Self-Supervised Learning for Speaker Recognition

Apprentissage Auto-Supervisé pour la Reconnaissance du Locuteur

Ph.D. Thesis Defense • Paris, 13 February 2026

Théo LEPAGE

Advisors

Réda DEHAK (LRE)

Thierry GÉRAUD (LRE)

Committee (R) Reviewer

Benjamin LECOUTEUX (LIG) (R)

Driss MATROUF (LIA) (R)

Jean-François BONASTRE (AMIAD)

Irina ILLINA (LORIA/Inria)

Anthony LARCHER (LIUM)

Douglas REYNOLDS (MIT LL)

Self-Supervised Learning for Speaker Recognition

Introduction

🗣️ **Voice as a biometric trait** (natural, remote, non-intrusive)

🛡️ **Critical applications** (authentication & forensics)

🎭 **Rising societal challenges** (bias, privacy, deep fakes)

Self-Supervised Learning for Speaker Recognition

Introduction

🗣️ Voice as a biometric trait (natural, remote, non-intrusive)

🛡️ Critical applications (authentication & forensics)

🎭 Rising societal challenges (bias, privacy, deep fakes)

Self-Supervised Learning for Speaker Recognition

🌍 Massive amounts of speech available online

🧠 Learning from data without human annotation

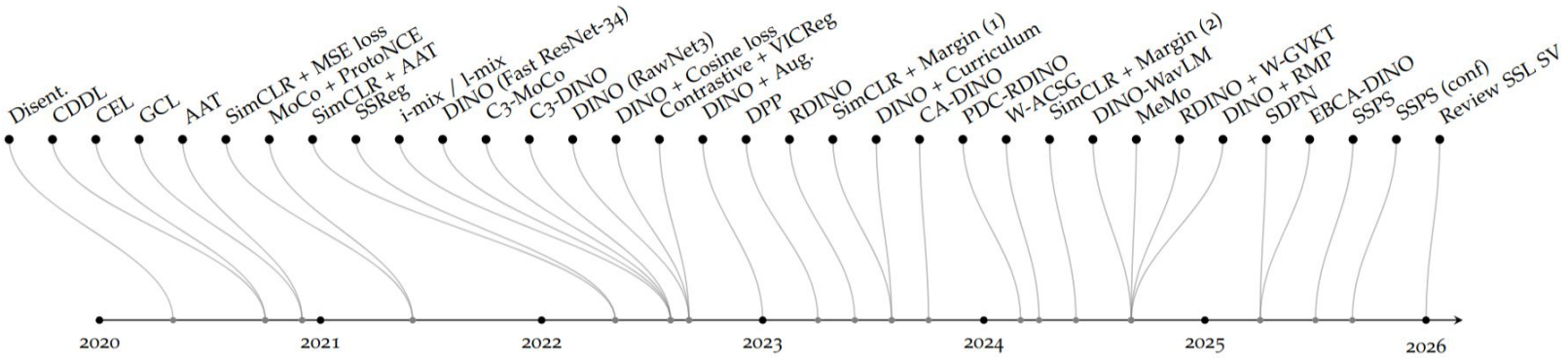
🚀 Proven success in vision, language, and speech

Introduction

Question: *How to learn reliable speaker representations from speech without labels?*

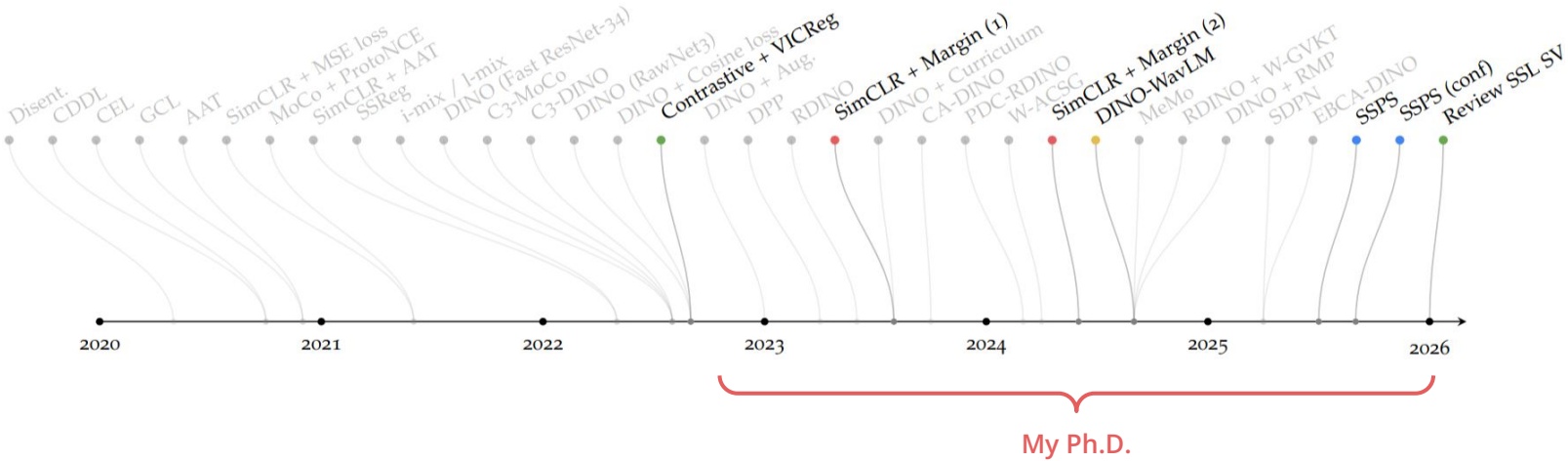
Introduction

Question: *How to learn reliable speaker representations from speech without labels?*



Introduction

Question: *How to learn reliable speaker representations from speech without labels?*



➔ This thesis contributes to this fast-evolving paradigm for Speaker Recognition, toward greater generalization and reduced reliance on labeled data.

Background

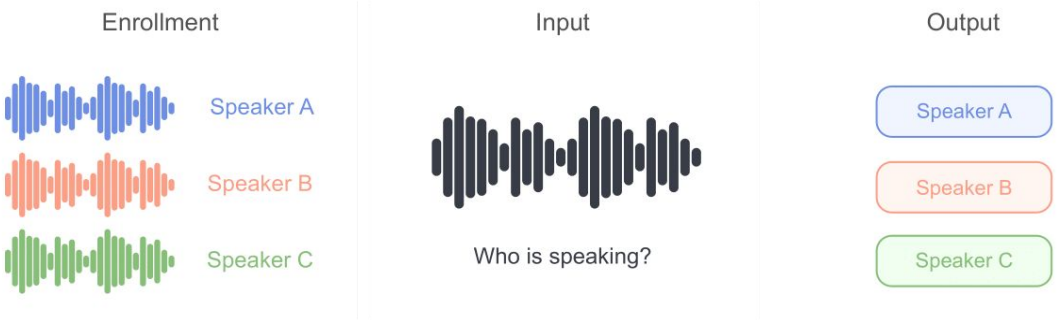
Speaker Recognition, Deep Learning & Self-Supervised Learning

Background

Speaker Recognition (SR)

Speaker Identification (SID)

… assigns an unknown voice to one of several enrolled speakers.

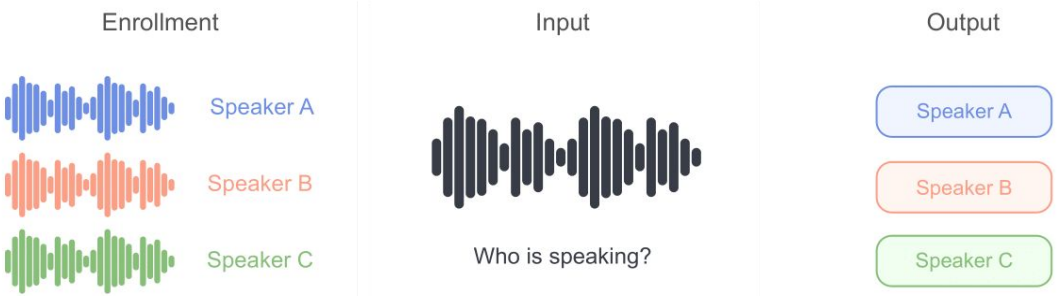


Background

Speaker Recognition (SR)

Speaker Identification (SID)

... assigns an unknown voice to one of several enrolled speakers.



Speaker Verification (SV)

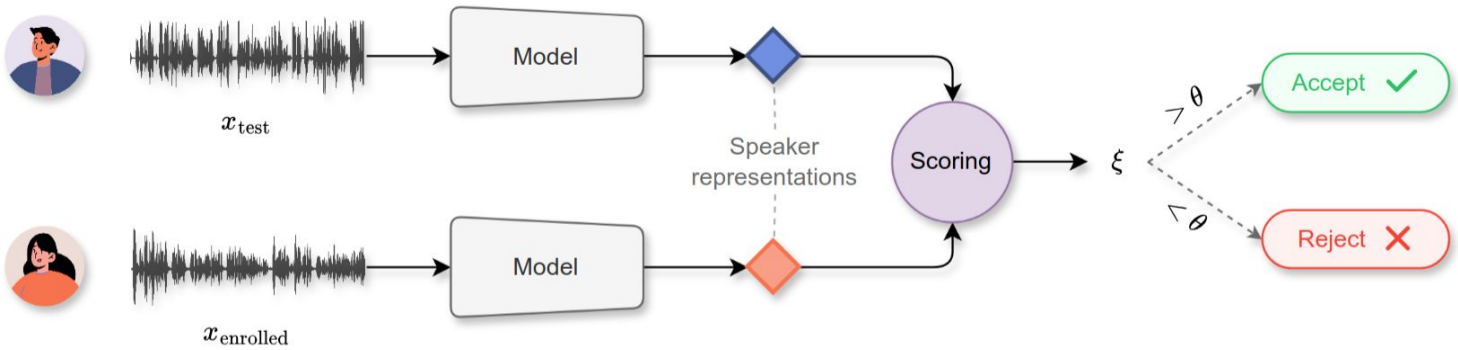
... determines whether an unknown voice matches a claimed identity.

Background

Speaker Verification (SV)

The modern **Speaker Verification (SV)** framework operates by extracting **speaker representations** from two speech samples (***test*** and ***enrolled***) and computing a **similarity score** ξ.

The score reflects the likelihood that both samples originate from the same speaker and is compared to a **decision threshold** θ to either **accept or reject the identity claim** [1].

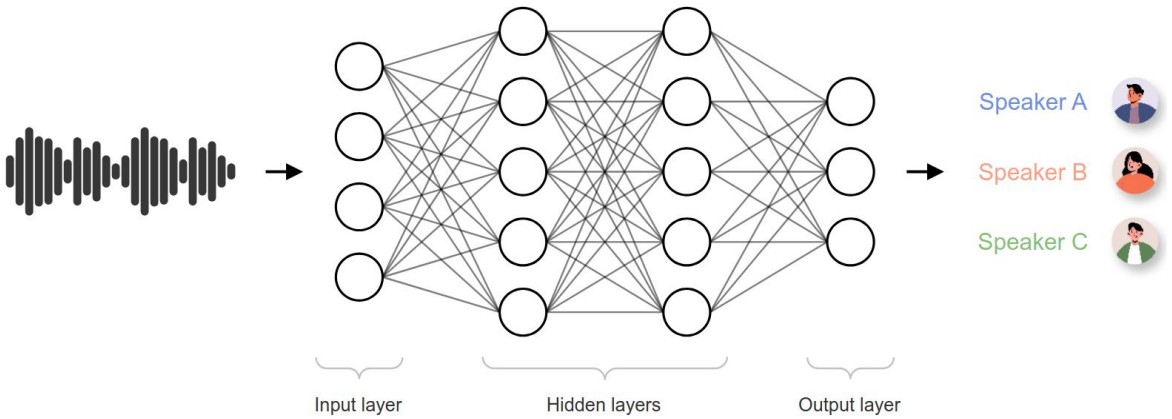


[1] D. A. Reynolds. *An Overview of Automatic Speaker Recognition Technology*. ICASSP, 2002.

Background

Deep Learning for Speaker Verification

Recent SV systems learn speaker representations by training **Deep Neural Networks (DNN)** in a supervised setting, typically using a speaker classification objective [1,2,3].

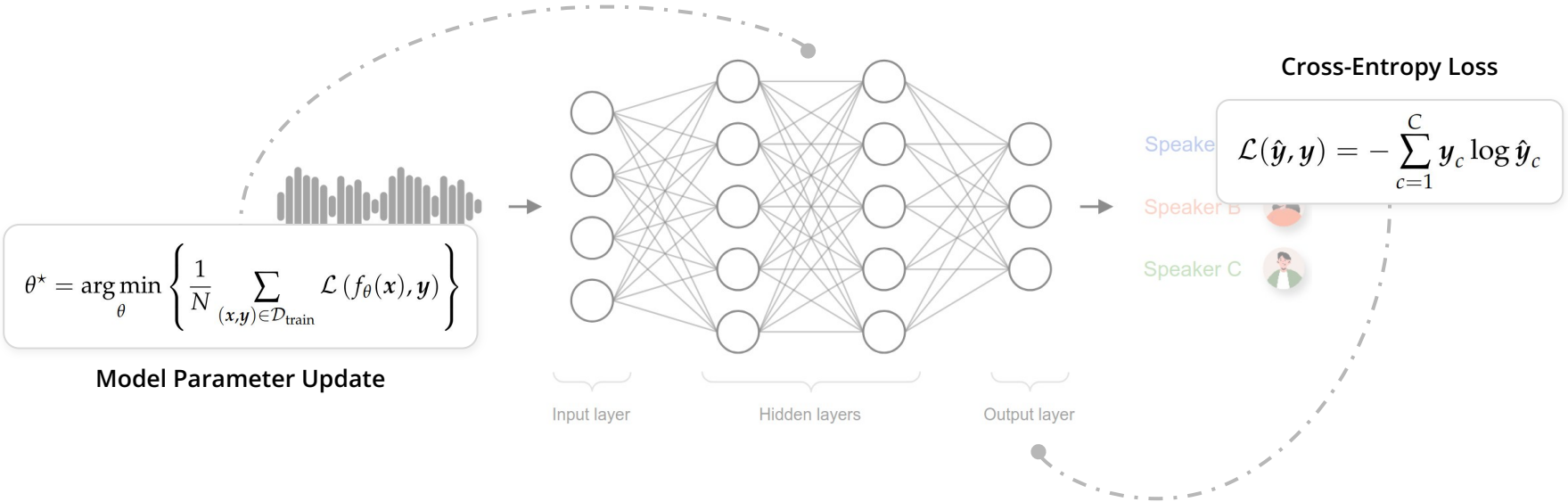


[1] D. Snyder et al. *X-Vectors: Robust DNN Embeddings for Speaker Recognition*. ICASSP, 2018.
[2] B. Desplanques et al. *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. Interspeech, 2020.
[3] Z. Bai et al. *Speaker recognition based on deep learning: An overview*. Neural Networks, 2021.

Background

Deep Learning for Speaker Verification

Recent SV systems learn speaker representations by training **Deep Neural Networks (DNN)** in a supervised setting, typically using a speaker classification objective [1,2,3].



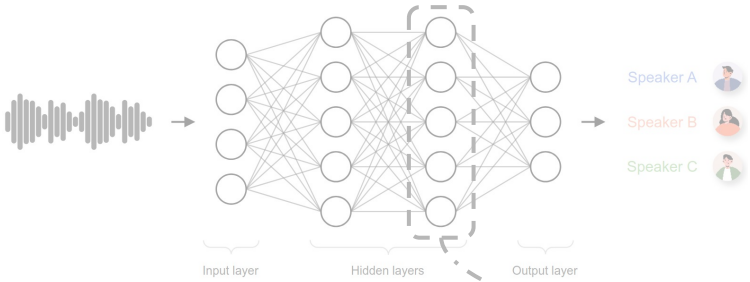
[1] D. Snyder et al. *X-Vectors: Robust DNN Embeddings for Speaker Recognition*. ICASSP, 2018.
[2] B. Desplanques et al. *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*. Interspeech, 2020.
[3] Z. Bai et al. *Speaker recognition based on deep learning: An overview*. Neural Networks, 2021.

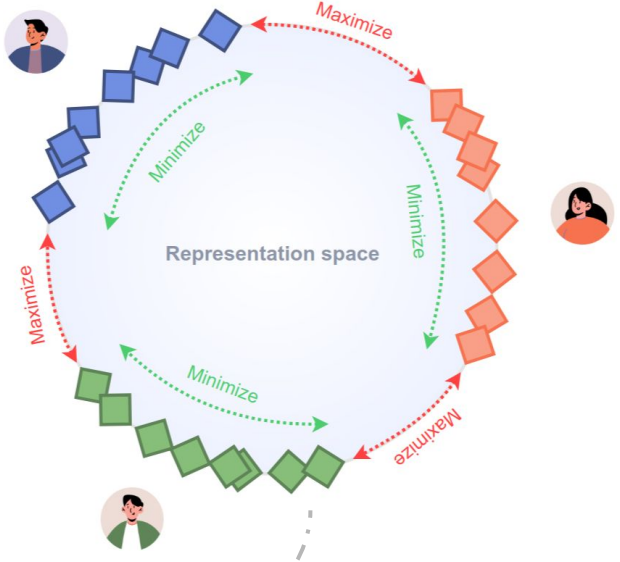
Background

Representation Learning for Speaker Verification

An optimal speaker representation space should exhibit the following properties [1]:

1. High inter-speaker separation ↗
2. Low intra-speaker variance ↘
3. Invariance to factors extrinsic to speaker identity
   (*e.g., noise, environment, speaking style, recording conditions, ...*)





[1] P. Rose. *Forensic Speaker Identification*. 2002.

Background

Limitations of Supervised Learning

Deep Learning is inherently dependent on large labeled datasets as performance scales with data, resulting in several important limitations [1].

- **Cost of Annotations**
  → Expensive and impractical at scale for speech
- **Human-Induced Biases**
  → Subjective and inconsistent labels lead to bias
- **Constrained Generalization**
  → Predefined and specific tasks limit transferability
- **Toward Autonomous Intelligence**
  → Less supervision = closer to human cognition

[1] R. Balestriero et al. *A Cookbook of Self-Supervised Learning*. arXiv preprint, 2023.

Background

Self-Supervised Learning (SSL)

Self-Supervised Learning (SSL) exploits supervisory signals naturally present in the data, enabling DNNs to learn from large amounts of unlabeled data.

Input Target

$$\mathcal{L}_{\text{Sup}}\left(f_\theta(\boldsymbol{x}), \boldsymbol{y}\right)$$

The target is provided through human annotation

Input Target

$$\mathcal{L}_{\text{SSL}}\left(f_\theta(\boldsymbol{x}), g(\boldsymbol{x})\right)$$

The target is automatically derived from the data

Unsupervised learning for DNNs

SSL has proven effective across vision [1], speech [2], and language [3], often matching or outperforming supervised approaches. 🚀

[1] M. Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. arXiv preprint, 2023.
[2] S. Chen et al. *WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing*. IEEE JSTSP, 2022.
[3] J. Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT, 2019.

Background

Self-Supervised Learning (SSL)

Self-Supervised Learning (SSL) exploits supervisory signals naturally present in the data, enabling DNNs to learn from large amounts of unlabeled data.

Input Target

$$\mathcal{L}_{\text{Sup}}\left(f_\theta(\boldsymbol{x}), \boldsymbol{y}\right)$$

The target is provided through human annotation

Input Target

$$\mathcal{L}_{\text{SSL}}\left(f_\theta(\boldsymbol{x}), g(\boldsymbol{x})\right)$$

The target is automatically derived from the data

Unsupervised learning for DNNs

SSL has proven effective across vision [1], speech [2], and language [3], often matching or outperforming supervised approaches. 🚀

[1] M. Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. arXiv preprint, 2023.
[2] S. Chen et al. *WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing*. IEEE JSTSP, 2022.
[3] J. Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT, 2019.

Background

Self-Supervised Learning (SSL)

Self-Supervised Learning (SSL) exploits supervisory signals naturally present in the data, enabling DNNs to learn from large amounts of unlabeled data.

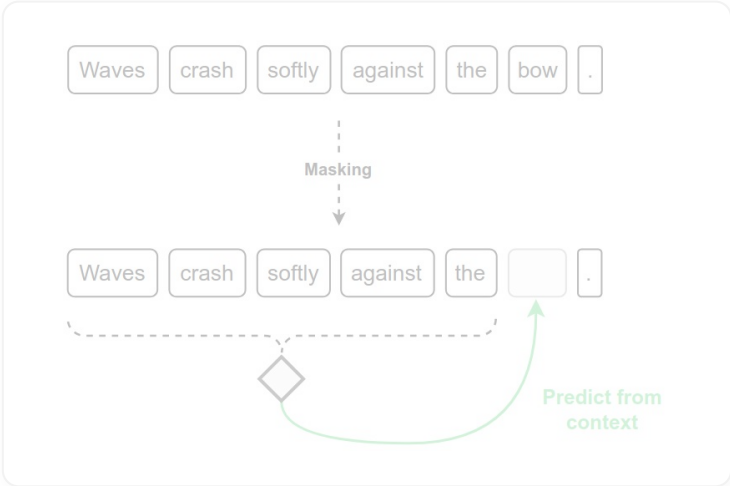
SSL has proven effective across vision [1], speech [2], and language [3], often matching or outperforming supervised approaches. 🚀

[1] M. Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. arXiv preprint, 2023.
[2] S. Chen et al. *WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing*. IEEE JSTSP, 2022.
[3] J. Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT, 2019.

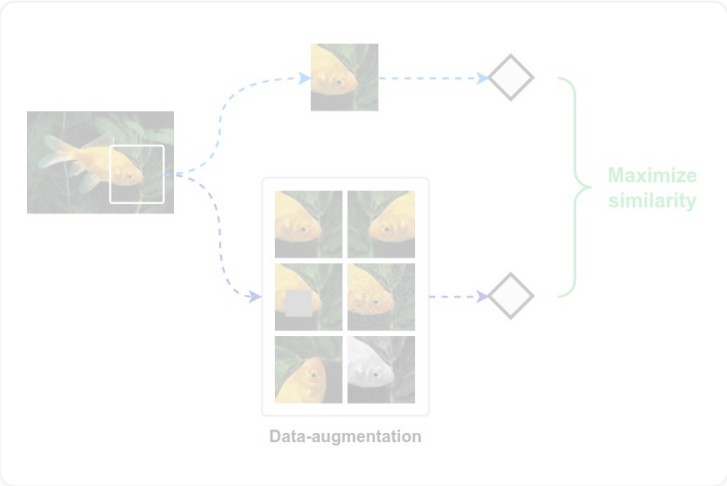
Background

Self-Supervised Learning (SSL)

SSL pre-trains a model on an *unsupervised* pretext task to learn *(transferable)* representations for downstream task(s) [1].



Generative



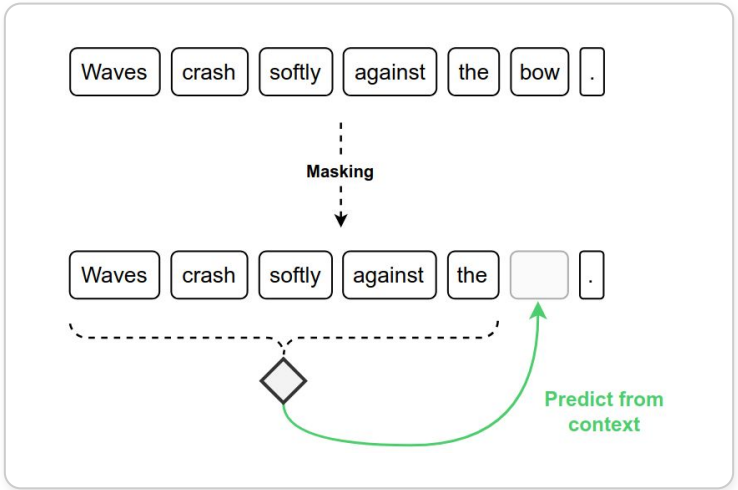
Instance-Invariance

[1] R. Balestriero et al. *A Cookbook of Self-Supervised Learning*. arXiv preprint, 2023.

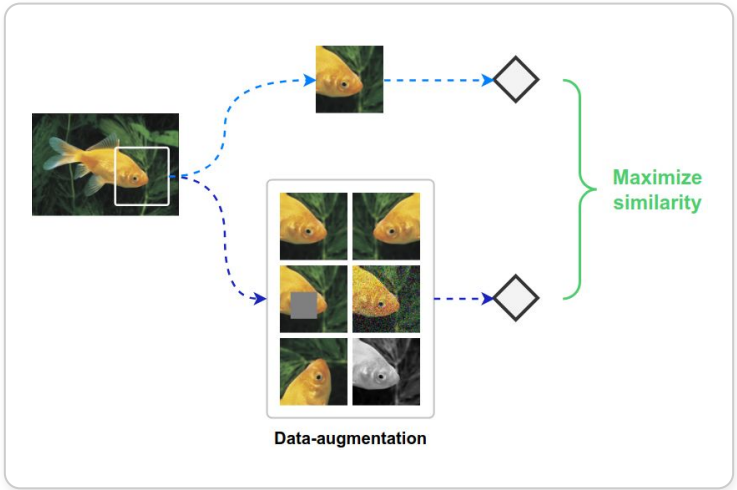
Background

Self-Supervised Learning (SSL)

SSL pre-trains a model on an *unsupervised* pretext task to learn *(transferable)* representations for downstream task(s) [1].



Generative



Instance-Invariance

[1] R. Balestriero et al. *A Cookbook of Self-Supervised Learning*. arXiv preprint, 2023.

Outline

Part I – Self-Supervised Learning for Speaker Verification SSLSV

Part II – Learning Discriminative Speaker Representations Margins

Part III – Self-Supervised Positive Sampling from Latent Space SSPS

Part IV – Leveraging Speech Foundation Models Foundation

Outline

Part I – Self-Supervised Learning for Speaker Verification SSLSV

Part II – Learning Discriminative Speaker Representations Margins

Part III – Self-Supervised Positive Sampling from Latent Space SSPS

Part IV – Leveraging Speech Foundation Models Foundation

Outline

Part I – Self-Supervised Learning for Speaker Verification SSLSV

Part II – Learning Discriminative Speaker Representations Margins

Part III – Self-Supervised Positive Sampling from Latent Space SSPS

Part IV – Leveraging Speech Foundation Models Foundation

Outline

Part I – Self-Supervised Learning for Speaker Verification SSLSV

Part II – Learning Discriminative Speaker Representations Margins

Part III – Self-Supervised Positive Sampling from Latent Space SSPS

Part IV – Leveraging Speech Foundation Models Foundation

SSLSV

Part I

Self-Supervised Learning for Speaker Verification

Related Publications

Label-Efficient Self-Supervised Speaker Verification With Information Maximization and Contrastive Learning
Interspeech, 2022

Self-Supervised Learning for Speaker Recognition: A study and review
Speech Communication, vol. 176, 2026

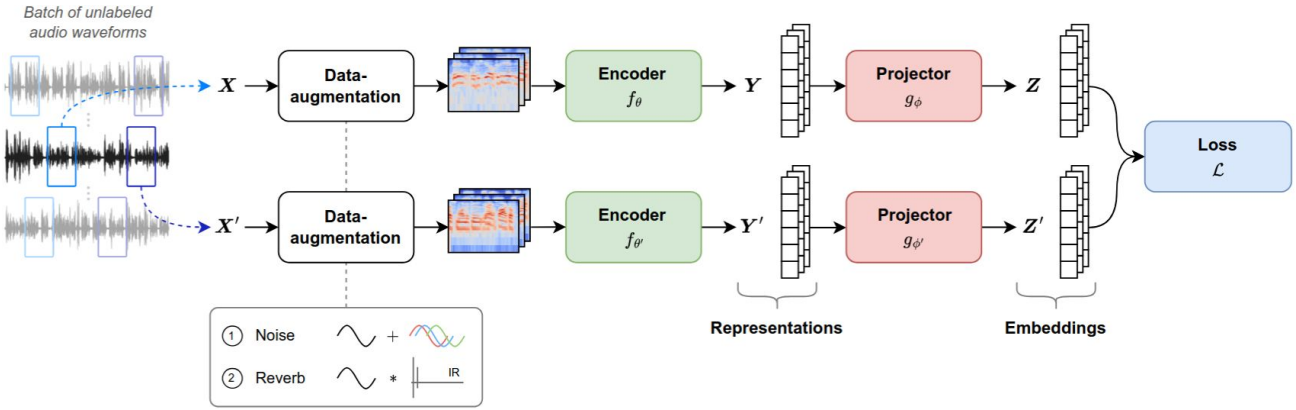
I – Self-Supervised Learning for Speaker Verification SSLSV

Introduction

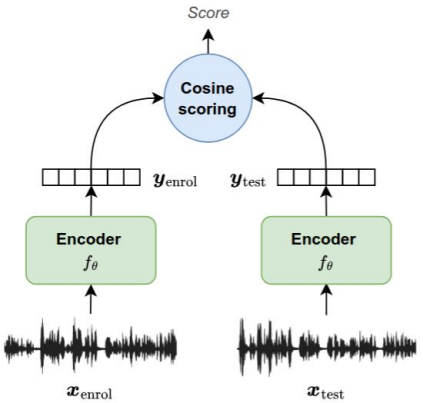
The SSL instance-invariance framework uses a joint-embedding architecture where an anchor-positive pair is generated from the same speech sample using distinct transformations.

Inputs are mapped to representations *(downstream task)* by an encoder, then to embeddings *(pretext task)* by a projector.

Training ***(pretext)***



Evaluation ***(downstream)***



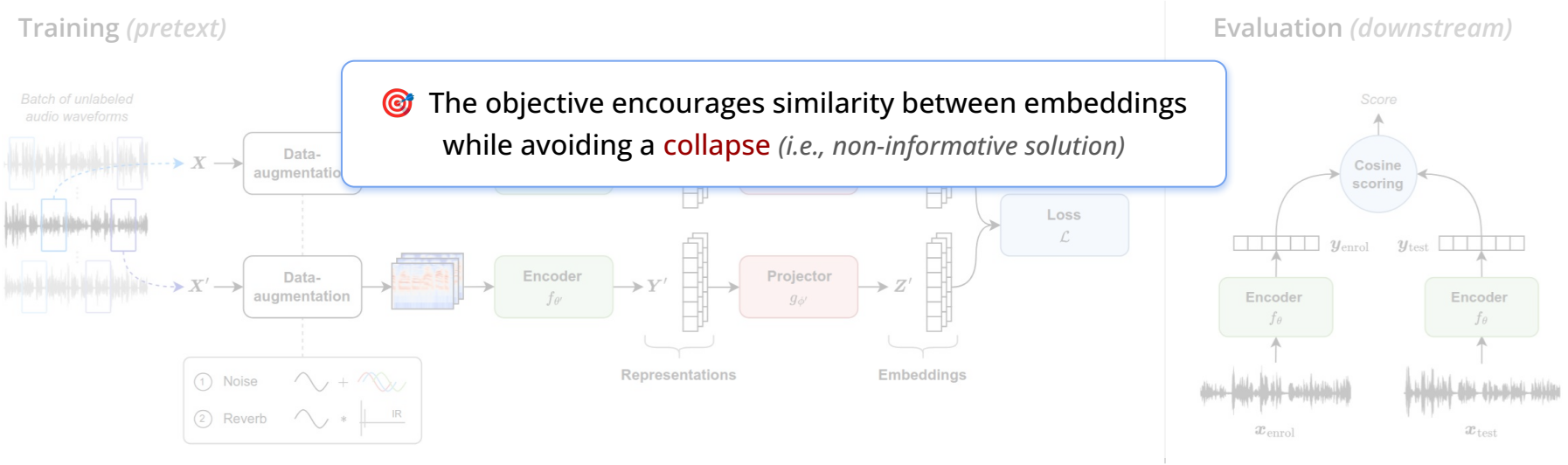
I – Self-Supervised Learning for Speaker Verification

SSLSV

Introduction

The SSL instance-invariance framework uses a joint-embedding architecture where an anchor-positive pair is generated from the same speech sample using distinct transformations.

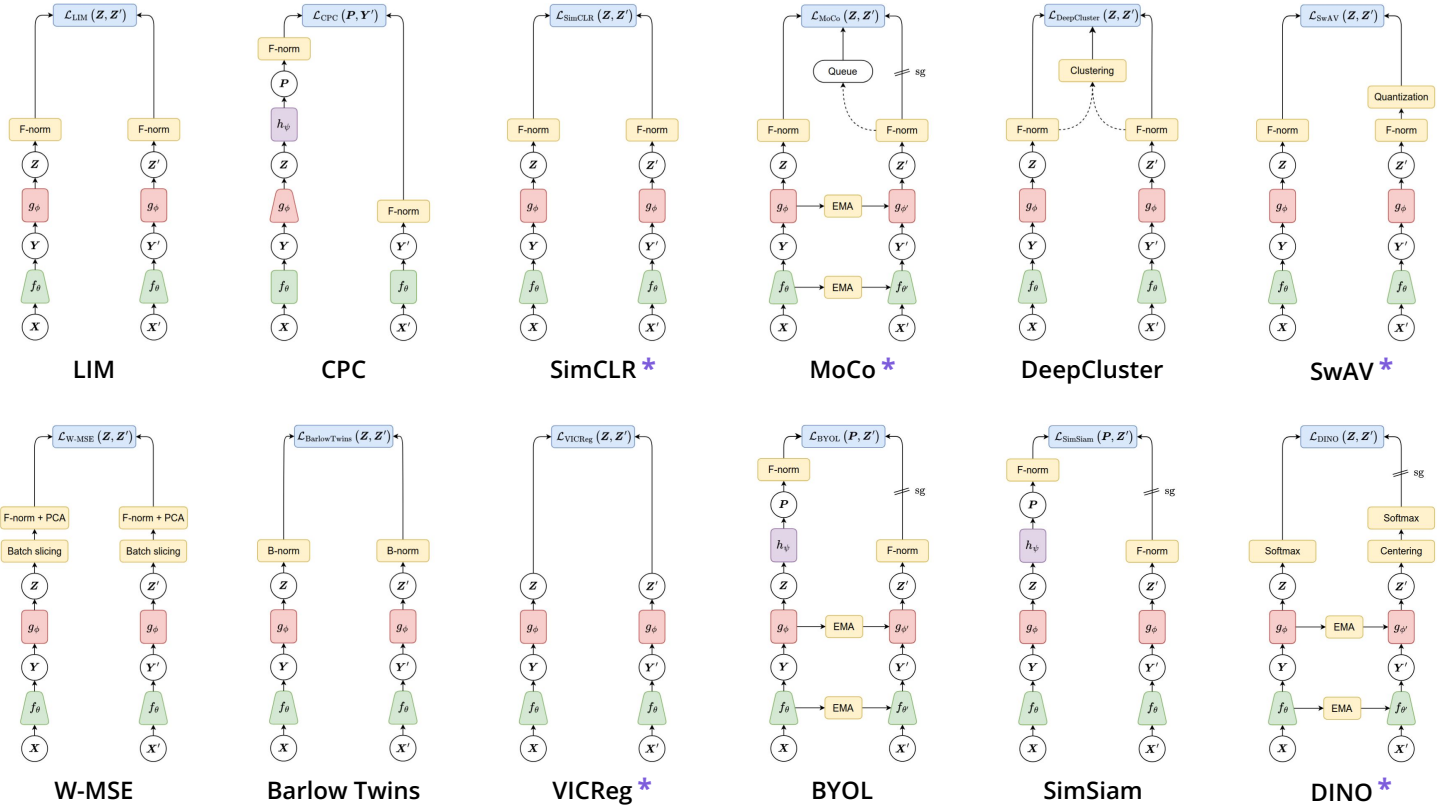
Inputs are mapped to representations *(downstream task)* by an encoder, then to embeddings *(pretext task)* by a projector.



I – Self-Supervised Learning for Speaker Verification

SSLSV

Overview of Instance-Invariance Frameworks

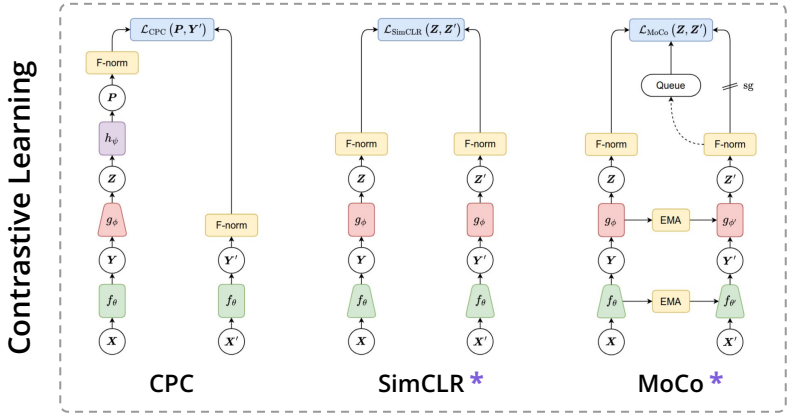


Manuscript: Section 5.2

I – Self-Supervised Learning for Speaker Verification

SSLSV

Overview of Instance-Invariance Frameworks



DeepCluster SwAV *

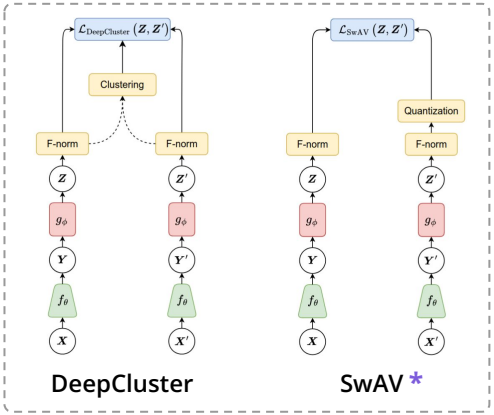
W-MSE Barlow Twins VICReg * BYOL SimSiam DINO *

Manuscript: Section 5.2

I – Self-Supervised Learning for Speaker Verification

SSLSV

Overview of Instance-Invariance Frameworks



Clustering

CPC

SimCLR *

MoCo *

W-MSE

Barlow Twins

VICReg *

BYOL

SimSiam

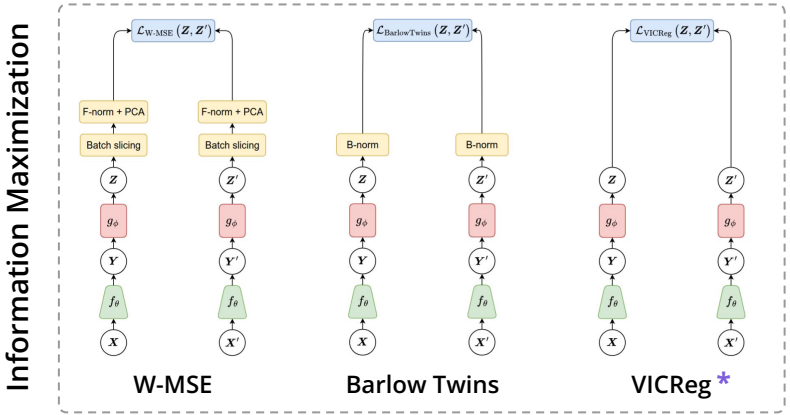
DINO *

Manuscript: Section 5.2

I – Self-Supervised Learning for Speaker Verification

SSLSV

Overview of Instance-Invariance Frameworks



CPC SimCLR * MoCo * DeepCluster SwAV *

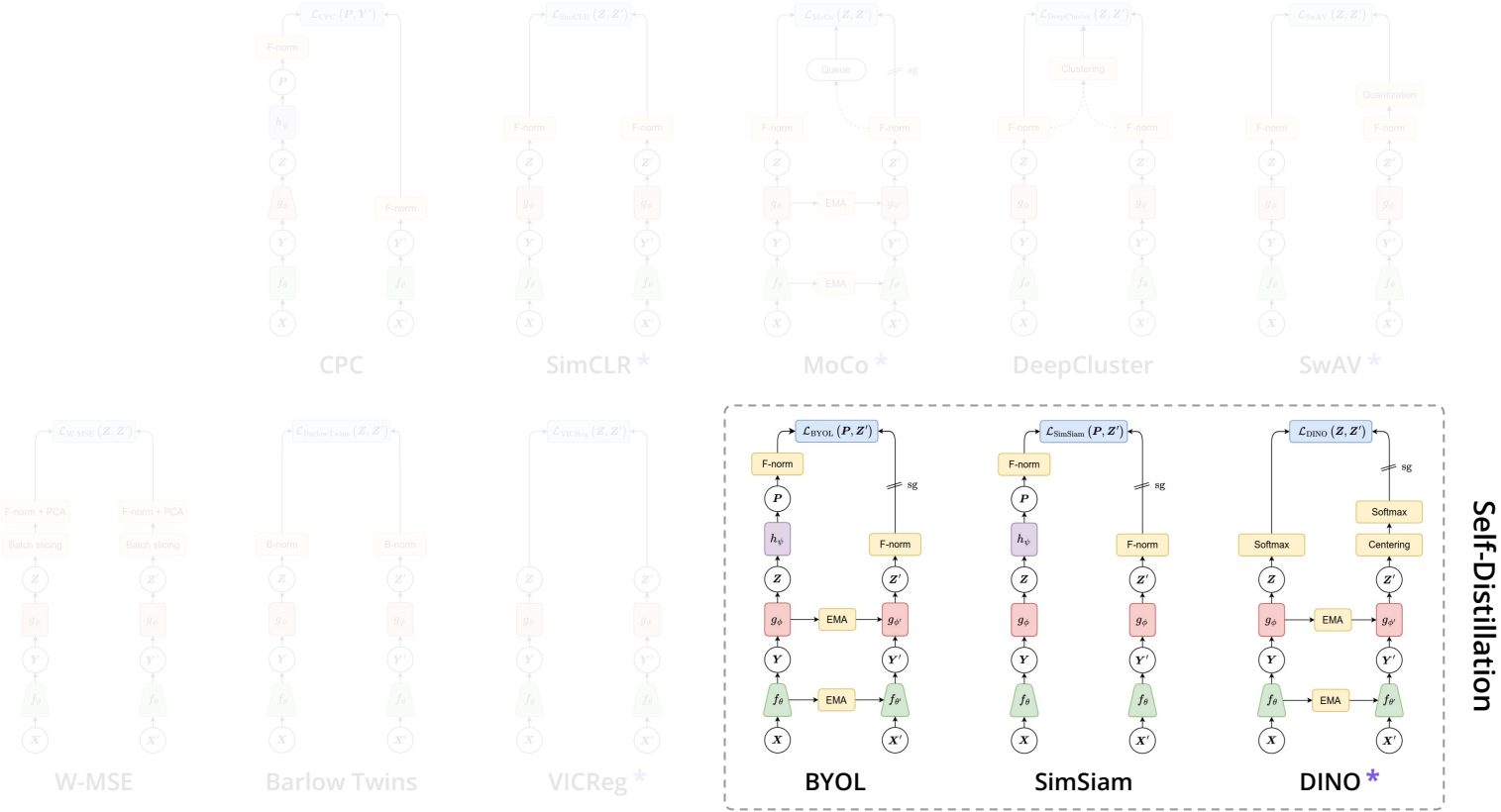
BYOL SimSiam DINO *

Manuscript: Section 5.2

I – Self-Supervised Learning for Speaker Verification

SSLSV

Overview of Instance-Invariance Frameworks

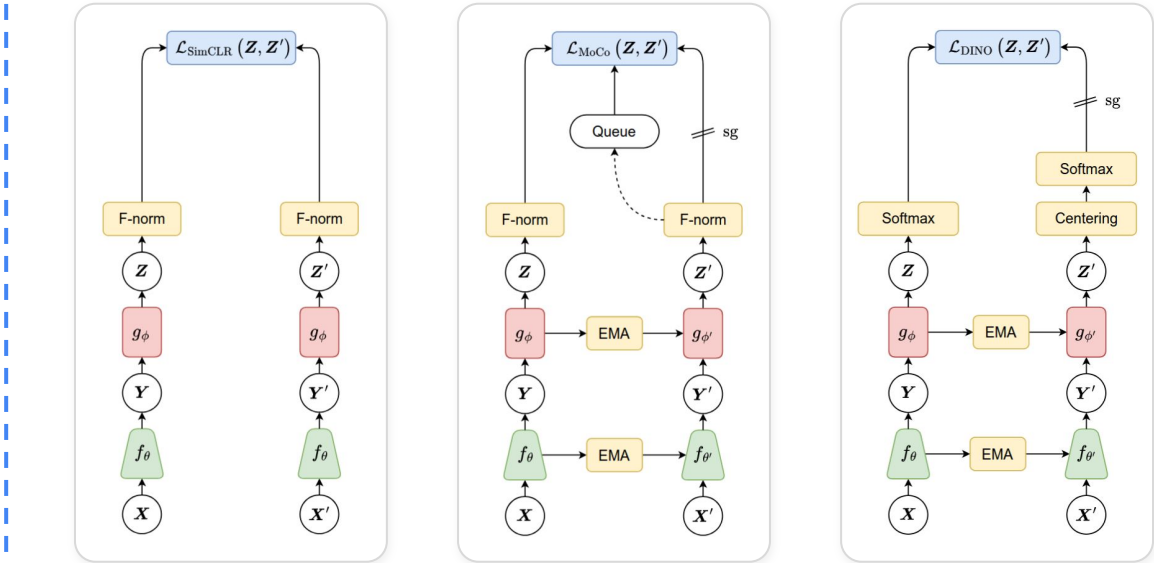


Manuscript: Section 5.2

I – Self-Supervised Learning for Speaker Verification

SSLSV

Overview of Instance-Invariance Frameworks



I – Self-Supervised Learning for Speaker Verification

SSLSV

Contrastive Learning

Contrastive learning aims at **maximizing** the similarity of **anchor-positive** pairs while **minimizing** the similarity of **anchor-negative** pairs, with **randomly sampled negatives** assumed to yield negligible **class collisions** *(false negatives)*.

I – Self-Supervised Learning for Speaker Verification

SSLSV

Contrastive Learning – SimCLR

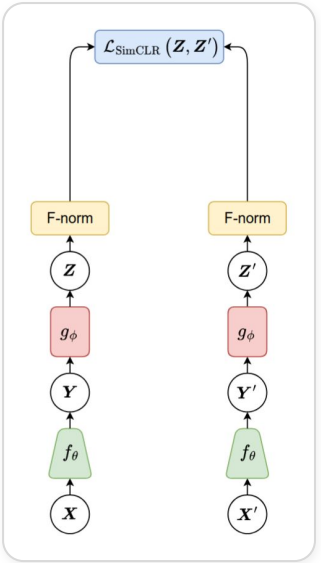
Contrastive learning aims at maximizing the similarity of anchor-positive pairs while minimizing the similarity of anchor-negative pairs, with randomly sampled negatives assumed to yield negligible class collisions *(false negatives)*.

SimCLR [1] *(symmetrical)* samples negatives from the current batch of training embeddings.

$$\mathcal{L}_{\text{SimCLR}} = -\frac{1}{B}\sum_{i\in\mathcal{B}} \log \frac{\ell\left(\boldsymbol{z}_i, \boldsymbol{z}'_i\right)}{\sum_{j\in\mathcal{B}} \ell\left(\boldsymbol{z}_i, \boldsymbol{z}'_j\right)}$$

⚠ Collapse avoided via negatives

$$\ell(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(\text{sim}\left(\boldsymbol{a}, \boldsymbol{b}\right)/\tau\right)$$



SimCLR

[1] T. Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. ICML, 2020.

I – Self-Supervised Learning for Speaker Verification

SSLSV

Contrastive Learning – MoCo

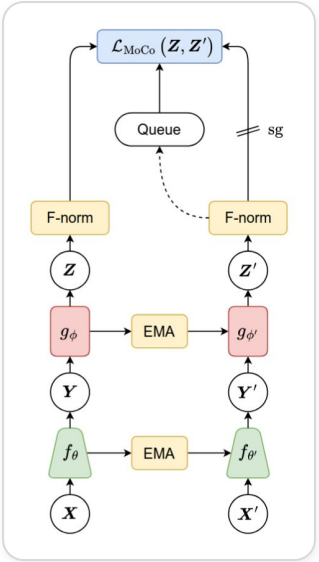
Contrastive learning aims at **maximizing** the similarity of **anchor-positive** pairs while **minimizing** the similarity of **anchor-negative** pairs, with **randomly sampled negatives** assumed to yield negligible **class collisions** *(false negatives)*.

MoCo [2] *(asymmetrical)* samples negatives **from a larger set of latest training embeddings**.

$$\mathcal{L}_{\text{MoCo}} = -\frac{1}{B}\sum_{i\in\mathcal{B}} \log \frac{\ell\left(\boldsymbol{z}_i, \boldsymbol{z}_i'\right)}{\ell\left(\boldsymbol{z}_i, \boldsymbol{z}_i'\right) + \sum_{j\in\mathcal{J}} \ell\left(\boldsymbol{z}_i, \boldsymbol{q}_j\right)}$$

⚠ Collapse avoided via negatives

$$\ell(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(\text{sim}\left(\boldsymbol{a}, \boldsymbol{b}\right)/\tau\right)$$



MoCo

[2] K. He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. CVPR, 2020.

I – Self-Supervised Learning for Speaker Verification

SSLSV

Self-Distillation

Self-Distillation extends knowledge distillation by training a student model to predict the output of a teacher model, whose supervision is bootstrapped from the data and progressively refined throughout the training.

I – Self-Supervised Learning for Speaker Verification

SSLSV

Self-Distillation – DINO

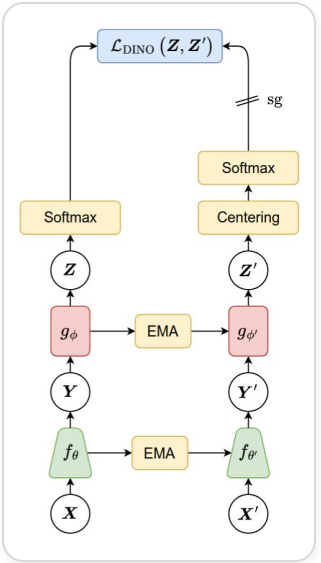
Self-Distillation extends knowledge distillation by training a student model to predict the output of a teacher model, whose supervision is bootstrapped from the data and progressively refined throughout the training.

DINO [1] *(asymmetrical)* uses a larger set of augmented segments and normalizes the teacher outputs (sharpening & centering).

$$\mathcal{L}_{\text{DINO}} = \frac{1}{B} \sum_{i \in \mathcal{B}} \sum_{t=1}^{G} \sum_{\substack{s=1 \\ s \neq t}}^{G+L} H\left(\frac{z'_{i,t} - c}{\tau_t}, \frac{z_{i,s}}{\tau_s} \right)$$

⚠ Collapse avoided via norm. & asymmetry

$$H(a, b) = -\operatorname{softmax}(a) \log\left(\operatorname{softmax}(b)\right)$$



DINO

[1] M. Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. ICCV, 2021.

I – Self-Supervised Learning for Speaker Verification SSLSV

Experimental Setup

Datasets

VoxCeleb [1] is a large-scale audio dataset consisting of speech extracted from interview videos posted on YouTube.

Train data

- VoxCeleb2 (1,092,009 utterances from 5,994 speakers)
- Speaker labels are discarded
- 2 seconds audio and 40-d log-mel spectrogram features
- Augmentations: reverberation + noise

Test data

- VoxCeleb1 (148,642 utterances from 1,211 speakers)
- Trials: Original (O), Extended (E) and Hard (H)

Models & Training

Encoder: Fast ResNet-34 [2] / ECAPA-TDNN (C=1024) [3]

Training: Adam, batch size of 256, 100 epochs

Toolkit: sslsv (https://github.com/theolepage/sslsv)

GPUs: 2× and 4× NVIDIA V100 32GB *(Jean Zay, CNRS/IDRIS)*

Evaluation

Scoring: Cosine similarity of representations

Metrics

↓ : lower is better

- EER (Equal Error Rate) ↓
- minDCF (minimum Detection Cost Function) ↓
  ↳ $P_{\text{target}} = 0.01,\ C_{\text{fa}} = 1,\ C_{\text{miss}} = 1$

[1] J. S. Chung et al. *VoxCeleb2: Deep Speaker Recognition*. Interspeech, 2018.
[2] J. S. Chung et al. *In Defence of Metric Learning for Speaker Recognition*. Interspeech, 2020.
[3] B. Desplanques et al. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech, 2020.

I – Self-Supervised Learning for Speaker Verification SSLSV

Experimental Setup

Datasets

VoxCeleb [1] is a large-scale audio dataset consisting of speech extracted from interview videos posted on YouTube.

Train data

- VoxCeleb2 (1,092,009 utterances from 5,994 speakers)
- Speaker labels are discarded
- 2 seconds audio and 40-d log-mel spectrogram features
- Augmentations: reverberation + noise

Test data

- VoxCeleb1 (148,642 utterances from 1,211 speakers)
- Trials: Original (O), Extended (E) and Hard (H)

Models & Training

Encoder: Fast ResNet-34 [2] / ECAPA-TDNN (C=1024) [3]

Training: Adam, batch size of 256, 100 epochs

Toolkit: sslsv (https://github.com/theolepage/sslsv)

GPUs: 2× and 4× NVIDIA V100 32GB *(Jean Zay, CNRS/IDRIS)*

Evaluation

Scoring: Cosine similarity of representations

Metrics

↓ : lower is better

- EER (Equal Error Rate) ↓
- minDCF (minimum Detection Cost Function) ↓
  ↳ $P_{\text{target}} = 0.01,\ C_{\text{fa}} = 1,\ C_{\text{miss}} = 1$

[1] J. S. Chung et al. *VoxCeleb2: Deep Speaker Recognition*. Interspeech, 2018.
[2] J. S. Chung et al. *In Defence of Metric Learning for Speaker Recognition*. Interspeech, 2020.
[3] B. Desplanques et al. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech, 2020.

I – Self-Supervised Learning for Speaker Verification

SSLSV

Experimental Setup

Datasets

VoxCeleb [1] is a large-scale audio dataset consisting of speech extracted from interview videos posted on YouTube.

Train data

- VoxCeleb2 (1,092,009 utterances from 5,994 speakers)
- Speaker labels are discarded
- 2 seconds audio and 40-d log-mel spectrogram features
- Augmentations: reverberation + noise

Test data

- VoxCeleb1 (148,642 utterances from 1,211 speakers)
- Trials: Original (O), Extended (E) and Hard (H)

Models & Training

Encoder: Fast ResNet-34 [2] / ECAPA-TDNN (C=1024) [3]

Training: Adam, batch size of 256, 100 epochs

Toolkit: sslsv (https://github.com/theolepage/sslsv)

GPUs: 2× and 4× NVIDIA V100 32GB *(Jean Zay, CNRS/IDRIS)*

Evaluation

Scoring: Cosine similarity of representations

Metrics

↓ : lower is better

- EER (Equal Error Rate) ↓
- minDCF (minimum Detection Cost Function) ↓
  ↳ $P_{\text{target}} = 0.01,\ C_{\text{fa}} = 1,\ C_{\text{miss}} = 1$

[1] J. S. Chung et al. *VoxCeleb2: Deep Speaker Recognition*. Interspeech, 2018.
[2] J. S. Chung et al. *In Defence of Metric Learning for Speaker Recognition*. Interspeech, 2020.
[3] B. Desplanques et al. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech, 2020.

I – Self-Supervised Learning for Speaker Verification

SSLSV

Study of Self-Supervised Mechanisms

- **Data-Augmentation:** Independent anchor-positive augmentation is fundamental for SSL as it enables modeling intra-speaker variability.
- **Training Distribution:** MoCo benefits from more speakers *(inter-speaker modeling)*, while DINO benefits from more samples per speaker *(intra-speaker modeling)*.
- **Projector:** The projector mitigates misalignment between pretext and downstream objectives, benefiting non-contrastive frameworks.
- **Positive Sampling:** Positive sampling critically shapes the information encoded in learned representations by defining invariant factors.

Framework	*No data-aug.* ✗		*Data-aug.* ✓	
	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	23.11	0.8224	9.05	0.6364
MoCo	20.27	0.8084	8.49	0.5990
SwAV	30.29	0.8285	11.82	0.7177
VICReg	30.02	0.8362	11.33	0.6658
DINO	28.91	0.8306	6.04	0.4526
Supervised	2.33	0.2092	2.95	0.3122

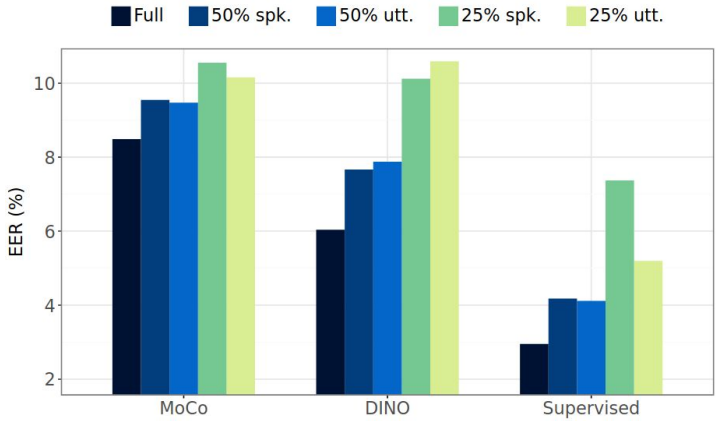
Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

I – Self-Supervised Learning for Speaker Verification

SSLSV

Study of Self-Supervised Mechanisms

- **Data-Augmentation:** Independent anchor-positive augmentation is fundamental for SSL as it enables modeling intra-speaker variability.
- **Training Distribution:** MoCo benefits from more speakers *(inter-speaker modeling)*, while DINO benefits from more samples per speaker *(intra-speaker modeling)*.
- **Projector:** The projector mitigates misalignment between pretext and downstream objectives, benefiting non-contrastive frameworks.
- **Positive Sampling:** Positive sampling critically shapes the information encoded in learned representations by defining invariant factors.



Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

I – Self-Supervised Learning for Speaker Verification

SSLSV

Study of Self-Supervised Mechanisms

- **Data-Augmentation:** Independent anchor-positive augmentation is fundamental for SSL as it enables modeling intra-speaker variability.
- **Training Distribution:** MoCo benefits from more speakers *(inter-speaker modeling)*, while DINO benefits from more samples per speaker *(intra-speaker modeling)*.
- **Projector:** The projector mitigates misalignment between pretext and downstream objectives, benefiting non-contrastive frameworks.
- **Positive Sampling:** Positive sampling critically shapes the information encoded in learned representations by defining invariant factors.

Framework	Positive sampling	*Without proj.* EER (%)	*With proj.* EER (%)	Δ
SimCLR	SSL	**9.05**	10.91	−1.86
MoCo		**8.49**	12.46	−3.97
SwAV		13.07	**11.82**	1.25
VICReg		20.53	**11.33**	9.20
DINO		14.23	**6.04**	8.19
SimCLR	Supervised	**3.66**	6.19	−2.53
MoCo		**3.73**	6.72	−2.99
SwAV		9.78	**8.17**	1.61
VICReg		13.46	**6.63**	6.83
DINO		8.90	**4.71**	4.19

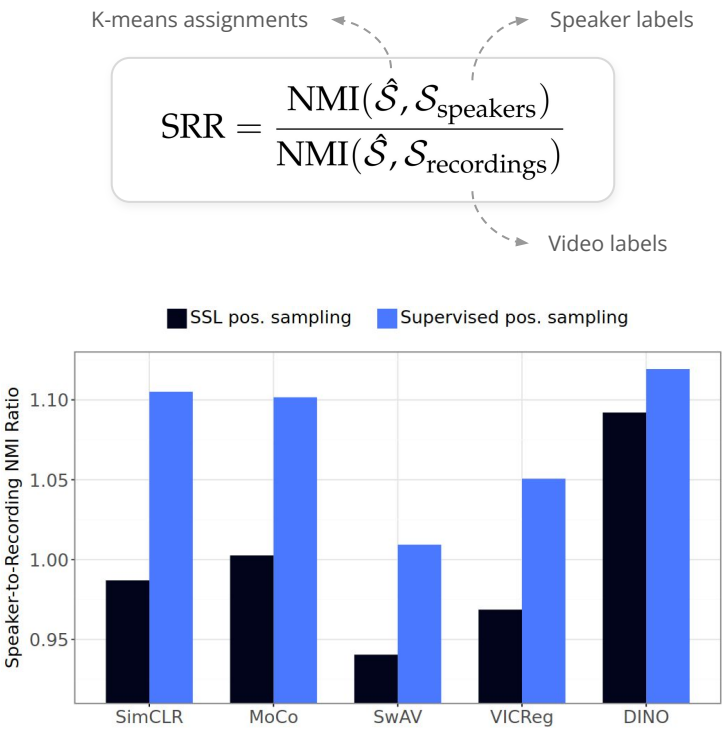
Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

I – Self-Supervised Learning for Speaker Verification

SSLSV

Study of Self-Supervised Mechanisms

- **Data-Augmentation:** Independent anchor-positive augmentation is fundamental for SSL as it enables modeling intra-speaker variability.
- **Training Distribution:** MoCo benefits from more speakers *(inter-speaker modeling)*, while DINO benefits from more samples per speaker *(intra-speaker modeling)*.
- **Projector:** The projector mitigates misalignment between pretext and downstream objectives, benefiting non-contrastive frameworks.
- **Positive Sampling:** Positive sampling critically shapes the information encoded in learned representations by defining invariant factors.



Clustering on VoxCeleb1 with K=5994 • **Encoder:** Fast ResNet-34

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Frameworks

Framework	Encoder	FLOPs (G)	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Random			42.05	0.9969	42.38	0.9994	43.94	0.9997
LIM	Fast ResNet-34	1.82	16.13	0.9015	16.67	0.9778	25.41	0.9931
CPC		1.92	12.77	0.8033	13.75	0.8526	21.47	0.9048
SimCLR		1.80	9.05	0.6364	9.79	0.6769	15.21	0.7664
MoCo		3.66	8.49	0.5990	9.21	0.6588	14.25	0.7503
DeepCluster		1.84	15.16	0.8193	16.53	0.8941	23.40	0.9437
SwAV		1.84	11.82	0.7177	13.25	0.8301	20.19	0.9100
W-MSE		1.80	14.62	0.8506	15.98	0.9342	24.83	0.9810
Barlow Twins		1.82	13.22	0.7658	14.31	0.8424	20.54	0.9080
VICReg		1.82	11.33	0.6658	12.75	0.7581	18.92	0.8577
BYOL		3.66	13.99	0.7509	14.94	0.8498	21.85	0.9181
SimSiam		1.84	28.94	0.9984	27.80	0.9997	41.53	0.9997
DINO		10.90	**6.04**	**0.4526**	**6.92**	**0.5272**	**10.54**	**0.6456**
Supervised		0.91	2.95	0.3122	3.01	0.3475	5.45	0.4993
SimCLR	ECAPA-TDNN	14.94	6.41	0.5160	6.91	0.5616	11.06	0.6708
MoCo		29.95	6.48	0.5372	6.77	0.5635	11.01	0.6700
SwAV		14.98	8.12	0.6148	9.12	0.7119	15.54	0.8282
VICReg		14.97	7.42	0.5659	8.75	0.6970	15.01	0.8518
DINO		90.87	**2.82**	**0.3463**	**3.03**	**0.3923**	**5.93**	**0.5756**
Supervised		7.48	1.34	0.1521	1.49	0.1736	2.84	0.2887
Fusion			**2.60**	**0.2886**	**2.88**	**0.3426**	**5.19**	**0.4997**

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Frameworks

Framework	Encoder	FLOPs (G)	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Random			42.05	0.9969	42.38	0.9994	43.94	0.9997
LIM	Fast ResNet-34	1.82	16.13	0.9015	16.67	0.9778	25.41	0.9931
CPC		1.92	12.77	0.8033	13.75	0.8526	21.47	0.9048
SimCLR		1.80	9.05	0.6364	9.79	0.6769	15.21	0.7664
MoCo		3.66	8.49	0.5990	9.21	0.6588	14.25	0.7503
DeepCluster		1.84	15.16	0.8193	16.53			
SwAV		1.84	11.82	0.7177	13.25			
W-MSE		1.80	14.62	0.8506	15.98			
Barlow Twins		1.82	13.22	0.7658	14.31			
VICReg		1.82	11.33	0.6658	12.75	0.7581	18.92	0.8577
BYOL		3.66	13.99	0.7509	14.94	0.8498	21.85	0.9181
SimSiam		1.84	28.94	0.9984	27.80	0.9997	41.53	0.9997
DINO		10.90	**6.04**	**0.4526**	6.92	0.5272	10.54	0.6456
Supervised		0.91	2.95	0.3122	3.01	0.3475	5.45	0.4993
SimCLR	ECAPA-TDNN	14.94	6.41	0.5160	6.91	0.5616	11.06	0.6708
MoCo		29.95	6.48	0.5372				
SwAV		14.98	8.12	0.6148				
VICReg		14.97	7.42	0.5659				
DINO		90.87	**2.82**	**0.3463**				
Supervised		7.48	1.34	0.1521	1.49	0.1736	2.84	0.2887
Fusion			2.60	0.2886	2.88	0.3426	5.19	0.4997

Scaling the encoder capacity improves downstream performance, especially for DINO

SimCLR and MoCo stand out due to their competitive results and lower computational requirements

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Frameworks

Framework	Encoder	FLOPs (G)	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$minDCF_{0.01}$	EER (%)	$minDCF_{0.01}$	EER (%)	$minDCF_{0.01}$
Random			42.05	0.9969	42.38	0.9994	43.94	0.9997
LIM	Fast ResNet-34	1.82	16.13	0.9015	16.67	0.9778	25.41	0.9931
CPC		1.92	12.77	0.8033	13.75	0.8526	21.47	0.9048
SimCLR		1.80	9.05	0.6364	9.79	0.6769	15.21	0.7664
MoCo		3.66	8.49	0.5990	9.21	0.6588	14.25	0.7503
DeepCluster		1.84	15.16	0.8193	16.53	0.8941	23.40	0.9437
SwAV		1.84	11.82	0.7177	13.25	0.8301	20.19	0.9100
W-MSE		1.80	14.62	0.8506	15.98	0.9342	24.83	0.9810
Barlow Twins		1.82	13.22	0.7658	14.31	0.8424	20.54	0.9080
VICReg		1.82	11.33	0.6658	12.75	0.7581	18.92	0.8577
BYOL		3.66	13.99	0.7509	14.94	0.8498	21.85	0.9181
SimSiam		1.84	28.94	0.9984	27.80	0.9997	41.53	0.9997
DINO		10.90	6.04	0.4526	6.92	0.5272	10.54	0.6456
Supervised		0.91	2.95	0.3122	3.01	0.3475	5.45	0.4993
SimCLR	ECAPA-TDNN	14.94	6.41	0.5160	6.91	0.5616	11.06	0.6708
MoCo		29.95	6.48	0.5372	6.77	0.5635	11.01	0.6700
SwAV		14.98	8.12	0.6148	9.12	0.7119	15.54	0.8282
VICReg		14.97	7.42	0.5659	8.75	0.6970	15.01	0.8518
DINO		90.87	2.82	0.3463	**3.03**	**0.3923**	**5.93**	**0.5756**
Supervised		7.48	1.34	0.1521	1.49	0.1736	2.84	0.2887
Fusion			2.60	0.2886	2.88	0.3426	5.19	0.4997

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Frameworks

Framework	Encoder	FLOPs (G)	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Random			42.05	0.9969	42.38	0.9994	43.94	0.9997
LIM	Fast ResNet-34	1.82	16.13	0.9015	16.67	0.9778	25.41	0.9931
CPC		1.92	12.77	0.8033	13.75	0.8526	21.47	0.9048
SimCLR		1.80	9.05	0.6364	9.79	0.6769	15.21	0.7664
MoCo		3.66	8.49	0.5990	9.21	0.6588	14.25	0.7503
DeepCluster		1.84	15.16	0.8193	16.53	0.8941	23.40	0.9437
SwAV		1.84	11.82	0.7177	13.25	0.8301	20.19	0.9100
W-MSE		1.80	14.62	0.8506	15.98	0.9342	24.83	0.9810
Barlow Twins		1.82	13.22	0.7658	14.31	0.8424	20.54	0.9080
VICReg		1.82	11.33	0.6658	12.75	0.7581	18.92	0.8577
BYOL		3.66	13.99	0.7509	14.94	0.8498	21.85	0.9181
SimSiam		1.84	28.94	0.9984	27.80	0.9997	41.53	0.9997
DINO		10.90	6.04	0.4526	6.92	0.5272	10.54	0.6456
Supervised		0.91	2.95	0.3122	3.01	0.3475	5.45	0.4993
SimCLR	ECAPA-TDNN	14.94	6.41	0.5160	6.91	0.5616	11.96	0.6708
MoCo		29.95	6.48	0.5372	6.[illegible]			
SwAV		14.98	8.12	0.6148	9.[illegible]			
VICReg		14.97	7.42	0.5659	8.[illegible]			
DINO		90.87	**2.82**	**0.3463**	3.[illegible]			
Supervised		7.48	1.34	0.1521	1.49	0.1736	2.84	0.2887
Fusion			**2.60**	**0.2886**	2.88	0.3426	5.19	0.4997

Consistent ranking on out-of-domain benchmarks (SITW, VOiCES)

Manuscript: Table 5.13

Notable complementarity of SSL systems

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Frameworks

Framework	Encoder	FLOPs (G)	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Random			42.05	0.9969	42.38	0.9994	43.94	0.9997
LIM	Fast ResNet-34	1.82	16.13	0.9015	16.67	0.9778	25.41	0.9931
CPC		1.92	12.77	0.8033	13.75	0.8526	21.47	0.9048
SimCLR		1.80	9.05	0.6364	9.79			
MoCo		3.66	8.49	0.5990	9.21			
DeepCluster		1.84	15.16	0.8193	16.53			
SwAV		1.84	11.82	0.7177	13.25			
W-MSE		1.80	14.62	0.8506	15.98			
Barlow Twins		1.82	13.22	0.7658	14.31			
VICReg		1.82	11.33	0.6658	12.75			
BYOL		3.66	13.99	0.7509	14.94			
SimSiam		1.84	28.94	0.9984	27.80	0.9997	41.53	0.9997
DINO		10.90	6.04	0.4526	6.92	0.5272	10.54	0.6456
Supervised		0.91	2.95	0.3122	3.01	0.3475	5.45	0.4993
SimCLR	ECAPA-TDNN	14.94	6.41	0.5160	6.91	0.5616	11.06	0.6708
MoCo		29.95	6.48	0.5372	6.77	0.5635	11.01	0.6700
SwAV		14.98	8.12	0.6148	9.12	0.7119	15.54	0.8282
VICReg		14.97	7.42	0.5659	8.75	0.6970	15.01	0.8518
DINO		90.87	**2.82**	**0.3463**	3.03	0.3923	5.93	0.5756
Supervised		7.48	1.34	0.1521	1.49	0.1736	2.84	0.2887
Fusion			**2.60**	**0.2886**	2.88	0.3426	5.19	0.4997

Reference points

- Random baseline: ~42% EER
- Crowd-sourced annotators: ~27% EER
- Group of SR researchers: ~16% EER

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Frameworks

Framework	Encoder	FLOPs (G)	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Random			42.05	0.9969	42.38	0.9994	43.94	0.9997
LIM	Fast ResNet-34	1.82	16.13	0.9015	16.67	0.9778	25.41	0.9931
CPC		1.92	12.77	0.8033	13.75	0.8526	21.47	0.9048
SimCLR		1.80	9.05	0.6364	9.79			
MoCo		3.66	8.49	0.5990	9.21			
DeepCluster		1.84	15.16	0.8193	16.53			
SwAV		1.84	11.82	0.7177	13.25			
W-MSE		1.80	14.62	0.8506	15.98			
Barlow Twins		1.82	13.22	0.7658	14.31			
VICReg		1.82	11.33	0.6658	12.75			
BYOL		3.66	13.99	0.7509	14.94			
SimSiam		1.84	28.94	0.9984	27.80	0.9997	41.53	0.9997
DINO		10.90	6.04	0.4526	6.92	0.5272	10.54	0.6456
Supervised		0.91	2.95	0.3122	3.01	0.3475	5.45	0.4993
SimCLR	ECAPA-TDNN	14.94	6.41	0.5160	6.91	0.5616	11.06	0.6708
MoCo		29.95	6.48	0.5372				
SwAV		14.98	8.12	0.6148				
VICReg		14.97	7.42	0.5659				
DINO		90.87	**2.82**	**0.3463**				
Supervised		7.48	1.34	0.1521	1.49	0.1736	2.84	0.2887
Fusion			**2.60**	**0.2886**	2.88	0.3426	5.19	0.4997

Reference points

- Random baseline: ~42% EER
- Crowd-sourced annotators: ~27% EER
- Group of SR researchers: ~16% EER

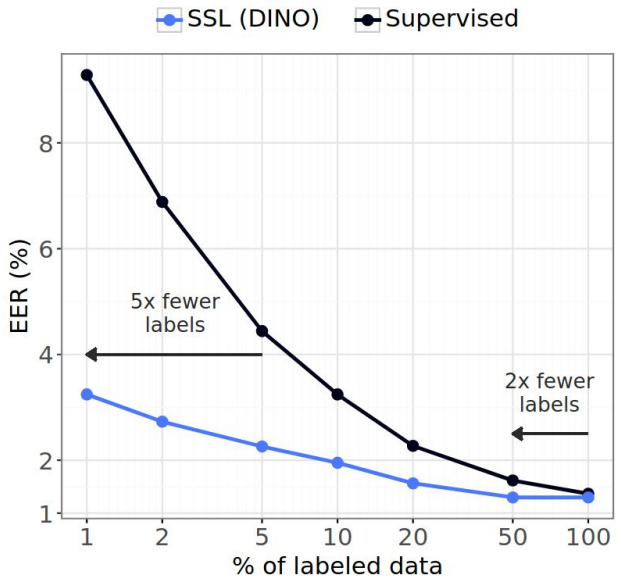
Despite the promising performance of SSL, supervised pre-training still yields the best performance

I – Self-Supervised Learning for Speaker Verification

SSLSV

Label-Efficiency Study

- SSL (DINO) consistently outperforms supervised training across all proportions of labeled data on VoxCeleb1-O.
- Comparable performance achieved with 2× to 5× fewer labels, demonstrating robustness in low-resource scenarios.
- Pre-training on unlabeled speech substantially reduces reliance on manual annotations for SV.



Benchmark: VoxCeleb1-O • Encoder: ECAPA-TDNN

I – Self-Supervised Learning for Speaker Verification SSLSV

Conclusions

- Best downstream performance is achieved by DINO, SimCLR and MoCo on VoxCeleb1-O, although a gap with fully supervised system persists.
- Data-augmentation and positive sampling are fundamental to learning speaker-relevant information, while the projector mitigates misalignment between pretext and downstream objectives.
- Contrastive methods primarily encode inter-speaker variability, whereas self-distillation methods focuses on intra-speaker variability and are more sensitive to training conditions.
- Label-efficient evaluations confirms the promise of SSL, matching supervised performance with 2× fewer labels.

Margins

Part II

Learning Discriminative Speaker Representations

Related Publications

Experimenting with Additive Margins for Contrastive Self-Supervised Speaker Verification
Interspeech, 2023

Additive Margin in Contrastive Self-Supervised Frameworks to Learn Discriminative Speaker Representations
Odyssey Workshop, 2024

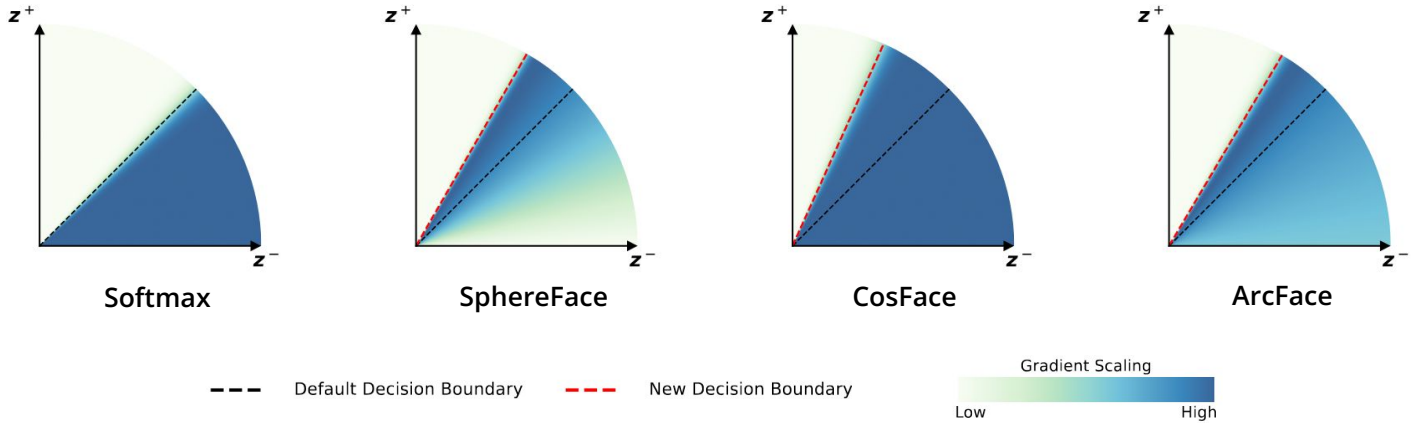
II – Learning Discriminative Speaker Representations

Margins

Introduction

Margin-based losses enforce class compactness and separation by constraining classification decision boundaries.

Initially developed for face recognition (e.g., CosFace [1], ArcFace [2]), they have been successfully transferred to SR [3].



[1] H. Wang et al. *CosFace: Large Margin Cosine Loss for Deep Face Recognition*. CVPR, 2018.
[2] J. Deng et al. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*. CVPR, 2019.
[3] Y. Liu et al. *Large Margin Softmax Loss for Speaker Verification*. Interspeech, 2019.

II – Learning Discriminative Speaker Representations

Margins

Overview of the Method

The self-supervised contrastive loss is reformulated by decoupling positive and negative similarities, enabling the integration of SphereFace [1], CosFace [2], ArcFace [3], CurricularFace [4], MagFace [5], and AdaFace [6].

$$\mathcal{L}_{\text{SimCLR}} = -\frac{1}{2B}\sum_{i\in\hat{\mathcal{B}}}\log\frac{\exp\left(\ell^{+}(z_i, z_i')\right)}{\exp\left(\ell^{+}(z_i, z_i')\right) + \sum_{j\in\hat{\mathcal{J}}(i)}\exp\left(\ell^{-}(z_i, z_j)\right)}$$

Without any margin *(default)*

$$\ell^{+}(a,b) = \ell^{-}(a,b) = \cos\left(\theta_{a,b}\right)/\tau$$

$$\mathcal{L}_{\text{MoCo}} = -\frac{1}{B}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\ell^{+}(z_i, z_i')\right)}{\exp\left(\ell^{+}(z_i, z_i')\right) + \sum_{j\in\mathcal{J}}\exp\left(\ell^{-}(z_i, q_j)\right)}$$

With CosFace

AM

$$\mathcal{L}^{(\text{Cos})}_{\text{SimCLR}},\ \mathcal{L}^{(\text{Cos})}_{\text{MoCo}}: \begin{cases} \ell^{+}(a,b) = \left(\cos\left(\theta_{a,b}\right) - m\right)/\tau \\ \ell^{-}(a,b) = \cos\left(\theta_{a,b}\right)/\tau \end{cases}$$

[1] W. Liu et al. *SphereFace: Deep Hypersphere Embedding for Face Recognition*. CVPR, 2017.
[2] H. Wang et al. *CosFace: Large Margin Cosine Loss for Deep Face Recognition*. CVPR, 2018.
[3] J. Deng et al. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*. CVPR, 2019.
[4] Y. Huang et al. *CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition*. CVPR, 2020.
[5] Q. Meng et al. *MagFace: A Universal Representation for Face Recognition and Quality Assessment*. CVPR, 2021.
[6] M. Kim et al. *AdaFace: Quality Adaptive Margin for Face Recognition*. CVPR, 2022.

Manuscript: Section 6.3

II – Learning Discriminative Speaker Representations

Margins

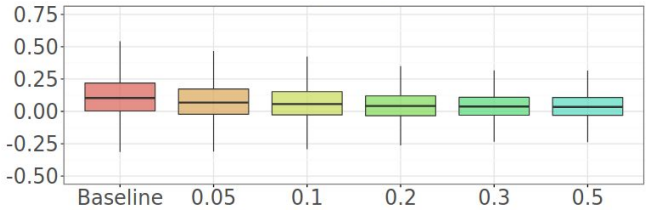
Effect of Margins on Performance & Representations

CosFace

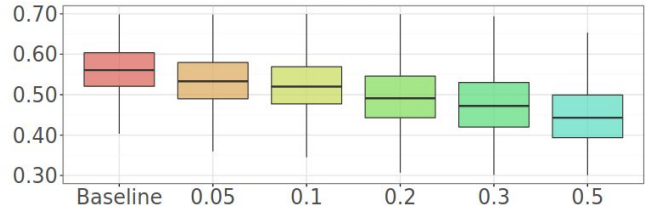
EER (%)

Margin (m)	0	0.05	0.1	0.2	0.3	0.5
EER (%)	9.05	8.79	8.10	**7.80**	8.33	9.16

Inter-speaker similarity

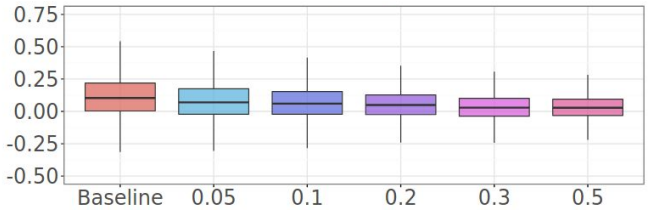


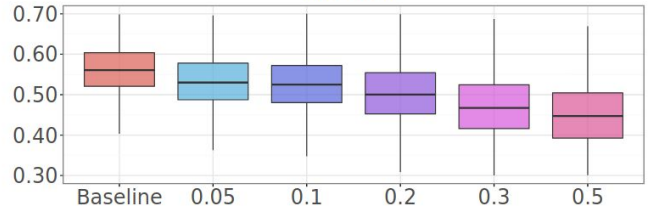
Intra-speaker similarity



AdaFace

Margin (m)	0	0.05	0.1	0.2	0.3	0.5
EER (%)	9.05	8.44	8.07	**7.70**	8.00	8.63





Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

II – Learning Discriminative Speaker Representations

Margins

Effect of Margins on Performance & Representations

	CosFace						AdaFace					
Margin (m)	0	0.05	0.1	0.2	0.3	0.5	0	0.05	0.1	0.2	0.3	0.5
EER (%)	9.05	8.79	8.10	**7.80**	8.33	9.16	9.05	8.44	8.07	**7.70**	8.00	8.63

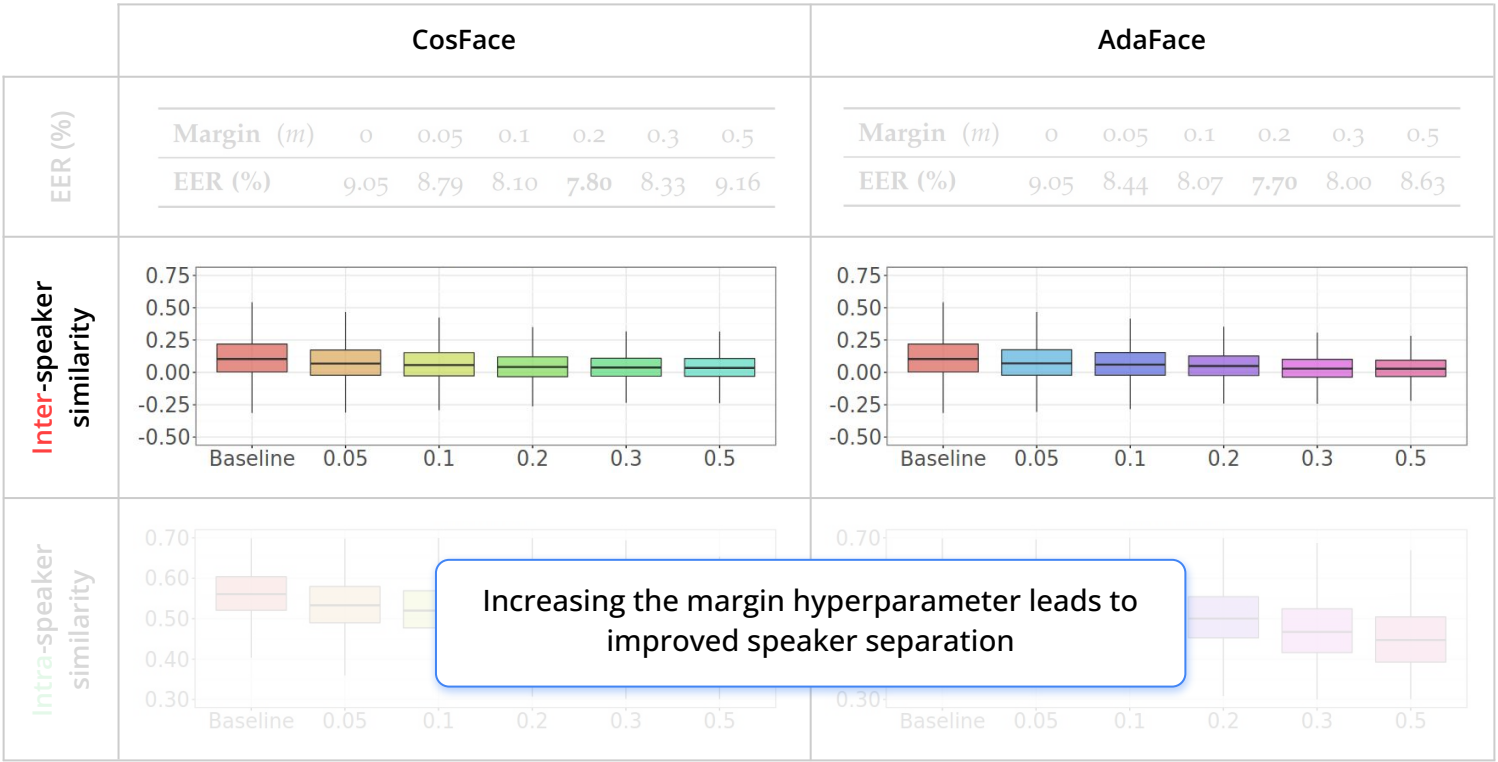
Applying margins to SimCLR (w/ CosFace & AdaFace) consistently improves downstream performance

Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

II – Learning Discriminative Speaker Representations

Margins

Effect of Margins on Performance & Representations



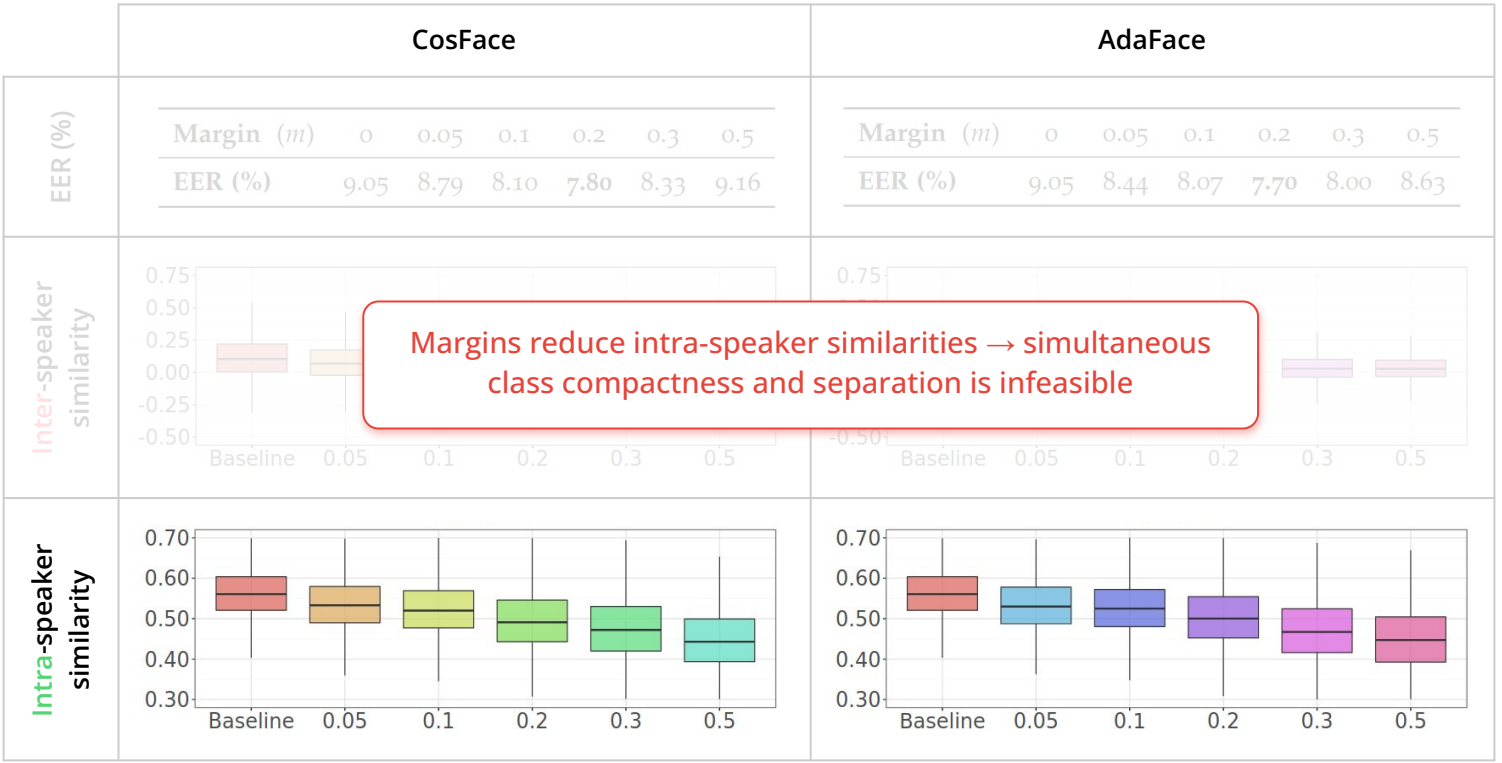
Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

II – Learning Discriminative Speaker Representations

Margins

Effect of Margins on Performance & Representations

	CosFace						AdaFace					
Margin (m)	0	0.05	0.1	0.2	0.3	0.5	0	0.05	0.1	0.2	0.3	0.5
EER (%)	9.05	8.79	8.10	**7.80**	8.33	9.16	9.05	8.44	8.07	**7.70**	8.00	8.63



Margins reduce intra-speaker similarities → simultaneous class compactness and separation is infeasible

Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

II – Learning Discriminative Speaker Representations

Margins

Conclusions

- Margin-based losses can be successfully integrated into self-supervised contrastive frameworks for SV, improving the discriminative capacity of speaker representations.
- SimCLR consistently benefits from margins with CosFace and AdaFace, yielding notable improvements in inter-speaker separation and verification performance.
- The learned representations exhibit high intra-speaker variability, motivating the development of strategies for more effective speaker modeling.

SSPS

Part III

Self-Supervised Positive Sampling from Latent Space

Related Publications

Self-Supervised Frameworks for Speaker Recognition via Bootstrapped Positive Sampling
IEEE TASLP, vol. 33, 2025

SSPS: Self-Supervised Positive Sampling for Robust Self-Supervised Speaker Verification
Interspeech, 2025

III – Self-Supervised Positive Sampling from Latent Space SSPS

Introduction

SSL frameworks encode a significant amount of channel and recording-related information *(e.g., VoxCeleb videos collected "in the wild")* even with data-augmentation because of the standard same-utterance positive sampling.

Framework	Positive sampling	VoxCeleb1-O		Δ (%)
		EER (%)	$\text{minDCF}_{0.01}$	
SimCLR	SSL	6.30	0.5286	72.62
	Supervised	**1.72**	**0.2395**	
MoCo	SSL	6.20	0.5501	71.77
	Supervised	**1.75**	**0.2547**	
SwAV	SSL	7.97	0.6097	45.06
	Supervised	**4.38**	**0.4837**	
VICReg	SSL	7.70	0.5883	41.30
	Supervised	**4.52**	**0.4993**	
DINO	SSL	3.07	0.3616	23.14
	Supervised	**2.36**	**0.2712**	
Supervised		1.34	0.1521	

Benchmark: VoxCeleb1-O • Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Introduction

SSL frameworks encode a significant amount of channel and recording-related information *(e.g., VoxCeleb videos collected "in the wild")* even with data-augmentation because of the standard same-utterance positive sampling.

Framework	Positive sampling	VoxCeleb1-O		Δ (%)
		EER (%)	$\text{minDCF}_{0.01}$	
SimCLR	SSL	6.30	0.5286	72.62
	Supervised	**1.72**	**0.2395**	
MoCo	SSL	6.20	0.5501	71.77
	Supervised	**1.75**	**0.2547**	
SwAV	SSL	7.97	0.6097	45.06
	Supervised	**4.38**	**0.4837**	
VICReg	SSL	7.70	0.5883	41.30
	Supervised	**4.52**	**0.4993**	
DINO	SSL	3.07	0.3616	23.14
	Supervised	**2.36**	**0.2712**	
Supervised		1.34	0.1521	

Up to 70% EER reduction with a supervised positive sampling strategy

Benchmark: VoxCeleb1-O • Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Related Work

- AAT [1] – Adversarial loss penalizing the encoding of undesired transformation-related information
- i-mixup [2] – Mix utterances to create diverse synthetic training samples
- DPP [3] – Cross-references speech and face data to select diverse positives
- CA-DINO [4] – Cluster speaker representations to determine appropriate positives

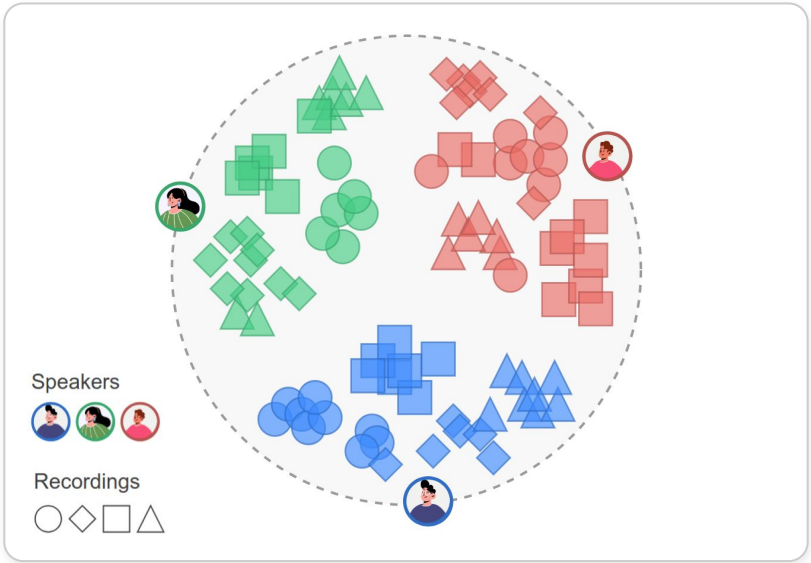
[1] J. Huh et al. *Augmentation Adversarial Training for Self-Supervised Speaker Representation Learning*. NeurIPS Workshop, 2020.
[2] W. H. Kang et al. *Robust Self-Supervised Speaker Representation Learning Via Instance Mix Regularization*. ICASSP, 2022.
[3] R. Tao et al. *Self-Supervised Training of Speaker Encoder With Multi-Modal Diverse Positive Pairs*. IEEE TASLP, 2023.
[4] B. Han et al. *Self-Supervised Learning With Cluster-Aware-DINO for High-Performance Robust Speaker Verification*. IEEE TASLP, 2024.

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Assumption on Latent Space

Assumption: The standard same-utterance positive sampling group samples of the same recording before modeling speaker identities in latent space.

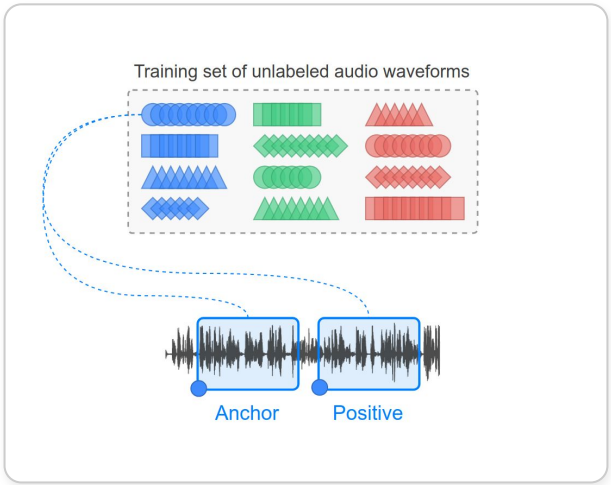


Latent Space of Speaker Representations

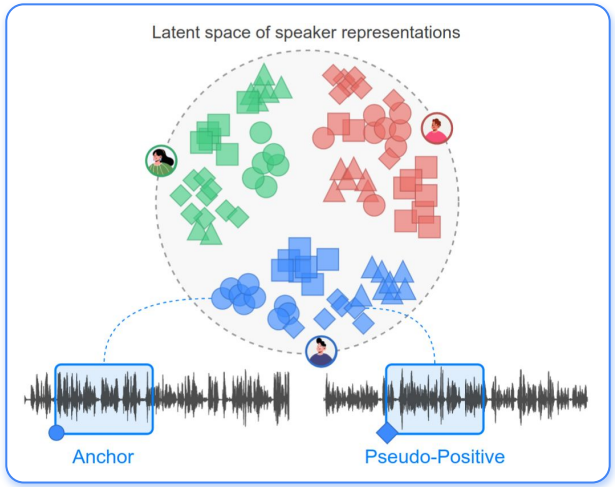
III – Self-Supervised Positive Sampling from Latent Space SSPS

Overview of the Method

Method: Sample positives from the neighborhood of the anchor in latent space to create anchor-positive pairs from different recordings of the same speaker.



Same-Utterance Positive Sampling
(standard)



SSPS: Self-Supervised Positive Sampling
(proposed)

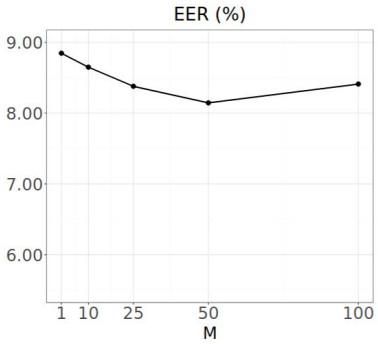
III – Self-Supervised Positive Sampling from Latent Space

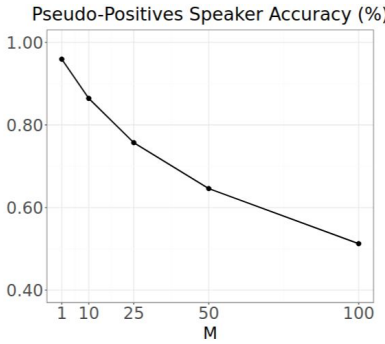
SSPS

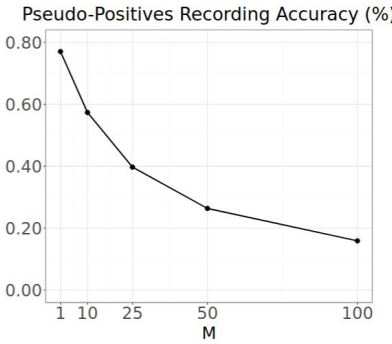
Sampling with Nearest Neighbors & Clustering

Nearest Neighbors (SSPS-NN)

Select a *pseudo-positive* from the M nearest samples from the anchor in latent space.

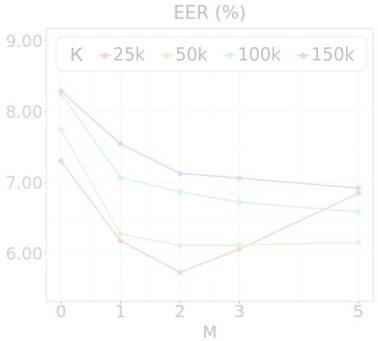


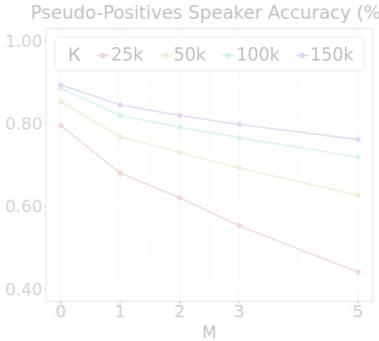


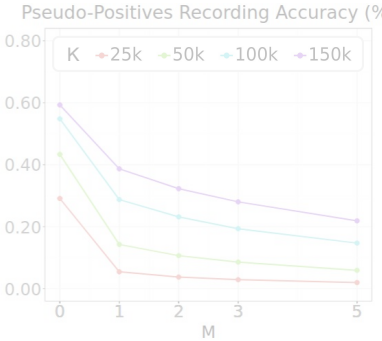


Clustering (SSPS-Clustering)

Cluster representations in K classes and select a *pseudo-positive* from the M nearest clusters around the anchor in latent space.







Baseline at 9.41% EER • **Benchmark:** VoxCeleb1-O • **Encoder:** Fast ResNet-34

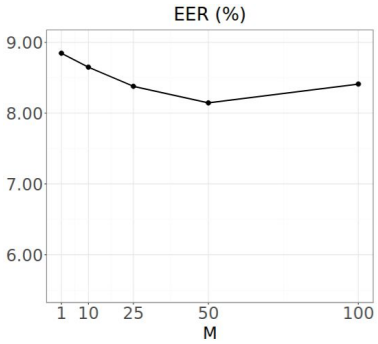
III – Self-Supervised Positive Sampling from Latent Space

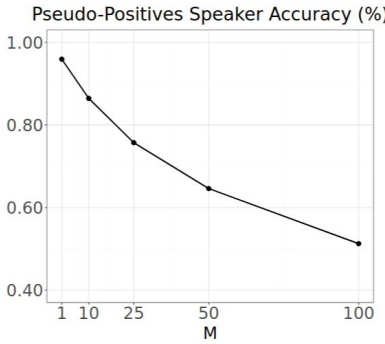
SSPS

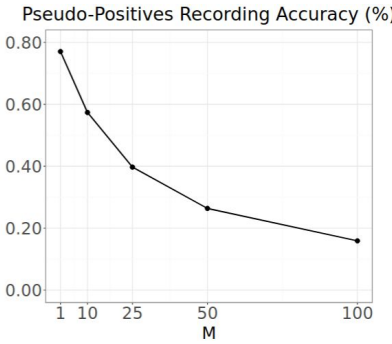
Sampling with Nearest Neighbors & Clustering

Nearest Neighbors (SSPS-NN)

Select a *pseudo-positive* from the M nearest samples from the anchor in latent space.

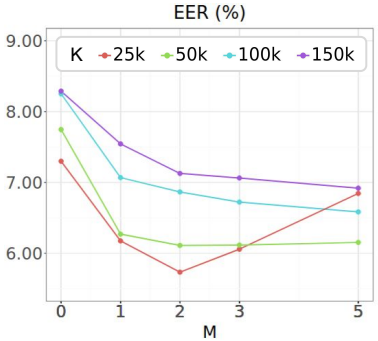


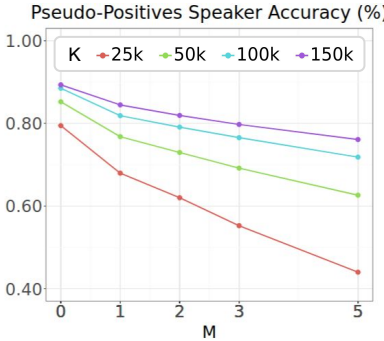


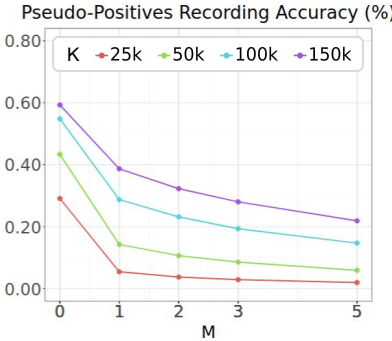


Clustering (SSPS-Clustering)

Cluster representations in K classes and select a *pseudo-positive* from the M nearest clusters around the anchor in latent space.







Baseline at 9.41% EER • **Benchmark:** VoxCeleb1-O • **Encoder:** Fast ResNet-34

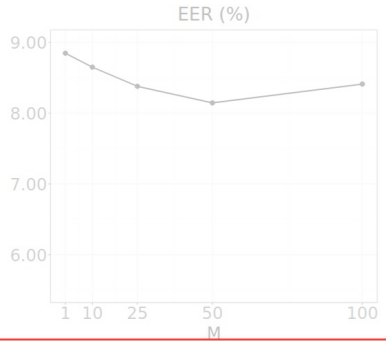
III – Self-Supervised Positive Sampling from Latent Space

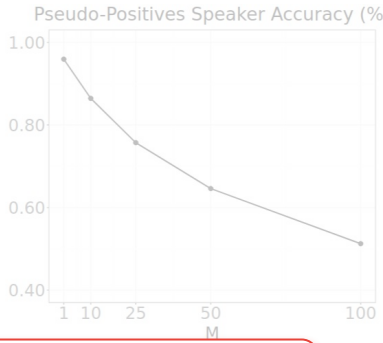
SSPS

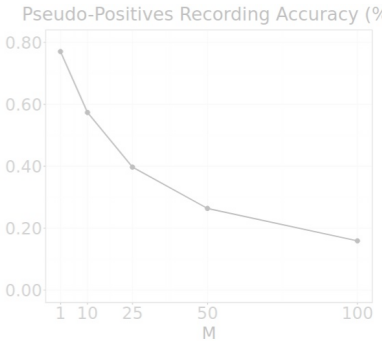
Sampling with Nearest Neighbors & Clustering

Nearest Neighbors (SSPS-NN)

Select a *pseudo-positive* from the M nearest samples from the anchor in latent space.



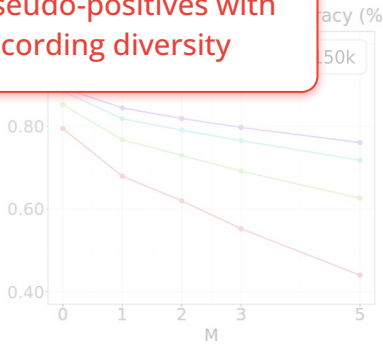


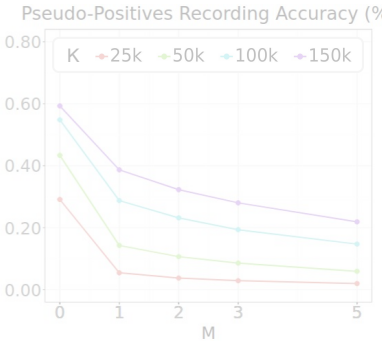


SSPS-Clustering allows sampling pseudo-positives with higher speaker accuracy and recording diversity

Clustering (SSPS-Clustering)

Cluster representations in K classes and select a *pseudo-positive* from the M nearest clusters around the anchor in latent space.





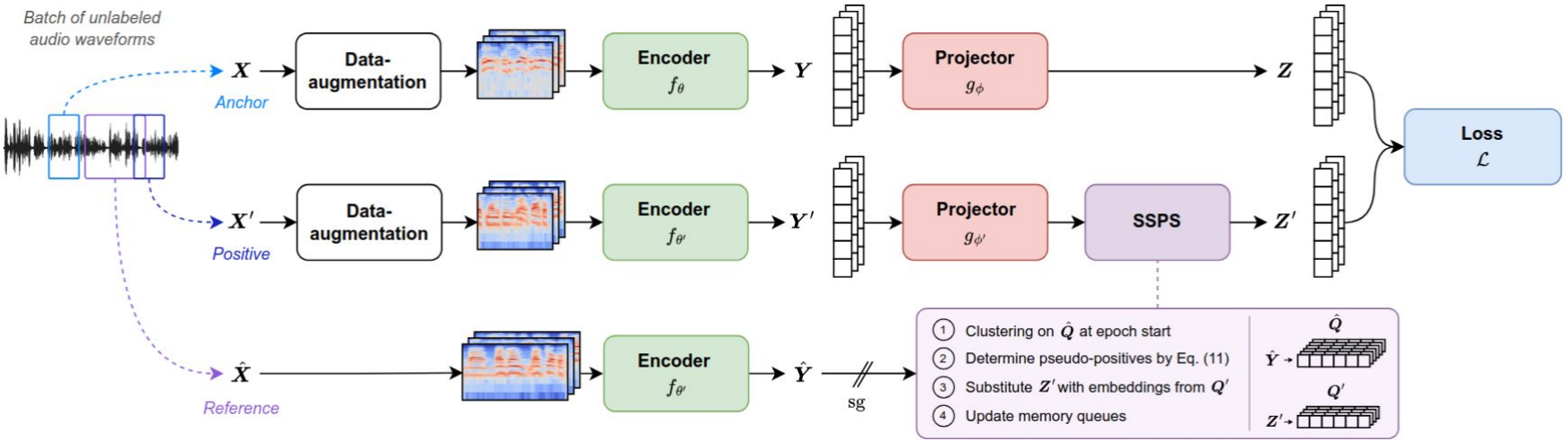
Baseline at 9.41% EER • Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

III – Self-Supervised Positive Sampling from Latent Space SSPS

Overview of the Training Framework

Compared to the standard SSL framework, the positive is substituted by a *pseudo-positive*, which is retrieved from a memory queue of latest embeddings based on the anchor's clustering assignment.

Clustering is performed at the beginning of each training epoch on reference representations *(longer and clean segments)*.



Manuscript: Section 7.3

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Evaluation

Framework	Pos. sampling	Hyper-params.		VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		K	M	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	SSL			6.30	0.5286	6.86	0.5599	10.98	0.6692
	SSPS	6000	0	2.90	0.3206	3.38	0.3292	6.13	0.4887
	SSPS	25 000	1	**2.57**	**0.3033**	**3.11**	**0.3125**	**5.56**	**0.4638**
	Supervised			1.72	0.2395	1.88	0.2314	3.66	0.3641
SwAV	SSL			7.97	0.6097	8.87	0.7052	15.15	0.8273
	SSPS	6000	0	7.07	0.5847	7.96	0.6803	13.93	0.8149
	SSPS	25 000	1	**6.50**	**0.5687**	**7.35**	**0.6507**	**13.03**	**0.8014**
	Supervised			4.38	0.4837	4.92	0.5538	9.49	0.7254
VICReg	SSL			7.70	0.5883	9.05	0.7170	15.25	0.8431
	SSPS	6000	0	7.45	0.5513	8.56	0.6926	14.15	0.8343
	SSPS	25 000	1	**6.95**	**0.5262**	**8.16**	**0.6759**	**13.51**	**0.8263**
	Supervised			4.52	0.4993	5.43	0.6343	10.37	0.7908
DINO	SSL			3.07	0.3616	3.32	0.4003	6.20	0.5731
	SSPS	6000	0	2.78	0.3140	3.07	0.3456	5.66	0.5158
	SSPS	25 000	1	**2.53**	**0.2843**	**2.55**	**0.3150**	**4.93**	**0.4632**
	Supervised			2.36	0.2712	2.41	0.2986	4.64	0.4378
Supervised				1.34	0.1521	1.49	0.1736	2.84	0.2887

Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Evaluation

Framework	Pos. sampling	Hyper-params.		VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		K	M	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	SSL			6.30	0.5286	6.86	0.5599	10.98	0.6692
	SSPS	6000	0	2.90	0.3206	3.38	0.3292	6.13	0.4887
	SSPS	25 000	1	**2.57**	**0.3033**	3.11	0.3125	5.56	0.4638
	Supervised			1.72	0.2395	1.88	0.2314	3.66	0.3641
SwAV	SSL			7.97	0.6097	8.87	0.7052	15.15	0.8273
	SSPS	6000	0	7.07	0.5847	7.96	0.6803	13.93	0.8149
	SSPS	25 000	1	**6.50**	**0.5687**	7.35	0.6507	13.03	0.8014
	Supervised			4.38	0.4837				
VICReg	SSL			7.70	0.5883				
	SSPS	6000	0	7.45	0.5513	8.56	0.6926	14.15	0.8343
	SSPS	25 000	1	**6.95**	**0.5262**	8.16	0.6759	13.51	0.8263
	Supervised			4.52	0.4993	5.43	0.6343	10.37	0.7908
DINO	SSL			3.07	0.3616	3.32	0.4003	6.20	0.5731
	SSPS	6000	0	2.78	0.3140	3.07	0.3456	5.66	0.5158
	SSPS	25 000	1	**2.53**	**0.2843**	2.55	0.3150	4.93	0.4632
	Supervised			2.36	0.2712	2.41	0.2986	4.64	0.4378
Supervised				1.34	0.1521	1.49	0.1736	2.84	0.2887

SSPS outperforms the standard same-utterance positive sampling across all SSL frameworks

Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Evaluation

Framework	Pos. sampling	Hyper-params.		VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		K	M	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	SSL			6.30	0.5286	6.86	0.5599		
	SSPS	6000	0	2.90	0.3206	3.38	0.3292		
	SSPS	25 000	1	**2.57**	**0.3033**	3.11	0.3125		
	Supervised			1.72	0.2395	1.88	0.2314		
SwAV	SSL			7.97	0.6097	8.87	0.7052		
	SSPS	6000	0	7.07	0.5847	7.96	0.6803	13.93	0.8149
	SSPS	25 000	1	**6.50**	**0.5687**				
	Supervised			4.38	0.4837				
VICReg	SSL			7.70	0.5883				
	SSPS	6000	0	7.45	0.5513				
	SSPS	25 000	1	**6.95**	**0.5262**	8.16	0.6759	13.51	0.8263
	Supervised			4.52	0.4993	5.43	0.6343	10.37	0.7908
DINO	SSL			3.07	0.3616	3.32	0.4003	6.20	0.5731
	SSPS	6000	0	2.78	0.3140	3.07	0.3456	5.66	0.5158
	SSPS	25 000	1	**2.53**	**0.2843**	2.55	0.3150	4.93	0.4632
	Supervised			2.36	0.2712	2.41	0.2986	4.64	0.4378
Supervised				1.34	0.1521	1.49	0.1736	2.84	0.2887

VoxCeleb2
- 5,994 speakers
- 145,569 recordings / videos

Sampling from neighboring clusters consistently improves performance → confirms the initial assumption on the latent space

Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Evaluation

Framework	Pos. sampling	Hyper-params.		VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		K	M	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	SSL			6.30	0.5286				
	SSPS	6000	0	2.90	0.3206				
	SSPS	25 000	1	**2.57**	**0.3033**				
	Supervised			1.72	0.2395				
SwAV	SSL			7.97	0.6097	8.87	0.7052	15.15	0.8273
	SSPS	6000	0	7.07	0.5847	7.96	0.6803	13.93	0.8149
	SSPS	25 000	1	6.50	0.5687	7.35	0.6507	13.03	0.8014
	Supervised			4.38	0.4837	4.92	0.5538	9.49	0.7254
VICReg	SSL			7.70	0.5883	9.05	0.7170	15.25	0.8431
	SSPS	6000	0	7.45	0.5513	8.56	0.6926	14.15	0.8343
	SSPS	25 000	1	6.95	0.5262	8.16	0.6759	13.51	0.8263
	Supervised			4.52	0.4993	5.43	0.6343	10.37	0.7908
DINO	SSL			3.07	0.3616				
	SSPS	6000	0	2.78	0.3140				
	SSPS	25 000	1	**2.53**	**0.2843**				
	Supervised			2.36	0.2712				
Supervised				1.34	0.1521	1.49	0.1736	2.84	0.2887

SimCLR-SSPS achieves a 58% reduction in EER, while supervised pos. sampling demonstrates further potential

DINO is outperformed by SimCLR-SSPS, despite relying on a more complex training framework

Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Evaluation

Framework	Pos. sampling	Hyper-params.		VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		K	M	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	SSL			6.30	0.5286	6.86	0.5599	10.98	0.6692
	SSPS	6000	0	2.90	0.3206	3.38	0.3292	6.13	0.4887
	SSPS	25 000	1	2.57	0.3033	**3.11**	**0.3125**	**5.56**	**0.4638**
	Supervised			1.72	0.2395	1.88	0.2314	3.66	0.3641
SwAV	SSL			7.97	0.6097	8.87	0.7052	15.15	0.8273
	SSPS	6000	0	7.07	0.5847	7.96	0.6803	13.93	0.8149
	SSPS	25					.6507	13.03	0.8014
	Supervised						.5538	9.49	0.7254
VICReg	SSL						.7170	15.25	0.8431
	SSPS	6000	0	7.45	0.5513	8.56	0.6926	14.15	0.8343
	SSPS	25 000	1	6.95	0.5262	8.16	0.6759	13.51	0.8263
	Supervised			4.52	0.4993	5.43	0.6343	10.37	0.7908
DINO	SSL			3.07	0.3616	3.32	0.4003	6.20	0.5731
	SSPS	6000	0	2.78	0.3140	3.07	0.3456	5.66	0.5158
	SSPS	25 000	1	2.53	0.2843	**2.55**	**0.3150**	**4.93**	**0.4632**
	Supervised			2.36	0.2712	2.41	0.2986	4.64	0.4378
Supervised				1.34	0.1521	1.49	0.1736	2.84	0.2887

SimCLR-SSPS and DINO-SSPS achieve competitive results on VoxCeleb1-E & VoxCeleb1-H

Encoder: ECAPA-TDNN

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Comparison to State-of-the-Art Methods

Method	VoxCeleb1-O	
	EER (%)	$\text{minDCF}_{0.01}$
Contrastive		
AP + AAT (Huh et al. 2020)	8.65	
SimCLR + MSE loss (Haoran Zhang et al. 2021)	8.28	0.6100
MoCo + ProtoNCE (W. Xia et al. 2021)	8.23	0.5900
C3-MoCo (C. Zhang et al. 2022)	6.40	
DPP (Tao et al. 2023)	2.89	
SimCLR-SSPS	**2.57**	**0.3033**
Self-distillation		
DINO + Cosine loss (Han et al. 2022)	6.16	0.5240
DINO (Jaejin Cho et al. 2022b)	4.83	0.4630
CA-DINO (Han et al. 2024)	3.59	0.3529
RDINO (Y. Chen et al. 2023a)	3.29	
DINO-SSPS	**2.53**	**0.2843**
DINO + Aug. (Zhengyang Chen et al. 2022b)	2.51	
C3-DINO (C. Zhang et al. 2022)	2.50	

- SSPS outperforms contrastive-based methods and demonstrates competitive performance among self-distillation methods.
- This validates the effectiveness of the proposed positive sampling strategy, which addresses a key limitation of existing approaches.

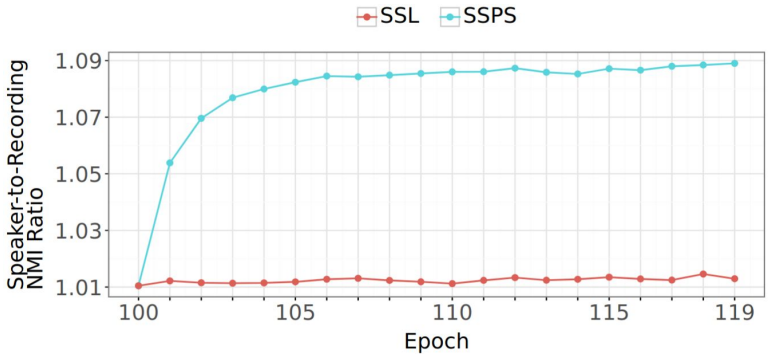
Benchmark: VoxCeleb1-O

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Effect on Learned Representations

- With the standard positive sampling, representations encode comparable amounts of speaker and recording-related information *(SRR ≈ 1.01)*.
- With SSPS, the balance shifts toward speaker identity, yielding an ≈ 8% relative increase in SRR.
- SSPS learns representations that are more robust to extrinsic variability.



K-means assignments

Speaker labels

$$\mathrm{SRR} = \frac{\mathrm{NMI}(\hat{\mathcal{S}}, \mathcal{S}_{\text{speakers}})}{\mathrm{NMI}(\hat{\mathcal{S}}, \mathcal{S}_{\text{recordings}})}$$

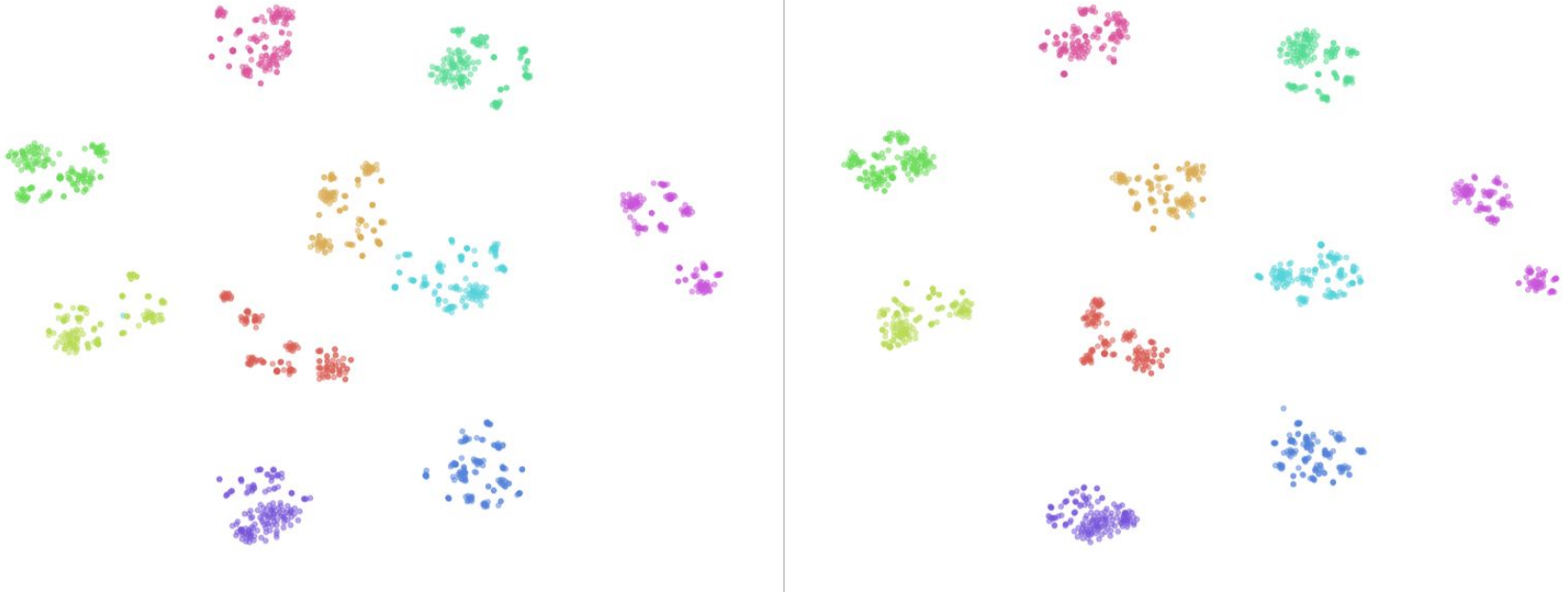
Video labels

Clustering on VoxCeleb2 with K=6000 • **Encoder:** ECAPA-TDNN • **Model:** SimCLR

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Visualisation of Speaker Representations

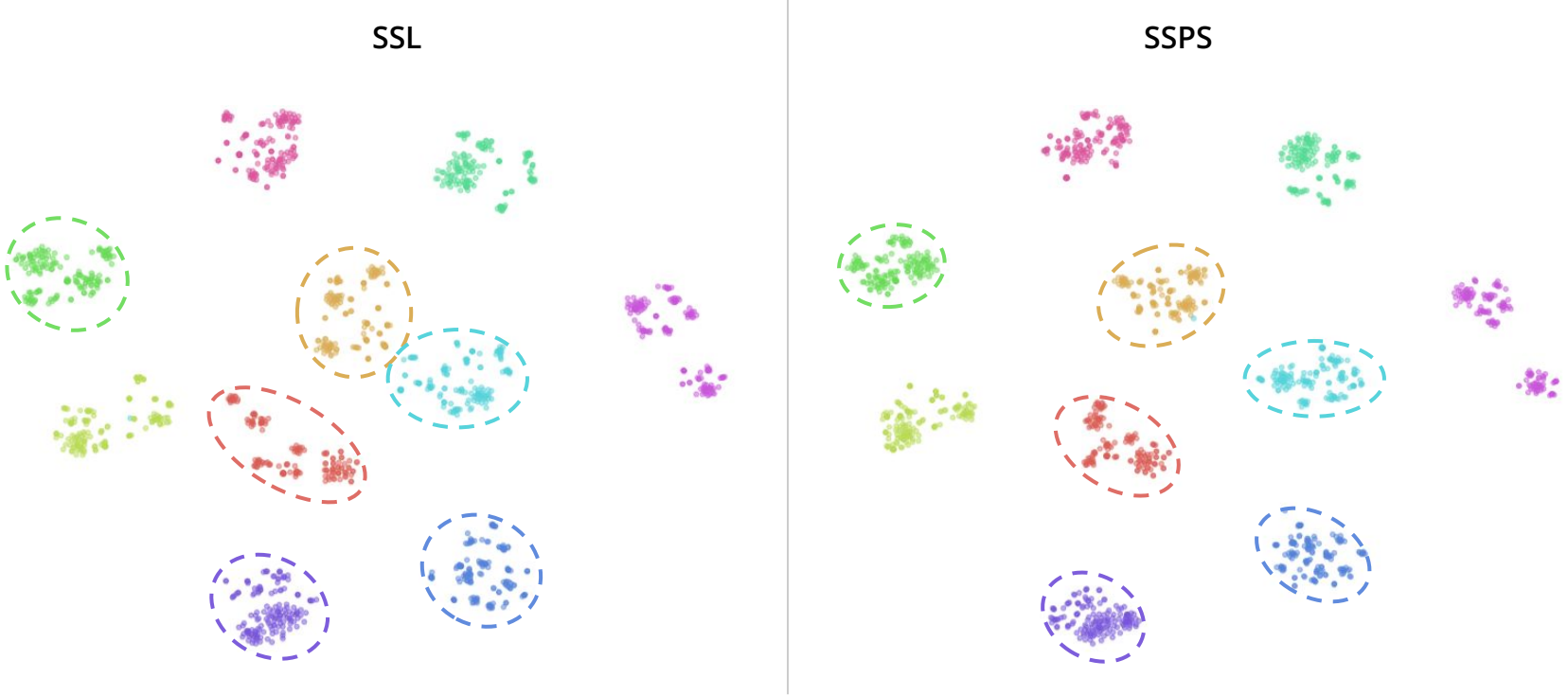


t-SNE on 10 speakers from VoxCeleb1 • **Encoder:** ECAPA-TDNN • **Model:** SimCLR

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Visualisation of Speaker Representations



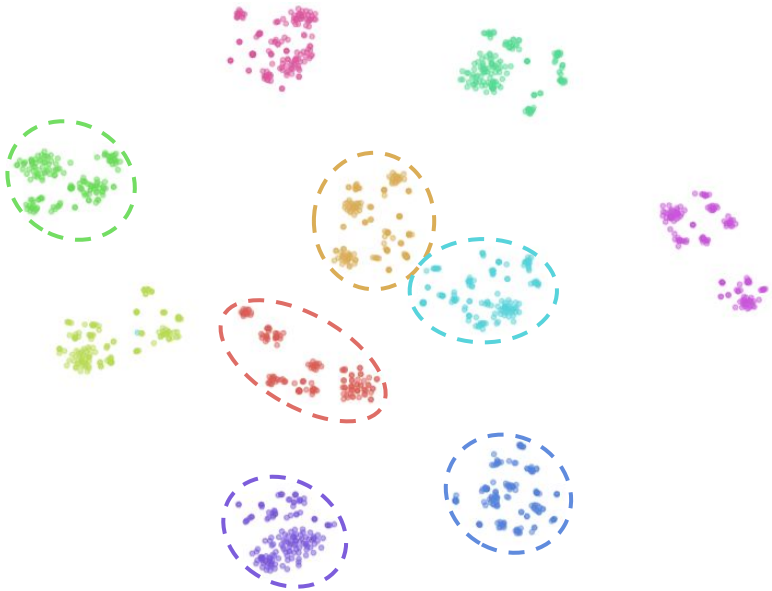
t-SNE on 10 speakers from VoxCeleb1 • **Encoder:** ECAPA-TDNN • **Model:** SimCLR

III – Self-Supervised Positive Sampling from Latent Space

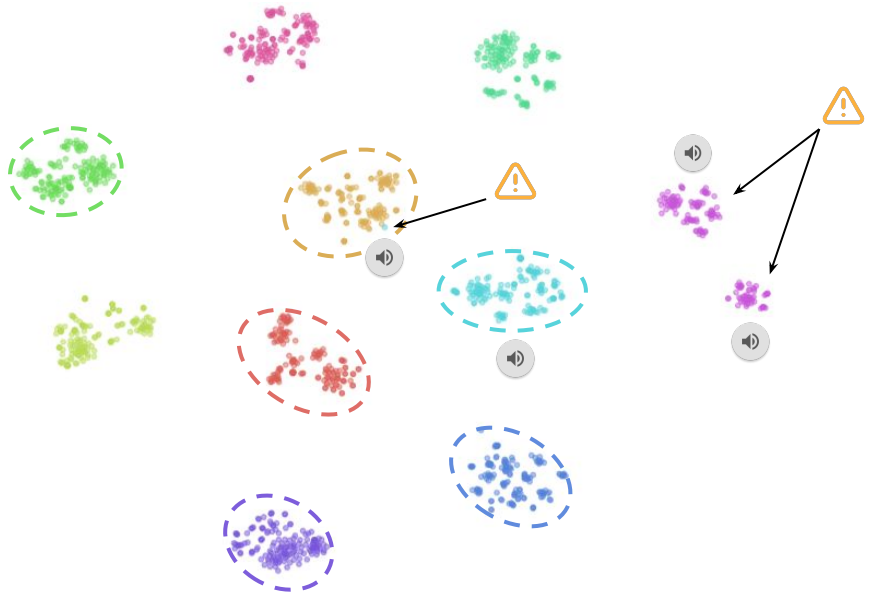
SSPS

Visualisation of Speaker Representations

SSL



SSPS

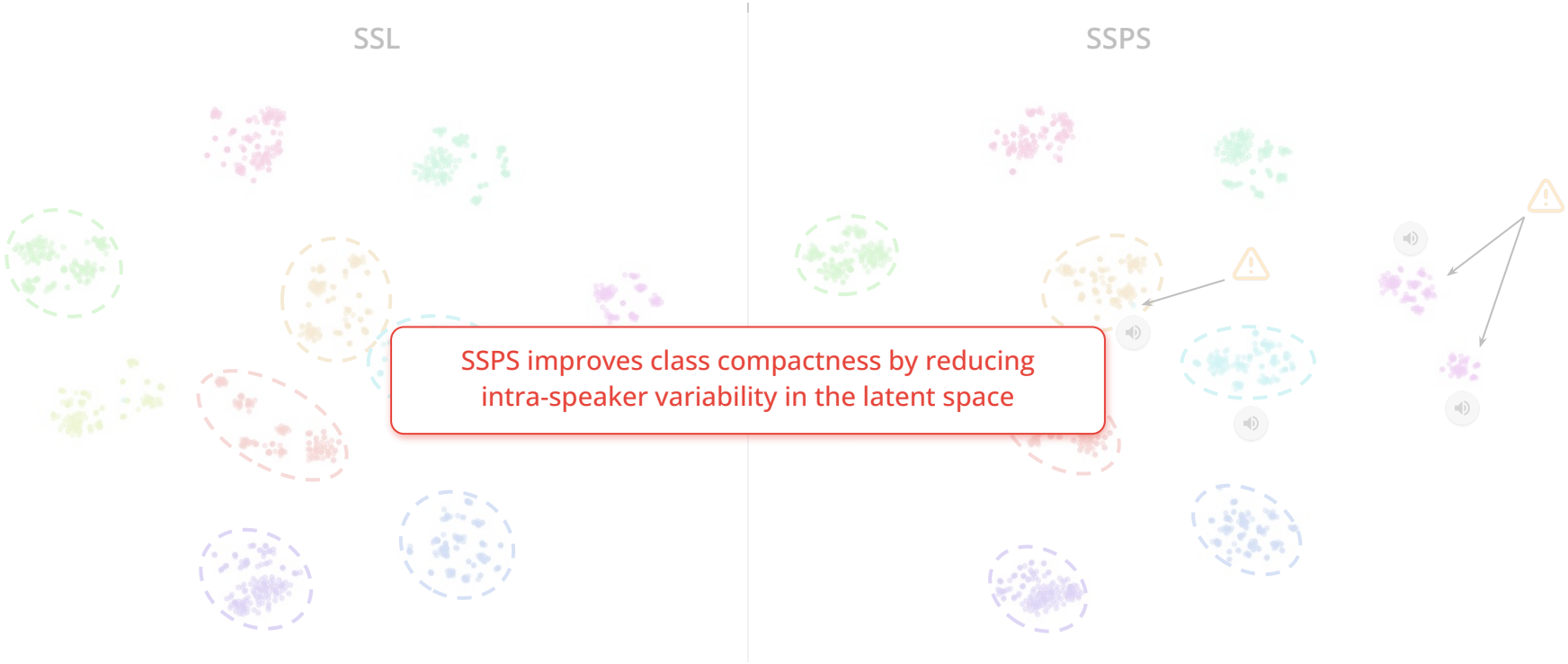


t-SNE on 10 speakers from VoxCeleb1 • **Encoder:** ECAPA-TDNN • **Model:** SimCLR

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Visualisation of Speaker Representations



t-SNE on 10 speakers from VoxCeleb1 • Encoder: ECAPA-TDNN • Model: SimCLR

III – Self-Supervised Positive Sampling from Latent Space SSPS

Robustness to the Absence of Data-Augmentation

Positive sampling	Data-aug.	VoxCeleb1-O	
		EER (%)	$\text{minDCF}_{0.01}$
SSL	✓	**6.30**	**0.5286**
	✗	15.00	0.7575
SSPS	✓	**2.57**	0.3033
	✗	2.77	**0.2840**

- Standard SSL strongly relies on data-augmentation, with significant degradation when removed *(15.00% EER)*.
- SSPS remains robust without data-augmentation, maintaining a low EER *(2.77%)* and even improving the minDCF *(0.3033 → 0.2840)*.
- Effective positive sampling reduces dependence on augmentation, exploiting natural variability from "in-the-wild" training sets.

Benchmark: VoxCeleb1-O • Encoder: ECAPA-TDNN • Model: SimCLR

III – Self-Supervised Positive Sampling from Latent Space SSPS

Conclusions

- SSPS overcomes the main limitation of SSL frameworks by mitigating intra-speaker-variability caused by the standard same-utterance positive sampling.
- The best performance is obtained with SimCLR and DINO (2.57% and 2.53% EER on VoxCeleb1-O).
- SimCLR-SSPS results in a 58% EER improvement, motivating a re-examination of self-supervised contrastive frameworks for the task of SV.
- Improved class compactness and reduced recording-related information are observed in the learned speaker representation space.

Foundation

Part IV

Leveraging Speech Foundation Models

Related Publication

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

Interspeech, 2024

IV – Leveraging Speech Foundation Models

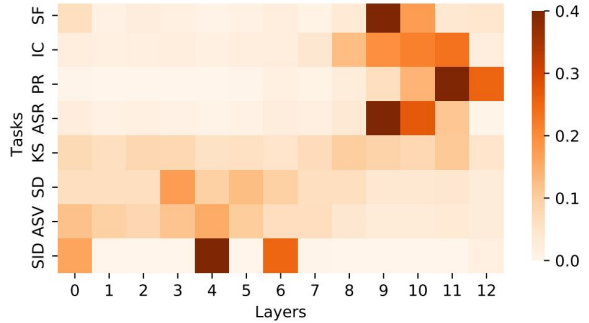
Foundation

Introduction

SSL has led to the emergence of speech foundation models [1,2,3], pre-trained on massive unlabeled data, originally for speech recognition.

Notable properties

1. Transferable representations, supporting a wide range of speech-related downstream tasks;
2. Hierarchical abstraction of information, from acoustic cues *(lower layers)* to phonetic and linguistic content *(higher layers)*.



WavLM weight analysis per layer when fine-tuning for different tasks of the SUPERB Benchmark [3].

[1] S. Schneider et al. *wav2vec:Unsupervised Pre-Training for Speech Recognition*. Interspeech, 2019.
[2] W. Hsu et al. *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*. IEEE JSTSP, 2022.

IV – Leveraging Speech Foundation Models

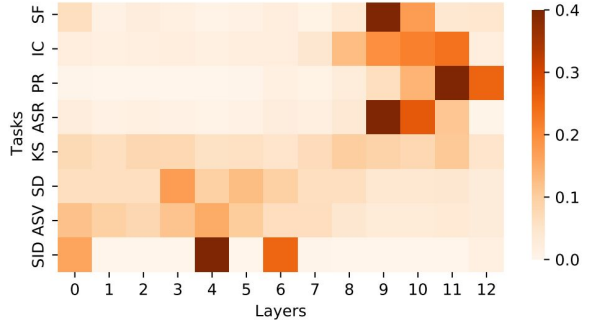
Foundation

Introduction

SSL has led to the emergence of speech foundation models [1,2,3], pre-trained on massive unlabeled data, originally for speech recognition.

Notable properties

1. Transferable representations, supporting a wide range of speech-related downstream tasks;
2. Hierarchical abstraction of information, from acoustic cues *(lower layers)* to phonetic and linguistic content *(higher layers)*.



WavLM weight analysis per layer when fine-tuning for different tasks of the SUPERB Benchmark [3].

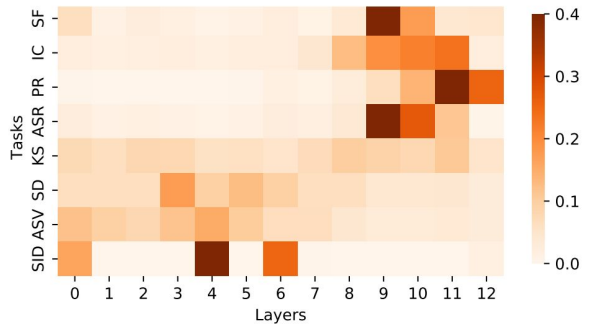
[1] S. Schneider et al. *wav2vec:Unsupervised Pre-Training for Speech Recognition*. Interspeech, 2019.
[2] W. Hsu et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. IEEE TASLP, 2021.
[3] S. Chen et al. *WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing*. IEEE JSTSP, 2022.

IV – Leveraging Speech Foundation Models

Foundation

Introduction

SSL has led to the emergence of speech foundation models [1,2,3], pre-trained on massive unlabeled data, originally for speech recognition.



WavLM weight analysis per layer when fine-tuning for different tasks of the SUPERB Benchmark [3].

Notable properties

1. Transferable representations, supporting a wide range of speech-related downstream tasks;
2. Hierarchical abstraction of information, from acoustic cues *(lower layers)* to phonetic and linguistic content *(higher layers)*.

🏆 WavLM achieves SOTA performance for SV when fine-tuned in a supervised setting [3]

[1] S. Schneider et al. *wav2vec:Unsupervised Pre-Training for Speech Recognition*. Interspeech, 2019.
[2] W. Hsu et al. *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*. IEEE JSTSP, 2022.

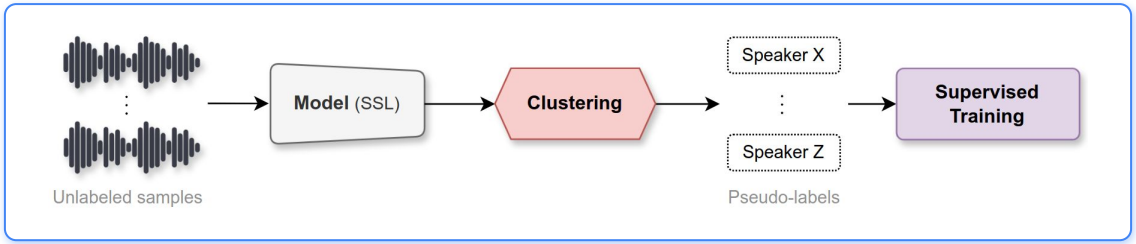
IV – Leveraging Speech Foundation Models

Foundation

Infeasibility of Self-Supervised Fine-Tuning (1)

Front-end	VoxCeleb1-O	
	EER (%)	$\text{minDCF}_{0.01}$
Self-supervised *(SimCLR)*		
Mel-scaled spectrogram	9.69	0.6457
WavLM Base+ (frozen)	13.89	0.7405
WavLM Base+ (fine-tuned)	13.87	0.7428
Supervised *(AAM-Softmax)* ▷ pseudo-labels		
Mel-scaled spectrogram	4.13	0.4127
WavLM Base+ (frozen)	3.63	0.4133
WavLM Base+ (fine-tuned)	3.30	0.3974

Self-supervised fine-tuning *(SimCLR loss)* is ineffective, while supervised fine-tuning *(AAM-Softmax loss)* with pseudo-labels effectively exploits the front-end



Supervised Training with Pseudo-Labels

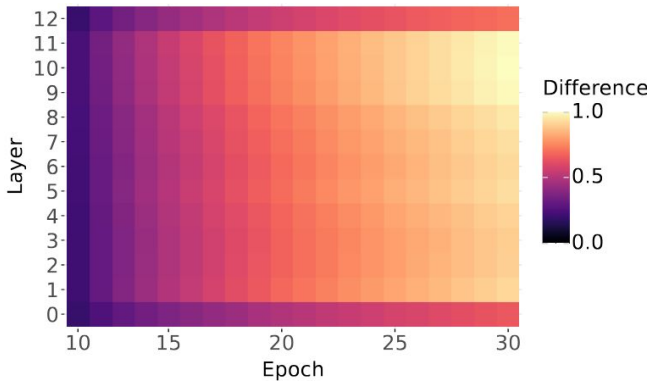
Benchmark: VoxCeleb1-O

IV – Leveraging Speech Foundation Models

Foundation

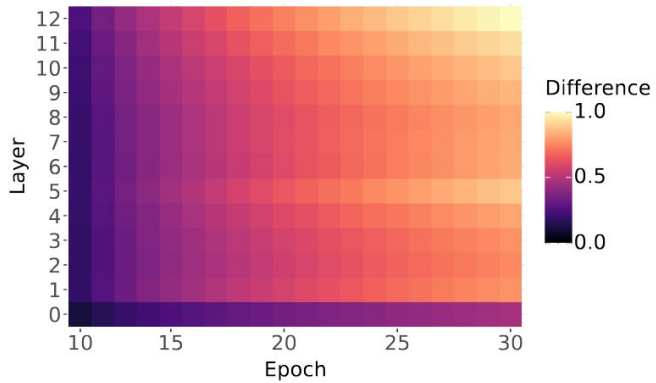
Infeasibility of Self-Supervised Fine-Tuning (2)

Self-Supervised Fine-Tuning



… induces large changes in lower layers, indicating an over-reliance on low-level acoustic features.

Supervised Fine-Tuning

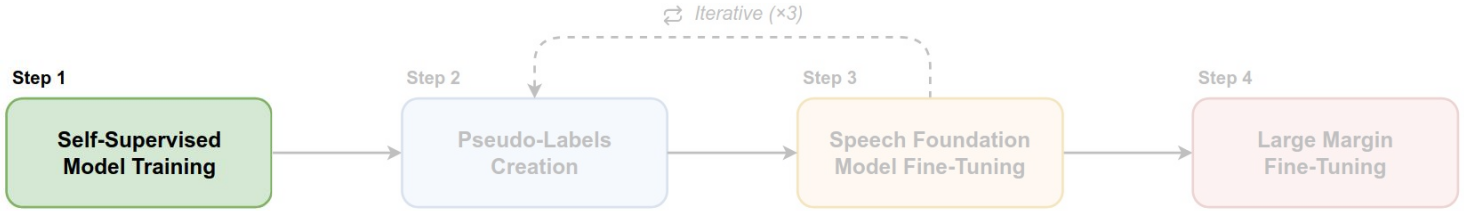


… preserves low and mid layers while progressively adapting higher layers to learn speaker-discriminative representations.

IV – Leveraging Speech Foundation Models

Foundation

Overview of the Training Framework



The proposed training framework leverages speech foundation models without ground-truth speaker labels by:

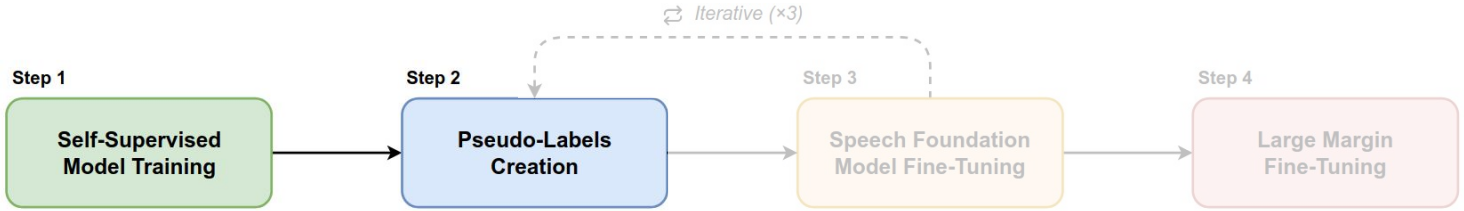
- (Step 1) self-supervised training of DINO;
- (Step 2) generating pseudo-labels by clustering training representations;
- (Step 3) supervised fine-tuning of WavLM + MHFA using the pseudo-labels;
- (Step 2 & Step 3) repeating this process multiple times to refine pseudo-labels;
- (Step 4) and performing an optional Large Margin Fine-Tuning (LMFT).

Manuscript: Section 8.3

IV – Leveraging Speech Foundation Models

Foundation

Overview of the Training Framework



The proposed training framework leverages speech foundation models without ground-truth speaker labels by:

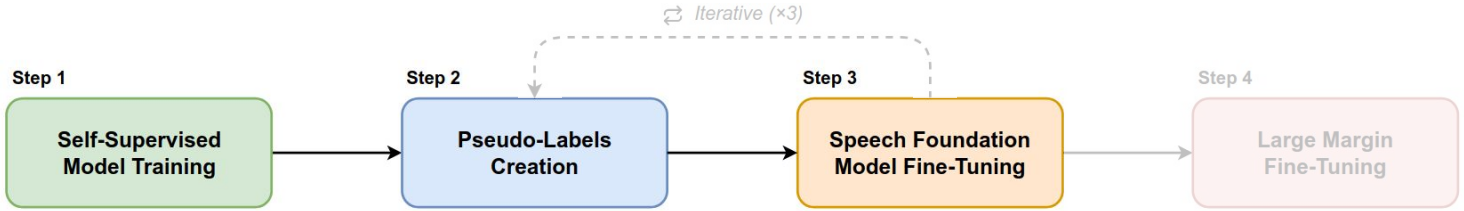
- (Step 1) self-supervised training of DINO;
- (Step 2) generating pseudo-labels by clustering training representations;
- (Step 3) supervised fine-tuning of WavLM + MHFA using the pseudo-labels;
- (Step 2 & Step 3) repeating this process multiple times to refine pseudo-labels;
- (Step 4) and performing an optional Large Margin Fine-Tuning (LMFT).

Manuscript: Section 8.3

IV – Leveraging Speech Foundation Models

Foundation

Overview of the Training Framework



The proposed training framework leverages speech foundation models without ground-truth speaker labels by:

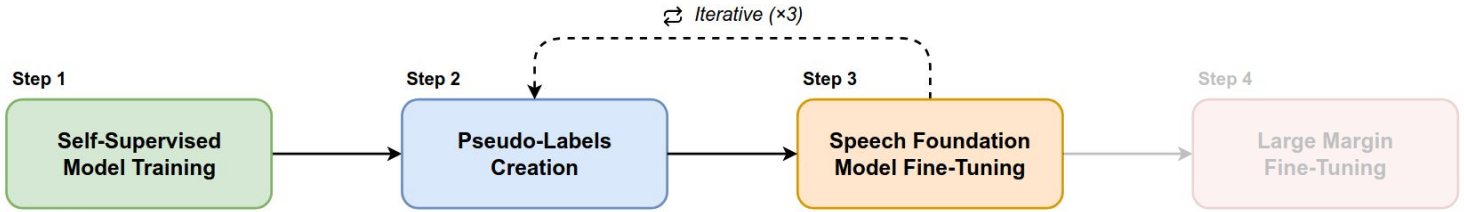
- (Step 1) self-supervised training of DINO;
- (Step 2) generating pseudo-labels by clustering training representations;
- (Step 3) supervised fine-tuning of WavLM + MHFA using the pseudo-labels;
- (Step 2 & Step 3) repeating this process multiple times to refine pseudo-labels;
- (Step 4) and performing an optional Large Margin Fine-Tuning (LMFT).

Manuscript: Section 8.3

IV – Leveraging Speech Foundation Models

Foundation

Overview of the Training Framework



The proposed training framework leverages speech foundation models without ground-truth speaker labels by:

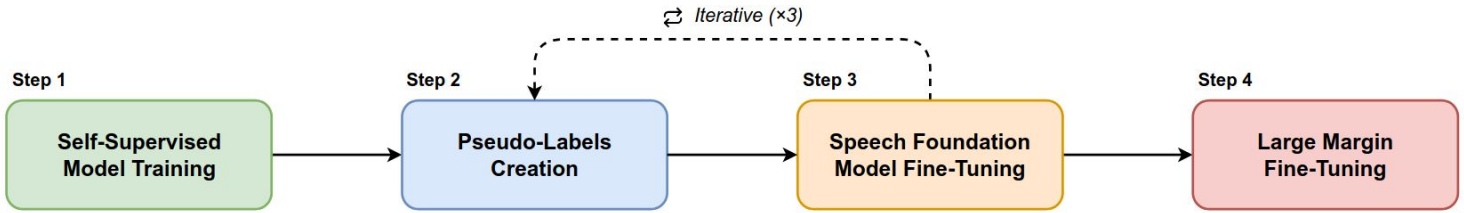
- (Step 1) self-supervised training of DINO;
- (Step 2) generating pseudo-labels by clustering training representations;
- (Step 3) supervised fine-tuning of WavLM + MHFA using the pseudo-labels;
- (Step 2 & Step 3) repeating this process multiple times to refine pseudo-labels;
- (Step 4) and performing an optional Large Margin Fine-Tuning (LMFT).

Manuscript: Section 8.3

IV – Leveraging Speech Foundation Models

Foundation

Overview of the Training Framework



The proposed training framework leverages speech foundation models without ground-truth speaker labels by:

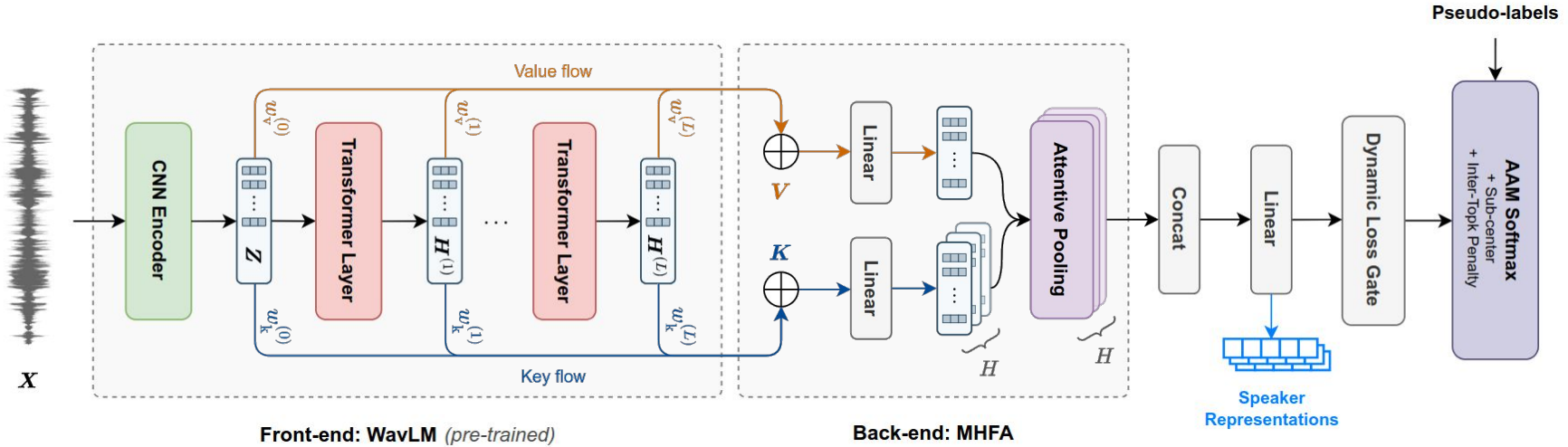
- (Step 1) self-supervised training of DINO;
- (Step 2) generating pseudo-labels by clustering training representations;
- (Step 3) supervised fine-tuning of WavLM + MHFA using the pseudo-labels;
- (Step 2 & Step 3) repeating this process multiple times to refine pseudo-labels;
- (Step 4) and performing an optional Large Margin Fine-Tuning (LMFT).

Manuscript: Section 8.3

IV – Leveraging Speech Foundation Models

Foundation

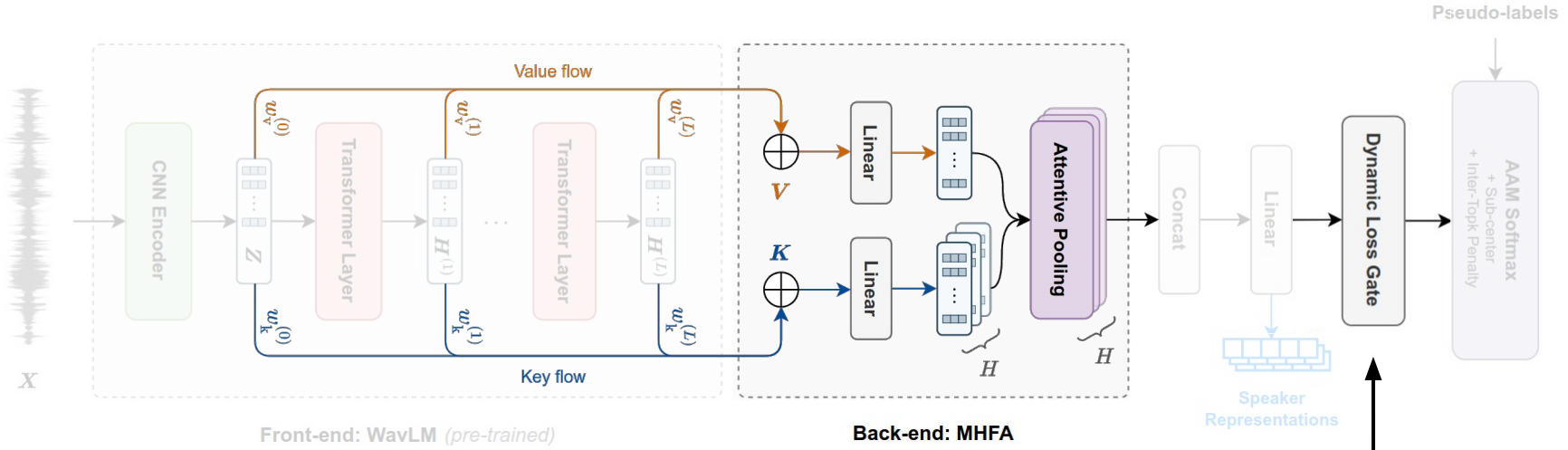
Overview of the Fine-Tuning Stage



IV – Leveraging Speech Foundation Models

Foundation

Overview of the Fine-Tuning Stage

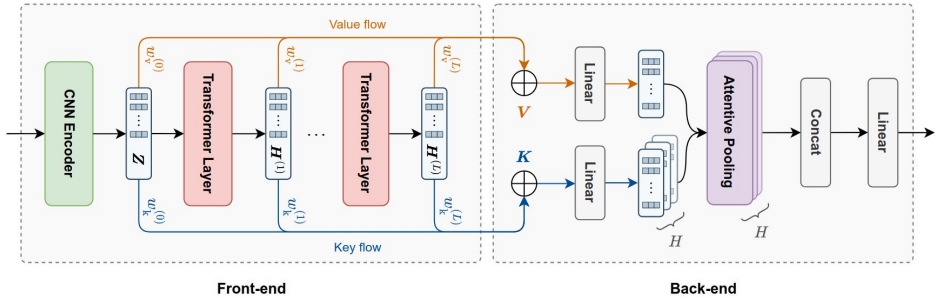


IV – Leveraging Speech Foundation Models

Foundation

Fine-Tuning Stage: MHFA Back-End

Multi-Head Factorized Attention (MHFA) [1] aggregates intermediate representations over time and layers via attention to disentangle phonetic variability (keys) from speaker-discriminative information (values).



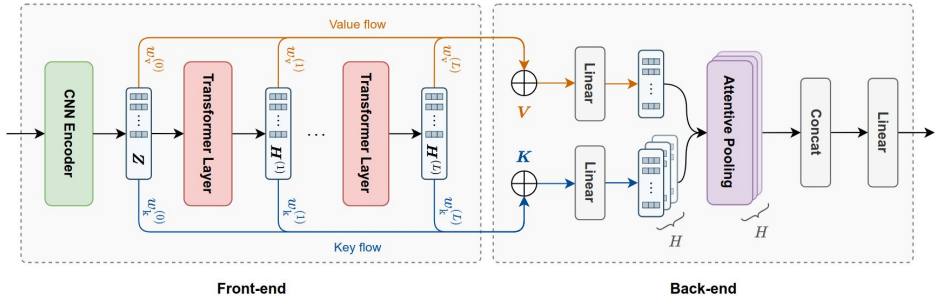
[1] J. Peng et al. *An Attention-Based Backend Allowing Efficient Fine-Tuning of Transformer Models for Speaker Verification*. IEEE SLT, 2022.

IV – Leveraging Speech Foundation Models

Foundation

Fine-Tuning Stage: MHFA Back-End

Multi-Head Factorized Attention (MHFA) [1] aggregates intermediate representations over time and layers via attention to disentangle phonetic variability (keys) from speaker-discriminative information (values).



Back-end	# params.	*Pseudo-labels*	
		EER (%)	$\text{minDCF}_{0.01}$
ASP *(final layer)*	93.4 k	4.28	0.5468
SAP *(final layer)*	93.4 k	4.22	0.5064
ASP *(all layers)*	93.4 k	2.40	0.2377
SAP *(all layers)*	93.4 k	2.46	0.2314
MHFA *(all layers)*	2.3 M	**2.07**	**0.1874**
TDNN *(all layers)*	5.4 M	2.76	0.2708
ECAPA-TDNN *(all layers)*	18.4 M	2.30	0.2161

Benchmark: VoxCeleb1-O • Front-end: WavLM Base+

MHFA enables effective speaker modeling with few additional training parameters

[1] J. Peng et al. *An Attention-Based Backend Allowing Efficient Fine-Tuning of Transformer Models for Speaker Verification*. IEEE SLT, 2022.

IV – Leveraging Speech Foundation Models

Foundation

Fine-Tuning Stage: Pseudo-Label Filtering

Pseudo-label filtering is implemented to discard unreliable pseudo-labels by modeling the loss distribution with a two-component GMM and dynamically thresholding high-loss samples, following Dynamic Loss-Gate [1].

$$p(\mathcal{L}) = \pi_1 \mathcal{N}(\mathcal{L} \mid \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mathcal{L} \mid \mu_2, \sigma_2^2)$$

[1] B. Han et al. *Self-Supervised Speaker Verification Using Dynamic Loss-Gate and Label Correction*. Interspeech, 2022.

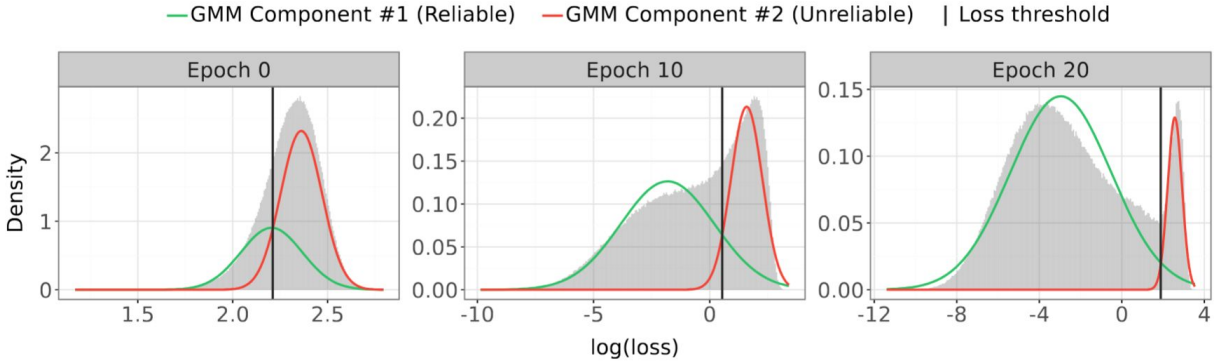
IV – Leveraging Speech Foundation Models

Foundation

Fine-Tuning Stage: Pseudo-Label Filtering

Pseudo-label filtering is implemented to discard unreliable pseudo-labels by modeling the loss distribution with a two-component GMM and dynamically thresholding high-loss samples, following Dynamic Loss-Gate [1].

$$p(\mathcal{L}) = \pi_1 \mathcal{N}(\mathcal{L} \mid \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mathcal{L} \mid \mu_2, \sigma_2^2)$$



Pseudo-label filtering improves robustness by limiting error propagation during fine-tuning.

[1] B. Han et al. *Self-Supervised Speaker Verification Using Dynamic Loss-Gate and Label Correction*. Interspeech, 2022.

IV – Leveraging Speech Foundation Models

Foundation

Evaluation

Method	# params.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Supervised							
wav2vec 2.0 Large (TDNN) (Novoselov et al. 2022)	317 M + 3 M	0.84	0.0580				
WavLM Base+ (TDNN) (S. Chen et al. 2022a)	94 M + 6 M	0.84		0.92		1.75	
WavLM Base+ (MHFA) (Peng et al. 2022)	94 M + 2 M	0.66	0.0740	0.89	0.0970	1.90	0.1900
HuBERT Large (TDNN) (Zhengyang Chen et al. 2022)	316 M + 6 M	0.59		0.65		1.22	
WavLM Large (MHFA) (Peng et al. 2022)	316 M + 2 M	0.49	0.0810	0.70	0.0910	1.70	0.1770
WavLM Large (TDNN) (S. Chen et al. 2022a)	316 M + 6 M	**0.38**		**0.48**		**0.98**	
WavLM Base+ (MHFA)	94 M + 2 M	0.74	0.1020	0.71	0.0772	1.43	0.1379
WavLM Large (MHFA)	316 M + 2 M	**0.58**	**0.0675**	**0.58**	**0.0590**	**1.18**	**0.1209**
Self-supervised							
JHU *(VoxSRC-21)* (Cho et al. 2021)	25 M	1.89					
DKU-DukeECE *(VoxSRC-21)* (Cai et al. 2021a)	1 M	1.81		2.06		3.80	
SNU-HIL *(VoxSRC-21)* (Mun et al. 2021)	20 M	1.66					
LGL (Tao et al. 2022)	20 M	1.66		2.18		3.76	
DLG-LC (Han et al. 2022)	20 M	1.47		1.78		3.19	
DPP (Tao et al. 2023)	20 M	1.44		1.77		3.27	
CA-DINO + DLG-LC (Han et al. 2024)	20 M	1.29		1.57		2.80	
Sub-PTM (Zhicong Chen et al. 2023)	316 M + 20 M	1.25					
IPL (Aldeneh et al. 2025)	20 M	**1.06**		**2.16**		**3.61**	
DINO ▷ WavLM Base+ (MHFA)	94 M + 2 M	1.18	0.1108	1.85	0.1149	2.89	0.1926
DINO ▷ WavLM Base+ (MHFA) + LMFT	94 M + 2 M	1.16	0.1159	1.63	0.1070	2.60	0.1853
DINO ▷ WavLM Large (MHFA)	316 M + 2 M	1.07	**0.0938**	1.40	0.0951	2.35	0.1528
DINO ▷ WavLM Large (MHFA) + LMFT	316 M + 2 M	**1.06**	0.0939	**1.37**	**0.0935**	**2.31**	**0.1499**

IV – Leveraging Speech Foundation Models

Foundation

Evaluation

Method	# params.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Supervised							
wav2vec 2.0 Large (TDNN) (Novoselov et al. 2022)	317 M + 3 M	0.84	0.0580				
WavLM Base+ (TDNN) (S. Chen et al. 2022a)	94 M + 6 M	0.84		0.92		1.75	
WavLM Base+ (MHFA) (Peng et al. 2022)	94 M + 2 M	0.66	0.0740	0.89	0.0970	1.90	0.1900
HuBERT Large (TDNN) (Zhengyang Chen et al. 2022)	316 M + 6 M	0.59		0.65		1.22	
WavLM Large (MHFA) (Peng et al. 2022)	316 M + 2 M	0.49	0.0810	0.70	0.0910	1.70	0.1770
WavLM Large (TDNN) (S. Chen et al. 2022a)	316 M + 6 M	**0.38**		**0.48**		**0.98**	
WavLM Base+ (MHFA)	94 M + 2 M	0.74	0.1020	0.71	0.0772	1.43	0.1379
WavLM Large (MHFA)	316 M + 2 M	0.58	0.0675	0.58	0.0590	1.18	0.1209
Self-supervised							
JHU *(VoxSRC-21)* (Cho et al. 2021)	25 M	1.89					
DKU-DukeECE *(VoxSRC-21)* (Cai et al. 2021a)	1 M	1.81		2.06		3.80	
SNU-HIL *(VoxSRC-21)* (Mun [illegible])							
LGL (Tao et al. 2022)						3.76	
DLG-LC (Han et al. 2022)						3.19	
DPP (Tao et al. 2023)						3.27	
CA-DINO + DLG-LC (Han et al. 2024)	20 M	1.29		1.57		2.80	
Sub-PTM (Zhicong Chen et al. 2023)	316 M + 20 M	1.25					
IPL (Aldeneh et al. 2025)	20 M	1.06		2.16		3.61	
DINO ▷ WavLM Base+ (MHFA)	94 M + 2 M	1.18	0.1108	1.85	0.1149	2.89	0.1926
DINO ▷ WavLM Base+ (MHFA) + LMFT	94 M + 2 M	1.16	0.1159	1.63	0.1070	2.60	0.1853
DINO ▷ WavLM Large (MHFA)	316 M + 2 M	1.07	**0.0938**	1.40	0.0951	2.35	0.1528
DINO ▷ WavLM Large (MHFA) + LMFT	316 M + 2 M	**1.06**	0.0939	**1.37**	**0.0935**	**2.31**	**0.1499**

The proposed method achieves 1.18% EER with WavLM Base+ and 1.07% EER with WavLM Large on VoxCeleb1-O

IV – Leveraging Speech Foundation Models

Foundation

Evaluation

Method	# params.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Supervised							
wav2vec 2.0 Large (TDNN) (Novoselov et al. 2022)	317 M + 3 M	0.84	0.0580				
WavLM Base+ (TDNN) (S. Chen et al. 2022a)	94 M + 6 M	0.84		0.92		1.75	
WavLM Base+ (MHFA) (Peng et al. 2022)	94 M + 2 M	0.66	0.0740	0.89	0.0970	1.90	0.1900
HuBERT Large (TDNN) (Zhengyang Chen et al. 2022)	316 M + 6 M	0.59		0.65		1.22	
WavLM Large (MHFA) (Peng et al. 2022)	316 M + 2 M	0.49	0.0810	0.70	0.0910	1.70	0.1770
WavLM Large (TDNN) (S. Chen et al. 2022a)	316 M + 6 M	**0.38**		**0.48**		**0.98**	
WavLM Base+ (MHFA)	94 M + 2 M	0.74	0.1020	0.71	0.0772	1.43	0.1379
WavLM Large (MHFA)	316 M + 2 M	0.58	0.0675	0.58	0.0590	1.18	0.1209
Self-supervised							
JHU *(VoxSRC-21)* (Cho et al. 2021)	25 M	1.89					
DKU-DukeECE *(VoxSRC-21)* (Cai et al. 2021a)	1 M	1.81		2.06		3.80	
SNU-HIL *(VoxSRC-21)* (Mun et al. 20							
LGL (Tao et al. 2022)						3.76	
DLG-LC (Han et al. 2022)						3.19	
DPP (Tao et al. 2023)						3.27	
CA-DINO + DLG-LC (Han et al. 2024)	20 M	1.29		1.57		2.80	
Sub-PTM (Zhicong Chen et al. 2023)	316 M + 20 M	1.25					
IPL (Aldeneh et al. 2025)	20 M	**1.06**		2.16		3.61	
DINO ▷ WavLM Base+ (MHFA)	94 M + 2 M	1.18	0.1108	1.85	0.1149	2.89	0.1926
DINO ▷ WavLM Base+ (MHFA) + LMFT	94 M + 2 M	1.16	0.1159	1.63	0.1070	2.60	0.1853
DINO ▷ WavLM Large (MHFA)	316 M + 2 M	1.07	**0.0938**	1.40	0.0951	2.35	0.1528
DINO ▷ WavLM Large (MHFA) + LMFT	316 M + 2 M	**1.06**	0.0939	**1.37**	**0.0935**	**2.31**	**0.1499**

Applying Large Margin Fine-Tuning (LMFT) yields slight but consistent improvements

IV – Leveraging Speech Foundation Models

Foundation

Evaluation

Method	# params.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Supervised							
wav2vec 2.0 Large (TDNN) (Novoselov et al. 2022)	317 M + 3 M	0.84	0.0580				
WavLM Base+ (TDNN) (S. Chen et al. 2022a)	94 M + 6 M	0.84		0.92		1.75	
WavLM Base+ (MHFA) (Peng et al. 2022)	94 M + 2 M	0.66	0.0740	0.89	0.0970	1.90	0.1900
HuBERT Large (TDNN) (Zhengyang Chen et al. 2022)	316 M + 6 M	0.59		0.65		1.22	
WavLM Large (MHFA) (Peng et al. 2022)	316 M + 2 M	0.49	0.0810	0.70	0.0910	1.70	0.1770
WavLM Large (TDNN) (S. Chen et al. 2022a)	316 M + 6 M	**0.38**		**0.48**		**0.98**	
WavLM Base+ (MHFA)	94 M + 2 M	0.74	0.1020	0.71	0.0772	1.43	0.1379
WavLM Large (MHFA)	316 M + 2 M	0.58	0.0675	0.58	0.0590	1.18	0.1209
Self-supervised							
JHU *(VoxSRC-21)* (Cho et al. 2021)	25 M	1.89					
DKU-DukeECE *(VoxSRC-21)* (Cai et al. 2021a)	1 M	1.81		2.06		3.80	
SNU-HIL *(VoxSRC-21)* (Mun [illegible])	[illegible]						
LGL (Tao et al. 2022)	[illegible]					3.76	
DLG-LC (Han et al. 2022)	[illegible]					3.19	
DPP (Tao et al. 2023)	[illegible]					3.27	
CA-DINO + DLG-LC (Han et al. 2024)	20 M	1.29		1.57		2.80	
Sub-PTM (Zhicong Chen et al. 2023)	316 M + 20 M	1.25					
IPL (Aldeneh et al. 2025)	20 M	**1.06**		2.16		3.61	
DINO ▷ WavLM Base+ (MHFA)	94 M + 2 M	1.18	0.1108	1.85	0.1149	2.89	0.1926
DINO ▷ WavLM Base+ (MHFA) + LMFT	94 M + 2 M	1.16	0.1159	1.63	0.1070	2.60	0.1853
DINO ▷ WavLM Large (MHFA)	316 M + 2 M	1.07	**0.0938**	1.40	0.0951	2.35	0.1528
DINO ▷ WavLM Large (MHFA) + LMFT	316 M + 2 M	**1.06**	0.0939	**1.37**	**0.0935**	**2.31**	**0.1499**

The best performance is achieved with WavLM Large + LMFT, reaching 1.06% EER on VoxCeleb1-O

IV – Leveraging Speech Foundation Models

Foundation

Evaluation

Method	# params.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Supervised							
wav2vec 2.0 Large (TDNN) (Novoselov et al. 2022)	317 M + 3 M	0.84	0.0580				
WavLM Base+ (TDNN) (S. Chen et al. 2022a)	94 M + 6 M	0.84		0.92		1.75	
WavLM Base+ (MHFA) (Peng et al. 2023)	94 M + 2 M	0.66	0.0749	0.89	0.0970	1.90	0.1900
HuBERT Large (TDNN) (Zh						1.22	
WavLM Large (MHFA) (Per						1.70	0.1770
WavLM Large (TDNN) (S. C						0.98	
WavLM Base+ (MHFA)	94 M + 2 M	0.74	0.1020	0.71	0.0772	1.43	0.1379
WavLM Large (MHFA)	316 M + 2 M	0.58	0.0675	0.58	0.0590	1.18	0.1209
Self-supervised							
JHU *(VoxSRC-21)* (Cho et al. 2021)	25 M	1.89					
DKU-DukeECE *(VoxSRC-21)* (Cai et al. 2021a)	1 M	1.81		2.06		3.80	
SNU-HIL *(VoxSRC-21)* (Mun et al. 2021)	20 M	1.66					
LGL (Tao et al. 2022)	20 M	1.66		2.18		3.76	
DLG-LC (Han et al. 2022)	20 M	1.47		1.78		3.19	
DPP (Tao et al. 2023)	20 M	1.44		1.77		3.27	
CA-DINO + DLG-LC (Han et al. 2024)	20 M	1.29		1.57		2.80	
Sub-PTM (Zhicong Chen et al. 2023)	316 M + 20 M	1.25					
IPL (Aldeneh et al. 2025)	20 M	**1.06**		**2.16**		**3.61**	
DINO ▷ WavLM Base+ (MHFA)	94 M + 2 M	1.18	0.1108	1.85	0.1149	2.89	0.1926
DINO ▷ WavLM Base+ (MHFA) + LMFT	94 M + 2 M	1.16	0.1159	1.63	0.1070	2.60	0.1853
DINO ▷ WavLM Large (MHFA)	316 M + 2 M	1.07	**0.0938**	1.40	0.0951	2.35	0.1528
DINO ▷ WavLM Large (MHFA) + LMFT	316 M + 2 M	**1.06**	0.0939	**1.37**	**0.0935**	**2.31**	**0.1499**

New SOTA among self-supervised methods on VoxCeleb 🏆

IV – Leveraging Speech Foundation Models

Foundation

Evaluation

Method	# params.	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Supervised							
wav2vec 2.0 Large (TDNN) (Novoselov et al. 2022)	317 M + 3 M	0.84	0.0580				
WavLM Base+ (TDNN) (S. Chen et al. 2022a)	94 M + 6 M	0.84		0.92		1.75	
WavLM Base+ (MHFA) (Peng et al. 2022)	94 M + 2 M	0.66	0.0740	0.89	0.0970	1.90	0.1900
HuBERT Large (TDNN) (Zhengyang Chen et al. 2022)	316 M + 6 M	0.59		0.65		1.22	
WavLM Large (MHFA) (Peng et al. 2022)	316 M + 2 M	0.49	0.0810	0.70	0.0910	1.70	0.1770
WavLM Large (TDNN) (S. Chen et al. 2022a)	316 M + 6 M	**0.38**		**0.48**		**0.98**	
WavLM Base+ (MHFA)	94 M + 2 M	0.74	0.1020	0.71	0.0772	1.43	0.1379
WavLM Large (MHFA)	316 M + 2 M	**0.58**	**0.0675**	**0.58**	**0.0590**	**1.18**	**0.1209**
Self-supervised							
JHU *(VoxSRC-21)* (Cho et al. 2021)	25 M	1.89					
DKU-DukeECE *(VoxSRC-21)* (Cai et al. 2021a)	1 M	1.81		2.06		3.80	
SNU-HIL *(VoxSRC-21)* (Mun et al. 2021)							
LGL (Tao et al. 2022)						3.76	
DLG-LC (Han et al. 2022)						3.19	
DPP (Tao et al. 2023)						3.27	
CA-DINO + DLG-LC (Han et al. 2024)	20 M	1.29		1.57		2.80	
Sub-PTM (Zhicong Chen et al. 2023)	316 M + 20 M	1.25					
IPL (Aldeneh et al. 2025)	20 M	1.06		2.16		3.61	
DINO ▷ WavLM Base+ (MHFA)	94 M + 2 M	1.18	0.1108	1.85	0.1149	2.89	0.1926
DINO ▷ WavLM Base+ (MHFA) + LMFT	94 M + 2 M	1.16	0.1159	1.63	0.1070	2.60	0.1853
DINO ▷ WavLM Large (MHFA)	316 M + 2 M	1.07	**0.0938**	1.40	0.0951	2.35	0.1528
DINO ▷ WavLM Large (MHFA) + LMFT	316 M + 2 M	**1.06**	0.0939	**1.37**	**0.0935**	**2.31**	**0.1499**

The proposed method substantially narrows the performance gap with supervised systems 🚀

IV – Leveraging Speech Foundation Models

Foundation

Conclusions

- Speech foundation models can be adapted to SV in a fully self-supervised manner using pseudo-labels, initially derived from a DINO-based model and iteratively refined.

- Direct self-supervised fine-tuning is ineffective since SSL objectives tend to exploit low-level acoustic information and fail to adapt high-level representations to speaker discrimination.

- The proposed iterative pseudo-labeling framework based on WavLM Large establishes a new SOTA on VoxCeleb *(1.06% EER on VoxCeleb1-O)*, moving SSL closer to supervised performance.

Conclusions

Application and Study of SSL for SV SSLSV

Benchmark and study SSL frameworks (e.g., SimCLR, MoCo, DINO) on SV under controlled conditions
→ Identify the role and limitations of positive sampling in modeling intra-speaker variability

Margins in Self-Supervised Contrastive Frameworks Margins

Integrate CosFace, ArcFace, AdaFace, and other margin-based constraints into SimCLR and MoCo
→ Improve speaker separability in fully self-supervised settings

Self-Supervised Positive Sampling (SSPS) SSPS

Exploit latent-space proximity to sample cross-recording pseudo-positives
→ Reduce intra-speaker variability and improve SV performance across frameworks (-58% EER for SimCLR)

Speech Foundation Models for SV without Labels Foundation

Develop an iterative pseudo-labeling approach to enable WavLM fine-tuning from a DINO-based model
→ 1.06% EER on VoxCeleb1-O, setting a new SOTA and approaching supervised performance

sslsv: Open-Source PyTorch Toolkit for Self-Supervised SV

Release a PyTorch toolkit to support reproducibility and future research → https://github.com/theolepage/sslsv

Conclusions

Application and Study of SSL for SV SSLSV

Benchmark and study SSL frameworks (e.g., SimCLR, MoCo, DINO) on SV under controlled conditions
→ Identify the role and limitations of positive sampling in modeling intra-speaker variability

Margins in Self-Supervised Contrastive Frameworks Margins

Integrate CosFace, ArcFace, AdaFace, and other margin-based constraints into SimCLR and MoCo
→ Improve speaker separability in fully self-supervised settings

Self-Supervised Positive Sampling (SSPS) SSPS

Exploit latent-space proximity to sample cross-recording pseudo-positives
→ Reduce intra-speaker variability and improve SV performance across frameworks (-58% EER for SimCLR)

Speech Foundation Models for SV without Labels Foundation

Develop an iterative pseudo-labeling approach to enable WavLM fine-tuning from a DINO-based model
→ 1.06% EER on VoxCeleb1-O, setting a new SOTA and approaching supervised performance

sslsv: Open-Source PyTorch Toolkit for Self-Supervised SV

Release a PyTorch toolkit to support reproducibility and future research → https://github.com/theolepage/sslsv

Conclusions

Application and Study of SSL for SV SSLSV

Benchmark and study SSL frameworks (e.g., SimCLR, MoCo, DINO) on SV under controlled conditions
→ Identify the role and limitations of positive sampling in modeling intra-speaker variability

Margins in Self-Supervised Contrastive Frameworks Margins

Integrate CosFace, ArcFace, AdaFace, and other margin-based constraints into SimCLR and MoCo
→ Improve speaker separability in fully self-supervised settings

Self-Supervised Positive Sampling (SSPS) SSPS

Exploit latent-space proximity to sample cross-recording pseudo-positives
→ Reduce intra-speaker variability and improve SV performance across frameworks (-58% EER for SimCLR)

Speech Foundation Models for SV without Labels Foundation

Develop an iterative pseudo-labeling approach to enable WavLM fine-tuning from a DINO-based model
→ 1.06% EER on VoxCeleb1-O, setting a new SOTA and approaching supervised performance

sslsv: Open-Source PyTorch Toolkit for Self-Supervised SV

Release a PyTorch toolkit to support reproducibility and future research → https://github.com/theolepage/sslsv

Conclusions

Application and Study of SSL for SV SSLSV

Benchmark and study SSL frameworks (e.g., SimCLR, MoCo, DINO) on SV under controlled conditions
→ Identify the role and limitations of positive sampling in modeling intra-speaker variability

Margins in Self-Supervised Contrastive Frameworks Margins

Integrate CosFace, ArcFace, AdaFace, and other margin-based constraints into SimCLR and MoCo
→ Improve speaker separability in fully self-supervised settings

Self-Supervised Positive Sampling (SSPS) SSPS

Exploit latent-space proximity to sample cross-recording pseudo-positives
→ Reduce intra-speaker variability and improve SV performance across frameworks (-58% EER for SimCLR)

Speech Foundation Models for SV without Labels Foundation

Develop an iterative pseudo-labeling approach to enable WavLM fine-tuning from a DINO-based model
→ 1.06% EER on VoxCeleb1-O, setting a new SOTA and approaching supervised performance

sslsv: Open-Source PyTorch Toolkit for Self-Supervised SV

Release a PyTorch toolkit to support reproducibility and future research → https://github.com/theolepage/sslsv

Conclusions

Application and Study of SSL for SV SSLSV

Benchmark and study SSL frameworks (e.g., SimCLR, MoCo, DINO) on SV under controlled conditions
→ Identify the role and limitations of positive sampling in modeling intra-speaker variability

Margins in Self-Supervised Contrastive Frameworks Margins

Integrate CosFace, ArcFace, AdaFace, and other margin-based constraints into SimCLR and MoCo
→ Improve speaker separability in fully self-supervised settings

Self-Supervised Positive Sampling (SSPS) SSPS

Exploit latent-space proximity to sample cross-recording pseudo-positives
→ Reduce intra-speaker variability and improve SV performance across frameworks (-58% EER for SimCLR)

Speech Foundation Models for SV without Labels Foundation

Develop an iterative pseudo-labeling approach to enable WavLM fine-tuning from a DINO-based model
→ 1.06% EER on VoxCeleb1-O, setting a new SOTA and approaching supervised performance

sslsv: Open-Source PyTorch Toolkit for Self-Supervised SV

Release a PyTorch toolkit to support reproducibility and future research → https://github.com/theolepage/sslsv

Conclusions

➔ SSL is a promising alternative to supervised learning, enabling robust representation learning from unlabeled speech and paving the way toward next-generation SR systems.

Perspectives

Scaling Training Data and Models

Do larger datasets and larger encoders improve performance, as observed in other domains?

Leveraging Multimodal Information

How to use speech-face cues from unlabeled video under limited audio quality?

Encoding of Extrinsic Information

How to further disentangle speaker identity from recording variability in representations?

Learning with Class Prototypes

Can prototype-based objectives mirror supervised training for better inter-speaker separation?

Investigating the Projector

How does the projector shape information transfer to the SV downstream task?

Exploring Other Speaker-Related Tasks

How to extend representation learning beyond identity verification to language and emotion recognition?

Publications

* First author

Label-Efficient Self-Supervised Speaker Verification With Information Maximization and Contrastive Learning *
Interspeech • 2022 — SSLSV

Experimenting with Additive Margins for Contrastive Self-Supervised Speaker Verification *
Interspeech • 2023 — Margins

Additive Margin in Contrastive Self-Supervised Frameworks to Learn Discriminative Speaker Representations *
The Speaker and Language Recognition Workshop (Odyssey) • 2024 — Margins

Exploring WavLM Back-ends for Speech Spoofing and Deepfake Detection
The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof) • 2024 — Spoof

Towards Supervised Performance on Speaker Verification with SSL by Leveraging Large-Scale ASR Models
Interspeech • 2024 — Foundation

Self-Supervised Frameworks for Speaker Recognition via Bootstrapped Positive Sampling *
IEEE Transactions on Audio, Speech, and Language Processing (TASLP) • vol. 33 • 2025 — SSPS

SSPS: Self-Supervised Positive Sampling for Robust Self-Supervised Speaker Verification *
Interspeech • 2025 — SSPS

Self-Supervised Learning for Speaker Recognition: A study and review *
Speech Communication • vol. 176 • 2026 — SSLSV

Questions?

Additional Resources

I – Self-Supervised Learning for Speaker Verification

SSLSV

Hyperparameters Search – MoCo

Num. negs. (Q_{MoCo})	*Default (SSL)*		*No false-negs. (Sup.)*	
	EER (%)	**minDCF$_{0.01}$**	**EER (%)**	**minDCF$_{0.01}$**
256	9.35	0.6513	9.17	0.6426
1024	9.30	0.6604	8.69	0.6111
16 384	8.92	0.6235	8.33	**0.5790**
32 768	**8.49**	**0.5990**	**7.92**	0.5904
65 536	8.71	0.6044	8.50	0.6197

Temperature (τ)	EER (%)	minDCF$_{0.01}$	Entropy
0.01	**8.25**	0.6218	0.9
0.03	8.49	**0.5990**	2.2
0.05	8.44	0.6124	4.1
0.07	8.80	0.6356	6.4
0.1	9.63	0.6935	8.4

Momentum (m)	EER (%)	minDCF$_{0.01}$
0.9	24.36	0.9997
0.99	8.37	0.6050
0.996	**8.20**	0.6169
0.999	8.49	**0.5990**
1.0	24.47	0.9839

I – Self-Supervised Learning for Speaker Verification

SSLSV

Hyperparameters Search – DINO

Method	EER (%)	minDCF$_{0.01}$
DINO	19.24	0.9781
+ Weight decay (wd = $5e^{-5}$)	20.01	0.9456
+ LR sched. (warmup + cosine)	10.47	0.6620
+ Encoder pooling (ASP)	8.85	0.6160
+ Data-aug. (clean/rir/noise/both)	7.00	0.5449
+ Input dim. ($D_{\text{inp}} = 80$)	**5.94**	0.4903
+ Optimizer (SGD, lr = 0.2)	6.04	**0.4526**

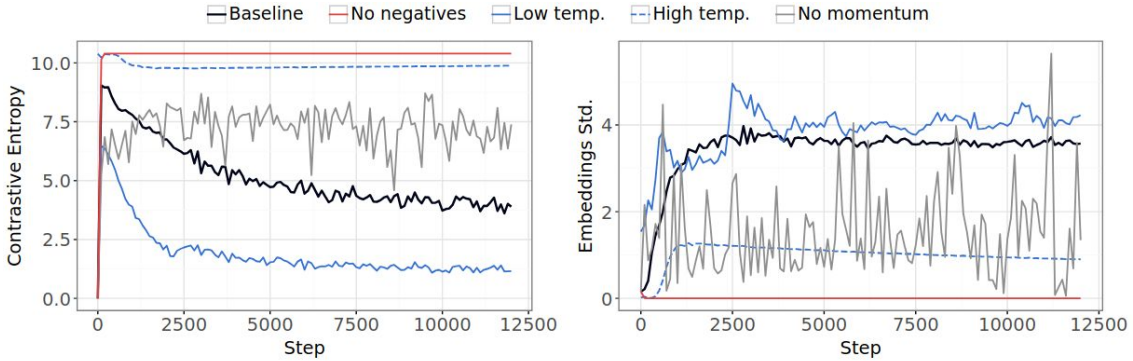
Teacher temp. (τ_t)	EER (%)	minDCF$_{0.01}$	Entropy
0.01	23.18	0.9989	0.02
0.03	7.42	0.4948	0.04
0.04	6.04	**0.4526**	0.07
0.05	6.38	0.4753	0.28
0.07	**5.61**	0.4567	3.18
0.04 → 0.07	6.25	0.4858	3.90

Global views (G)	Local views (L)	EER (%)	minDCF$_{0.01}$
1 × 2 s	1 × 2 s	14.53	0.7636
2 × 4 s	4 × 2 s	**6.04**	**0.4526**
2 × 3 s	4 × 2 s	6.61	0.5195
2 × 2 s	4 × 2 s	8.33	0.6543
1 × 2 s	4 × 2 s	9.31	0.6793
2 × 4 s	2 × 2 s	6.43	0.4855

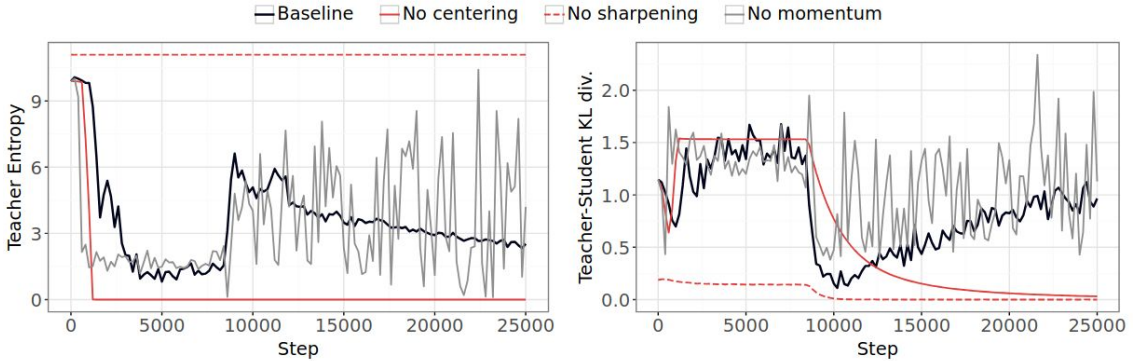
Head output dim. (D_{emb})	EER (%)	minDCF$_{0.01}$
2048	7.59	0.5189
16 384	6.53	0.4848
32 768	6.71	0.4921
65 536	**6.04**	**0.4526**

I – Self-Supervised Learning for Speaker Verification SSLSV

Study of Collapse



(a) **MoCo**



(b) **DINO**

I – Self-Supervised Learning for Speaker Verification

SSLSV

Evaluation of Single-Stage Methods

Category	Method	Encoder	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER (%)	$\text{minDCF}_{0.01/0.05}$	EER (%)	$\text{minDCF}_{0.01/0.05}$	EER (%)	$\text{minDCF}_{0.01/0.05}$
Contrastive	Disent.* (Nagrani et al. 2020)	VGG-M-40	22.09	-	-	-	-	-
	CDDL* (S.-W. Chung et al. 2020)	VGG-M-40	17.52	-	-	-	-	-
	GCL (Inoue et al. 2020)	ResNet-18	15.26	-	-	-	-	-
	AP + i-mix (Kang et al. 2022)	ECAPA-TDNN	10.63	-	-	-	-	-
	AP + l-mix (Kang et al. 2022)	ECAPA-TDNN	10.49	-	-	-	-	-
	AP + AAT (Huh et al. 2020)	Fast ResNet-34	8.65	0.4540	-	-	-	-
	Contrastive + VICReg (Lepage et al. 2022)	Thin ResNet-34	8.47	0.6400	-	-	-	-
	SimCLR + MSE loss (Haoran Zhang et al. 2021)	Thin ResNet-34	8.28	0.6100	-	-	-	-
	MoCo + ProtoNCE (W. Xia et al. 2021)	TDNN	8.23	0.5900	-	-	-	-
	CEL (Mun et al. 2020)	Fast ResNet-34	8.01	-	-	-	-	-
	SimCLR + Margin (Lepage et al. 2024)	Fast ResNet-34	7.85	0.6168	-	-	-	-
	SimCLR + AAT (Tao et al. 2022)	ECAPA-TDNN	7.36	-	7.90	-	12.32	-
	W-ACSG (Gan et al. 2024)	ECAPA-TDNN	7.02	0.3900	-	-	-	-
	C3-MoCo (C. Zhang et al. 2022)	ECAPA-TDNN	6.40	-	-	-	-	-
	DPP* (Tao et al. 2023)	ECAPA-TDNN	2.89	-	3.17	-	6.27	-
	SimCLR-SSPS (Lepage et al. 2025b)	ECAPA-TDNN	**2.57**	**0.3033**	**3.11**	**0.3125**	**5.56**	**0.4638**
Self-distillation	SSReg (Sang et al. 2022)	Thin ResNet-34	6.99	0.4340	-	-	-	-
	DINO + Cosine loss (Han et al. 2022)	Thin-ResNet34	6.16	0.5240	-	-	-	-
	DINO (Jung et al. 2022)	RawNet3	5.40	0.3396	-	-	-	-
	DINO (Jaejin Cho et al. 2022b)	Fast ResNet-34	4.83	0.4630	-	-	-	-
	DINO + Curriculum (H.-S. Heo et al. 2023)	ECAPA-TDNN	4.47	0.3057	-	-	-	-
	CA-DINO (Han et al. 2024)	ECAPA-TDNN	3.59	0.3529	3.85	0.4182	6.92	0.5743
	RDINO (Y. Chen et al. 2023a)	ECAPA-TDNN	3.29	0.2470	-	-	-	-
	MeMo (Jin et al. 2024a)	ECAPA-TDNN	3.10	0.2290	3.53	0.2970	7.04	0.5690
	EBCA-DINO (Hao et al. 2025)	ECAPA-TDNN	2.94	-	2.99	-	5.14	-
	RDINO + W-GVKT (Jin et al. 2024b)	ECAPA-TDNN	2.89	0.3330	-	-	-	-
	PDC-RDINO (Z. Zhao et al. 2024)	ECAPA-TDNN	2.80	0.3150	-	-	-	-
	DINO + RMP (J.-h. Kim et al. 2024)	BA-Transformer	2.62	0.1736	2.84	0.1936	5.14	0.3036
	DINO-SSPS (Lepage et al. 2025b)	ECAPA-TDNN	2.53	0.2843	2.55	0.3150	4.93	0.4632
	DINO + Aug. (Zhengyang Chen et al. 2022b)	ECAPA-TDNN	2.51	0.1626	2.47	-	4.79	-
	C3-DINO (C. Zhang et al. 2022)	ECAPA-TDNN	2.50	-	-	-	-	-
	SDPN (Y. Chen et al. 2025a)	ECAPA-TDNN	**1.80**	0.1390	**1.99**	0.1310	**3.62**	0.2190

I – Self-Supervised Learning for Speaker Verification SSLSV

Evaluation of Multi-Stage Methods

Method	Encoder	Base model	Clustering	# iters.	LF/LC	VoxCeleb1-O EER (%)	VoxCeleb1-E EER (%)	VoxCeleb1-H EER (%)
DKU-DukeECE (VoxSRC-20) (Weiqing Wang et al. 2020)	Fast ResNet-34	SimCLR	K-M (6k)	2	✓	5.42	6.22	9.60
DukeECE (D. Cai et al. 2021b)	Fast ResNet-34	SimCLR	K-M (6k)	5	✓	3.45	4.02	6.57
CAMSAT (Fathan et al. 2024)	ECAPA-TDNN	i-vector	CAMSAT	-	✗	3.06	-	-
AdaptiveDrop (Fathan et al. 2025)	ECAPA-TDNN	i-vector	CAMSAT	-	✓	2.41	-	-
IDLAB (VoxSRC-20) (Thienpondt et al. 2020)	ECAPA-TDNN	MoCo	K-M (50k); AHC (7.5k)	7	✗	2.10	-	-
JHU (VoxSRC-21) (Jejin Cho et al. 2021)	Res2Net-50	DINO	K-M (50k); AHC (7.5k)	4	✗	1.89	-	-
DKU-DukeECE* (VoxSRC-21) (D. Cai et al. 2021a; D. Cai et al. 2022)	Fast ResNet-34	SimCLR	K-M (6k)	4	✓	1.81	2.06	3.80
SNU-HIL (VoxSRC-21) (Mun et al. 2021)	ECAPA-TDNN	CEL	K-M (50k); AHC (7.5k)	8	✗	1.66	-	-
LGL (Tao et al. 2022)	ECAPA-TDNN	SimCLR	K-M (6k)	5	✓	1.66	2.18	3.76
DLG-LC (Han et al. 2022)	ECAPA-TDNN	DINO	K-M (7.5k)	5	✓	1.47	1.78	3.19
DPP* (Tao et al. 2023)	ECAPA-TDNN	SimCLR	K-M (6k)	4	✗	1.44	1.77	3.27
AT + HT (Z. Zhou et al. 2024)	ECAPA-TDNN	DINO	K-M (6k)	5	✓	1.35	-	-
BDS-BPLC (J. Wang et al. 2025)	ECAPA-TDNN	DINO	K-M (7.5k)	5	✓	1.33	1.56	2.78
CA-DINO + DLG-LC* (Han et al. 2024)	ECAPA-TDNN	CA-DINO	K-M (7.5k)	2	✓	1.29	1.57	2.80
Co-Meta* (H. Chen et al. 2023)	ECAPA-TDNN	SimCLR	K-M (6k)	7	✗	1.27	1.82	3.31
Sub-PTM (Zhicong Chen et al. 2023)	WavLM Large	WavLM	Infomap	5	✓	1.25	-	-
IPL (Aldeneh et al. 2025)	MFA-Conformer	i-vector	K-M (25k); AHC (7.5k)	8	✗	1.06	2.16	3.61
SSRL (D. Cai et al. 2025)	WavLM Large	DINO	K-M (8k)	1	✓	1.04	-	-
DINO-WavLM (Miara et al. 2024)	WavLM Base+	DINO	K-M (50k); AHC (7.5k)	3	✓	**0.99**	**1.21**	**2.35**

I – Self-Supervised Learning for Speaker Verification

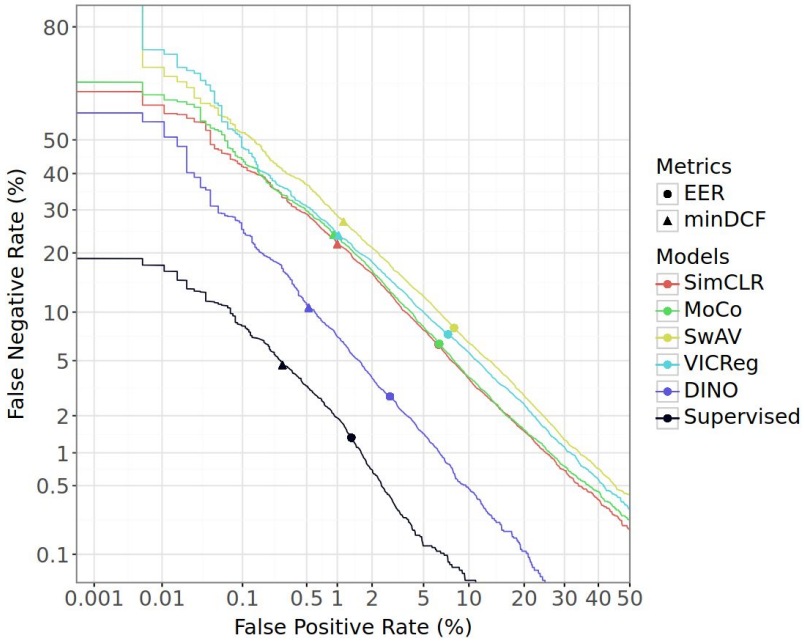
SSLSV

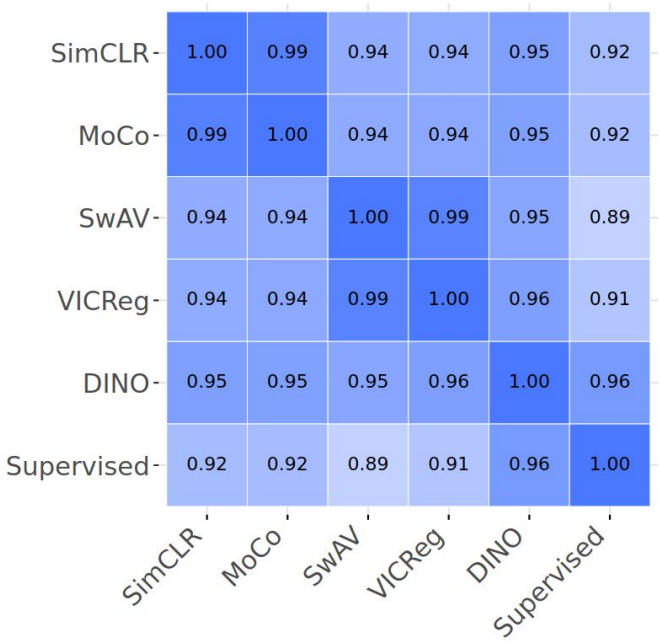
Evaluation on Out-of-Domain Benchmarks

Framework	SITW		VOiCES	
	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
SimCLR	10.09	0.8010	3.92	0.4172
MoCo	10.01	0.7779	3.96	0.4212
SwAV	13.68	0.8709	5.79	0.5954
VICReg	13.59	0.8453	5.59	0.5798
DINO	**4.92**	**0.5123**	**1.85**	**0.2656**
Supervised	2.05	0.2032	1.15	0.1437

I – Self-Supervised Learning for Speaker Verification SSLSV

DET Curves & Correlation Matrix

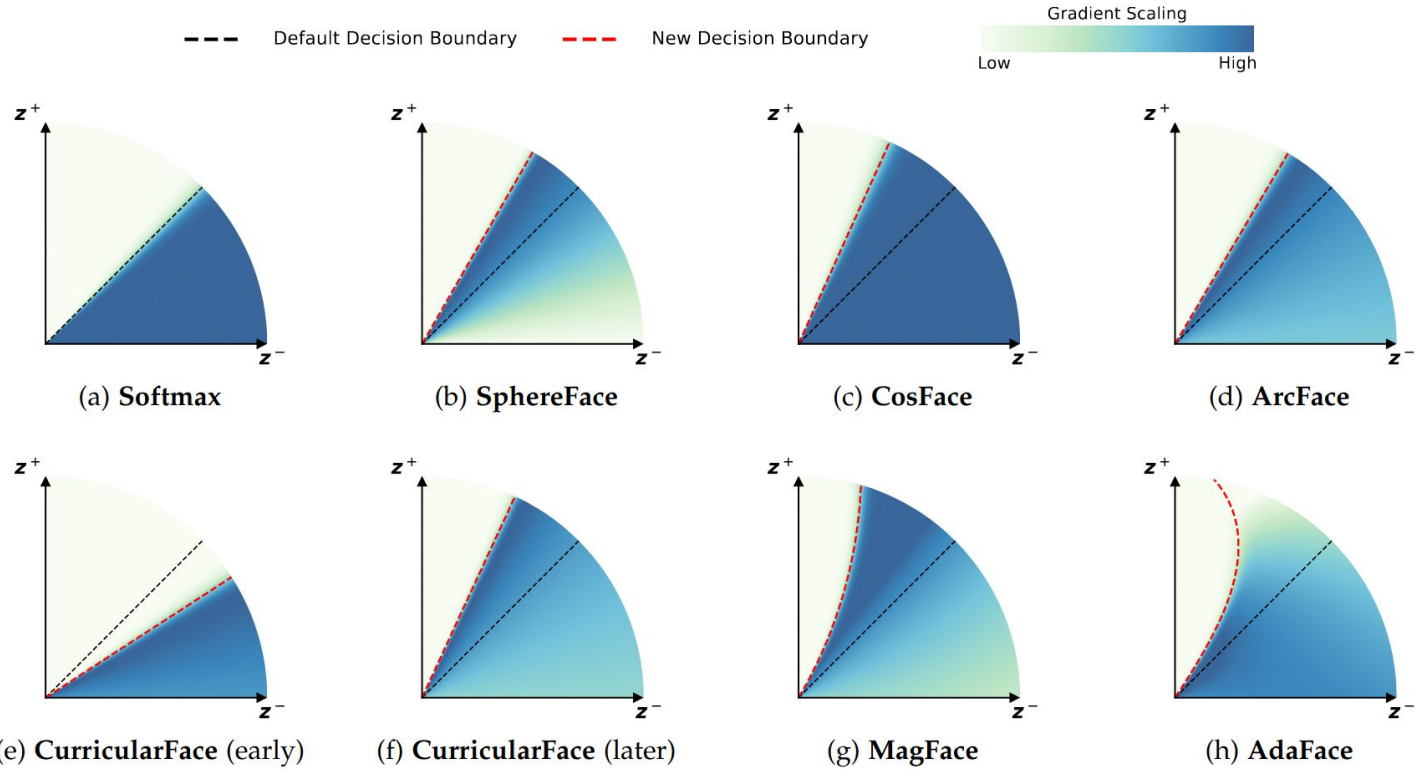




II – Learning Discriminative Speaker Representations

Margins

Visualization of Margin-Based Approaches



II – Learning Discriminative Speaker Representations

Margins

Definition of the Margin-Based Losses (1)

The following standard margins are integrated into SimCLR and MoCo: SphereFace [1], CosFace [2], and ArcFace [3].

CosFace [2] introduces an Additive Margin (AM) by subtracting a fixed margin hyperparameter from the anchor-positive cosine similarity.

$$\mathcal{L}^{(\text{Cos})}_{\text{SimCLR}},\ \mathcal{L}^{(\text{Cos})}_{\text{MoCo}} : \begin{cases} \ell^{+}(\boldsymbol{a},\boldsymbol{b}) = \left(\cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}}\right) - \underset{\text{AM}}{m}\right)/\tau \\ \ell^{-}(\boldsymbol{a},\boldsymbol{b}) = \cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}}\right)/\tau \end{cases}$$

ArcFace [3] applies an Additive Angular Margin (AAM) by shifting the anchor-positive angle by a fixed margin hyperparameter.

$$\mathcal{L}^{(\text{Arc})}_{\text{SimCLR}},\ \mathcal{L}^{(\text{Arc})}_{\text{MoCo}} : \begin{cases} \ell^{+}(\boldsymbol{a},\boldsymbol{b}) = \cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}} + \underset{\text{AAM}}{m}\right)/\tau \\ \ell^{-}(\boldsymbol{a},\boldsymbol{b}) = \cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}}\right)/\tau \end{cases}$$

[1] W. Liu et al. *SphereFace: Deep Hypersphere Embedding for Face Recognition*. CVPR, 2017.
[2] H. Wang et al. *CosFace: Large Margin Cosine Loss for Deep Face Recognition*. CVPR, 2018.
[3] J. Deng et al. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*. CVPR, 2019.

II – Learning Discriminative Speaker Representations

Margins

Definition of the Margin-Based Losses (2)

The following adaptive margins are integrated into SimCLR and MoCo: CurricularFace [4], MagFace [5], and AdaFace [6].

AdaFace [6] jointly adapts the AM and AAM based on sample quality, which is inferred from the feature norm.

$$\widehat{\|\boldsymbol{z}_i\|} = \left\lfloor \frac{\mathrm{sg}\left(\|\boldsymbol{z}_i\|\right) - \mu_z}{\sigma_z / h} \right\rceil_{-1}^{1}$$

$$\mathcal{L}_{\text{SimCLR}}^{(\text{Ada})},\ \mathcal{L}_{\text{MoCo}}^{(\text{Ada})} : \begin{cases} \ell^{+}(\boldsymbol{a},\boldsymbol{b}) = \left(\cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}} + m_{\text{angle}}\right) - m_{\text{add}}\right)/\tau \\ \ell^{-}(\boldsymbol{a},\boldsymbol{b}) = \cos\left(\theta_{\boldsymbol{a},\boldsymbol{b}}\right)/\tau \end{cases}$$

AAM ↓ AM ↓

$$m_{\text{angle}} = -m \cdot \widehat{\|\boldsymbol{z}_i\|}$$

$$m_{\text{add}} = m \cdot \widehat{\|\boldsymbol{z}_i\|} + m$$

[4] Y. Huang et al. *CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition*. CVPR, 2020.
[5] Q. Meng et al. *MagFace: A Universal Representation for Face Recognition and Quality Assessment*. CVPR, 2021.
[6] M. Kim et al. *AdaFace: Quality Adaptive Margin for Face Recognition*. CVPR, 2022.

II – Learning Discriminative Speaker Representations

Margins

Performance of Margin-Based Losses

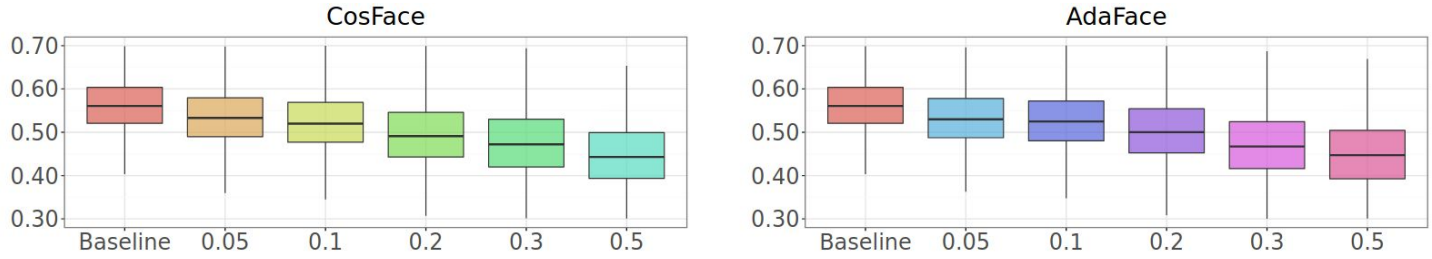
Method	Hyper-params.	SimCLR		MoCo	
	m	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Baseline		9.05	0.6364	8.49	0.5990
SphereFace	2	8.87	**0.6179**	8.93	0.6378
	3	**8.80**	0.6236	8.62	0.6293
	4	9.09	0.6203	8.91	0.6615
CosFace	0.05	8.79	0.6232	8.72	0.6143
	0.1	8.10	0.6085	8.78	0.6260
	0.2	**7.80**	**0.6065**	8.54	0.6090
	0.3	8.33	0.6084	8.58	0.6123
	0.5	9.16	0.6673	8.59	0.6144
ArcFace	0.005	8.82	0.6276	8.52	0.6238
	0.01	**8.88**	**0.6235**	8.67	0.6147
	0.05	49.99	1.0000	8.91	0.6065
	0.1	49.99	1.0000	9.34	0.6305
	0.2	49.99	1.0000	49.99	1.0000

Method	Hyper-params.	SimCLR		MoCo	
	m	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
Baseline		9.05	0.6364	8.49	0.5990
CurricularFace	0.005	8.24	**0.5965**	8.86	0.6178
	0.01	8.31	0.5979	8.95	0.6332
	0.05	8.22	0.6155	9.00	0.5959
	0.1	**7.97**	0.6183	8.80	0.6464
	0.2	49.99	1.0000	9.48	0.6564
MagFace	[0.01; 0.05]	**8.51**	**0.5934**	8.97	0.6177
	[0.05; 0.10]	49.99	1.0000	8.86	0.6364
AdaFace	0.05	8.44	0.6232	8.88	0.6295
	0.1	8.07	0.6192	8.79	0.6117
	0.2	**7.70**	0.6067	8.55	0.6064
	0.3	8.00	**0.5883**	8.78	0.6261
	0.5	8.63	0.6292	8.69	0.6188

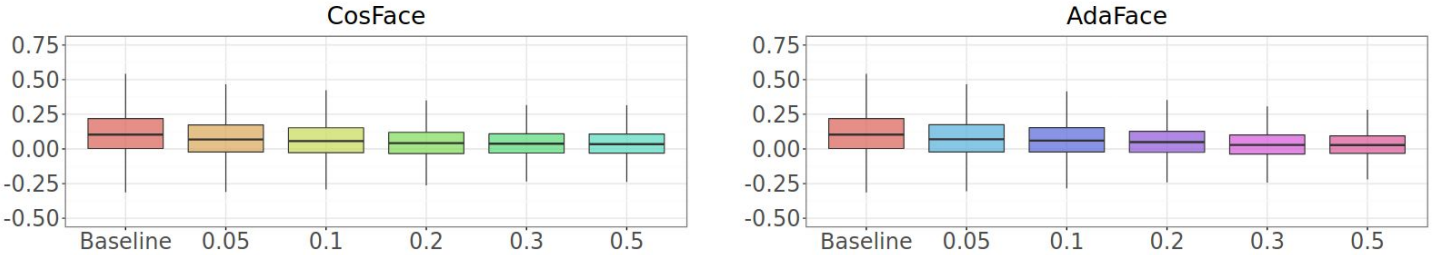
II – Learning Discriminative Speaker Representations

Margins

Intra & Inter-Speaker Similarity



(a) **Intra-speaker similarity**

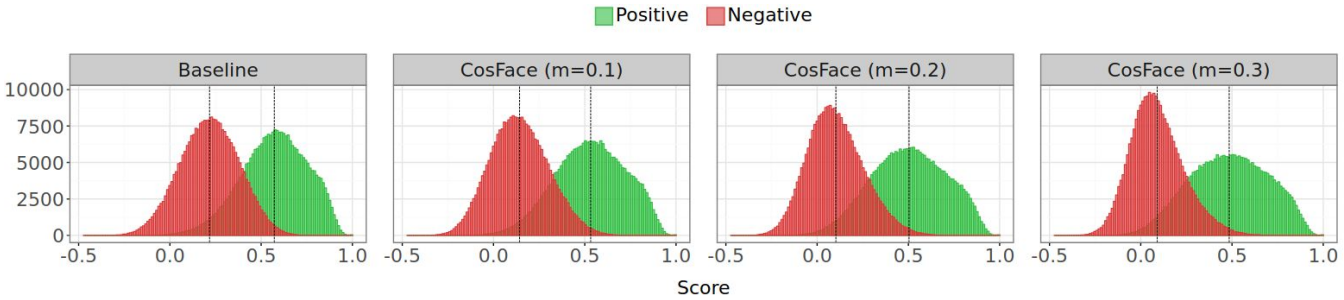


(b) **Inter-speaker similarity**

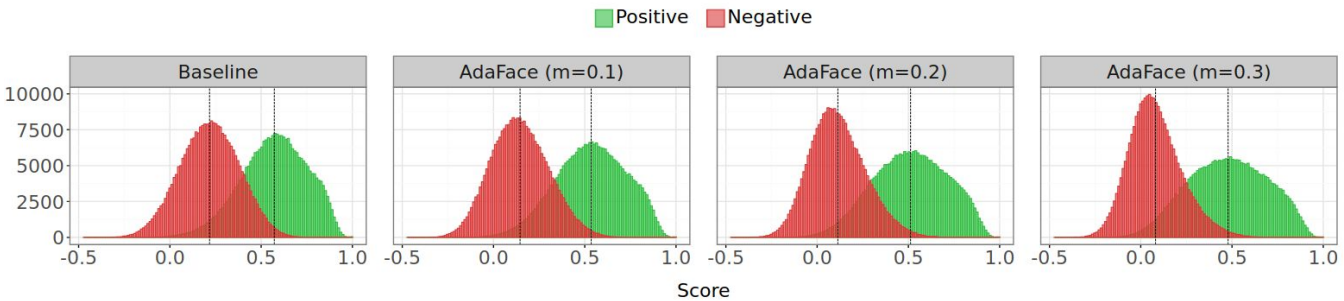
II – Learning Discriminative Speaker Representations

Margins

Score Distributions



(a) **CosFace**



(b) **AdaFace**

II – Learning Discriminative Speaker Representations

Margins

Impact of the Self-Supervised Training

- **Experiment A & B:** Margins remain effective across varying number of negatives and speaker counts in the training set.
- **Experiment C:** Margins provide comparable relative improvements in supervised and self-supervised settings.
- **Experiment D:** Class collisions and class imbalance have negligible impact, while supervised positive sampling significantly boosts performance, identifying positive pair construction as a fundamental bottleneck.

Configuration	*Baseline*		*CosFace* ($m = 0.2$)	
	EER (%)	**minDCF$_{0.01}$**	**EER (%)**	**minDCF$_{0.01}$**
Experiment A: *Batch size (B)*				
$B = 128$	9.26	0.6489	7.71	0.6166
$B = 256$ ●	**9.05**	**0.6364**	**7.80**	**0.6065**
$B = 512$	8.81	0.6474	7.81	0.6198
Experiment B: *Number of speakers (C)*				
$C = 1000$	11.61	0.7148	10.59	0.6663
$C = 3000$	9.49	0.6548	8.83	0.6271
$C = 6000$ ●	**9.05**	**0.6364**	**7.80**	**0.6065**
Experiment C: *Supervised vs. SSL*				
SSL ●	9.05	0.6364	7.80	0.6065
Supervised	**3.22**	**0.3520**	**3.07**	**0.3156**
Experiment D: *Training setup*				
Baseline ●	9.05	0.6364	7.80	0.6065
Balanced training set	9.17	0.6140	8.01	0.5981
No class collisions	8.84	0.6371	7.83	0.5963
Supervised positive sampling	**3.36**	**0.3905**	**3.38**	**0.4367**

Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

II – Learning Discriminative Speaker Representations

Margins

Impact of the Self-Supervised Training

- **Experiment A & B:** Margins remain effective across varying number of negatives and speaker counts in the training set.
- **Experiment C:** Margins provide comparable relative improvements in supervised and self-supervised settings.
- **Experiment D:** Class collisions and class imbalance have negligible impact, while supervised positive sampling significantly boosts performance, identifying positive pair construction as a fundamental bottleneck.

Configuration	*Baseline*		*CosFace* ($m = 0.2$)	
	EER (%)	**minDCF$_{0.01}$**	**EER (%)**	**minDCF$_{0.01}$**
Experiment A: *Batch size (B)*				
$B = 128$	9.26	0.6489	7.71	0.6166
$B = 256$ ●	**9.05**	**0.6364**	**7.80**	**0.6065**
$B = 512$	8.81	0.6474	7.81	0.6198
Experiment B: *Number of speakers (C)*				
$C = 1000$	11.61	0.7148	10.59	0.6663
$C = 3000$	9.49	0.6548	8.83	0.6271
$C = 6000$ ●	**9.05**	**0.6364**	**7.80**	**0.6065**
Experiment C: *Supervised vs. SSL*				
SSL ●	9.05	0.6364	7.80	0.6065
Supervised	**3.22**	**0.3520**	**3.07**	**0.3156**
Experiment D: *Training setup*				
Baseline ●	9.05	0.6364	7.80	0.6065
Balanced training set	9.17	0.6140	8.01	0.5981
No class collisions	8.84	0.6371	7.83	0.5963
Supervised positive sampling	**3.36**	**0.3905**	**3.38**	**0.4367**

Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

II – Learning Discriminative Speaker Representations

Margins

Impact of the Self-Supervised Training

- **Experiment A & B:** Margins remain effective across varying number of negatives and speaker counts in the training set.
- **Experiment C:** Margins provide comparable relative improvements in supervised and self-supervised settings.
- **Experiment D:** Class collisions and class imbalance have negligible impact, while supervised positive sampling significantly boosts performance, identifying positive pair construction as a fundamental bottleneck.

Configuration	*Baseline*		*CosFace* ($m = 0.2$)	
	EER (%)	**minDCF$_{0.01}$**	**EER (%)**	**minDCF$_{0.01}$**
Experiment A: *Batch size (B)*				
$B = 128$	9.26	0.6489	7.71	0.6166
$B = 256$ ●	**9.05**	**0.6364**	**7.80**	**0.6065**
$B = 512$	8.81	0.6474	7.81	0.6198
Experiment B: *Number of speakers (C)*				
$C = 1000$	11.61	0.7148	10.59	0.6663
$C = 3000$	9.49	0.6548	8.83	0.6271
$C = 6000$ ●	**9.05**	**0.6364**	**7.80**	**0.6065**
Experiment C: *Supervised vs. SSL*				
SSL ●	9.05	0.6364	7.80	0.6065
Supervised	3.22	0.3520	3.07	0.3156
Experiment D: *Training setup*				
Baseline ●	9.05	0.6364	7.80	0.6065
Balanced training set	9.17	0.6140	8.01	0.5981
No class collisions	8.84	0.6371	7.83	0.5963
Supervised positive sampling	**3.36**	**0.3905**	**3.38**	**0.4367**

Benchmark: VoxCeleb1-O • Encoder: Fast ResNet-34

Evaluation

Method	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$	EER (%)	$\text{minDCF}_{0.01}$
AP (Huh et al. 2020)	9.56					
AP + AAT (Huh et al. 2020)	8.65					
Contrastive + VICReg (Lepage et al. 2022)	8.47	0.6400				
SimCLR + MSE loss (Haoran Zhang et al. 2021)	8.28	0.6100				
MoCo + ProtoNCE (W. Xia et al. 2021)	8.23	0.5900				
CEL (Mun et al. 2020)	8.01					
SimCLR	9.05	0.6364	9.79	0.6769	15.21	0.7664
SimCLR + CosFace ($m = 0.2$)	7.80	0.6065	**8.80**	**0.6597**	**13.72**	0.7676
SimCLR + AdaFace ($m = 0.2$)	**7.70**	0.6067	8.90	0.6613	13.76	**0.7648**
SimCLR + AdaFace ($m = 0.3$)	8.00	**0.5883**	9.35	0.6697	14.08	0.7649

III – Self-Supervised Positive Sampling from Latent Space SSPS

Maths & Pseudocode

SSPS-NN

$$\mathcal{N}_i \triangleq \underset{j \neq i}{\text{top k}} \left(\left\{ \text{sim} \left(\hat{\boldsymbol{q}}_i, \hat{\boldsymbol{q}}_j \right), \forall j \in \mathcal{I} \right\}; M \right)$$

$$pos(i) = \text{sample} \left(\mathcal{N}_i \right)$$

SSPS-Clustering

$$\mathcal{C}_k \triangleq \underset{j \neq k}{\text{top k}} \left(\{ \text{sim} \left(\boldsymbol{m}_k, \boldsymbol{m}_j \right), \forall j \in [1, K] \}; M \right)$$

$$pos(i) = \text{sample} \left(\mathcal{S}_{\hat{c}_i} \right)$$

$$\mathcal{S}_k \triangleq \{ j \in \mathcal{I} \text{ s.t. } c_j = k \}$$

Algorithm 7.1 SSPS-Clustering: Epoch Initialization

Inputs: K, M, $\hat{\boldsymbol{Q}}$

1: Run k-means on $\hat{\boldsymbol{Q}}$ with K clusters
2: Determine $\{\mathcal{C}_k\}_{k=1,\ldots,K}$ by Equation (7.5)
3: Determine $\{\mathcal{S}_k\}_{k=1,\ldots,K}$ by Equation (7.7)
4: **Return** $\{\mathcal{C}_k\}$, $\{\mathcal{S}_k\}$

Algorithm 7.2 SSPS-Clustering: Training Iteration

Inputs: $\mathcal{B}$, $\hat{\boldsymbol{Y}}$, $\boldsymbol{Z}$, $\boldsymbol{Z}'$, $\mathcal{L}$, $\boldsymbol{Q}'$

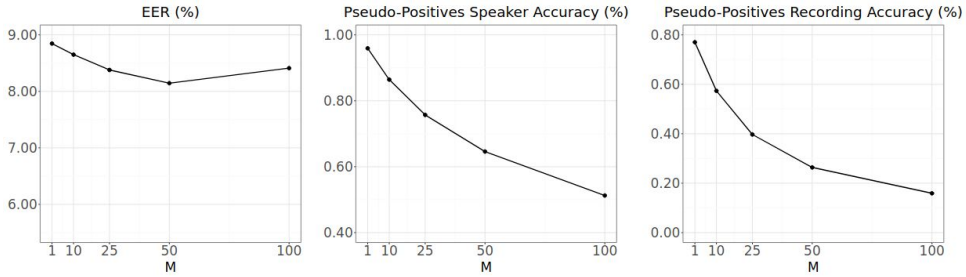
1: $\boldsymbol{Z}'_{\text{SSPS}} \leftarrow \left\{ \begin{array}{ll} \boldsymbol{q}'_{pos(i)}, & \text{if } pos(i) \neq \varnothing, \\ \boldsymbol{z}'_i, & \text{otherwise} \end{array} \right\}_{i \in \mathcal{B}}$
2: Optimize model with $\mathcal{L} \left(\boldsymbol{Z}, \boldsymbol{Z}'_{\text{SSPS}} \right)$
3: Insert $\hat{\boldsymbol{Y}}$ into $\hat{\boldsymbol{Q}}$
4: Insert $\boldsymbol{Z}'$ into $\boldsymbol{Q}'$

III – Self-Supervised Positive Sampling from Latent Space

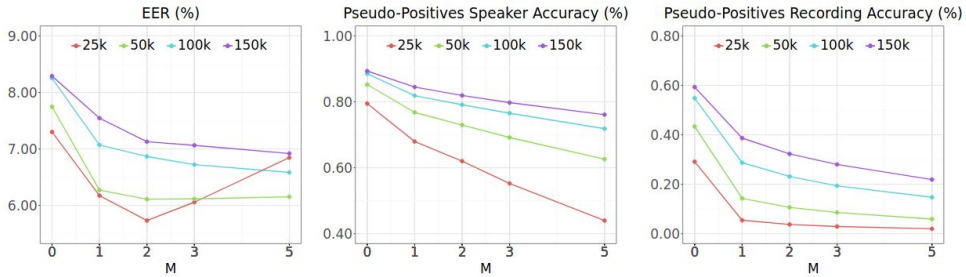
SSPS

Hyperparameters Search

Positive sampling	Hyper-params.		VoxCeleb1-O	
	K	M	EER (%)	$\text{minDCF}_{0.01}$
SSL			9.41	0.6378
SSPS-NN		1	8.85	0.6240
		10	8.65	0.6186
		25	8.38	0.6189
		50	**8.15**	**0.6145**
		100	8.41	0.6164
SSPS-Clustering	6000	0	6.63	0.5493
	10 000	0	6.82	0.5629
	25 000	0	7.30	0.5805
		1	5.80	**0.5250**
		2	**5.73**	0.5258
		3	6.06	0.5410
		5	6.85	0.5672
	50 000	5	6.15	0.5282
	100 000	5	6.58	0.5627
	150 000	5	6.92	0.5483
SSPS-Clustering (C)	6000	0	13.31	0.8125
	25 000	1	**11.04**	**0.7453**
Supervised			3.93	0.3900



(a) **SSPS-NN**

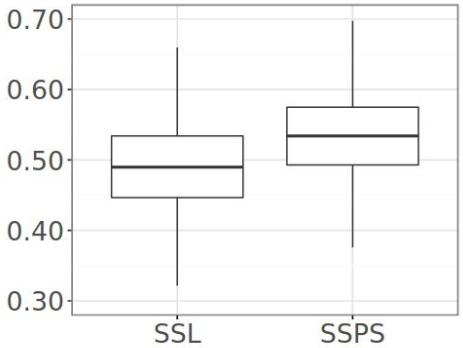


(b) **SSPS-Clustering**

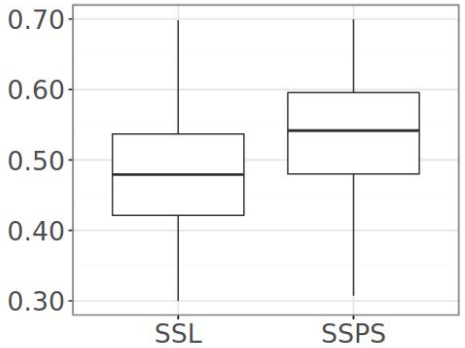
Additional Resources

III – Self-Supervised Positive Sampling from Latent Space SSPS

Intra-Speaker Similarity



(a) **Test data** (VoxCeleb1)

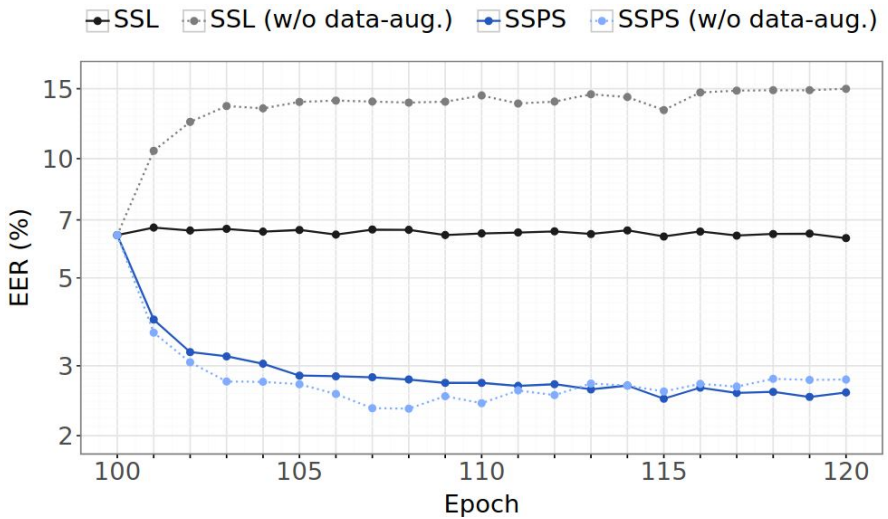


(b) **Train data** (VoxCeleb2)

III – Self-Supervised Positive Sampling from Latent Space

SSPS

Robustness to the Absence of Data-Augmentation



IV – Leveraging Speech Foundation Models

Foundation

Performance of Front-Ends and Back-Ends

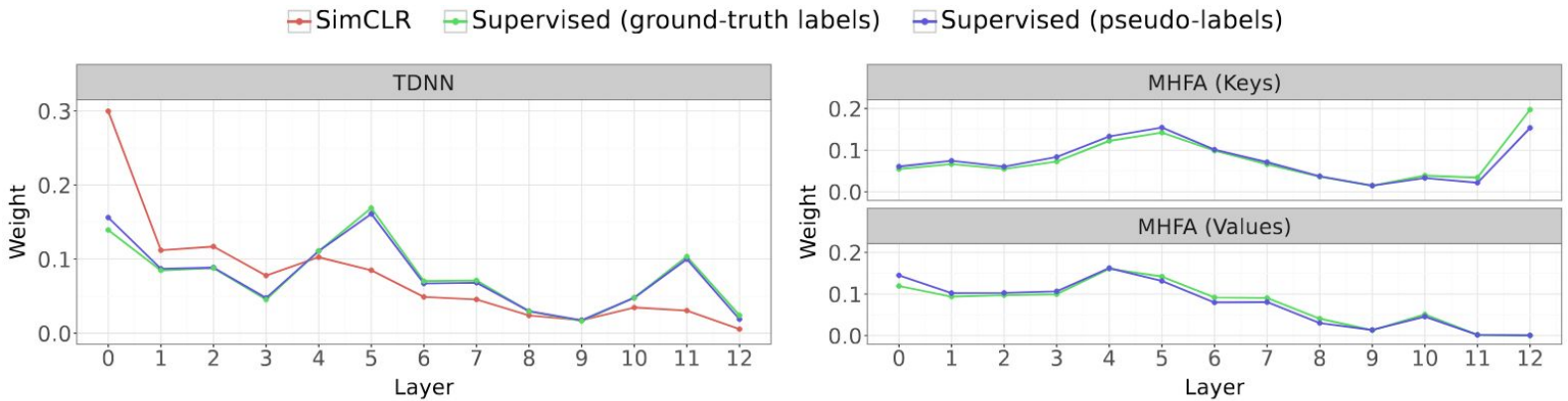
Front-end	# params.	*Pseudo-labels*		*Ground-truth labels*	
		EER (%)	**minDCF$_{0.01}$**	**EER (%)**	**minDCF$_{0.01}$**
wav2vec 2.0 Base	95 M	2.47	0.2142	0.85	0.1240
HuBERT Base	94 M	2.02	0.2085	0.75	0.1074
WavLM Base	94 M	1.99	0.1950	0.75	0.1110
WavLM Base+	94 M	2.07	0.1874	0.74	0.1020
WavLM Large	316 M	**1.66**	**0.1747**	**0.58**	**0.0675**

Back-end	# params.	*Pseudo-labels*		*Ground-truth labels*	
		EER (%)	**minDCF$_{0.01}$**	**EER (%)**	**minDCF$_{0.01}$**
ASP *(final layer)*	93.4 k	4.28	0.5468	•	•
SAP *(final layer)*	93.4 k	4.22	0.5064	•	•
ASP *(all layers)*	93.4 k	2.40	0.2377	0.91	0.1167
SAP *(all layers)*	93.4 k	2.46	0.2314	0.93	0.1286
MHFA *(all layers)*	2.3 M	**2.07**	**0.1874**	**0.74**	**0.1020**
TDNN *(all layers)*	5.4 M	2.76	0.2708	0.98	0.1315
ECAPA-TDNN *(all layers)*	18.4 M	2.30	0.2161	0.77	0.1093

IV – Leveraging Speech Foundation Models

Foundation

Layer-Wise Front-End Weights



(a) **TDNN** (b) **MHFA**

IV – Leveraging Speech Foundation Models

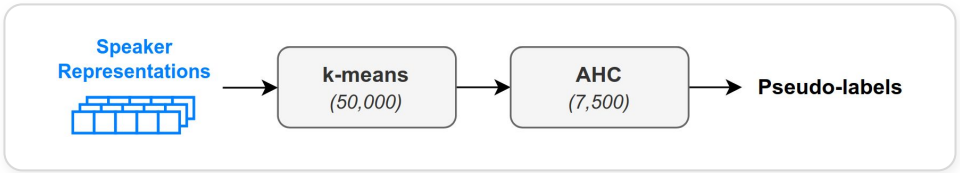
Foundation

Effect of Pseudo-Label Generation, Filtering & Refinement

Clustering	Filtering	Iteration	NMI (%)	VoxCeleb1-O	
				EER (%)	$\text{minDCF}_{0.01}$
k-means (7500)			92.95	2.45	0.2547
k-means (50 000) + AHC (5000)			94.95	**1.94**	0.2102
k-means (50 000) + AHC (6000)	✗		95.37	1.96	0.2044
k-means (50 000) + AHC (7500)			**95.47**	2.07	**0.1874**
k-means (50 000) + AHC (10 000)			95.12	2.17	0.2092
k-means (50 000) + AHC (7500)	✓			**1.63**	**0.1788**
k-means (50 000) + AHC (7500)	✓	1	96.76	1.42	0.1294
		2	97.21	1.27	0.1155
		3	97.43	**1.18**	**0.1108**
		4	**97.50**	1.31	0.1178
		5	97.47	1.33	0.1191

Benchmark: VoxCeleb1-O • Front-end: WavLM Base+ • Back-end: MHFA

- Over-segmentation and AHC improves pseudo-label quality, under high intra-speaker variability.
- Loss-based pseudo-label filtering effectively enhances robustness.
- Iterative pseudo-label refinement yields substantial gains: ~28% EER reduction after three iterations.



IV – Leveraging Speech Foundation Models

Foundation

Effect of Pseudo-Label Generation, Filtering & Refinement

Clustering	Filtering	Iteration	NMI (%)	VoxCeleb1-O	
				EER (%)	$\text{minDCF}_{0.01}$
k-means (7500)			92.95	2.45	0.2547
k-means (50 000) + AHC (5000)			94.95	**1.94**	0.2102
k-means (50 000) + AHC (6000)	✗	•	95.37	1.96	0.2044
k-means (50 000) + AHC (7500)			**95.47**	2.07	**0.1874**
k-means (50 000) + AHC (10 000)			95.12	2.17	0.2092
k-means (50 000) + AHC (7500)	✓	•	•	**1.63**	**0.1788**
		1	96.76	1.42	0.1294
		2	97.21	1.27	0.1155
k-means (50 000) + AHC (7500)	✓	3	97.43	**1.18**	**0.1108**
		4	**97.50**	1.31	0.1178
		5	97.47	1.33	0.1191

ⓘ Benchmark: VoxCeleb1-O • Front-end: WavLM Base+ • Back-end: MHFA

- Over-segmentation and AHC improves pseudo-label quality, under high intra-speaker variability.
- Loss-based pseudo-label filtering effectively enhances robustness.
- Iterative pseudo-label refinement yields substantial gains: ~28% EER reduction after three iterations.

IV – Leveraging Speech Foundation Models

Foundation

Effect of Pseudo-Label Generation, Filtering & Refinement

Clustering	Filtering	Iteration	NMI (%)	VoxCeleb1-O	
				EER (%)	$\text{minDCF}_{0.01}$
k-means (7500)			92.95	2.45	0.2547
k-means (50 000) + AHC (5000)			94.95	**1.94**	0.2102
k-means (50 000) + AHC (6000)	✗		95.37	1.96	0.2044
k-means (50 000) + AHC (7500)			**95.47**	2.07	**0.1874**
k-means (50 000) + AHC (10 000)			95.12	2.17	0.2092
k-means (50 000) + AHC (7500)	✓			**1.63**	**0.1788**
k-means (50 000) + AHC (7500)	✓	1	96.76	1.42	0.1294
		2	97.21	1.27	0.1155
		3	97.43	**1.18**	**0.1108**
		4	**97.50**	1.31	0.1178
		5	97.47	1.33	0.1191

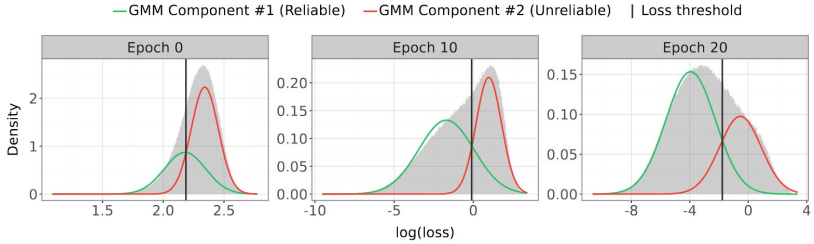
Benchmark: VoxCeleb1-O • Front-end: WavLM Base+ • Back-end: MHFA

- Over-segmentation and AHC improves pseudo-label quality, under high intra-speaker variability.
- Loss-based pseudo-label filtering effectively enhances robustness.
- Iterative pseudo-label refinement yields substantial gains: ~28% EER reduction after three iterations.

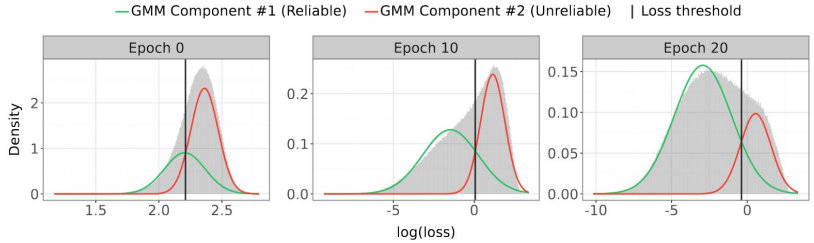
IV – Leveraging Speech Foundation Models

Foundation

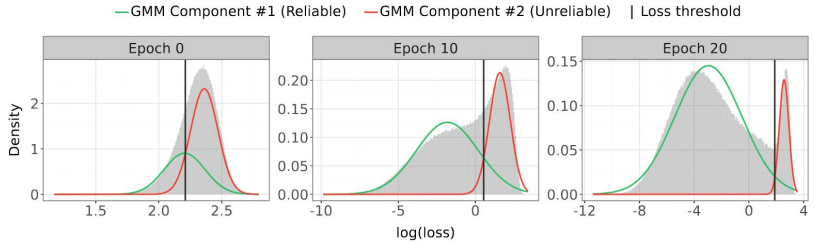
Loss Distributions



(a) **Ground-truth labels**



(b) **Pseudo-labels**



(c) **Pseudo-labels with filtering**

Appendix – Evaluation on Other Downstream Tasks

Results on Language, Emotion, and Gender Recognition

Task	Framework	*Linear*		*Fine-tuned*	
		Accuracy	**F1 score**	**Accuracy**	**F1 score**
Speaker Identification	SimCLR	99.65	99.64	99.62	99.61
	DINO	99.66	99.66	**99.85**	**99.85**
	Supervised	**99.85**	**99.85**	99.84	99.84
Language Recognition	SimCLR	55.81	65.78	79.99	86.08
	DINO	**69.79**	**76.85**	**86.82**	**91.19**
	Supervised	52.02	61.69	84.15	89.12
Emotion Recognition	SimCLR	61.84	60.92	73.59	73.52
	DINO	**69.57**	**69.37**	**79.23**	**79.29**
	Supervised	51.69	51.19	70.53	70.69
Gender Recognition	SimCLR	**97.83**	**97.83**	**98.25**	**98.25**
	DINO	97.81	97.81	97.51	97.50
	Supervised	95.33	95.33	98.11	98.11