INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

Experimenting with Additive Margins for Contrastive Self-Supervised Speaker Verification

Théo Lepage, Réda Dehak

Speaker and Language Recognition Group (ESLR),
Laboratoire de Recherche de l’EPITA, France

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

Outline

- Introduction
  - Learning embeddings for speaker verification
  - Self-supervised learning with a contrastive loss
  - Margin-based approaches for verification tasks
- Method
  - Symmetric contrastive loss formulation
  - Introducing Additive Margins in the contrastive loss
  - Introducing Additive Angular Margins in the contrastive loss
  - Overview of our self-supervised framework
- Experiments and results
  - Experimental setup
  - Effect of the symmetric loss and the margins
  - Study of the distribution of positive and negative scores
  - Final self-supervised results
  - *Additional results on VoxCeleb2*
- Conclusions

Learning embeddings for speaker verification

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

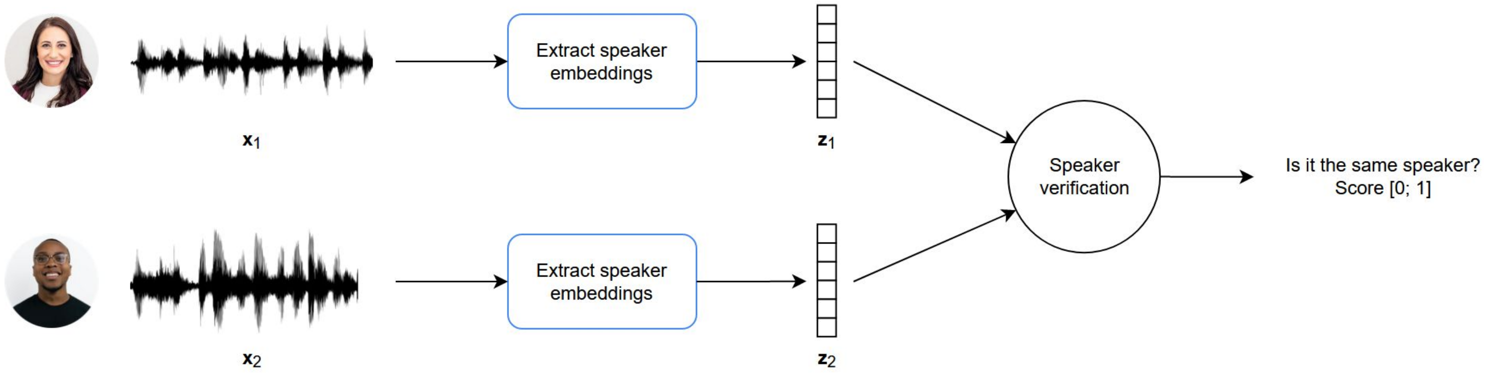


Figure 1. Learning speaker embeddings space for speaker verification systems.

State-of-the-art methods are based on deep learning models [1, 2], inherently dependent on some kind of human supervision, as they are trained on massive amounts of labeled data.

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

Self-supervised learning with a contrastive loss

Large human-labeled datasets are important for the development of deep learning as the capacity of recent models is constantly increasing.

→ However, labeling datasets is expensive, tedious, slow and not scalable to the data available online.

Self-supervised contrastive learning [1, 2, 3] learn embeddings directly from raw audio by assuming that each utterance in the mini-batch $\mathbf{Z} \in \mathbb{R}^{N \times D}$, and its augmented copy $\mathbf{Z}' \in \mathbb{R}^{N \times D}$, belongs to a unique speaker.

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{N} \sum_{i=1}^{N} - \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}'_i / \tau\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_j / \tau\right)}$$

[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.

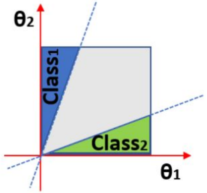
Margin-based approaches for verification tasks

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]

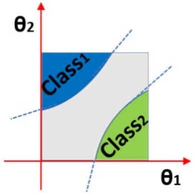
$$\cos(m\theta_{y_i}, i)$$



AM-Softmax

CosFace [4]

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



AAM-Softmax

ArcFace [5]

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

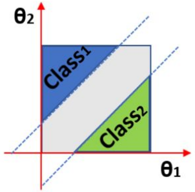


Figure 2. Overview of the different margin-based loss functions. Decision margins are from Figure 5 of [5].

[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.

Symmetric contrastive loss formulation

Normalized Temperature-scaled Cross Entropy loss (NT-Xent)

→ N positive pairs have N − 1 negatives

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

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

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

Symmetric NT-Xent loss (SNT-Xent) [1]

→ 2N positive pairs have 2(N − 1) negatives

$$\mathcal{L}_{\text{SNT-Xent}} = -\frac{1}{2N}\sum_{i\in \hat{I}}\log\frac{\ell\left(\boldsymbol{z}_i, \boldsymbol{z}_{j(i)}\right)}{\sum_{a\in A(i)}\ell\left(\boldsymbol{z}_i, \boldsymbol{z}_a\right)}$$

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

$j(i)$ is the index of augmented version of $\boldsymbol{z}_i$

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

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

Introducing Additive Margins in the contrastive loss

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

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

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

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

Additive Margin (CosFace) [1]

- 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.

- Decision boundary: $\cos\left(\theta_{\boldsymbol{z}_a, \boldsymbol{z}_p}\right) - m > \cos\left(\theta_{\boldsymbol{z}_a, \boldsymbol{z}_n}\right)$

[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.

Introducing Additive Angular Margins in the contrastive loss

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

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

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

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

Additive Angular Margin (ArcFace) [1]

- Introduce $m \geq 0$ in angle space which provides the exact correspondence to the geodesic distance.
- Decision boundary: $\cos\left(\theta_{\boldsymbol{z}_a, \boldsymbol{z}_p} + m\right) > \cos\left(\theta_{\boldsymbol{z}_a, \boldsymbol{z}_n}\right)$

[1] 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

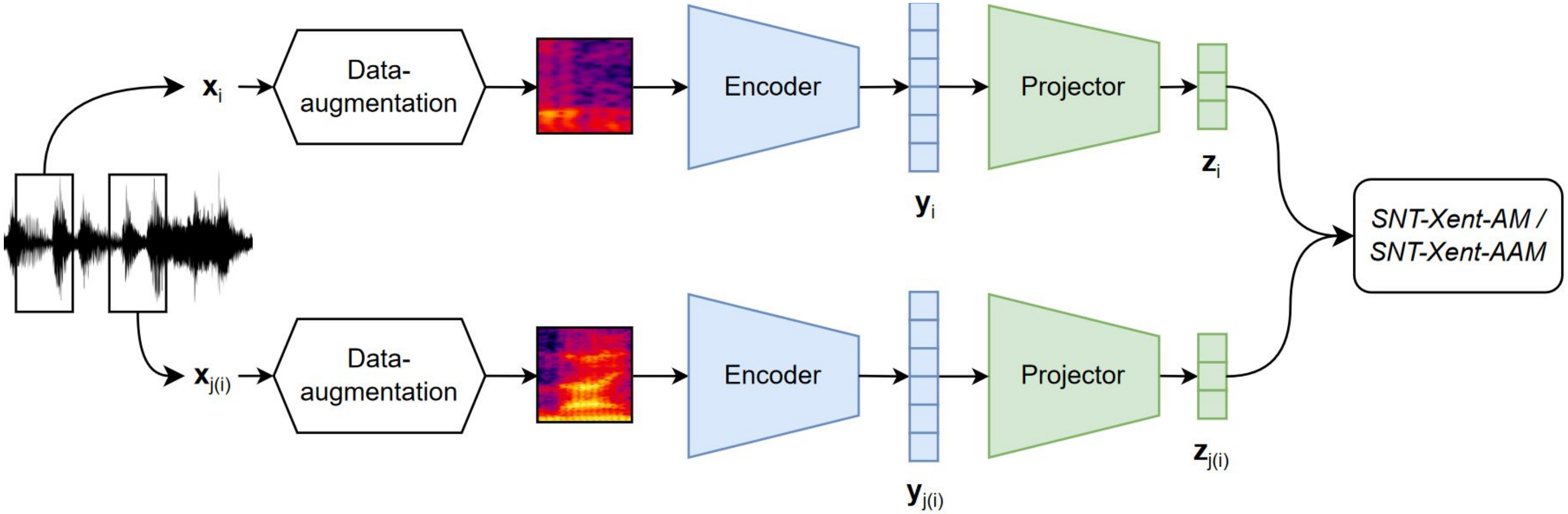


Figure 3. Overview of our self-supervised training framework.

Experimental setup

- Datasets and feature extraction
  - Training on VoxCeleb1 *dev set* [1]
  - Evaluation on VoxCeleb1 *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: Thin-ResNet34 / ResNet34
  - Projector: 2-layer MLP
  - By default τ is set to 0.2
  - Epochs: 200 / 300
  - Optimizer: Adam (no weight decay)
  - Batch size: 256
  - 2x NVIDIA Titan X (Pascal) 12 GB

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

[1] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in Interspeech, 2017.
[2] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” arXiv preprint arXiv:1510.08484, 2015.
[3] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in ICASSP, 2017.

Effect of the symmetric loss and the margins

- Using a symmetric contrastive loss improves the EER from 9.45% to 9.35%.

- Choosing the right margin factor is fundamental: trade-off between discriminative power and learning task complexity. Learning the margin value jointly with the model does not improve the performance.

- The best result is obtained with SNT-Xent-AAM which achieves 8.98% EER.

- The improvement does not translate on the minDCF as the model is trained to reduce the number of false positives and false negatives indistinctly.

Method	**EER(%)**	**minDCF**
Baseline	9.45	0.7094
Baseline w/o Data-augmentation	28.17	0.8656
Baseline w/o Projector	13.55	0.8435
Baseline w/ SNT-Xent	**9.35**	**0.6647**

Table 1. The effect of training components on SV results.

Loss	**Margin**	**EER(%)**	**minDCF**
SNT-Xent	-	9.35	0.6647
SNT-Xent-AM	0.1	9.30	0.7610
	0.2	9.01	0.6907
	0.3	8.93	0.6909
	0.4	**8.70**	**0.6873**
	0.5	8.87	0.7182
	Learnable	9.26	0.7093
SNT-Xent-AAM	0.05	8.92	0.7006
	0.1	**8.98**	**0.6742**
	0.2	9.22	0.6846
	0.3	*Exploding gradients*	
	Learnable	9.18	0.6717

Table 2. SV results when introducing margins in the self-supervised contrastive loss.

Study of the distribution of positive and negative scores

- The spread between the distribution of positive and negative scores is further when using SNT-Xent-AM (m=0.4).

- The difference between the mean of the two distributions is 0.259 without margins while it reaches 0.278 with margins.

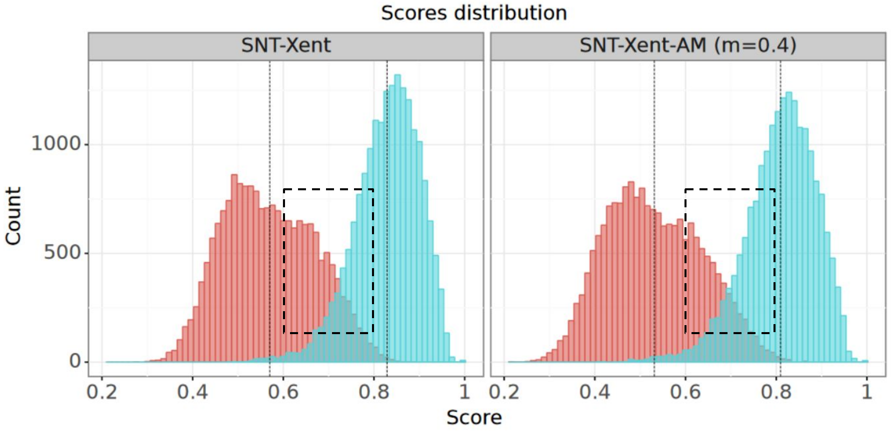


Figure 4. Positive (light blue) and negative (red) trials scores distribution obtained after training with SNT-Xent and SNT-Xent-AM (m = 0.4) losses. The mean of each distribution is represented by a vertical dashed line.

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

Final self-supervised results

Method	EER(%)	minDCF
AP+AAT [21]	8.65	–
SimCLR [10]	8.28	0.6100
MoCo [9]	8.23	0.5900
SNT-Xent	7.56	**0.5785**
SNT-Xent-AM ($m = 0.4$)	**7.50**	0.5804
SNT-Xent-AAM ($m = 0.01$)	7.56	0.6281

Table 3. Comparison of self-supervised contrastive methods for speaker verification.

- Training for more epochs and using a larger encoder, we reach 7.50% EER with SNT-Xent-AM.
- Our method outperforms other works based on contrastive learning for self-supervised speaker verification while using a smaller training set (VoxCeleb1).

➔ There is still considerable potential for improving self-supervised contrastive methods for speaker verification.

[9] 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.
[10] 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.

Additional results on VoxCeleb2

Method	Symmetric	Scale ($1/\tau$)	Margin	EER (%)	minDCF
AP-AAT [1]	✘	*learnable*	*learnable*	9.64	0.6598
NT-Xent	✘	30	0	8.98	0.6714
	✔	30	0	8.41	0.6235
SNT-Xent-AM	✔	30	0.1	<u>7.85</u>	<u>0.6168</u>
	✔	30	0.2	8.13	0.6211

➔ The effect of our improvements is more significant when training on a larger corpus (VoxCeleb2).

➔ We achieve a 18.6% relative reduction of the baseline [1] EER.

[1] 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

- ➔ Self-supervised contrastive frameworks can be further improved, notably with optimizations tailored for the downstream task.
  - Self-supervision. Providing additional positive and negative pairs results in a lower EER.
  - Speaker verification. Introducing margins in the contrastive loss function leads to a better speaker separability.

- ➔ Our improvements combined with a larger encoder model achieves 7.50% EER on VoxCeleb1 test set which is competitive with other equivalent approaches trained on VoxCeleb2.

- ➔ Early experiments on VoxCeleb2 are showing very promising results!