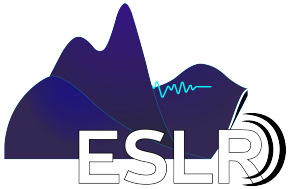

Label-Efficient Self-Supervised Speaker Verification With Information Maximization and Contrastive Learning

Théo Lepage, Réda Dehak

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

The code associated with this work is publicly available at https://github.com/theolepage/sslsv

Outline

- Introduction
  - Learning embeddings for speaker verification
  - Motivation behind self-supervised learning
  - Contrastive learning and limitations of existing approaches
- Method
  - Maximizing information with Variance-Invariance-Covariance Regularization
  - Overview of our self-supervised training framework
  - Combining contrastive learning and information maximization
- Experiments and results
  - Experimental setup
  - The role of variance, invariance and covariance coupled with data augmentation
  - Self-supervised results
  - Label-efficient evaluation
- Conclusion

Learning embeddings for speaker verification

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

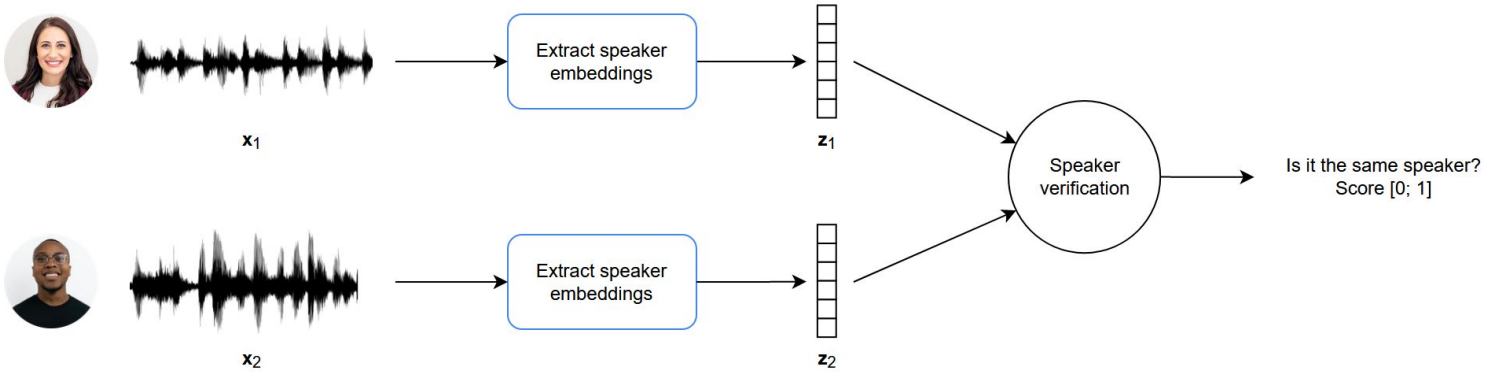

Figure 1. Overview of the use of speaker embeddings for speaker verification systems.

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.

Motivation behind self-supervised learning

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

Drawbacks of supervised learning:

1. Labeling datasets is expensive, tedious and slow.
2. Manual labeling is not scalable to the amount of data available today.
3. It could lead to biased models towards the considered problem.

➔ Supervised learning is a bottleneck for building intelligent speaker verification systems.

Contrastive learning and limitations of existing approaches

Contrastive learning [1, 2] 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.

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

Problem: They rely on a costly negative sampling process to avoid a *collapse* (non-informative embeddings). Alternative methods rely on complex architectures such as:

- a ”momentum” encoder (MoCo [3]);
- a stop gradient operation (SimSiam [4]);
- vector quantization (SwAV [5]);
- clustering (DeepCluster [6]).

[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.
[4] X. Chen and K. He, “Exploring Simple Siamese Representation Learning,” in CVPR, 2021.
[5] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments,” in NeurIPS, 2021.
[6] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep Clustering for Unsupervised Learning of Visual Features,” in ECCV, 2018.

Maximizing information with Variance-Invariance-Covariance Regularization [1]

$$\mathcal{L}_{\text{VICReg}} = \lambda\, s\left(\mathbf{Z}, \mathbf{Z}'\right) + \mu\left(v(\mathbf{Z}) + v\left(\mathbf{Z}'\right)\right) + \nu\left(c(\mathbf{Z}) + c\left(\mathbf{Z}'\right)\right)$$

- **Variance**

  Objective: Enforce the variance to reach 1 along the D dimensions.

  Intuition: Avoid the *collapse* problem.

$$v\left(\mathbf{Z}\right) = \frac{1}{D}\sum_{j=1}^{D} \max\left(0, 1 - \sqrt{\text{Var}(\mathbf{z}^j)}\right)$$

[1] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning,” arXiv preprint arXiv:2105.04906, 2021.

Maximizing information with Variance-Invariance-Covariance Regularization [1]

$$\mathcal{L}_{\text{VICReg}} = \lambda\, s\left(\mathbf{Z}, \mathbf{Z}'\right) + \mu\left(v(\mathbf{Z}) + v\left(\mathbf{Z}'\right)\right) + \nu\left(c(\mathbf{Z}) + c\left(\mathbf{Z}'\right)\right)$$

- **Invariance**

  Objective: Reduce the l_2 distance between two augmented copies.

  Intuition: Learn channel invariant representations.

$$s\left(\mathbf{Z}, \mathbf{Z}'\right) = \frac{1}{N} \sum_{i=1}^{N} \left\|\mathbf{z}_i - \mathbf{z}'_i\right\|_2^2$$

[1] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning,” arXiv preprint arXiv:2105.04906, 2021.

Maximizing information with Variance-Invariance-Covariance Regularization [1]

$$\mathcal{L}_{\mathrm{VICReg}} = \lambda\, s\left(\mathbf{Z}, \mathbf{Z}'\right) + \mu\left(v(\mathbf{Z}) + v\left(\mathbf{Z}'\right)\right) + \nu\left(c(\mathbf{Z}) + c\left(\mathbf{Z}'\right)\right)$$

- **Covariance**

  Objective: Make the off-diagonal coefficients of the covariance matrix to be close to 0.

  Intuition: Spread information across the dimensions.

$$c\left(\mathbf{Z}\right) = \frac{1}{D}\sum_{i \neq j}[C(\mathbf{Z})]_{i,j}^2$$

[1] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning,” arXiv preprint arXiv:2105.04906, 2021.

Overview of our self-supervised training framework (1)

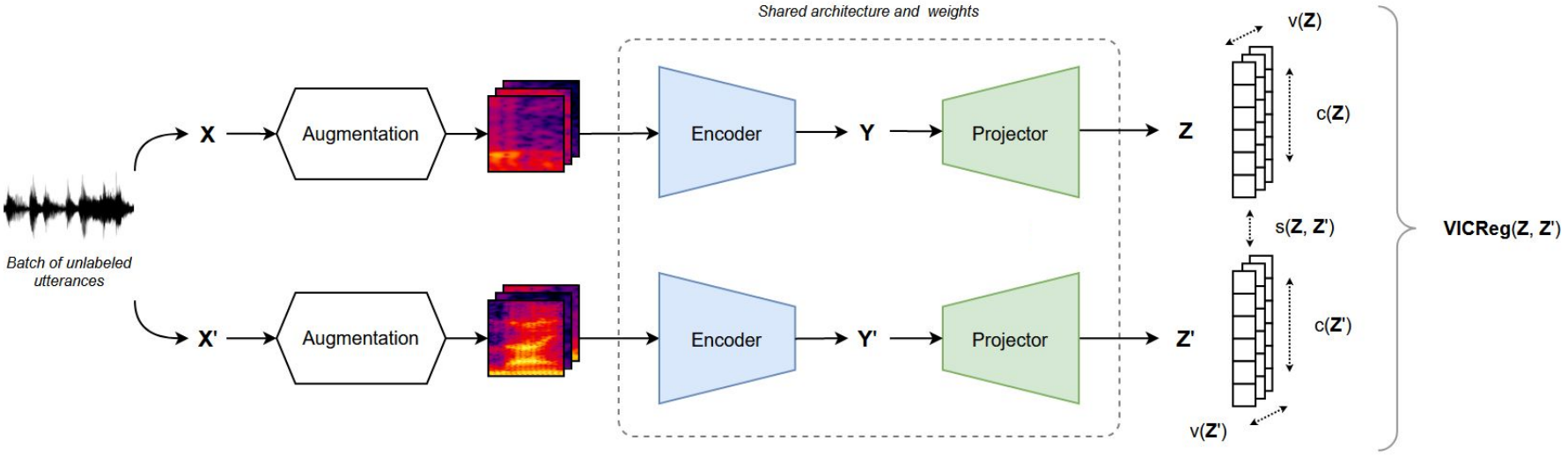

Figure 2. Overview of our self-supervised training framework.

- **X** and **X'** are input batches *(composed of different non-overlapping frames from the same utterances)*
- **Y** and **Y'** are referred to as **representations**
- **Z** and **Z'** are referred to as **embeddings**

Overview of our self-supervised training framework (2)

1. Self-supervised pre-training (100% samples - 0% labels)
2. Evaluation on speaker verification (2 utterances without labels)

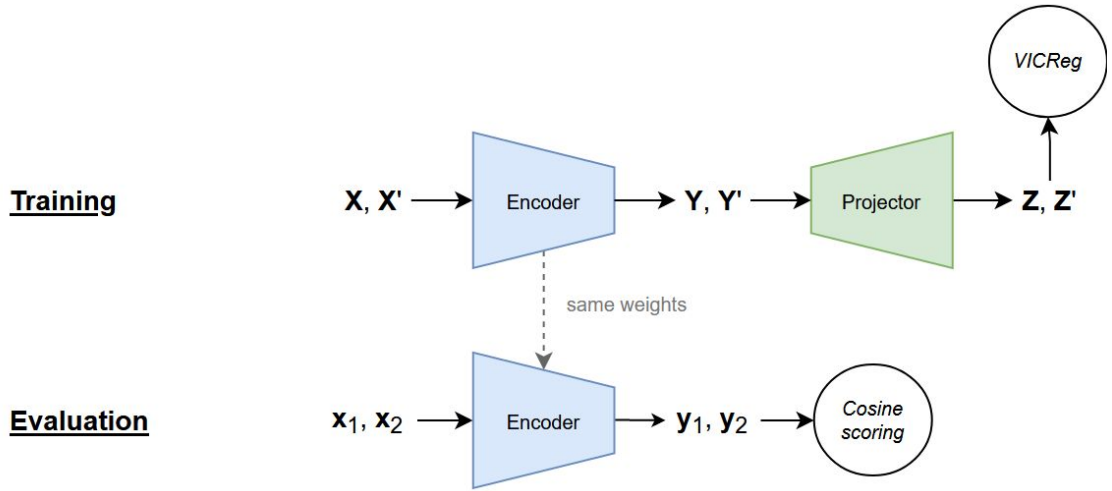

Figure 3. Overview of the forward pass of the training and evaluation steps.

Combining contrastive learning and information maximization

Optimize the model jointly with the two objective functions by:

- Using both losses at different stages of the neural network;

$$\mathcal{L}^{1}_{\text{comp}} = \mathcal{L}_{\text{VICReg}}\left(\mathbf{Y}, \mathbf{Y}'\right) + \mathcal{L}_{\text{InfoNCE}}\left(\mathbf{Z}, \mathbf{Z}'\right)$$
$$\mathcal{L}^{2}_{\text{comp}} = \mathcal{L}_{\text{InfoNCE}}\left(\mathbf{Y}, \mathbf{Y}'\right) + \mathcal{L}_{\text{VICReg}}\left(\mathbf{Z}, \mathbf{Z}'\right)$$

- Using information maximization as a regularization term.

$$\mathcal{L}^{\text{Y}}_{\text{reg}} = \mathcal{L}_{\text{InfoNCE}}\left(\mathbf{Y}, \mathbf{Y}'\right) + \alpha\, \mathcal{L}_{\text{VICReg}}\left(\mathbf{Y}, \mathbf{Y}'\right)$$
$$\mathcal{L}^{\text{Z}}_{\text{reg}} = \mathcal{L}_{\text{InfoNCE}}\left(\mathbf{Z}, \mathbf{Z}'\right) + \alpha\, \mathcal{L}_{\text{VICReg}}\left(\mathbf{Z}, \mathbf{Z}'\right)$$

Experimental setup

- Datasets and feature extraction
  - VoxCeleb1 *dev* and *test* sets
  - Speaker labels are discarded
  - 2 seconds audio chunks
  - 40-dimensional log-mel spectrogram input features
- Model architecture and training
  - Encoder: Thin-ResNet34
  - Projector: 3-layer MLP
  - 500 epochs with Adam optimizer
  - Batch size of 256
  - 2x NVIDIA Titan X (Pascal) 12 GB
- Data augmentation
  - Music, speech and babble background noises from the MUSAN corpus
  - Reverberation from the Simulated Room Impulse Response Database
- Evaluation protocol
  - Scoring with cosine similarity
  - Equal Error Rate (EER)
  - minimum Detection Cost Function (minDCF) with p=0.01

The role of *variance*, *invariance* and *covariance* coupled with data augmentation

- Convergence is determined by the weight of the *variance* (μ) component relatively to the other terms.

- *Invariance* (λ) is fundamental to learn representations that have small intra-speaker distances.

  Data augmentation is the key to achieve this objective as applying multiple transformations leads to a 62.7% relative improvement of the EER.

- *Covariance* (ν) is also necessary to produce meaningful representations (53.6% improvement of the EER between the 1st and 3rd configuration).

λ	μ	ν	EER	minDCF
1	1	0	24.00	0.9964
1	0.5	0.1	15.71	0.8554
1	1	0.04	**11.14**	**0.6843**
1	1	0.1	11.87	0.7101

Table 1. SV results with different scaling factors for VICReg loss components. λ: Invariance, μ: Variance, ν: Covariance.

Method	EER	minDCF
No augmentation	29.87	0.8833
Musan	21.22	0.8388
RIR	22.28	0.8525
Musan + RIR	**11.14**	**0.6843**

Table 2. SV results with different data augmentation strategies.

Self-supervised results

Method	Loss	EER	minDCF
NPC [21]	Cross-entropy	15.54	0.8700
SimCLR [6]	InfoNCE	9.87	0.6760
Ours	$\mathcal{L}_{\text{InfoNCE}}$	10.42	**0.6276**
	$\mathcal{L}_{\text{BarlowTwins}}$	13.46	0.8473
	$\mathcal{L}_{\text{VICReg}}$	**9.25**	0.6432
Ours (Section 2.3)	$\mathcal{L}^1_{\text{comp}}$	13.14	0.6950
	$\mathcal{L}^2_{\text{comp}}$	**8.47**	**0.6400**
	$\mathcal{L}^{\text{Y}}_{\text{reg}}$	9.09	0.6894
	$\mathcal{L}^{\text{Z}}_{\text{reg}}$	10.38	0.6913

Table 3. Final results on SV (VoxCeleb1 test).

- Our method, VICReg, achieves 9.25% EER and outperforms the contrastive baseline, InfoNCE, and the other unsupervised approaches.

- We reach the best performance with $\mathcal{L}^2_{\text{comp}}$ objective function as it achieves **8.47% EER.**

$$\mathcal{L}^2_{\text{comp}} = \mathcal{L}_{\text{InfoNCE}}\left(\mathbf{Y}, \mathbf{Y}'\right) + \mathcal{L}_{\text{VICReg}}\left(\mathbf{Z}, \mathbf{Z}'\right)$$

➔ Combining information maximization and contrastive learning at different stages of the model is very effective.

Label-efficient evaluation

Use a limited amount of annotated utterances to improve the performance of our method:

1. Train a linear classifier on top of the frozen self-supervised representations;
2. Fine-tune the whole pre-trained neural network.

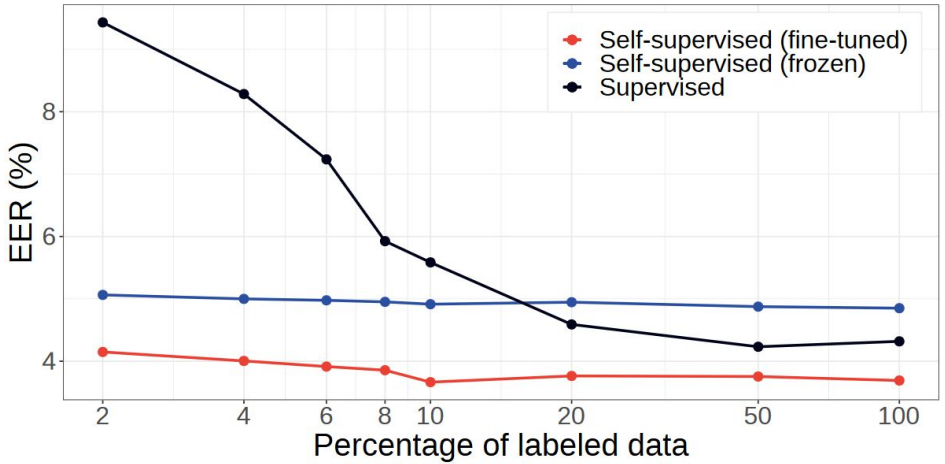

Figure 4. SV results with different percentage of labeled data used during training.

➔ Fine-tuning with only 2% of labeled data is sufficient to outperform the supervised baseline.

Conclusions

- ➔ Information maximization methods allow learning robust speaker representations without contrastive samples.

- ➔ Our self-supervised training framework can reach better results when combining contrastive learning and information maximization.

- ➔ Our method outperforms its supervised counterpart when fine-tuned with only 2% of labeled utterances, which is a step toward label-efficient speaker verification systems.