ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096915

PUSHING THE LIMITS OF SELF-SUPERVISED SPEAKER VERIFICATION USING REGULARIZED DISTILLATION FRAMEWORK

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen

Speech Lab of DAMO Academy, Alibaba Group

{chenyafeng.cyf, zsq174630, tongmu.wh}@alibaba-inc.com

ABSTRACT

Training robust speaker verification systems without speaker labels has long been a challenging task. Previous studies observed a large performance gap between self-supervised and fully supervised methods. In this paper, we apply a non-contrastive self-supervised learning framework called DIstillation with NO labels (DINO) and propose two regularization terms applied to embeddings in DINO. One regularization term guarantees the diversity of the embeddings, while the other regularization term decorrelates the variables of each embedding. The effectiveness of various data augmentation techniques are explored, on both time and frequency domain. A range of experiments conducted on the VoxCeleb datasets demonstrate the superiority of the regularized DINO framework in speaker verification. Our method achieves the state-of-the-art speaker verification performance under a single-stage self-supervised setting on VoxCeleb.

Index Terms— Speaker verification, self-supervised learning, DINO, regularization term

1. INTRODUCTION

Deep learning methods have achieved significant performance gains on speaker verification (SV) task. The availability of large labeled datasets and data augmentation methods have spurred remarkable improvements. However, collecting large amounts of labeled data in the real world is laborious and expensive. Therefore, self-supervised learning (SSL), that requires only unlabeled data, provides an alternative solution for learning representations from speech.

Different unsupervised and self-supervised frameworks have been proposed to obtain meaningful speaker representations as in [1–10]. More recently, some self-supervised SV systems propose two-stage SSL training [2, 4, 8]. In stage I, a speaker encoder is trained to obtain speech representations in a completely self-supervised manner. In stage II, a clustering algorithm is adopted to generate pseudo-labels for each utterance based on the previous learned representations. Then a new model is trained based on the estimated pseudo-labels iteratively in supervised learning. However, the second stage requires a "good enough" estimate of the number of speakers in the training data, such as M = 6000 in [2, 4] when using the development portions of VoxCeleb2 [11]. Inaccurate settings will lead to significant performance degradation. This prior assumption restrains us from training SSL models from numerous unlabeled speech data involving a large but unknown number of speakers, which is self-contradictory with the original purpose of using SSL. Therefore, we stick to the single-stage approaches in self-supervised SV, without making assumptions on the overall distribution of training data.

Single-stage methods can be roughly divided into contrastive and non-contrastive ones. Xia et al. [1] applies SimCLR framework [12] to extract robust speaker embeddings by directly maximizing the similarity between augmented positive pairs and minimizing the similarity of negative pairs via a contrastive InfoNCE loss [13, 14]. Such frameworks require large number of negative samples and huge batch size during training. He et al. tried to remove the batch size limitations by introducing Momentum Contrast (MoCo) framework [15]. MoCo uses a dynamic queue to store negative samples and compute the InfoNCE loss.

Most contrastive methods either require a huge batch size or maintain a dynamic queue, both of which are computationally expensive. Furthermore, contrastive methods in speaker verification select negative samples through random sampling, which may result in false negative samples. It will cause instability in network training and degrade system performance since the contrastive loss may push the potential positive samples further to each other.

Non-contrastive methods are free of such worries because the negative samples are not required in the training process. In fact, these methods [3, 5–8] have shown comparable or better performance compared to contrastive counterparts. Among non-contrastive frameworks, BYOL [16] and DINO [17] are the more attractive ones proposed in computer vision. BYOL is composed of online and target networks. The online network predicts the target network representation of the same utterance under a different augmented view. In [7], a self-supervised regularization term is proposed in speaker verification inspired by BYOL. DINO is a self-distillation framework which contains a teacher and a student network with an identical architecture but different

parameters. It compares multiple different views generated from a single utterance and employs a self-distillation method which is widely used in speaker verification [3, 5, 6, 8].

Due to the lack of negative samples, model collapse has been a common problem for non-contrastive methods. The network is more inclined to map positive pairs to the same or similar positions in the embedding space, resulting in trivial solutions. To avoid model collapse, [7] introduces the regularization MLP structure which is applied on online network. DINO applies sharpening and centering techniques to the teacher output and uses exponential moving average (EMA) update strategy [6].

Inspired by [18], we propose two regularization terms applied to embeddings to further alleviate this problem in speaker verification task. One regularization term guarantees the diversity of the embeddings within a batch, while the other regularization term decorrelates the variables of each embedding. In order to effectively capture the utterance-dependent variability into the embedding, three kinds of augmentation strategies - WavAugment [19], SpecAugment [20] and Acoustic feature shuffling [21] are investigated. Experimental results indicate that the proposed Regularized DINO framework (RDINO) can significantly boost the performance of self-supervised speaker verification.

2. DINO APPLIED IN SV

Inspired by [17], we apply DINO to speaker verification. DINO is a self-distillation framework where the outputs of a teacher network are used as ground truth to optimize a student network in parallel. Various types of cropped and augmented views are constructed from an utterance, divided into local and global views depending on the length of segments. Global views go through teacher network and all views are fed into student network. The global information learned by teacher guides the training of student, therefore encouraging "local-to-global" correspondences. Both networks share the same architecture g with different sets of parameters.

Take teacher module as example, it is composed of a backbone f (ECAPA-TDNN), and of a projection head $h : g = f \circ h$. The speaker embedding is the backbone f output. ECAPA-TDNN is the most commonly used network structure in speaker verification, with powerful feature extraction capability. The projection head h consists of three fully connected (FC) layers with hidden dimension 2048-2048-256 followed by $L2$ normalization and a weight normalized FC layer with K dimensions. Cross-entropy loss is calculated to minimize the probability distribution as follows.

$$L_{CE} = \sum_{\mathbf{x} \in \mathbf{X}_l} \sum_{\substack{\mathbf{x}' \in \mathbf{X}_l \cup \mathbf{X}_s \\ \mathbf{x}' \neq \mathbf{x}}} H(P^{tea}(\mathbf{x}) | P^{stu}(\mathbf{x}')) \tag{1}$$

where $H(a|b) = -a * logb$ is cross-entropy, $\mathbf{X}_l = \{\mathbf{x}_{l_1}, \mathbf{x}_{l_2}\}$ stands for two long segments and $\mathbf{X}_s = \{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, \mathbf{x}_{s_3}, \mathbf{x}_{s_4}\}$ stands for four short segments. P^{tea}, P^{stu} represent the output probability distributions of teacher network g_θ^{tea} and student network g_θ^{stu} respectively. Mathematically,

$$P^{tea}(\mathbf{x}) = Softmax\left(g_\theta^{tea}(\mathbf{x})/\tau_t\right) \tag{2}$$

τ_t is a temperature parameter that controls the sharpness of the teacher's output distribution, and a similar formula holds for P^{stu} with temperature τ_s.

To avoid model collapse, teacher network is updated by EMA of the student's parameters. In addition, sharpening and centering techniques are applied to the teacher output. Beyond that, we add diversity regularization and redundancy elimination regularization to conquer the model collapse problem, which will be introduced in the following section.

3. PROPOSED METHOD

In order to further alleviate the model collapse problem and improve the robustness of the speaker embeddings in DINO framework, we propose two regularization terms called diversity regularization and redundancy elimination regularization.

3.1. Diversity regularization

The diversity regularization term guarantees the diversity of the embeddings within a batch. It forces the embeddings of utterances to be different and prevents trivial solutions. The diversity regularization loss can be calculated as shown in Eq. 3. $\mathbf{z}_j^{tea}$ and $\mathbf{z}_j^{stu}$ stand for the teacher's and student's embeddings of dimension j respectively. We calculate the standard deviation of each dimension in embeddings within one batch. ϵ is a small scalar preventing numerical instabilities.

$$\begin{aligned} L_{DR} = & \frac{1}{d}\sum_{j=1}^{d} \max\left(0, 1 - \sqrt{\mathrm{Var}(\mathbf{z}_j^{tea}) + \epsilon}\right) \quad + \\ & \frac{1}{d}\sum_{j=1}^{d} \max\left(0, 1 - \sqrt{\mathrm{Var}(\mathbf{z}_j^{stu}) + \epsilon}\right) \end{aligned} \tag{3}$$

3.2. Redundancy elimination regularization

The redundancy elimination regularization term decorrelates the variables of each embedding while minimizing the redundancy. It attracts the covariances of each dimension in all speaker embeddings within a batch towards zero and prevents an informational collapse in which the variables would be highly correlated. The redundancy elimination regularization loss is calculated as follows.

$$L_{RER} = \sum_i \sum_{j \neq i} C_{ij}^2 \tag{4}$$

where C is the cross-correlation matrix computed between the global outputs of teacher network and student network along

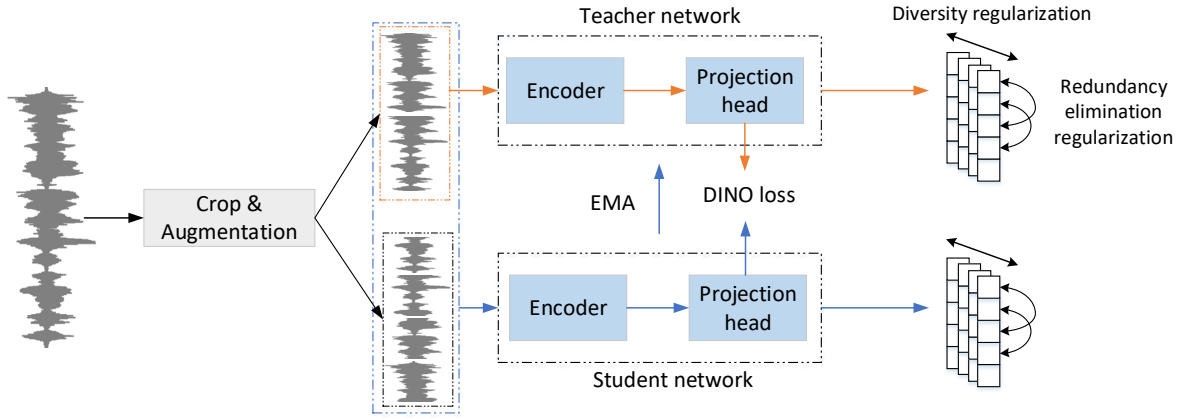

Fig. 1. Overview of the proposed regularized DINO framework.

the batch dimension, and C_{ij} is defined as Eq. 5:

$$C_{ij} = \frac{\sum_b z^{tea}_{b,i} z^{stu}_{b,j}}{\sqrt{\sum_b \left(z^{tea}_{b,i}\right)^2} \sqrt{\sum_b \left(z^{stu}_{b,j}\right)^2}} \tag{5}$$

where b indexes batch samples and i, j index the embedding dimension. The redundancy elimination term decorrelates the different vector components of the embeddings by trying to equate the off-diagonal elements of the cross-correlation matrix to 0.

3.3. Regularized DINO framework

The overview of regularized DINO framework is depicted in Fig. 1. First of all, we sample two long segments $\mathbf{X}_l = \{\mathbf{x}_{l_1}, \mathbf{x}_{l_2}\}$ and four short segments $\mathbf{X}_s = \{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, \mathbf{x}_{s_3}, \mathbf{x}_{s_4}\}$ from an utterance with a multi-crop strategy and different data augmentation strategies. Then $\mathbf{X}_l$ are first encoded by f^{tea}_{ϑ} and f^{stu}_{ϑ} respectively into their representations regarded as speaker embeddings $\mathbf{E}^{tea}_l$ and $\mathbf{E}^{stu}_l$. Then $\mathbf{E}^{tea}_l$ and $\mathbf{E}^{stu}_l$ are mapped by four FC layers with hidden dimension $2048, 2048, 8192, 256$ and a weight normalized FC layer with 65536 dimensions. We denote $\mathbf{Z}^{tea}_l = [\mathbf{z}^{tea}_{l_1}, \mathbf{z}^{tea}_{l_2}]$ and $\mathbf{Z}^{stu}_l = [\mathbf{z}^{stu}_{l_1}, \mathbf{z}^{stu}_{l_2}]$ as the teacher's output and student's output of dimension 8192. The diversity regularization and redundancy elimination regularization loss are computed in $\mathbf{Z} = \mathbf{Z}^{tea}_l \cup \mathbf{Z}^{stu}_l$ shown as Eq. 3 and Eq. 4. The speaker embedding network is jointly trained with the L_{CE}, L_{DR} and L_{RER}. The overall loss is calculated as Eq. 6, the hyperparameter λ controls the balance of all losses.

$$Loss = L_{CE} + \lambda(L_{DR} + L_{RER}) \tag{6}$$

4. EXPERIMENTS AND ANALYSIS

4.1. Experimental settings

4.1.1. Datasets and evaluation metrics

To investigate the effectiveness of the proposed method, we conduct experiments on the VoxCeleb datasets. The development portions of VoxCeleb2 [11] are used for training. It comprises 1,092,009 utterances among 5,994 speakers. Performance of all systems are evaluated on the test set of VoxCeleb1 [22]. No speaker labels are used during training in all experiments.

The results are reported in terms of two metrics, namely, the equal error rate (EER) and the minimum of the normalized detection cost function (MinDCF) with the settings of P_{target} = 0.05 and $C_{fa} = C_{miss}$ = 1.

4.1.2. Input features

For each utterance, we use the multi-crop strategy for RDINO training in which 4s segments and 2s segments regarded as global views and local views respectively. The acoustic features used in the experiments are 80-dimensional Filter Bank (FBank) with 25ms windows and 10ms shift. Speech activity detection (SAD) is not performed as training data consists mostly of continuous speech. Mean and variance normalization is applied using instance normalization on FBank features.

4.2. Data augmentation

Data augmentation has been proven to be crucial for both supervised and self-supervised representation learning. Therefore, we explore the impact of data augmentation in our regularized DINO framework. WavAugment, SpecAugment and Acoustic feature shuffling are considered.

WavAugment: MUSAN corpus with SNR between 0 to 15 for additive noise and Room Impulse Response (RIR) for reverberation are applied to each long segment and short segment randomly.

SpecAugment: Apply time masking and frequency masking at acoustic feature level. The time masking length is 0 to 15 frames and frequency masking length is 0 to 6 dimensions. Each time we apply one randomly selected time mask and frequency mask on the FBank features.

Acoustic feature shuffling: To learn the sequential order invariant speaker embeddings, we shuffle the time order of acoustic feature frames at the segment scale of 50 frames. Detailed steps are discribed in [21].

4.3. Model configurations and implementation details

We exploit the ECAPA-TDNN with attentive statistical pooling as the encoder model f, followed by a 512-d FC layer. The projection head h consists of four FC layers with hidden size of $2048, 2048, 8192, 256$ and a weight normalized FC layer with 65536 dimensions. We train the model 60 epochs using the stochastic gradient descent (SGD) optimizer with momentum of 0.9. The learning rate is linearly ramped up during the first 10 epochs to 0.325. After this warmup, we decay the learning rate with a cosine schedule. The temperature τ_t and τ_s are set to 0.04 and 0.1 respectively.

4.4. Results and analysis

We investigate the performance of regularized DINO framework and evaluate it on the VoxCeleb1 test set. We compare our method to [1,3,4,6–8,23,24] which are recently proposed self-supervised learning architectures as shown in Table 1.

Table 1. Comparison with self-supervised learning models

Framework	EER(%)	MinDCF
AP+AAT [23]	8.65	0.454
MoCo + WavAug [1]	8.23	0.590
CEL [24]	8.01	N / R
Contrastive [4]	7.36	N / R
SSReg [7]	6.99	0.434
DINO + Cosine loss [8]	6.16	0.524
DINO [3]	4.83	N / R
DINO + CL [6]	4.47	0.306
RDINO (Ours)	**3.29**	**0.247**

It can be observed that the RDINO system decreases the EER to 3.29%, which outperforms the previous self-supervised methods (4.47% EER) by a significant +26.4% relative improvement. This finding verifies the effectiveness of the two regularization terms.

Table 2. The effect of the weight λ in RDINO

Weight	EER(%)	MinDCF
$\lambda = 0$	3.62	0.262
$\lambda = 0.1$	3.40	0.259
λ **= 0.2**	**3.24**	**0.252**
λ **= 0.3**	**3.29**	**0.247**
$\lambda = 0.4$	3.37	0.253
$\lambda = 0.5$	3.52	0.251

Moreover, we also conduct experiments to investigate the effect of weight λ in RDINO system as shown in Table 2. We observe that applying two regularization terms outperforms the DINO system with even a small weight λ. It achieves 3.24% EER with $\lambda = 0.2$ and 0.247 MinDCF with $\lambda = 0.3$. The results show superiority of the proposed method in self-supervised speaker verification task.

Additionally, we study the impact of different data augmentation strategies including WavAugment, SpecAugment and Acoustic feature shuffling on the training data. The experimental results are as follows.

Table 3. The impact of data augmentation in RDINO

Augmentation	EER(%)	MinDCF
No Augment	20.9	0.777
WavAugment	3.29	0.247
+ SpecAugment	5.35	0.369
+ Acoustic feature shuffling	3.45	0.250

It is observed that WavAugment is the most efficient strategy in the RDINO framework. If no augmentation applied in training process, the whole network was difficult to converge due to DINO's inherent property. We find that SpecAugment worsened the verification results, which is different from our empirical observation for supervised speaker verification. SpecAugment uses erasing operation on the acoustic feature level to improve the model generalization. But the Voxceleb1 test data may not contain enough variabilities in the spectral domain. Acoustic feature shuffling strategy generates no gain in RDINO system, which may be caused by the operation in short length of the local view.

5. CONCLUSIONS

In this paper, we introduce the DINO framework with different augmentation strategies for self-supervised speaker verification. In order to address the model collapse problem and further improve the system performance, we propose two regularization terms in DINO which achieve excellent performance over the conventional self-supervised models.

6. REFERENCES

[1] Wei Xia, Chunlei Zhang, Chao Weng, Meng Yu, and Dong Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *ICASSP*. 2021, pp. 6723–6727, IEEE.

[2] Danwei Cai, Weiqing Wang, and Ming Li, "An iterative framework for self-supervised deep speaker representation learning," in *ICASSP*. 2021, pp. 6728–6732, IEEE.

[3] Jejin Cho, Jesus Villalba, and Najim Dehak, "The jhu submission to voxsrc-21: Track 3," *arXiv preprint arXiv:2109.13425*, 2021.

[4] Ruijie Tao, Kong Aik Lee, Rohan Kumar Das, Ville Hautamäki, and Haizhou Li, "Self-supervised speaker recognition with loss-gated learning," in *ICASSP*. 2022, pp. 6142–6146, IEEE.

[5] Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung, "Pushing the limits of raw waveform speaker recognition," in *Interspeech*. 2022, pp. 2228–2232, ISCA.

[6] Hee-Soo Heo, Jee-weon Jung, Jingu Kang, Youngki Kwon, You Jin Kim, and Bong-Jin Lee abd Joon Son Chung, "Self-supervised curriculum learning for speaker verification," *arXiv preprint arXiv:2203.14525*, 2022.

[7] Mufan Sang, Haoqi Li, Fang Liu, Andrew O. Arnold, and Li Wan, "Self-supervised speaker verification with simple siamese network and self-supervised regularization," in *ICASSP*. 2022, pp. 6127–6131, IEEE.

[8] Bing Han, Zhengyang Chen, and Yanmin Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," in *Interspeech*. 2022, pp. 4780–4784, ISCA.

[9] Siqi Zheng, Gang Liu, Hongbin Suo, and Yun Lei, "Autoencoder-based semi-supervised curriculum learning for out-of-domain speaker verification," in *Interspeech*. 2019, pp. 4360–4364, ISCA.

[10] Siqi Zheng, Gang Liu, Hongbin Suo, and Yun Lei, "Towards a fault-tolerant speaker verification system: A regularization approach to reduce the condition number," in *Interspeech*. 2019, pp. 4065–4069, ISCA.

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*. 2018, pp. 1086–1090, ISCA.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*. 2020, vol. 119, pp. 1597–1607, PMLR.

[13] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[14] Siqi Zheng, Hongbin Suo, and Qian Chen, "PRISM: pre-trained indeterminate speaker representation model for speaker diarization and speaker verification," in *Interspeech*. 2022, pp. 1431–1435, ISCA.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*. 2020, pp. 9726–9735, Computer Vision Foundation / IEEE.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, and et al., "Bootstrap your own latent - A new approach to self-supervised learning," in *NIPS*, 2020.

[17] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," in *ICCV*. 2021, pp. 9630–9640, IEEE.

[18] Adrien Bardes, Jean Ponce, and Yann LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *ICLR*. 2022, OpenReview.net.

[19] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*. 2018, pp. 5329–5333, IEEE.

[20] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*. 2019, pp. 2613–2617, ISCA.

[21] Jin Li, Xin Fang, Fan Chu, Tian Gao, Yan Song, and Rong Li Dai, "Acoustic feature shuffling network for text-independent speaker verification," in *Interspeech*. 2022, pp. 4790–4794, ISCA.

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*. 2017, pp. 2616–2620, ISCA.

[23] Jaesung Huh, Hee Soo Heo, Jingu Kang, Shinji Watanabe, and Joon Son Chung, "Augmentation adversarial training for unsupervised speaker recognition," in *Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS*, 2020.

[24] Sung Hwan Mun, Woo Hyun Kang, Min Hyun Han, and Nam Soo Kim, "Unsupervised representation learning for speaker recognition via contrastive equilibrium learning," *CoRR*, vol. abs/2010.11433, 2020.