Odyssey 2024
The Speaker and Language Recognition Workshop

Additive Margin in Contrastive Self-Supervised Frameworks to Learn Discriminative Speaker Representations

Theo Lepage and Reda Dehak

EPITA Research Laboratory (LRE), France

Code: https://github.com/theolepage/sslsv

Outline

- Introduction
  - Learning representations for speaker verification
  - Self-supervised contrastive learning
  - Margin-based approaches for verification tasks

- Method
  - Overview of our self-supervised training framework
  - Revision of SimCLR contrastive objective function
  - Revision of MoCo contrastive objective function
  - Introduction of Additive Margin in the contrastive loss

- Experiments & Results
  - Experimental setup
  - Effect of our improvements to the self-supervised contrastive loss
  - Impact of class collisions from the SSL training
  - Comparison to other self-supervised contrastive methods for SV

- Conclusions

Learning representations for speaker verification

State-of-the-art **Speaker Verification** (SV) methods compute the **similarity** between two **speaker representations** extracted from **Deep Neural Networks** (DNN) pre-trained on speaker classification [1, 2].

Speaker representations should:

- maximize **inter-speaker** distances ;
- minimize **intra-speaker** variance ;
- discard **extrinsic variabilities** (*e.g. channel, noise, environment, …)*.

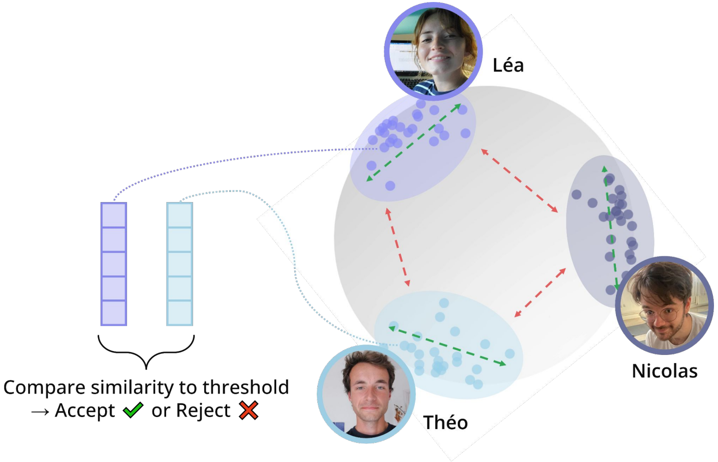


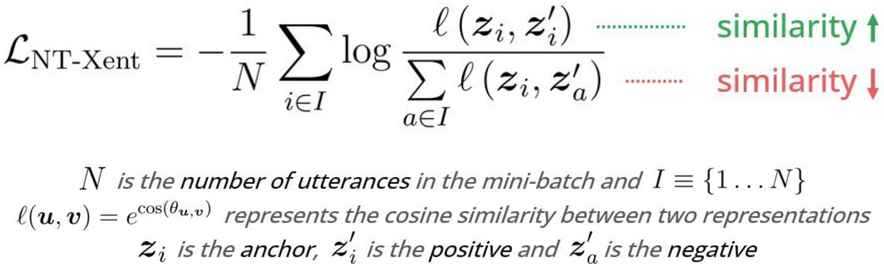
Figure 1. Learning speaker embeddings space for speaker verification systems.

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

Self-supervised contrastive learning

Deep Learning models are inherently dependent on some kind of human supervision as they are trained on large human-labeled datasets → complex and expensive process for speech-related tasks.

Self-supervised contrastive learning [1, 2, 3] learn embeddings directly from raw audio by assuming that two utterances randomly sampled from the training data belong to different speakers.



$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{N}\sum_{i\in I}\log\frac{\ell\left(\boldsymbol{z}_i, \boldsymbol{z}'_i\right)}{\sum_{a\in I}\ell\left(\boldsymbol{z}_i, \boldsymbol{z}'_a\right)}$$

similarity ↑

similarity ↓

N *is the number of utterances in the mini-batch and* $I \equiv \{1\ldots N\}$

$\ell(\boldsymbol{u}, \boldsymbol{v}) = e^{\cos(\theta_{\boldsymbol{u},\boldsymbol{v}})}$ *represents the cosine similarity between two representations*

$\boldsymbol{z}_i$ *is the anchor,* $\boldsymbol{z}'_i$ *is the positive and* $\boldsymbol{z}'_a$ *is the negative*

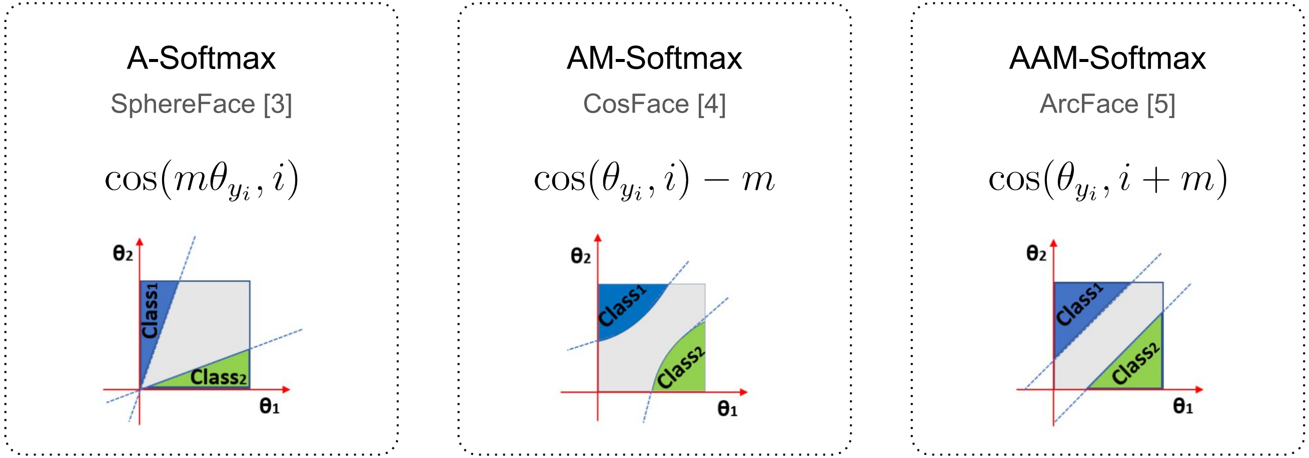
From a given training sample (anchor):

- ❑ the positive is created by applying data-augmentation on the anchor ;
- ❑ the negative is randomly sampled from the mini-batch or a memory queue.

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv preprint arXiv:1807.03748, 2019.
[2] H. Zhang, Y. Zou, and H. Wang, "Contrastive Self-Supervised Learning for Text-Independent Speaker Verification," in ICASSP, 2021.
[3] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised Text-independent Speaker Verification using Prototypical Momentum Contrastive Learning," in ICASSP, 2021.

Margin-based approaches for verification tasks

In supervised settings, margins have been successfully applied to the Softmax classification loss for face and speaker recognition [1, 2] with the aim of producing more discriminative embeddings.



A-Softmax

SphereFace [3]

$$\cos(m\theta_{y_i}, i)$$

AM-Softmax

CosFace [4]

$$\cos(\theta_{y_i}, i) - m$$

AAM-Softmax

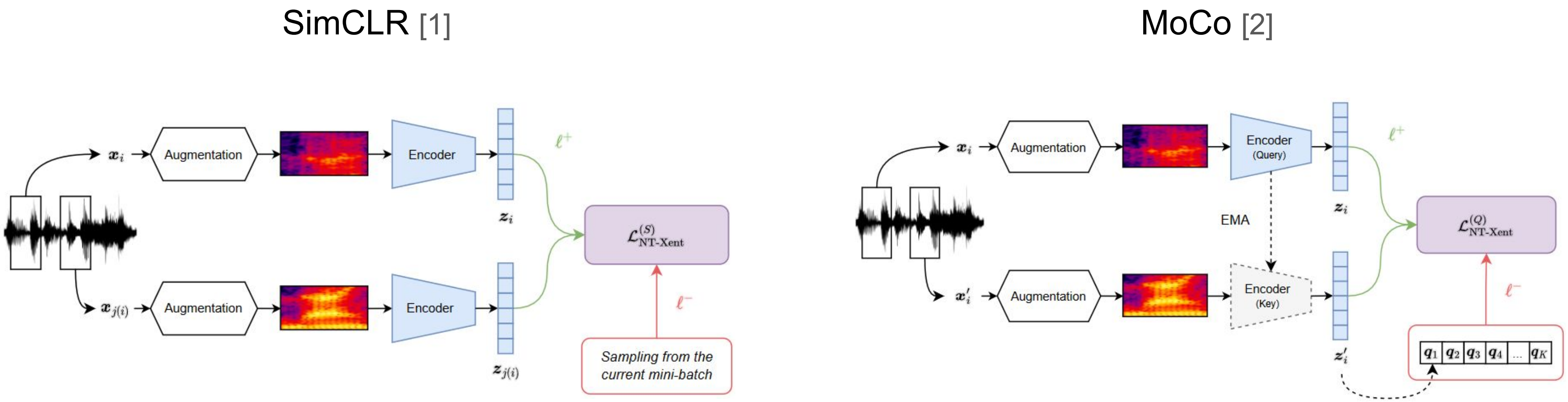
ArcFace [5]

$$\cos(\theta_{y_i}, i + m)$$

Figure 2. Overview of the different margin-based loss functions. Decision margins are from Figure 5 of [5].

[1] Y. Liu, L. He, and J. Liu, “Large Margin Softmax Loss for Speaker Verification,” in Interspeech, 2019.
[2] Y.-Q. Yu, L. Fan, and W.-J. Li, “Ensemble additive margin softmax for speaker verification,” in ICASSP, 2019.
[3] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song, “SphereFace: Deep Hypersphere Embedding for Face Recognition,” in CVPR, 2017.
[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in CVPR, 2018.
[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in CVPR, 2019.

Overview of our self-supervised training framework



SimCLR [1]

MoCo [2]

Figure 3. Diagram of our contrastive self-supervised training framework to learn speaker representations.

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in ICML, 2020.
[2] K. He, H. Fan, Y. Wu, S. Xie, R. Girschick, “Momentum Contrast for Unsupervised Visual Representation Learning,” in CVPR, 2020.

Revision of SimCLR contrastive objective function

SimCLR [1] framework samples negatives from the current mini-batch.

- ➔ Adopt the symmetric formulation of the NT-Xent loss to provide more supervision.
  - ◆ NT-Xent: N positive pairs with N−1 negatives
  - ◆ Symmetric NT-Xent: 2N positive pairs with 2(N-1) negatives

- ➔ Compute the similarity of positive and negative pairs differently to introduce additive margin.

$$\mathcal{L}^{(S)}_{\text{NT-Xent}} = -\frac{1}{2N}\sum_{i\in\hat{I}} \log \frac{\ell^{+}(\boldsymbol{z}_i,\boldsymbol{z}_{j(i)})}{\ell^{+}(\boldsymbol{z}_i,\boldsymbol{z}_{j(i)}) + \sum\limits_{a\in\hat{A}(i)} \ell^{-}(\boldsymbol{z}_i,\boldsymbol{z}_a)}$$

$i \in \hat{I} \equiv \{1 \dots 2N\}$

$j(i)$ is the index of the positive

$\hat{A}(i) \equiv \hat{I} \setminus \{i, j(i)\}$

$\ell^{+}(\mathbf{u},\mathbf{v}) = \ell^{-}(\mathbf{u},\mathbf{v}) = e^{\cos(\theta_{\boldsymbol{u},\boldsymbol{v}})/\tau}$

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in ICML, 2020.

Revision of MoCo contrastive objective function

MoCo [1] framework samples negatives from a memory queue of previous embeddings.

- ➔ Using a large queue of negatives alleviate the need of a symmetric loss.
- ➔ Compute the similarity of positive and negative pairs differently to introduce additive margin.

$$\mathcal{L}^{(Q)}_{\text{NT-Xent}} = -\frac{1}{N}\sum_{i \in I} \log \frac{\ell^{+}\left(\boldsymbol{z}_i, \boldsymbol{z}'_i\right)}{\ell^{+}\left(\boldsymbol{z}_i, \boldsymbol{z}'_i\right) + \sum_{b \in B} \ell^{-}\left(\boldsymbol{z}_i, \boldsymbol{q}_b\right)}$$

$$i \in I \equiv \{1 \dots N\}$$

$$B \equiv \{1 \dots K\}$$

$\boldsymbol{q}_i$ is the i-th element of the queue

$$\ell^{+}(\mathbf{u}, \mathbf{v}) = \ell^{-}(\mathbf{u}, \mathbf{v}) = e^{\cos(\theta_{\boldsymbol{u},\boldsymbol{v}})/\tau}$$

[1] K. He, H. Fan, Y. Wu, S. Xie, R. Girschick, “Momentum Contrast for Unsupervised Visual Representation Learning,” in CVPR, 2020.

Introduction of Additive Margin in the contrastive loss

The contrastive loss aims to penalize classification errors instead of producing discriminative representations relevant to the context of speaker verification.

Inspired by Additive Margin (CosFace) [1], we introduce $m \geq 0$ in cosine space to force the cosine similarity of positive pairs to be above a specific threshold and thus improve speaker separability.

$$\ell^{+}(\mathbf{u},\mathbf{v}) = e^{(\cos(\theta_{\mathbf{u},\mathbf{v}})-m)/\tau}$$

$$\ell^{-}(\mathbf{u},\mathbf{v}) = e^{\cos(\theta_{\boldsymbol{u},\boldsymbol{v}})/\tau}$$

➔ This creates a stringent constraint as the positive similarity has to be at least greater than the maximal negative similarity plus the margin constant: $\cos\left(\theta_{\boldsymbol{z}_a,\boldsymbol{z}_p}\right) - m > \cos\left(\theta_{\boldsymbol{z}_a,\boldsymbol{z}_n}\right)$.

[1] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in CVPR, 2018.

Experimental setup

- Datasets and feature extraction
  - Training on VoxCeleb2 *dev set* [1]
  - Evaluation on VoxCeleb1 ‘original’ *test* set
  - Speaker labels are discarded
  - 2 seconds audio chunks
  - 40-dimensional log-mel spectrogram input features

- Data augmentation
  - Music, speech and babble background noises from the MUSAN [2]
  - Reverberation from the Simulated Room Impulse Response Database [3]

- Model architecture and training
  - Encoder: Fast ResNet-34 [4]
  - Projector: none
  - By default τ is set to 1/30 ≈ 0.0333
  - Epochs: 150
  - Optimizer: Adam (no weight decay)
  - Batch size: 200
  - 2x NVIDIA Tesla V100 16 GB

- Evaluation protocol
  - Scoring with cosine similarity
  - Equal Error Rate (EER)
  - minimum Detection Cost Function (minDCF) with p=0.01

[1] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” in Interspeech, 2018.
[2] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” arXiv preprint arXiv:1510.08484, 2015.
[3] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in ICASSP, 2017.
[4] J.S. Chung, J. Huh, S. Mun, M. Lee, H.S. Heo, S. Chloe, C. Ham, S.W. Jung, B.J. Lee, and I. Han, “In Defence of Metric Learning for Speaker Recognition”, in Interspeech, 2020.

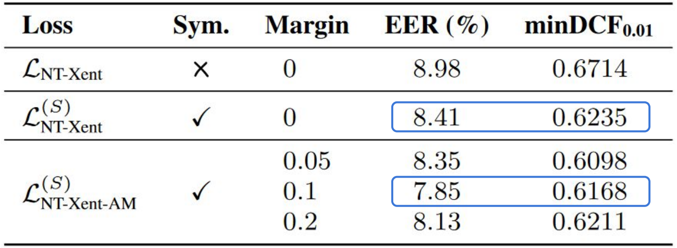
Effect of our improvements to the self-supervised contrastive loss

- Using a symmetric loss for SimCLR and additive margin for both frameworks improve downstream performance.

- The best result is obtained with SimCLR (m=0.1) which achieves 7.85% EER.

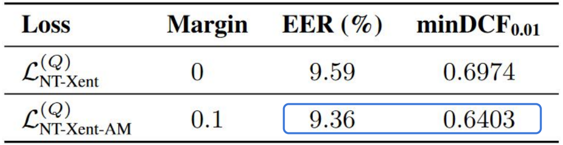
  This margin value corresponds to the value often used for AM-Softmax supervised training.

- Choosing the right margin factor is fundamental: trade-off between discriminative capacity and learning task complexity.



Loss	Sym.	Margin	EER (%)	minDCF$_{0.01}$
$\mathcal{L}_{\text{NT-Xent}}$	×	0	8.98	0.6714
$\mathcal{L}^{(S)}_{\text{NT-Xent}}$	✓	0	8.41	0.6235
$\mathcal{L}^{(S)}_{\text{NT-Xent-AM}}$	✓	0.05	8.35	0.6098
		0.1	7.85	0.6168
		0.2	8.13	0.6211

Table 1. The effect of the symmetric contrastive loss and additive margin on SimCLR self-supervised training performance on SV.

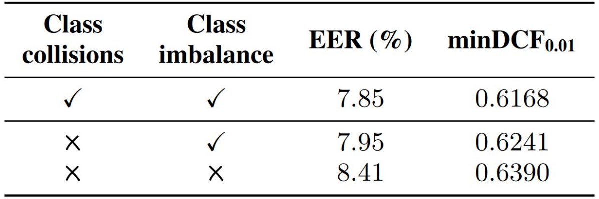


Loss	Margin	EER (%)	minDCF$_{0.01}$
$\mathcal{L}^{(Q)}_{\text{NT-Xent}}$	0	9.59	0.6974
$\mathcal{L}^{(Q)}_{\text{NT-Xent-AM}}$	0.1	9.36	0.6403

Table 2. The effect of additive margin on MoCo self-supervised training performance on SV.

Impact of class collisions from the SSL training

➔ Removing class collisions does not result in better downstream performance. The probability of class collisions is too small as VoxCeleb2 contains many speakers compared to the batch size.



Class collisions	Class imbalance	EER (%)	$\text{minDCF}_{0.01}$
✓	✓	7.85	0.6168
×	✓	7.95	0.6241
×	×	8.41	0.6390

Table 3. The impact of class collisions and class imbalance, which stems from SSL, on SV results.

➔ Our method successfully separate positive from negative scores even further which is consistent with the improvement of the Equal Error Rate.

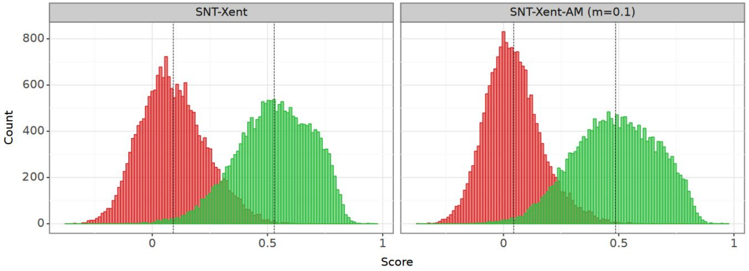
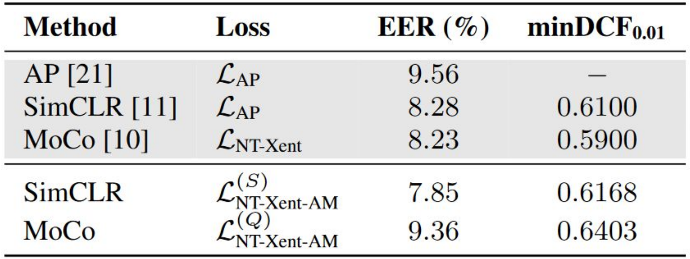


Figure 5. Positive (green) and negative (red) trials scores distribution obtained after training with and without additive margin. SNT-Xent-AM (m = 0.1) losses. The mean of each distribution is represented by a vertical dashed line.

Comparison to other self-supervised contrastive methods for SV



Method	Loss	EER (%)	minDCF$_{0.01}$
AP [21]	$\mathcal{L}_{\text{AP}}$	9.56	–
SimCLR [11]	$\mathcal{L}_{\text{AP}}$	8.28	0.6100
MoCo [10]	$\mathcal{L}_{\text{NT-Xent}}$	8.23	0.5900
SimCLR	$\mathcal{L}^{(S)}_{\text{NT-Xent-AM}}$	7.85	0.6168
MoCo	$\mathcal{L}^{(Q)}_{\text{NT-Xent-AM}}$	9.36	0.6403

Table 4. Final results of different self-supervised contrastive methods on speaker verification.

➔ We outperform other self-supervised contrastive methods based on objective functions equivalent to NT-Xent / Angular Prototypical (without margins).

➔ Self-supervised contrastive frameworks can be further improved with optimizations tailored for the training setup (additional positive and negative pairs) and the downstream task (margins).

[10] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, “Self-supervised Text-independent Speaker Verification using Prototypical Momentum Contrastive Learning,” in ICASSP, 2021.
[11] H. Zhang, Y. Zou, and H. Wang, “Contrastive Self-Supervised Learning for Text-Independent Speaker Verification,” in ICASSP, 2021.
[21] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. S. Chung, “Augmentation adversarial training for unsupervised speaker recognition,” in Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS, 2020.

Conclusions

- Our method achieves 7.85% EER on VoxCeleb1-O test set which is competitive with other equivalent approaches and shows that margins improve the discriminative capacity of representations learned with SSL even with the existence of class collisions.

- Perspectives:
  - Reduce the overfitting on channel characteristics caused by the same-utterance positive sampling → margins would be more effective
  - Experiment with other margin-based loss functions
  - Assess the effectiveness on other tasks and modalities (CV, LID, …)