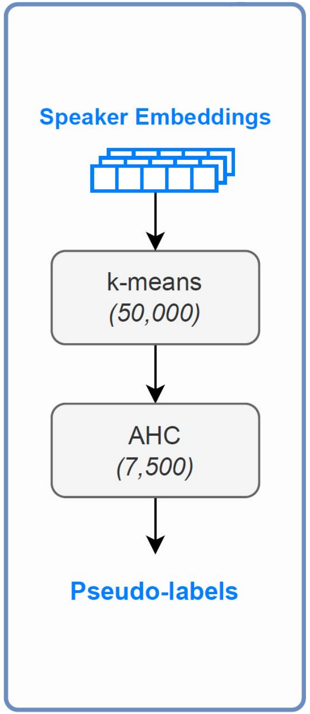
Method – Fine-tuning by leveraging pseudo-labels



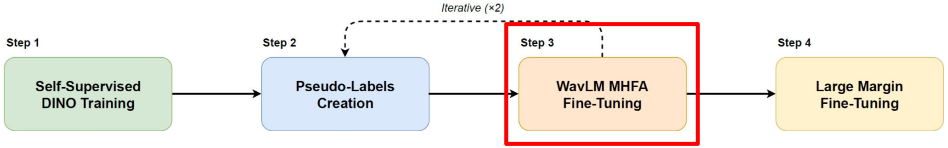
- What are **pseudo-labels** ?
  - Label inferred from data using a pre-trained model

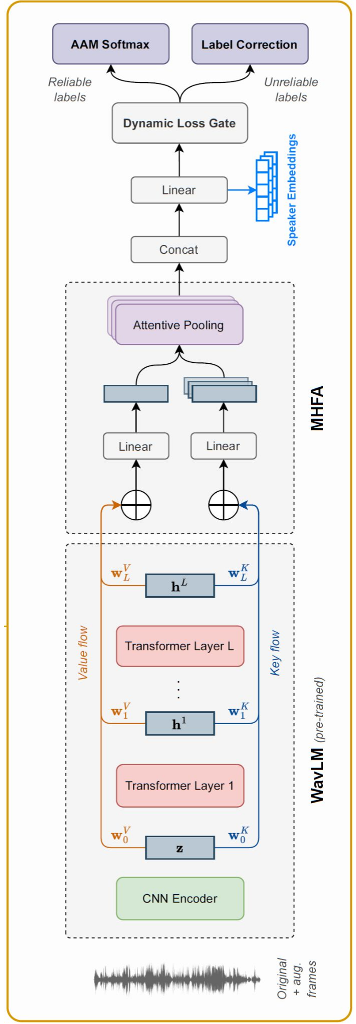
- **Clustering** training samples embeddings
  - k-means (50,000 clusters)
  - AHC (7,500 clusters)

- **Iterative** refining of pseudo-labels



Method – WavLM-based speaker recognition





- Multi-Head Factorized Attention (MHFA) [1]
  - Light-weight layer-wise **attentive pooling**
  - Efficiently capture valuable information from intermediate representations
- Dynamic Loss Gate + Label Correction (DLG-LC) [2]
  - Dealing with **unreliable pseudo-labels**

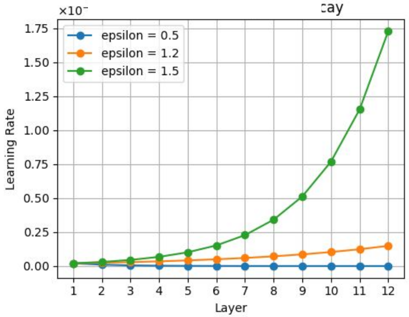
[1] J. Peng, O. Plchot, T. Stafylakis, L. Mosner, L. Burget, and J. Cernocky, "An Attention-Based Backend Allowing Efficient ´Fine-Tuning of Transformer Models for Speaker Verification," in IEEE SLT, 2022
[2] H. Bing, C. Zhengyang, and Q. Yanmin, "Self-Supervised Speaker Verification Using Dynamic Loss-Gate and Label Correction," in Interspeech, 2022

MHFA Training Stabilization

- **L2 regularization** towards **initial weights**
  - WavLM over-parameterized for supervised dataset
  - Avoid over-fitting + stabilize fine-tuning

$$\mathcal{L}_p = \sum_{j=1}^{|\Theta|} \left(\theta^j - \theta_p^j\right)^2$$

- Layer-wise learning rate decay
  - Stabilize **speaker** information in early layers
  - Modify layers containing **speech** information



$$LR_l = LR_1 \times \xi^{l-1}$$

Figure 5. Impact of the epsilon parameter on the learning rate decay per layer

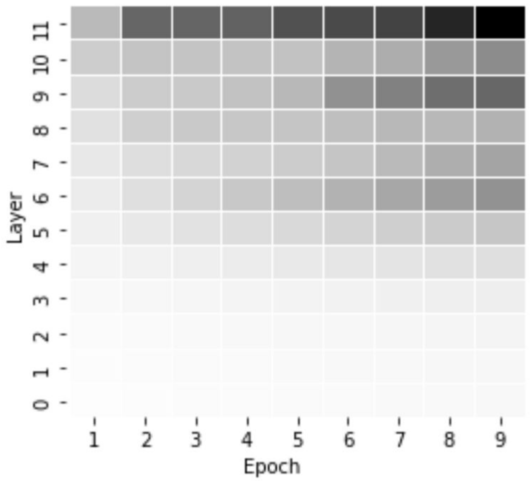


Figure 4. L2 distance between pre-trained and fined tuned weights of the WavLM at different layers and epochs

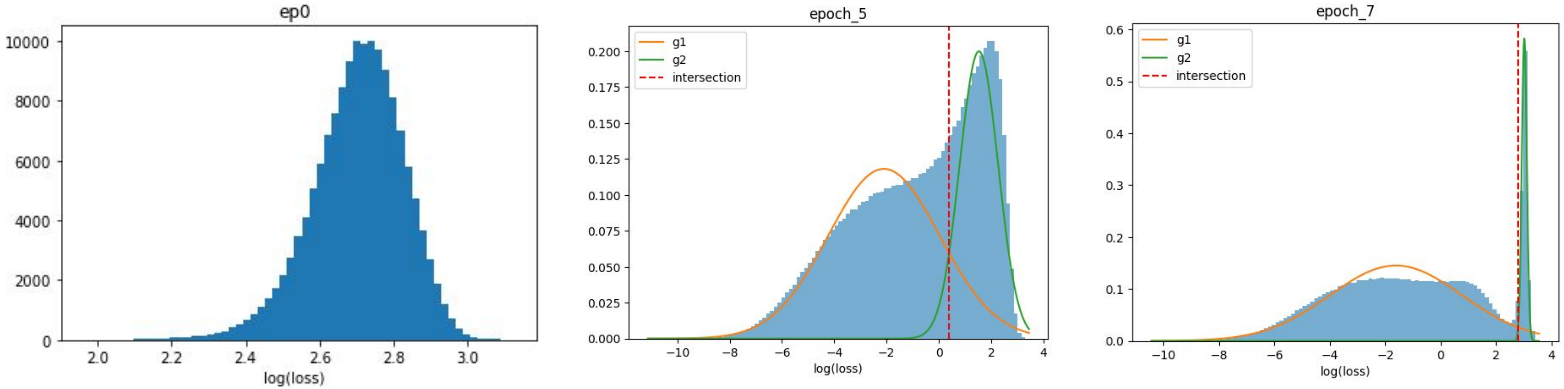
Method – Fine-tuning by leveraging pseudo-labels

- How to handle **incorrect** pseudo-labels?
- **Dynamic Loss-Gate** (DLG) [1]
  - **Higher loss** on **unreliable** samples
  - Ignore **unreliable** samples

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

- **Label Correction** (LC) [1]
  - Avoid discarding **unreliable** samples from training

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



[1] H. Bing, C. Zhengyang, and Q. Yanmin, “Self-Supervised Speaker Verification Using Dynamic Loss-Gate and Label Correction,” in Interspeech, 2022

Experimental setup

- Datasets
  - Training on VoxCeleb2 *dev set* [1]
  - Evaluation on VoxCeleb1 *test* set
  - Speaker labels are discarded
  - 2 seconds audio chunks (5s for LMFT)
  - Data augmentation
    - Music, speech and babble background noises from the MUSAN [2]
    - Reverberation from the Simulated Room Impulse Response Database [3]
- DINO
  - Encoder: ECAPA-TDNN
  - Same training setup as [4]
- WavLM MHFA
  - Pre-trained model: WavLM base+ [5]
  - Epochs: 15 (2 for LMFT)
  - Optimizer: AdamW
  - Batch size: 120
  - Loss: AAM Softmax (s=30, m=0.2)
  - 2x RTX Quadro 8000
- Evaluation protocol
  - Scoring with cosine similarity
  - Equal Error Rate (EER)
  - Minimum Detection Cost Function (minDCF) with p=0.01

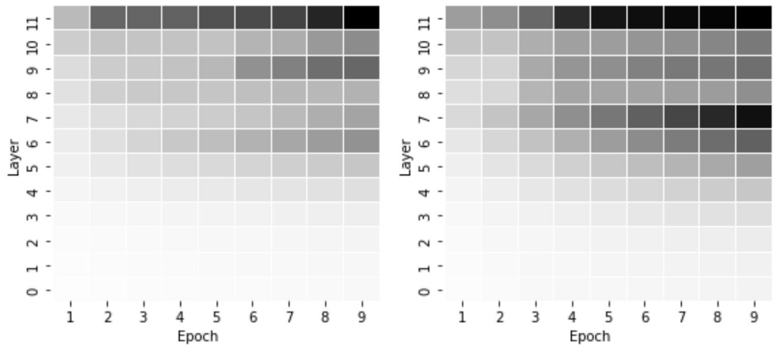
[1] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in Interspeech, 2017
[2] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” arXiv preprint arXiv:1510.08484, 2015
[3] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in ICASSP, 2017
[4] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, “Pushing the limits of self-supervised speaker verification using regularized distillation framework,” in ICASSP, 2023
[5] https://github.com/microsoft/unilm/tree/master/wavlm

Results – Infeasibility of end-to-end self-supervised fine-tuning

Self-supervised training (NT-Xent loss) does not converge to an optimal solution

- **Positive pairs** are extracted from **same utterances**: they share channel and noise characteristics
- Model focuses on learning **channel characteristics**
- Sampling **positive** pairs from **different utterances** improves significantly the performance

Positive pairs sampling	EER (%)	$\text{minDCF}_{0.01}$
Same audio files (SimCLR)	15.13	0.9586
Different audio files	10.81	0.9377



(a) *Supervised training* (b) *SSL contrastive training*

Figure 6. L2 distance between the pre-trained and fine-tuned weights of the WavLM at different layers and epochs.

Results – Incremental study of the components of our framework

- Fine-tuning **WavLM MHFA** on **DINO** pseudo-labels → **52.5%** relative EER reduction

- DLG + LC
  - handling **unreliable** pseudo-labels

- Pseudo-labels refinement
  - Adjusted Rand Index (ARI) : **0.81** → **0.90**
  - Normalized Mutual Information (NMI): **0.95** → **0.98**

Method	**EER (%)**	**minDCF$_{0.01}$**
DINO	3.16	0.2233
+ WavLM MHFA	1.50	0.1378
+ DLG	1.27	0.1401
+ LC	1.22	0.1531
+ IC (iter 1)	1.17	0.1351
+ IC (iter 2)	1.01	0.1399
+ IC (iter 3)	1.08	0.1340
+ LMFT	**0.99**	**0.0905**

Results – Evaluation of different self-supervised SV methods

- Achieving **state-of-the-art** performance on self-supervised **speaker verification**
- Closing the gap between **supervised** and **self-supervised** performance

Method	# of iterations	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
JHU [26]	4	1.89	-	-	-	-	-
DKU [35]	4	1.81	-	-	-	-	-
SNU [36]	4	1.66	-	-	-	-	-
LGL [30]	5	1.66	-	2.18	-	3.76	-
DLG-LC [25]	5	1.47	-	1.78	-	3.19	-
Ours	**3**	**0.99**	**0.0905**	**1.21**	**0.1263**	**2.35**	**0.2214**
Supervised	-	0.94	0.1179	0.93	0.1066	1.94	0.1919

Conclusions

- Our method consists in **fine-tuning a pre-trained ASR model** with the **MHFA backend** on **pseudo-labels** iteratively refined and initially extracted from a **DINO** SSL-based framework
- We achieve **0.99% EER** on VoxCeleb1-O, **without using any speaker label**, and outperform current state-of-the-art methods
- This contribution is a **step towards supervised performance** with self-supervised learning