Self-supervised curriculum learning for speaker verification

Hee-Soo Heo¹, Jee-Weon Jung¹, Jingu Kang¹, Youngki Kwon¹, You Jin Kim¹, Bong-Jin Lee¹, Joon Son Chung²

¹Naver Corporation, South Korea, ²KAIST, South Korea
heesoo.heo@navercorp.com

Abstract

Self-supervised learning is one of the emerging approaches to machine learning today, and has been successfully applied to vision, speech and natural processing tasks. There are a range of frameworks within self-supervised learning literature, but the speaker recognition literature has particularly adopted self-supervision via contrastive loss functions. Our work adapts the DINO framework for speaker recognition, which trains the model without exploiting negative utterance pairs. We introduce a curriculum learning strategy to the self-supervised framework, which guides effective training of speaker recognition models. Specifically, we propose two curriculum strategies where one gradually increases the number of speakers within a mini-batch, and the other gradually applies more aggressive augmentations as the training proceeds. A range of experiments conducted on the VoxCeleb1 evaluation protocol demonstrates the effectiveness of both the DINO framework on speaker verification and our proposed curriculum learning strategies where we report the state-of-the-art equal error rate of 4.47% with a single-phase training.

Index Terms: speaker recognition, speaker verification, self-supervised learning, curriculum learning

1. Introduction

Self-supervised training allows a model to map input data to a representative latent space without requiring human-annotated ground truth labels. Depending on the downstream task, models trained using self-supervision can serve as a pre-trained model or be used directly without further fine-tuning process [1–4]. In both cases, its effectiveness is receiving attention, and various frameworks are being studied [1]. The majority of self-supervised learning frameworks belong to one of the three categories: (i) contrastive-based, (ii) reconstruction-based, or (iii) non-negative-based [5–9].

The speaker verification literature has also embraced self-supervised learning and several pioneering works have been proposed [3, 4, 10, 11]. Several studies have employed a two-phase self-supervised learning strategy, namely iterative clustering. The first phase includes training the models via a contrastive learning framework. The second phase repeats two steps until the performance converges to the intended level. First, pseudo labels are generated using the trained model. Second, the model is trained once again in a supervised classification manner leveraging generated pseudo labels. Typically, the second phase is repeated, hence this strand of work is often referred to as iterative clustering. However, as far as we are concerned, only the contrastive-based frameworks have been adopted for speaker verification, although various frameworks are being rapidly developed in other domains.

To this end, this paper presents two main contributions to the self-supervised learning literature of speaker verification: (i) we adapt the DINO framework [7] which does not exploit negative pairs, and (ii) we propose curriculum learning methods to enable more effective self-supervised training. Our focus lies in exploring different frameworks and improving the self-supervised learning strategies, where we train a randomly initialised model. This line of research benefits the speaker verification literature in several aspects. First, it adopts single-phase training, saving a significant amount of time and not relying on the estimation of the number of clusters. Second, it coincides with developing other domains (e.g., image, natural language processing) where more advanced single-phase self-supervised frameworks are being proposed. Third, if required, our model can serve as the initial model used for generating pseudo labels in iterative clustering as our work corresponds to the first phase of it. Hence our improvements can complement the iterative clustering schemes.

We further focus on the potential of curriculum learning [12] and propose techniques to introduce it to self-supervised learning. We propose two sets of curriculum learning strategies: (i) gradually increasing the amount of training data and (ii) gradually making the training more difficult by adding noise and reverberation augmentation to increasing proportion of the data. The underlying assumption is that the model can benefit by first being trained with easier tasks and then gradually solving harder tasks, which is why both of the proposed strategies gradually increase difficulties the loss as training proceeds. As far as we are aware, this work is the first work that incorporates curriculum learning in a self-supervised training of deep neural networks.

The rest of this paper is organised as follows: Section 2 introduces the related works. The proposed approach, including model architecture, optimisation and modifications, is addressed in Section 3. The experiments and result analysis are presented in Section 4 and 5, respectively. The paper is concluded in Section 6.

2. Related Works

2.1. DINO framework

DINO [7] is a self-distillation framework based on the mean teacher [13] method. This framework employs a “local-to-global” distillation to guide the training of the student network. Precisely, various types of cropped and augmented views are constructed from an input, divided into local and global views depending on the resolution or the amount of information contained. Those with higher resolution or more information are the global views. This framework aims to minimise the difference between the output features when different views are digested by either the teacher or the student network. In particular, the student network digests all views and outputs local features, whereas the teacher network only digests global views, extracting global features. Based on these two types of features,
a loss $L_D$ that penalises the difference between them is defined and used to train the student network.

$$L_D = \sum_{x \in \{a_1^P, a_2^P\}} \sum_{\hat{x} \notin x} H(P_t(x), P_s(\hat{x})), \quad (1)$$

where $H(a, b) = -a \log b$, $a \neq b$ indicates the global view, $V$ is the set of views from an input and $P_t(\cdot)$ and $P_s(\cdot)$ are the output distribution of teacher and student network, respectively.

On the other hand, the teacher network is updated, adopting a momentum encoder of the student network without gradient-based training. In addition, sharpening and centering techniques are applied to the teacher output to avoid model collapse due to the DINO framework’s property, which involves only positive pairs.

2.2. Curriculum learning

Curriculum learning allows efficient training of deep neural networks by introducing training examples in increasing difficulty. Several configurations have been proved effective in the supervised learning field where performance improvements were observed with no additional overhead in terms of computations, and resources [12, 14].

In [15], authors first train the model under the text-dependent scenario, then extend to text-independent. Here, it has been shown that adjusting the training content through curriculum learning could teach the model to handle different textual content as well as make it robust in various acoustic environments. Other studies have also introduced methods of controlling training conditions. Specifically, when adopting the additive angular margin loss function, increasing the margin as training proceeds has become an essential setting [16, 17]. As such, it has been demonstrated that the curriculum learning applied to the training condition stabilised the training and improved the quality of the model. Based on these studies, we conclude that the curriculum learning is not limited to the data configuration but is also valid for hyper-parameter settings. However, to the best of our knowledge, there is no attempt to apply curriculum learning to train self-supervised speaker models.

3. Proposed Approach

This section addresses the model architecture (Section 3.1), the adaptation of DINO framework for speaker recognition (Section 3.2), and the proposed curriculum learning techniques (Section 3.3).

3.1. Model architecture

The model is based on the ECAPA-TDNN architecture that operates on a Mel-frequency cepstral coefficient input [18]. It comprises of a 1-dimensional convolution block followed by three Res2Net-based residual blocks with gradually increasing dilation values where a squeeze-excitation module is applied after each block. Up to the last residual block output, the sequence length remains because both pooling layers and stride size bigger than 1 are not included. Three residual block outputs are concatenated and fed to a convolution block. Then, a context and channel dependent statistical pooling layer aggregates frame representations into a single utterance representation. Finally, an affine transform derives the speaker embedding.

3.2. Adapted DINO framework

The DINO self-supervised framework addressed in Section 2.1 was originally proposed for the computer vision domain [7].

This section addresses our modifications to adapt it for speaker verification.

First, we change the global and local view to long and short crops of the same utterance with different augmentations. In particular, we construct two global views and five local views, resulting in seven views from each input. For augmentation, we use reverbération and noise from RIRs and Musan datasets [19, 20]. Then, we control the sharpness of the student and teacher output distributions by using temperatures of 0.1 and 0.04, respectively where the sharpness is controlled by dividing the output with temperature values before applying the softmax function of each network for minimising entropy. The aforementioned adaptation enables us to train the model using the DINO framework, where we followed other experimental settings proposed in the original paper. Figure 1 shows the overall process of the DINO framework adapted for speaker verification.

3.3. Proposed curriculum learning strategies

In typical curriculum learning strategies, the learning is designed to become gradually more difficult [12]. Various curriculum learning approaches in supervised learning leverage the ground truth label. However, because no labels are available in self-supervised learning, these strategies cannot be used. We hypothesise that we can leverage two factors to control the degree of difficulty based on our analysis: (i) the number of speakers and (ii) the extent of augmentation.

Number of speakers. A training dataset with more speakers and utterances with severe augmentation both tends to increase the train loss value, which we assume that makes training process more difficult. We assume that the number of speakers in a subset of the dataset has a positive correlation with the amount of data used. Although this assumption cannot always be guaranteed, a dataset that contains one million randomly collected utterances is likely have a greater number of different speakers compared to a dataset that contains one thousand utterances. At the start of training, only a small portion of the dataset is used, and the ratio of the data used for training gradually increases. The three following strategies are designed empirically,
and varies the degree of data limitation. Figure 2-(a) illustrates the detailed ratios of data used for each epoch. Note that these curriculum strategies are designed considering that one cycle is set to 16 epochs in stochastic gradient descent with warm restarts (SGDR) [21], a learning rate scheduler used for training.

Augmentation proportion. In the case of augmentation, we gradually increase the proportion of augmented utterances within each mini-batch. Figure 2-(b) depicts the two augmentation strategies we adopt. Note that the baseline augments all utterances within a mini-batch.

4. Experiments

This section introduces experimental settings to evaluate the proposed curriculum learning method. First, we demonstrate the effect of self-supervised learning. Then, using the pre-trained model via self-supervision, we conduct fine-tuning using a small amount of labelled data.

4.1. Dataset

Our experiments utilise the VoxCeleb 1 and 2 datasets for training and evaluating the models [22–24]. We use the development partition of the VoxCeleb2 dataset which includes over a million utterances from 5,994 speakers to train the model with self-supervision where we assume that the labels do not exist. The widely adopted equal error rate (EER) is used as the primary metric on the VoxCeleb1-O benchmark protocol.

For the experiment of fine-tuning phase, the development partition of the VoxCeleb1 dataset which includes 148,642 utterances from 1,211 speakers and CN-Celeb [25, 26] are used. We fine-tune the pre-trained model and evaluate using the corresponding test set using each data. CN-Celeb is a dataset consisting of the voices of Chinese celebrities. Among the total set, we utilise CN-Celeb1, of which 800 speakers are designated for training and 200 for evaluation. This dataset consisting only of speech from Chinese may be suitable for evaluating the adaptation performance of the pre-trained model in other domains.

4.2. Configurations

Table 1 describes the hyperparameters we use to train the model with the DINO framework. For the curriculum learning, we control the difficulty of training following settings shown in the Figure 2. Since the curriculum strategies of augmentation and the data partition are designed independently, they can be applied separately or together.

In the fine-tuning phase, the initial model is either randomly initialised, pre-trained using the DINO framework, or pre-trained via the DINO framework with proposed curriculum learning strategies. We use additive angular margin loss to optimise the model [27–29]. Most of the settings are used as the same settings as with self-supervised learning, but the total number of epochs is reduced to 50 and the learning rate scheduler is changed to SGDR without restart.

5. Results and analysis

5.1. Self-supervised learning

Table 2 compares our model trained via adapted DINO framework with existing works that adopt contrastive-based self-supervision frameworks. We found that DINO outperforms other self-supervised learning frameworks in speaker verification.

Table 3 addresses the effect of the two proposed curriculum learning strategies with the DINO framework. All curriculum strategies, regardless of combinations, consistently outperforms the baseline. Although both approaches are effective, the best performance was observed with CL_D3 which adopts data curriculum only. At most, applying the curriculum strategy to self-supervised learning brought 30% relative improvement over the baseline. When both curriculum strategies are applied at the same time, although there were some performance improvements depending on the combination, none exceeded the performance of the best individual strategy CL_D3. We interpret these results as that excessively lowering the level of difficulty of the training causes the underfitting of the model and decreases the generalization performance. Hence, it is required to construct a
supervised learning mainly lead the success of curriculum learning. Unlike supervised learning, which primarily obtains information from human-generated labels, self-supervised learning relies entirely on supervision from the model itself. For example, in the mean teacher framework applied to DINO, the student network is trained using the output of the teacher model as supervision, and the teacher is updated as the student’s momentum encoder. Hence, in the nature of self-supervised learning, it is crucial to rapidly increase the representation power of the model to improve the quality of supervision. In other words, by increasing the training speed, it is possible to reduce potential label errors. Then, the data curriculum would play a role in speeding up the training. This is because our curriculum strategy reduces data variance by decreasing the amount of data at the beginning of training.

**Validation.** The total variance of data is simply divided into between and within-speaker variances and mainly dominated by the former one. Hence, the key of data curriculum, which controls the number of speakers, would be highly correlated to the between-speaker variance. To valid this assumption, we design an additional experiment, which is less likely to occur under natural data collection pipelines, where we modify the CL_D3 to include all speakers regardless of the portion of train dataset used; in other words, we fix the between-speaker factor and only test the within-speaker factor since only the number of utterances per speaker increases in this experiment. This experiment resulted in an EER of 6.41%, in which the relative improvement over the baseline is 4.3%. Through this result, we hence conclude that the key property to successful data curriculum strategy is limiting the number of speakers (i.e., classes) in the early step of training.

**Insights and suggestions.** With the abovementioned results, it may be argued that the efficacy of the proposed method is limited due to the uncontrollable manipulation of the number of speakers across potential self-supervised learning scenarios. This paper assumes that the number of speakers correlates with the amount of data that may not hold in several cases. In such cases, simple k-means clustering can complement the application of the proposed curriculum strategy. In our internal experiments, where we test various numbers of clusters, we observed that involving more clusters as training proceeds leads to the successful application of the data curriculum technique. Based on these results, we expect that one can easily limit the number of speakers in the training process by using the simple clustering algorithm. Hence, we suggest future researchers cluster the training dataset and apply the proposed data curriculum approach, interpreting clusters as classes.

### 6. Conclusion

This work described our commitment to introduce the DINO framework, a new type of self-supervision framework based on only positive pairs, to speaker verification. We also initiated the use of curriculum learning scheme for the self-supervised learning field. Two types of strategies, designed in particular to enable curriculum learning under self-supervised learning framework, demonstrated consistent improvements across diverse configurations and datasets. Combining our proposals altogether, we established a new-state-of-the-art performance with an equal error rate of 4.47%.

---

**Table 2: Comparison with self-supervised learning models.**

| Framework          | EER(%) | minDCF |
|--------------------|--------|--------|
| Huh et al. [11]    | 8.65   | 0.4540 |
| Xia et al. [30]    | 8.23   | 0.5900 |
| Mun et al. [31]    | 8.01   | N/R    |
| Tao et al. [32]    | 7.36   | N/R    |
| Sang et al. [33]   | 6.99   | 0.4340 |

**Table 3: Speaker verification performance in EER (%) on the VoxCeleb1 test set.** Base indicates the results from the DINO framework without curriculum learning.

|        | Base | CL_D1 | CL_D2 | CL_D3 |
|--------|------|-------|-------|-------|
| Base   | 6.70 | 5.87  | 5.54  | 4.47  |
| CL_A1  | 6.35 | 6.08  | 5.10  | 4.69  |
| CL_A2  | 6.64 | 5.99  | 5.39  | 4.85  |

---

**Table 4: Speaker verification performances in EER (%) on the VoxCeleb1 and CN-Celeb1 test set. Each model is fine-tuned by using the corresponding development set.**

| Initial model | VoxCeleb1 | CN-Celeb1 |
|---------------|-----------|-----------|
| No pretraining| 2.32      | 12.46     |
| DINO          | 1.98      | 10.98     |
| DINO + CL     | **1.84**  | **10.65** |
7. References

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshokha, X. Xiao et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” arXiv preprint arXiv:2110.13900, 2021.

[2] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, “Unispeech: Unified speech representation learning with labeled and unlabeled data,” in International Conference on Machine Learning. PMLR, 2021, pp. 10,937–10,947.

[3] D. Cai and M. Li, “The dku-dukeece system for the self-supervision speaker verification task of the 2021 voxceleb speaker recognition challenge,” arXiv preprint arXiv:2109.02853, 2021.

[4] J. Cho, J. Villalba, and N. Dehak, “The jhu submission to voxsrc-21: Track 3,” arXiv preprint arXiv:2109.13425, 2021.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12,449–12,460, 2020.

[6] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “Data2vec: A general framework for self-supervised learning in speech, vision and language,” arXiv preprint arXiv:2202.03555, 2022.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” arXiv preprint arXiv:2111.06377, 2021.

[10] N. Vaessen and D. A. van Leeuwen, “Fine-tuning wav2vec2 for speaker recognition,” arXiv preprint arXiv:2109.15053, 2021.

[11] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. S. Chung, “Augmenting adversarial training for unsupervised speaker recognition,” in NearIPS workshop on Self-Supervised Learning for Speech and Audio Processing, 2020.

[12] Y. Liu, L. He, J. Liu, and J. Liu, “Large margin softmax loss for speaker verification,” arXiv preprint arXiv:1904.03479, 2019.

[13] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the voxceleb speaker recognition challenge 2020,” arXiv preprint arXiv:2009.14153, 2020.

[14] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7604–7608.

[15] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, “Cn-celeb: multi-genre speaker recognition,” Speech Communication, 2022.

[16] J. Deng, J. Guo, N. Xue, and S. Zaferiou, “Arface: Additive angular margin loss for deep face recognition,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.

[17] M. Sang, H. Li, F. Liu, A. O. Arnold, and L. Wan, “Self-supervised regularization,” arXiv preprint arXiv:2112.04459, 2021.