INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Curriculum learning for self-supervised speaker verification

Hee-Soo Heo[1], *Jee-weon Jung*[1], *Jingu Kang*[2], *Youngki Kwon*[1], *You Jin Kim*[1], *Bong-Jin Lee*[1], *Joon Son Chung*[3]

[1]NAVER Cloud Corporation, South Korea
[2]Bobidi, South Korea
[3]Korea Advanced Institute of Science and Technology, South Korea

heesoo.heo@navercorp.com

Abstract

The goal of this paper is to train effective *self-supervised* speaker representations without identity labels. We propose two curriculum learning strategies within a self-supervised learning framework. The first strategy aims to gradually increase the number of speakers in the training phase by enlarging the used portion of the train dataset. The second strategy applies various data augmentations to more utterances within a mini-batch as the training proceeds. A range of experiments conducted using the DINO self-supervised framework on the VoxCeleb1 evaluation protocol demonstrates the effectiveness of our proposed curriculum learning strategies. We report a competitive equal error rate of 4.47% with a single-phase training, and we also demonstrate that the performance further improves to 1.84% by fine-tuning on a small labelled dataset.

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

1. Introduction

Self-supervised learning (SSL) 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 [5–8].

The speaker verification literature has also adopted SSL and several works have been proposed [4, 9–15]. Specifically, few studies employed a two-phase self-supervised learning strategy, namely iterative clustering [3, 16]. The first phase includes training the models via a SSL 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.

Throughout this study, we focus on improving the self-supervised learning technique itself which involves randomly initialised models. 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 does not require 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 studied. 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 methods.

Curriculum learning, which gradually trains a model in a meaningful order, is widely adopted in diverse supervised learning tasks [17–19]. Even when the same dataset is used, training with an adequately configured curriculum can boost the performance of trained models. However, curriculum learning has not been investigated in conjunction with SSL, leaving the potential open. We thus focus on curriculum learning [18]. Two curriculum learning strategies for SSL that make the training more challenging are designed: (i) gradually increasing the amount of training data and (ii) gradually augmenting noise and reverberation to an increased proportion within each mini-batch. An underlying assumption is that SSL speaker verification will also benefit when the training becomes gradually difficult as it has been the case for a few preceding studies in supervised learning [20].

We conduct experiments on speaker verification with the ECAPA-TDNN model [21] under the DINO [22] SSL framework. Results demonstrate that SSL can also benefit from curriculum strategies. Both proposed curriculum learning techniques were effective, where we observed up to 33% improvement compared to a baseline. In addition, using models trained with SSL, we further explore a semi-supervised scenario where we fine-tune the model with a smaller set of data with labels.

The paper is organised as follows. Section 2 introduces conventional curriculum learning. The adopted SSL framework and model architecture, DINO, is addressed in Sections 3. The proposed curriculum approach and techniques are addressed in Section 4. Experiments and result analysis are presented in Section 5. Section 6 analyses and interprets the operation of the proposed technique in detail.

2. Curriculum learning

Curriculum learning, which defines the sequence of the training from the easy settings to the hard ones, allows efficient training of deep neural networks. 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 [17, 18]. In [19], the authors first trained the model under the text-dependent scenario, then extended 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 a common technique [20, 23]. 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.

10.21437/Interspeech.2023-1202

3. Adopted SSL framework: DINO

DINO [22] is a self-distillation framework based on the mean teacher [24] method, which was originally proposed for the computer vision domain. 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, then outputs local and global features, whereas the teacher network only digests global views, extracting global features. Based on these two types of features, a loss $\mathcal{L}_D$ that penalises the difference between them is defined and used to train the student network. The loss can be described as:

$$\mathcal{L}_D = \Sigma_{\boldsymbol{x}\in\{\boldsymbol{x}_1^g,\boldsymbol{x}_2^g\}} \Sigma_{\substack{\hat{\boldsymbol{x}}\in V \\ \hat{\boldsymbol{x}}\neq\boldsymbol{x}}} H(P_t(\boldsymbol{x}), P_s(\hat{\boldsymbol{x}})), \tag{1}$$

where $H(a,b) = -a\log b$, $\boldsymbol{x}_i^g$ 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. Different to the student which is trained with gradients, the teacher network's weight parameters are derived using an exponential moving average of the student. In addition, sharpening and centring techniques are applied to the teacher output. The purpose is to avoid model collapse which can easily occur in frameworks that only utilise positive pairs, including DINO.

We adapt DINO for speaker verification with a few modifications. Figure 1 illustrates the overall process of the DINO framework adapted for speaker verification. First, we change the global and local views 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 reverberation and noise from simulated RIRs and MUSAN datasets [25, 26]. Augmentation configurations follow that of [27]. 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. The aforementioned adaptation enables the speaker verification model training with the DINO framework. We followed the original paper for the other experimental configurations.

4. Proposed Approach

Existing curriculum learning strategies leverage the ground truth label. However, these strategies are not applicable for SSL because labels do not exist. Hence, we design two curriculum strategies that can be adopted in SSL without the ground truth label by increasing: (i) the size of the train set and (ii) the proportion of utterances within each mini-batch where data augmentation is applied. Both strategies tend to make the training progressively challenging.

4.1. Data curriculum

Finding a speaker discriminant latent space becomes more difficult as the number of speakers to represent increases. We wanted to gradually make the training more challenging by enlarging the number of speakers in the training dataset. However, because labels do not exist in SSL, we could not manually control the number of speakers. As an alternative, we assume that the number of speakers in a dataset will increase proportionally when the size of the dataset enlarges and therefore we control the number of speakers by limiting the size of the training dataset. Although this assumption cannot be guaranteed, a dataset that contains one million randomly collected utterances is likely to have a greater number of speakers compared to a dataset that contains one thousand utterances. In addition, in Section 6, we show another alternative where we adopt k-means clustering algorithm to select a subset of speakers' utterances.

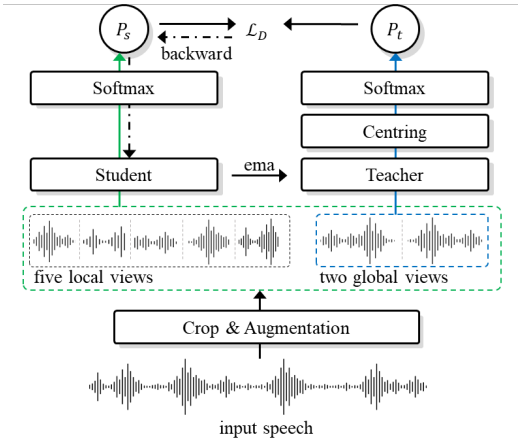

Figure 1: *Adapted DINO framework for speaker verification. Utterances with different durations and augmentations are fed instead of different-sized images. Local and global views are fed into the student network (green line), while the teacher inputs only global views (blue line).*

Three following curriculum courses are designed empirically and illustrated in Figure 2-(a) where it depicts the detailed ratios of data used for each epoch. For example, in the CL_D2 strategy, we use the following dataset for training: During the first 16'th epochs, half of the dataset is used. From 17'th to 32'nd epochs, 75% of the dataset is used. After the 32'nd epoch, entire dataset is used. These curriculum courses are designed considering that the learning rate is reset every 16 epochs in stochastic gradient descent with warm restarts (SGDR) [28] learning rate scheduler.

4.2. Data augmentation curriculum

We further propose a curriculum strategy which adjusts the frequency of data augmentation to control how difficult the training would be. We design two curriculum courses of augmentation by gradually increasing the proportion of augmented utterances within each mini-batch and illustrate specific curriculum courses in Figure 2-(b). Note that the baseline augments all utterances within a mini-batch from the beginning of the training phase.

5. Experiments

We first present two sets of experiments: (i) demonstration of the effectiveness of proposed two curriculum strategies and (ii) fine-tuning the SSL-trained model with a small amount of labelled data (i.e., semi-supervised scenario). Then, we compare our developed model's performance with the recent literature.

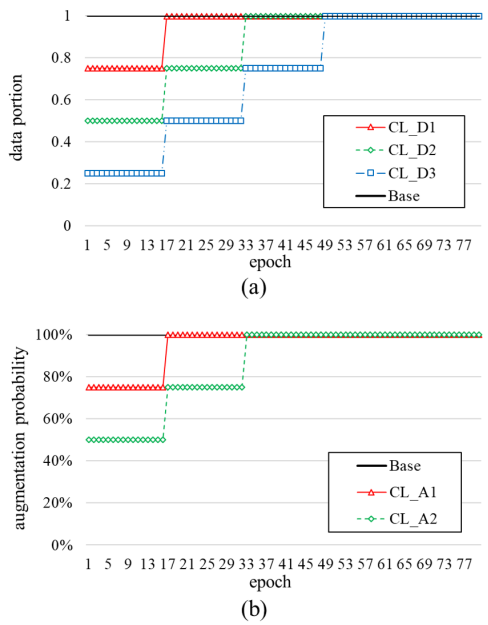

Figure 2: *Curriculum courses. Note that CL_D* stands for* ***C****urriculum* ***L****earning on* ***D****ataset, which controls the amount of dataset used. In a similar manner, in* ***C****urriculum* ***L****earning on* ***A****ugmentation (CL_A*), we control the ratio of augmented data. (a): three courses for exploiting different data portions (*CL_D1*,* CL_D2 *and* CL_D3*) (b): two courses for augmenting a mini-batch (*CL_A1 *and* CL_A2*).*

5.1. Dataset

Our experiments utilise the VoxCeleb1 and 2 datasets for training and evaluating the models [29–31]. 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 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 [32, 33] 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 entire set, we utilise CN-Celeb1, which includes 800 speakers for training and 200 speakers for evaluation. Since CN-Celeb only comprises Chinese, it is valuable for evaluating the scenario where SSL pre-training is done in another domain.

5.2. ECAPA-TDNN model

The model is based on the ECAPA-TDNN architecture that operates on mel-frequency cepstral coefficients input [21]. It comprises 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 one 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.

Table 1: *Hyperparameters of the DINO framework.*

Configuration	**Value**
optimizer	adam
initial learning rate	0.001
weight decay	5.00E-05
batch size	200
epoch	80
learning rate scheduler	SGDR [28] (restart period=16, decay=0.8)

Table 2: *Results of self-supervised learning 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

5.3. Configurations

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

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. The fine-tuning of the model is accomplished utilising an open-source trainer [1], with the following modifications. We use additive angular margin loss to optimise the model [27, 34, 35]. Most of the settings are used as the same 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.

5.4. Results and Analysis

Self-supervised learning. Table 2 addresses the effect of the two proposed curriculum learning strategies with the DINO framework. Both curriculum strategies, regardless of combinations, consistently outperform the baseline. However, although both approaches are effective when applied alone, they were not synergetic when applied simultaneously. The best performance was observed when only data curriculum was applied, CL_D3, where it brought 30% improvement over the baseline. We interpret these results that applying two kinds of curriculum at the same time lowers the difficulty of training beyond our expectation.

Semi-supervised learning. Table 3 shows the results of fine-tuning the best performing SSL-trained model, CL_D3, with two datasets. The first row reports the results from a ran-

[1] https://github.com/clovaai/voxceleb_trainer

Table 3: *Results of semi-supervised learning 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**

domly initialised model that is trained only using the small labelled dataset. In both datasets, replacing the random initialised model to SSL-trained model improved the performance, where 14.65% and 11.87% improvement were observed for VoxCeleb1 and CN-Celeb1. By adopting the initial model trained with SSL under proposed curriculum strategies, further improvement was made, where 7% and 3% additional improvement were achieved. Hence, we conclude that training in SSL with the proposed curriculum strategies are also effective for semi-supervised scenarios as well.

Comparison with recent literature. Table 4 compares the proposed models with existing works that adopt various self-supervision frameworks. Note that all these results show the performances of initial training without the iterative clustering step. First, we find that the mainstream of SSL in speaker verification is on its transition from contrastive-based [14–16, 36] to DINO [37, 38], which only leverages positive pairs. Performances differ in each study, but in general, DINO outperforms contrastive-based approaches by a large margin. Among studies that adopt DINO, the performance of our baseline model (6.70%) falls behind a bit. However, with the best-performing curriculum learning strategy (DINO+CL), EER is further reduced to 4.47%, which is competitive. Furthermore, since adapting DINO for speaker verification is subject to hyperparameter tuning, we argue that our improvements with curriculum strategies will further improve the performance of [37, 38].

6. Discussion

6.1. Analysis

In SSL, it has been reported that rapidly increasing the representation power of the model is crucial. This is more important for SSL frameworks such as DINO, which relies entirely on positive pairs because representation collapse occurs more often. We analyse that this may be the reason why curriculum learning was successful when applied to DINO SSL framework. Our two curriculum learning strategies both initiates the training with easy samples. Therefore, the train loss can decrease and the quality of supervision can improve more rapidly.

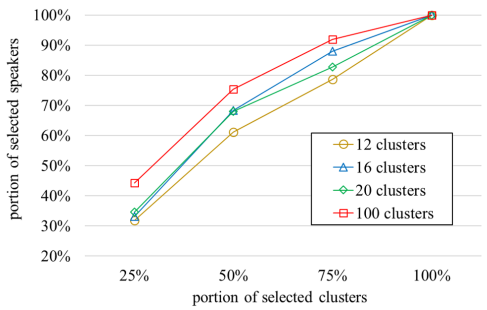

Figure 3: *The proportion of speakers selected according to the number of randomly selected clusters.*

Table 4: *Comparison with self-supervised learning models. minDCF is calculated with P_{target}=0.05 and $C_{false\_alarm} = C_{miss}$=1.*

	Framework	EER(%)	minDCF
Huh et al. [14]	AP+AAT	8.65	0.4540
Xia et al. [15]	MOCO+Wav-Aug(ProtoNCE)	8.23	0.5900
Mun et al. [36]	CEL	8.01	N/R
Tao et al. [16]	Contrastive	7.36	N/R
Sang et al. [39]	SSReg	6.99	0.4340
Han et al. [37]	DINO	6.16	N/R
Cho et al. [38]	DINO	4.83	N/R
Ours	DINO	6.70	0.4116
Ours	DINO+CL	4.47	0.3057

6.2. Ablation on CL_D3

We conduct an ablation experiment to analyse which aspect brought the success of CL_D3 in Table 2. We hypothesised that gradually increasing inter-speaker variance is the key, and we design an experiment to validate this idea by removing the inter-speaker factor and only increasing intra-speaker variance over time. To do so, we use labels to ensure that although the same proportion of the training data is fed to train, the number of speakers are identical throughout the whole training phase[2]. This experiment resulted in an EER of 6.41%, which is close to the EER of 6.70% from DINO without curriculum. We hence conclude that limiting inter-speaker variance at the initial phase of training is crucial to success.

6.3. Additional method for controlling the number of speakers

With the results of Section 6.2, a new concern can be made: the proposed data curriculum strategy may not hold if the number of speakers do not increase as the amount of collected data increase. To account for such scenarios, we additionally demonstrate an alternative approach. A simple k-means clustering with given number of clusters is first adopted to group the train dataset. Then, we select a proportion of clusters according to the curriculum strategy. Figure 3 shows the experimental results. Regardless of the number of total clusters we set, we could successfully control and increase the number of speakers by choosing more clusters. Based on this result, we conclude that even in the case where random selection of more data does not lead to the involvement of more speakers in the train dataset, we can apply a k-means algorithm alternatively and control the number of speakers. Note that label information is not required for this process.

7. Conclusion

In this paper, we proposed two curriculum learning strategies for self-supervised speaker recognition. The two strategies both demonstrated consistent improvements across a diverse range of experimental settings. We also showed that the proposed methods are valid in a semi-supervised scenario. In addition, in-depth analyses regarding the proposed curriculum strategies have been conducted.

[2]Labels are utilised for analysis purpose only.

8. References

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[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 *Proc. ICML*. 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] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. NeurIPS*, vol. 33, 2020, pp. 21 271–21 284.

[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," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 15 979–15 988.

[8] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[9] T. Stafylakis, J. Rohdin, O. Plchot, P. Mizera, and L. Burget, "Self-supervised speaker embeddings," in *Proc. Interspeech*, 2019.

[10] J. Cho, P. Żelasko, J. Villalba, S. Watanabe, and N. Dehak, "Learning speaker embedding from text-to-speech," in *Proc. Interspeech*, 2020.

[11] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," in *Proc. Interspeech*, 2019.

[12] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," in *Proc. ICASSP*. IEEE, 2021, pp. 6718–6722.

[13] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *Proc. ICASSP*, 2022, pp. 7967–7971.

[14] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. S. Chung, "Augmentation adversarial training for unsupervised speaker recognition," in *NeurIPS workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[15] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *Proc. ICASSP*, 2021, pp. 6723–6727.

[16] R. Tao, K. A. Lee, R. K. Das, V. Hautamäki, and H. Li, "Self-supervised speaker recognition with loss-gated learning," in *Proc. ICASSP*, 2022, pp. 6142–6146.

[17] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *arXiv preprint arXiv:2101.10382*, 2021.

[18] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009, pp. 41–48.

[19] E. Marchi, S. Shum, K. Hwang, S. Kajarekar, S. Sigtia, H. Richards, R. Haynes, Y. Kim, and J. Bridle, "Generalised discriminative transform via curriculum learning for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5324–5328.

[20] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[21] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 1–5.

[22] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, 2021, pp. 9650–9660.

[23] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *Proc. CVPR*, 2020, pp. 5901–5910.

[24] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Proc. NeurIPS*, vol. 30, 2017.

[25] 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 *Proc. ICASSP*. IEEE, 2017, pp. 5220–5224.

[26] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv*, 10 2015.

[27] 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.

[28] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[29] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[32] 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 *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.

[33] 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.

[34] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[35] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.

[36] S. H. Mun, W. H. Kang, M. H. Han, and N. S. Kim, "Unsupervised representation learning for speaker recognition via contrastive equilibrium learning," *arXiv preprint:2010.11433*, 2020.

[37] B. Han, Z. Chen, and Y. Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," in *Proc. Interspeech*, 2022.

[38] J. Cho, R. Pappagari, P. Żelasko, L. Moro-Velazquez, J. Villalba, and N. Dehak, "Non-contrastive self-supervised learning of utterance-level speech representations," in *Proc. Interspeech*, 2022.

[39] M. Sang, H. Li, F. Liu, A. O. Arnold, and L. Wan, "Self-supervised speaker verification with simple siamese network and self-supervised regularization," in *Proc. ICASSP*, 2022, pp. 6127–6131.