DISCRIMINATIVE SPEAKER REPRESENTATION VIA CONTRASTIVE LEARNING WITH CLASS-AWARE ATTENTION IN ANGULAR SPACE

Zhe LI [1], Man-Wai MAK[1], Helen Mei-Ling MENG[2]*

[1]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University
[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

ABSTRACT

The challenges in applying contrastive learning to speaker verification (SV) are that the softmax-based contrastive loss lacks discriminative power and that the hard negative pairs can easily influence learning. To overcome the first challenge, we propose a contrastive learning SV framework incorporating an additive angular margin into the supervised contrastive loss in which the margin improves the speaker representation's discrimination ability. For the second challenge, we introduce a class-aware attention mechanism through which hard negative samples contribute less significantly to the supervised contrastive loss. We also employed gradient-based multi-objective optimization to balance the classification and contrastive loss. Experimental results on CN-Celeb and Voxceleb1 show that this new learning objective can cause the encoder to find an embedding space that exhibits great speaker discrimination across languages.

Index Terms— Speaker verification; contrastive learning; additive angular margin; attention mechanism; multi-objective optimization

1. INTRODUCTION

Speaker verification (SV) requires learning a speaker representation through which utterances are encoded as fixed-sized embedding vectors. Similar speakers will have their embedding vectors close, whereas different speakers will have their embedding vectors far apart. For SV tasks, contrastive representation learning, such as InstDisc [1], SimCLR [2], and MoCo [3], are natural fits to this requirement. These methods aim to discover representations by maximizing the concordance among various augmentations or views of the same instance in a latent space.

Contrastive learning has attracted growing attention [4–6] in the SV community. The technique considers the augmented "views" of an utterance to be spoken by the same person and put them into positive pairs. It treats different utterances and their augmented views to be spoken by different speakers and
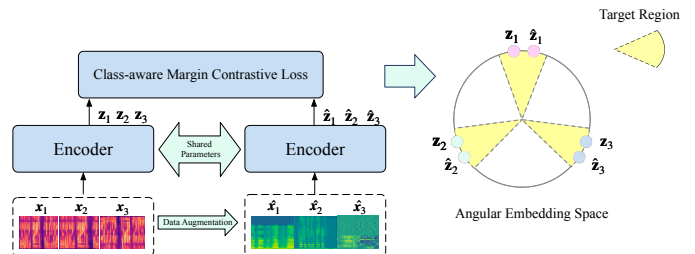

Fig. 1. Illustration of our basic idea. An embedding space is learned in which the same-speaker pairs stay close to each other while different-speaker pairs remain far apart. For simplicity, we only illustrate three speaker pairs.

puts them into negative pairs. The goal is to pull the embeddings in the positive pairs closer together and push those in the negative pairs apart. Many studies [7–9] used the NT-Xent loss, a variant of the cross-entropy loss with a softmax function, to achieve this goal. However, recent studies [10–12] have demonstrated that the traditional softmax-based loss is effective in optimizing the inter-class difference but not as effective in reducing the intra-class variation. As a result, the learned features are discriminative for closed-set classification but insufficient for open-set speaker recognition. Also, most contrastive approaches focus on determining the positive and negative pairs and improving the learning framework [13–15]. The optimization objective is seldom explored.

A hard sample is defined as one for which the deep learning model struggles to accurately predict the label. However, many sample mining methods used in prior studies [16] do not account for hard negative samples. Therefore, incorporating hard negative samples during training may lead to a higher training loss due to the presence of potential outliers.

To alleviate the above challenges, we propose a framework for learning discriminative speaker representations via **C**lass-**A**ware **A**ttention-based **Con**trastive learning with **Margin**, CAAMarginCon. We incorporate an additive angular margin into the supervised contrastive loss, which increases the discrimination capability of the speaker representation by maximizing the decision margin in the angular space. Meanwhile, class-aware attention (CAA) helps identify and

---

ignore very hard negatives. In addition, we employed a multi-objective optimization technique to balance the classification loss and contrastive loss. Fig. 1 illustrates our idea from a high-level perspective. Experiments conducted on CN-Celeb1&2 and VoxCeleb1 demonstrate the superiority of the proposed model over the standard ones. In addition, we conducted an ablation study to determine the contributions of different components of the proposed loss function.

2. METHODOLOGY

Our objective is to train an audio-based feature embedding network using labeled audio. Similar speakers' embedding vectors should be close, whereas those of dissimilar speakers should be far apart. We performed data augmentation on a batch of input samples to produce an augmented batch. Under various augmentations, the embedding vectors derived from the same instance should remain unchanged; on the other hand, the embeddings from different instances should be distinct. As illustrated in Fig. 2, an encoder network receives the spectrograms of both types of instances (the originals and their augmentations) as input and generates normalized embeddings as output. At the network's output, an attention-based class-aware contrastive loss with additive angular margin is computed.

2.1. Angular Margin-based Contrastive Learning

We incorporated the original and augmented speaker embeddings into a supervised contrastive loss [17, 18]:

$$L_{SupCon} = \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_p)/\tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_a)/\tau)}, \tag{1}$$

where $\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_p)$ is the cosine similarity. In Eq. 1, $\boldsymbol{z}_i$ is an anchor, $\boldsymbol{z}_a$ is a negative sample, $\mathcal{A}(i)$ is an index set of negative samples with respect to $\boldsymbol{z}_i$, $\boldsymbol{z}_p$ is a positive sample with respect to $\boldsymbol{z}_i$, and $\mathcal{P}(i)$ contains the indices of positive samples in the augmented batch (original + augmentation). $\tau \in \mathcal{R}^+$ is a scalar temperature parameter. Although the training objective attempts to pull the representations of similar speakers closer together and push the representations of different speakers apart, these representations may not be sufficiently discriminative or resilient to noise.

Let us denote the cosine similarity as

$$\cos \theta_{i,p} = \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_p}{\|\boldsymbol{z}_i\| \|\boldsymbol{z}_p\|}, \tag{2}$$

where $\theta_{i,p}$ is the angle between the embedding $\boldsymbol{z}_i$ and $\boldsymbol{z}_p$. The decision boundary of $\boldsymbol{z}_i$ for specific p and a is $\theta_{i,p} = \theta_{i,a}$, where p and a are indexes to positive and negative samples, respectively. A tiny perturbation of embedding vectors around the decision boundary may result in an incorrect conclusion if no decision margin is present. To overcome this

problem, we propose a new training objective for speaker representation learning by adding an additive angular margin m. We name it **Sup**ervised **Margin Con**trastive (SupMarginCon) loss, which can be formulated as follows:

$$L_{SupMarginCon} =$$
$$\sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\cos\left(\theta_{i,p} + m\right)/\tau\right)}{\sum_{a \in A(i)} \exp\left(\cos\left(\theta_{i,a}\right)/\tau\right)}. \tag{3}$$

In this loss, the decision boundary of $\boldsymbol{z}_i$ for specific p and a is $\theta_{i,p} + m = \theta_{i,a}$, which pushes $\boldsymbol{z}_i$ further towards the area where $\theta_{i,p}$ decreases and $\theta_{i,a}$ increases. Therefore, adding a margin can increase the compactness of same-speaker representations and the disparity between the different-speakers representations. This aid results in improved alignment and uniformity [19] – two quality measures fundamental to contrastive learning. These measures indicate how close positive-pair embeddings are to one another and how uniformly distributed the embeddings are. In addition, the decision boundary leaves an additional margin m from the boundary $\theta_{i,p} + m = \theta_{i,a}$, which is frequently employed during inference, making the boundary more noise-tolerant and resilient. All of these properties make the SupMarginCon loss more discriminative than the conventional loss, such as the SupCon loss (Eq. 1).

2.2. Class-Aware Attention

As indicated in [20, 21], hard samples can easily influence sample mining, resulting in a suboptimal local minimum for the trained model. Usually, we pay more attention to more challenging negatives; however, some of them may be outliers. To determine the relationship between speech embeddings and their classes, we compute the compatibility between an embedding vector and its class vector. The dot product of the two vectors can be used to measure their compatibility [22].

We apply a classification branch after the speech embeddings to learn the class vectors. The class vectors are trainable and can be updated by gradient descent. Denote $\{\boldsymbol{c}_k\}_{k=1}^{C}$ as the C class vectors. In our case, C is the number of speakers in a mini-batch. For utterances i and j, we compute their class-aware attention (CAA) score:

$$\alpha_{i,j} = \frac{\exp(\boldsymbol{z}_i^\top \boldsymbol{c}_{y_j})}{\sum_{k=1}^{C} \exp(\boldsymbol{z}_i^\top \boldsymbol{c}_k)}, \tag{4}$$

where y_j is the class label of $\boldsymbol{z}_j$. Through gradient descent, $\{\boldsymbol{c}_k\}_{k=1}^{C}$ will be able to represent the C classes.

2.3. Multi-objective Optimization for Weighted Loss

After finishing the optimization via the contrastive loss, fixing the encoder's parameters is typically required before training
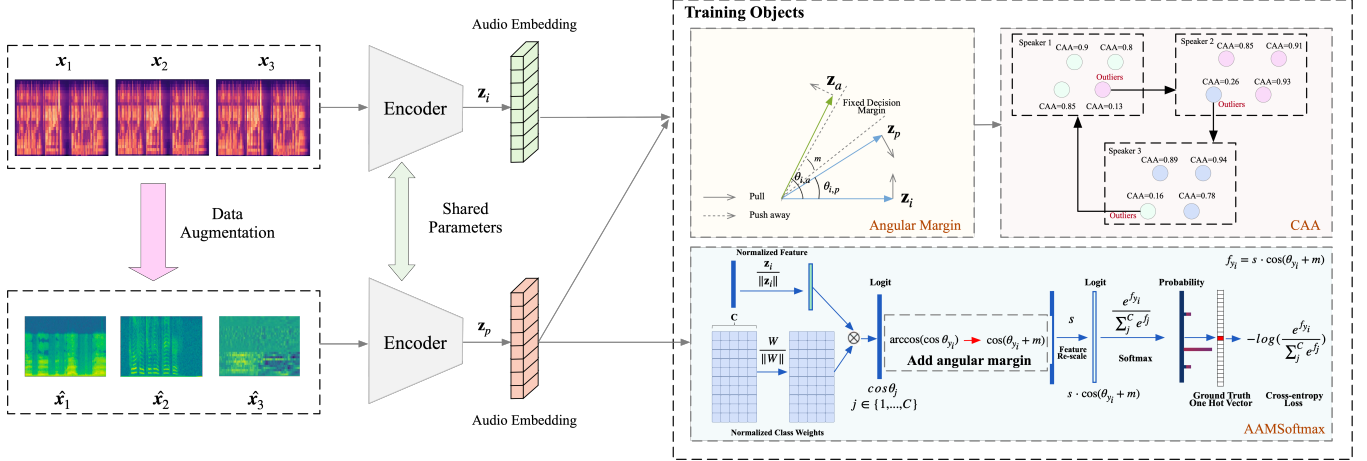
Fig. 2. The proposed architecture leverages additive angular margin and class-aware attention for supervised contrastive learning. CAA aims to address the impact of hard negative samples.

a linear classification layer. However, we advocate achieving both contrastive and classification tasks simultaneously. To this end, we introduce AAMSoftmax [11] to our classification task, which is optimized alongside the contrastive loss during training. We employed gradient-based multi-objective optimization [23] to perform this muti-task learning. This approach optimizes a proxy objective to minimize a weighted linear combination of losses.

Following AAMSoftmax [11], we express $\boldsymbol{w}_{y_i}^\top \boldsymbol{z}_i = \|\boldsymbol{w}_{y_i}\|\|\boldsymbol{z}_i\| \cos\theta_{y_i}$ as the dot product between the class weight of the y_i-th training speaker and the embedding $\boldsymbol{z}_i$ of the i-th utterance. Here, the class weight $\boldsymbol{w}_{y_i}$ is the weight vector corresponding to class y_i at the last layer of the classifier, y_i is the speaker label of $\boldsymbol{z}_i$, and θ_{y_i} is the angle between $\boldsymbol{w}_{y_i}$ and $\boldsymbol{z}_i$. We fix the norm of class weights and embedding vectors to 1 by l_2-normalization and re-scale it to s. We propose a class-aware attention margin contrastive softmax for supervised embedding learning by combining AAMSoftmax and SupMarginCon:

$$L_{CAAMarginCon} =$$
$$-\frac{\lambda_1}{N} \sum_{i=1}^{N} \log \frac{\exp(s \cdot \cos(\theta_{y_i} + m))}{\exp(s \cdot \cos(\theta_{y_i} + m)) + \sum_{j=1, j \neq y_i}^{C} \exp(s \cdot \cos\theta_j)}$$
$$+ \lambda_2 \sum_{i=1}^{n} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\cos(\theta_{i,p} + m) \cdot \alpha_{i,p}/\tau)}{\sum_{a \in A(i)} \exp(\cos(\theta_{i,a}) \cdot \alpha_{i,a}/\tau)}.$$
$$(5)$$

where $\alpha_{i,p}$ is the CAA score of pair $(\boldsymbol{z}_i, \boldsymbol{z}_p)$ and $\alpha_{i,a}$ is the CAA score of pair $(\boldsymbol{z}_i, \boldsymbol{z}_a)$. By considering the centroid of $\boldsymbol{z}_a$ instead of $\boldsymbol{z}_a$ itself, we can scale down the contribution of the hard negatives in Eq. 5 via the CAA scores in Eq. 4. This is because $\boldsymbol{c}_{y_a}$ is *less* hard when compared to $\boldsymbol{z}_a$. Using the same reasoning, Eq. 5 will pay less attention to the *easy* positive because $\alpha_{i,p}$ is less than 1.0 even if $\boldsymbol{z}_p$ is almost identical to $\boldsymbol{z}_i$.

3. EXPERIMENTS AND RESULTS

3.1. Implementation Details

We adopted 80-dimensional Fbank as the input features. We utilized SpecAugment [24] on the log-mel spectrograms. We augmented the original utterances with noise, music, chatter, and reverberation effects. Our training process was divided into three stages using 200, 400, and 600 frames. The Adam and SGD optimizers were used interchangeably. m is a scalar parameter, and we set the margin m to 0.2. We set the s to 30. The mini-batch size ranges between 512 and 4096. The contrastive learning temperature τ was set to 0.07.

3.2. Results and Analysis

We conducted experiments on CN-Celeb1&2 and Voxceleb1. For each metric, the best and the second-best are highlighted in bold and underline, respectively. The evaluation results are shown in Table 1, from which we can see that CAAMarginCon outperforms the previous approaches.

Table 1. The experimental results of the CAAMarginCon and conventional methods on the CN-Celeb evaluation set.

Network	Loss Function	EER(%)	minDCF
DisSpk-1 [25]	Cross Entropy	11.15	0.56
DisSpk-2 [25]	Cross Entropy	11.52	0.55
DisSpk-3 [25]	Cross Entropy	11.27	0.55
ETDNN [26]	Softmax	10.30	0.55
HNN [26]	Softmax	9.18	<u>0.50</u>
MSHNN [26]	Softmax	9.05	**0.48**
ENSEMBLE [26]	Softmax	8.94	**0.48**
TDNN-ASP [27]	BCE&GE2E [28]	9.90	0.56
ResNet34	RAM-Softmax [12]	11.05	N/A
ECAPA-TDNN-1 [27]	AAMSoftmax [11]	10.54	0.58
ECAPA-TDNN-2 [27]	AAMSoftmax [11]	9.51	0.56
ECAPA-TDNN-3 [27]	AAMSoftmax [11]	<u>8.93</u>	<u>0.50</u>
ECAPA-TDNN [29]	CAAMarginCon (ours)	**8.66**	**0.48**

Although ENSEMB and MSHNN [26] achieve the same minDCF as ours, our EER is lower than theirs. Zeng *et al.* [27] proposed an excellent attention-based backend by employing scaled-dot self-attention and feedforward self-attention networks that learn the relationships among the enrollment utterances. The different classification results of ECAPA-TDNN in [27] were caused by the fact that this method uses different backends for evaluating similarity. Our proposed model performs better than the discriminative speaker embedding with serialized multilayer multi-head attention in [25] because our model is equipped with a discriminative margin and pays less attention to the hard negative samples.

We also explored the performance induced by the CAA-MarginCon loss on VoxCeleb1. The results are shown in Table 2. We can see that CAAMarginCon enables our network to perform on par with or better than many state-of-the-art systems.

Table 2. The experimental results of the proposed CAA-MarginCon and conventional methods on the VoxCeleb1 evaluation set. The training set is VoxCeleb1-dev.

Network	Loss Function	EER(%)	minDCF
MACCIF-TDNN [30]	AAM-softmax [11]	3.60	0.36
ECAPA-TDNN [31]	Triplet(semi-hard) [20]	3.37	**0.22**
ECAPA-TDNNLite [32]	AAM-softmax [11]	3.00	0.32
ECAPA-TDNNLite [32]	AAM-softmax [11]	2.96	0.31
ECAPA-TDNN [29]	AAM-softmax [11]	<u>2.96</u>	0.30
ECAPA-TDNN [29]	CAAMargCon (ours)	**2.85**	<u>0.29</u>

3.3. Comparing Loss Functions

To verify the effectiveness of the proposed loss function, we used different loss functions but kept the encoder the same. As seen from Table 3, our proposed contrastive loss achieves the best performance, demonstrating the robustness of our proposed model across language. In addition to the cross-entropy loss, CAAMarginCon has a decision margin. It can be seen that the losses with a decision margin are much better than the ordinary softmax (cross-entropy) loss. Compared to AAMSoftmax, CAAMarginCon is superior because of the class-aware attention mechanism. This indicates that a better strategy for mining negative samples can improve the discriminative power of networks trained with contrastive objectives.

Table 3. The influence of loss functions on CN-Celeb.

Encoder	Loss Function	EER(%)	minDCF
	Cross Entropy	16.07	0.82
	AMSoftmax [10]	13.39	0.71
ECAPA-TDNN [29]	RAMSoftmax [12]	13.25	0.72
	AAMSoftmax [11]	<u>8.79</u>	<u>0.50</u>
	CAAMarginCon	**8.66**	**0.48**

Table 4 shows that CAAMarginCon is superior on Voxceleb1. The EER and minDCF of Softmax (cross-entropy) loss is the highest, suggesting that the additive angular margin contributes significantly to the performance improvement.

Table 4. The influence of loss functions on VoxCeleb1.

Encoder	Loss Function	EER(%)	minDCF
	Cross Entropy	7.72	0.70
	AMSoftmax [10]	5.03	0.48
ECAPA-TDNN [29]	RAMSoftmax [12]	5.18	0.47
	AAMSoftmax [11]	<u>2.96</u>	<u>0.30</u>
	CAAMarginCon	**2.85**	**0.29**

We explored the importance of CAA and margin by removing one component at a time from the CAAMarginCon loss. For simplicity, we set the batch size to 512 and ran for 300 epochs. Table 5 shows that adding a margin to SupCon can improve performance. Similarly, adding CAA can also improve performance; when the two are combined, the performance is the best.

Table 5. The influence of CAA and Margin on CAAMargin-Con. Results are based on CN-Celeb1&2 or Voxceleb1-dev for training and CN-Celeb1-test or Voxceleb1-test for evaluations.

Loss	CN-Celeb1-test		Voxceleb1-test	
	EER(%)	minDCF	EER(%)	minDCF
CAAMarginCon	10.31	0.61	4.06	0.46
w/o Margin	10.22	0.64	4.15	0.46
w/o CAA	10.39	0.62	4.11	0.46
w/o CAA and Margin	11.03	0.66	4.32	0.47

We also performed experiments on Voxceleb to explore the effect of the CAA and margin. As shown in Table 5, both can improve the model's performance. But unlike the performance on CN-Celeb, the effect of the CAA module is not so remarkable. We conjecture that Voxceleb has less noise than the CN-Celeb, i.e., fewer outliers. The CAA module has more contribution to performance under challenging conditions.

4. CONCLUSIONS

In this work, we propose CAAMarginCon, a supervised contrastive learning framework for learning discriminative speaker representations. CAAMarginCon incorporates an angular margin and a class-aware attention mechanism into the supervised contrastive loss. We optimized the classification and contrastive tasks using a gradient-based multi-objective optimization method. The experimental results of CN-Celeb and VoxCeleb1 demonstrate that both techniques bring substantial improvement.

5. REFERENCES

[1] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance discrimination," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. International Conference on Machine Learning*, 2020, pp. 1597–1607.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[4] Yanxi Tang, Jianzong Wang, Xiaoyang Qu, and Jing Xiao, "Contrastive learning for improving end-to-end speaker verification," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–7.

[5] Haoran Zhang, Yuexian Zou, and Helin Wang, "Contrastive self-supervised learning for text-independent speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6713–6717.

[6] Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao, "A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4892–4903.

[7] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu, "Consert: A contrastive framework for self-supervised sentence representation transfer," *arXiv preprint arXiv:2105.11741*, 2021.

[8] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 879–895.

[9] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6894–6910.

[10] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: 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.

[12] Lantian Li, Ruiqian Nai, and Dong Wang, "Real additive margin softmax for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7527–7531.

[13] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5834–5838.

[14] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6127–6131.

[15] Wei Xia, Chunlei Zhang, Chao Weng, Meng Yu, and Dong Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6723–6727.

[16] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. ECCV*, 2016, pp. 732–748.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[18] Zhe Li and Man-Wai Mak, "Speaker representation learning via contrastive loss with maximal speaker separability," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 962–967.

[19] Tongzhou Wang and Phillip Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. International Conference on Machine Learning*, 2020, pp. 9929–9939.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[21] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1153–1162.

[22] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr, "Learn to pay attention," in *Proc. International Conference on Learning Representations*, 2018.

[23] Ozan Sener and Vladlen Koltun, "Multi-task learning as multi-objective optimization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[24] 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 *Proc. Interspeech*, 2019, pp. 2613–2617.

[25] Hongning Zhu, Kong Aik Lee, and Haizhou Li, "Discriminative speaker embedding with serialized multi-layer multi-head attention," *Speech Communication*, vol. 144, pp. 89–100, 2022.

[26] Jahangir Alam, Woo Hyun Kang, and Abderrahim Fathan, "Hybrid neural network-based deep embedding extractors for text-independent speaker verification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 33–40.

[27] Chang Zeng, Xin Wang, Erica Cooper, Xiaoxiao Miao, and Junichi Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6717–6721.

[28] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[29] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[30] Fangyuan Wang, Zhigang Song, Hongchen Jiang, and Bo Xu, "MACCIF-TDNN: Multi aspect aggregation of channel and context interdependence features in tdnn-based speaker verification," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 214–219.

[31] Zhongxin Bai, Jianyu Wang, Xiao-Lei Zhang, and Jingdong Chen, "End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1330–1344, 2022.

[32] Qingjian Li, Lin Yang, Xuyang Wang, Xiaoyi Qin, Junjie Wang, and Ming Li, "Towards lightweight applications: Asymmetric enroll-verify structure for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7067–7071.