ASYMMETRIC CLEAN SEGMENTS-GUIDED SELF-SUPERVISED LEARNING FOR ROBUST SPEAKER VERIFICATION

Chong-Xin Gan, Man-Wai Mak and Weiwei Lin

Dept. of Electrical and Electronic Engineering
The Hong Kong Polytechnic University
Hong Kong SAR

Jen-Tzung Chien

Inst. of Electrical and Computer Engineering
National Yang Ming Chiao Tung University
Taiwan

ABSTRACT

Contrastive self-supervised learning (CSL) for speaker verification (SV) has drawn increasing interest recently due to its ability to exploit unlabeled data. Performing data augmentation on raw waveforms, such as adding noise or reverberation, plays a pivotal role in achieving promising results in SV. Data augmentation, however, demands meticulous calibration to ensure intact speaker-specific information, which is difficult to achieve without speaker labels. To address this issue, we introduce a novel framework by incorporating clean and augmented segments into the contrastive training pipeline. The clean segments are repurposed to pair with noisy segments to form additional positive and negative pairs. Moreover, the contrastive loss is weighted to increase the difference between the clean and augmented embeddings of different speakers. Experimental results on Voxceleb1 suggest that the proposed framework can achieve a remarkable 19% improvement over the conventional methods, and it surpasses many existing state-of-the-art techniques.

Index Terms— Speaker verification; contrastive learning; self-supervised learning; hard negative pairs; weighted contrastive loss

1. INTRODUCTION

Speaker verification (SV) aims to recognize if two utterances are spoken by the same person [1, 2, 3]. This technique can be integrated into various applications, including financial security [4] and forensic voice analysis [5]. Earlier work on SV trained a Gaussian mixture model on a population of speakers and aligned acoustic frames against the Gaussians in the model to extract utterance-level features called the i-vector [6]. The i-vectors of a target speaker and an unknown speaker were fed into a backend classifier, such as probabilistic linear discriminant analysis (PLDA) [7] for scoring. The classifiers effectively decoupled the speaker-specific information from potentially misleading channel factors.

This work was supported by the RGC of Hong Kong SAR, Grant No. PolyU 15210122, and the NSTC of Taiwan, Grant No. 112-2634-F-A49-006.

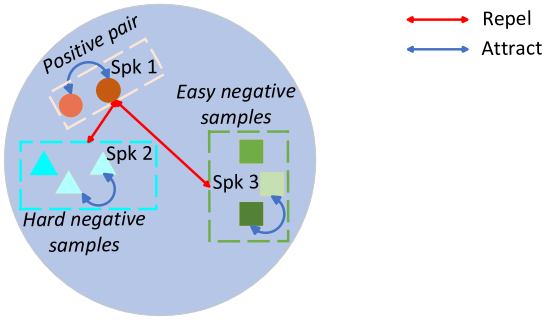

Fig. 1. Distribution of speakers in the embedding space. Easy negative samples are easily differentiated from the target speaker (Spk 1), whereas hard negative samples are confusable with the target speaker. The hard negatives may mislead the network to converge towards a suboptimal solution.

With the emerging trend of deep neural networks (DNNs), numerous studies [8, 9, 10, 11] demonstrated superior results by training networks with different deep architectures and loss functions [12, 13, 14, 15]. The significant accomplishments achieved so far heavily relied on a substantial amount of annotated data. Given the labeled data, a network can learn good speaker representations by minimizing the loss between the output probabilities and the ground truths. Labeled datasets, however, are difficult and expensive to collect. Thus, many researchers have shifted their focus to using self-supervised learning (SSL) for training speaker embedding networks. One category of SSL, which uses contrastive objectives, provides a practical solution. This category of approaches is inspired by some cutting-edge frameworks such as SimCLR [16] and MoCo [17]. These contrastive-based algorithms considered two non-overlapping segments from the same utterance as a positive pair and treated the segments from different utterances as negative pairs [18, 19, 20, 21]. The network was optimized with the objective of pulling positive samples (embeddings from the same speaker) closer and negative samples (embeddings from different speakers) apart.

Researchers commonly employ data augmentation tech-

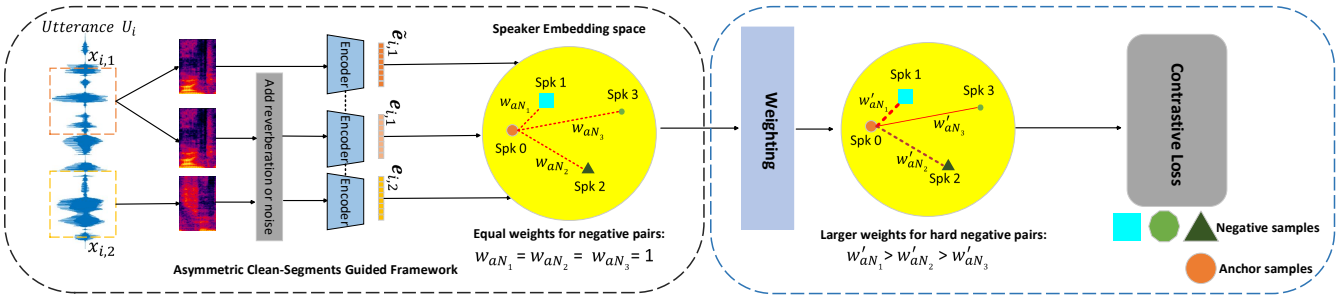

Fig. 2. Overview of asymmetric clean segments-guided self-supervised learning framework. One additional clean segment is fed into the speaker encoder to form extra positive and negative pairs. The weighting mechanism guides the network to pay attention to hard negative pairs. The proposed contrastive loss is used to optimize the speaker encoder.

niques to reduce non-speaker information in the positive pairs, e.g., adding different types of noise to two segments of the same utterance. Adding noise and reverberation has become a standard practice to enhance the robustness of a speaker encoder. However, researchers also argued [22, 23] that not all types of augmentation are good for learning representations. In particular, excessive noise and reverberation may lead to the loss of speaker information. On the other hand, if the augmentation types are too alike, the model may be biased towards a few acoustic conditions, causing poor performance in unseen conditions. Therefore, preventing an embedding network from encoding undesirable factors is crucial. To this end, the authors of [24, 25, 26] used adversarial training to suppress the channel effects in the embeddings. However, training the channel discriminator and the speaker classifier in an adversarial manner is challenging. In this study, we propose a simple solution from a different perspective. Specifically, we introduce clean segments without augmentation into the training process to form extra positive and negative pairs to mitigate speaker information loss.

Several studies [27, 28] have proposed strategies to identify and utilize hard negative pairs to strengthen the robustness of encoders. Fig. 1 shows some examples of hard negatives that are confusable with the target speaker. Paying more attention to these challenging negative pairs (i.e. the pairs formed by the target and the hard negatives) may help a speaker encoder learn robust representations with a large inter-class margin. However, this aspect has not been widely explored in the speech community. Recently, [29] proposed a class-aware attention mechanism that weights the positive and negative pairs according to the similarity of the embeddings with their class representatives. This approach effectively handled hard negatives but relies on speaker labels, which may not be available. Motivated by the same goal, we compute the weights that are dependent on the similarity scores of negative pairs. These weights can be dynamically adjusted according to the similarity between the target and negative samples. The method achieves impressive performance on the task of using Voxceleb dataset.

In summary, we proposed integrating clean segments into the training pipeline to alleviate the loss of speaker-dependent information caused by data augmentation. Meanwhile, we introduced a weighting mechanism to enlarge the inter-class margin by actively paying attention to hard negative pairs. Experimental results showed that the proposed method can achieve superior results, outperforming many existing approaches.

2. METHODOLOGY

To address the aforementioned challenges, we developed an **A**symmetric **C**lean **S**egments-**G**uided (ACSG) self-supervised contrastive learning framework for speaker verification. On top of the ACSG loss, we introduce the similarity-dependent weighting factors to weight across different negative pairs, which we refer to as W-ACSG. The weighting mechanism enhances the contribution of hard negative pairs in contrastive learning, making the speaker embedding network more robust. The architecture of the proposed method is illustrated in Fig. 2. To counteract any potential loss of speaker information caused by random augmentation, we feed additional clean segments alongside the two noisy segments into the speaker encoder.

2.1. Asymmetric Clean Segments-Guided Framework

Given a dataset, we apply data augmentation (adding noise and reverberation) to its utterances. We consider the original speech segments from these utterances as clean and treat those corrupted by data augmentation as noisy. A mini-batch comprising N utterances $\{U_1, \ldots, U_N\}$ are assumed to be spoken by N different speakers. This assumption is based on the low probability of false negative pairs, due to the small batch size and the large number of speakers in the dataset [18]. Two non-overlapping segments, denoted as $x_{i,1}$ and $x_{i,2}$, are randomly truncated from the utterance U_i. After performing data augmentation, two noisy segments are fed into the speaker encoder $f(.)$ to obtain the speaker embeddings $\boldsymbol{e}_{i,1}$ and $\boldsymbol{e}_{i,2}$:

$$\boldsymbol{e}_{i,j} = f(x_{i,j}), \quad j = 1, 2. \tag{1}$$

Segments from the same utterance should capture the same speaker information; thereby, they form the positive pairs. On the other hand, the segments from different utterances form the negative pairs. To pull the embeddings of positive pairs together while pushing the embeddings of negative pairs apart, we use the contrastive objective [16]:

$$\mathcal{L}_{i,j}^{\text{CSL}} = -\log \frac{\exp[\cos(\boldsymbol{e}_{i,1}, \boldsymbol{e}_{i,2})/\tau]}{\sum_{k=1}^{N}\sum_{l=1}^{2} \mathbb{1}_{\substack{k\neq i \\ j\neq l}} \exp[\cos(\boldsymbol{e}_{i,j}, \boldsymbol{e}_{k,l})/\tau]}, \tag{2}$$

where i indexes the samples in a mini-batch, N is the number of samples in the batch, and $\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$ is the cosine similarity. In Eq. 2, τ represents the contrastive temperature, which will be omitted in the following equations for simplicity, and the indicator function $\mathbb{1}$ outputs a 1 when $k \neq i$ and $l \neq j$; otherwise it outputs a 0.

Excessive augmentation could result in imperfect speaker representation because of the potential loss of speaker information in the augmentation process. We adopt a novel strategy to avoid such a situation. Specifically, we include the clean segments for training. To manage computational complexity, for each utterance, we only include the first clean segment in the training process. As a result, each utterance has three segments for contrastive learning: two noisy segments (augmented) and one clean (original) segment. For each mini-batch, $3N$ segments are fed into the speaker encoder. For each utterance, we form an additional positive pair by considering the first clean segment and the second noisy segment to prevent the speaker encoder from learning the contextual information. The negative pairs are formed by clean and noisy segments sampled from different utterances. We reformulated the contrastive loss by a new Asymmetric Clean Segments-Guided (ACSG) contrastive loss:

$$\mathcal{L}_{i,j}^{\text{ACSG}} = -\log \frac{\exp[\cos(\boldsymbol{e}_{i,1}, \boldsymbol{e}_{i,2}) + \cos(\tilde{\boldsymbol{e}}_{i,1}, \boldsymbol{e}_{i,2})]}{\sum_{k=1}^{N}\sum_{l=1}^{2} \mathbb{1}_{\substack{k\neq i \\ j\neq l}} \exp[\cos(\boldsymbol{e}_{i,j}, \boldsymbol{e}_{k,l}) + \cos(\tilde{\boldsymbol{e}}_{i,j}, \boldsymbol{e}_{k,l})]}, \tag{3}$$

where $\tilde{\boldsymbol{e}}_i$ denotes the embedding vector from the clean segment of U_i. The ACSG loss enhances the diversity of training samples by incorporating clean segments for network training. This approach efficiently utilizes the available samples.

2.2. Similarity-Dependent Weights for Hard Negatives

Negative pairs in contrastive learning generally can be divided into hard and easy negative pairs. Hard negative pairs refer to segments spoken by different speakers but still have high similarity scores. On the other hand, easy negative pairs exhibit significant dissimilarities between segments, enabling their embeddings to be easily distinguished. To tackle the challenge posed by hard negative pairs, we implement a weighting mechanism for the proposed ACSG contrastive loss.

Contrastive learning is enhanced by emphasizing those confusing pairs with hard negative samples. Treating contrastive learning as single-sample discrimination, the weighting scheme is inspired by employing a large-margin classifier for a contrastive learning problem. As indicated in Eq. 4 and Eq. 5, we compute two similarity-dependent weights, β_n and β_c, to amplify the contribution of negative pairs when their similarity scores are high and reduce their importance when their similarity is low. Here, we describe the method for a single clean segment. However, our approach can be extended to multiple clean segments.

$$\mathcal{L}_{i,j}^{\text{W-ACSG}} = -\log \frac{\exp[\cos(\boldsymbol{e}_{i,1}, \boldsymbol{e}_{i,2}) + \cos(\tilde{\boldsymbol{e}}_{i,1}, \boldsymbol{e}_{i,2})]}{\sum_{k=1}^{N}\sum_{l=1}^{2} \mathbb{1}_{\substack{k\neq i \\ j\neq l}} \exp[\beta_n \cos(\boldsymbol{e}_{i,j}, \boldsymbol{e}_{k,l}) + \beta_c \cos(\tilde{\boldsymbol{e}}_{i,j}, \boldsymbol{e}_{k,l})]}, \tag{4}$$

where

$$\beta_n = \exp(\cos(\boldsymbol{e}_{i,j}, \boldsymbol{e}_{k,l})) \text{ and } \beta_c = \exp(\cos(\tilde{\boldsymbol{e}}_{i,j}, \boldsymbol{e}_{k,l})). \tag{5}$$

The total loss for each mini-batch is yielded by:

$$\mathcal{L}^{\text{total}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2}\mathcal{L}_{i,j}^{\text{W-ACSG}}. \tag{6}$$

By optimizing the parameters of the speaker encoder with the reformulated contrastive loss, the network is capable of learning robust speaker representations.

3. EXPERIMENTAL SETUP

The training set used for training the speaker encoder is Voxceleb2 [30], which contains 1,092,009 utterances from 5,994 speakers. As most of these utterances are continuous speech, we do not apply voice activity detection (VAD) to filter out silent frames. We selected the emphasized channel attention, propagation and aggregation in time-delay neural network (ECAPA-TDNN) [10] as the speaker encoder. We set the channel size to 512. The duration of input audio is 1.8 seconds, and we used the 80-dim log mel-spectrograms extracted from the raw waveforms with a hamming window of 25ms and a frameshift of 10ms as the input features. The output of the speaker encoder is a 192-dim speaker embedding. During the training process, no speaker labels were used.

We evaluated the performance of the speaker encoder on the Voxceleb1 original test set [31], which consists of 4,874 utterances. We optimized the network with an Adam optimizer. During the testing stage, we calculated the cosine similarity between speaker embeddings of test pairs. The performance metrics are equal error rate (EER) and minimum detection cost function (MinDCF). We set the contrastive learning temperature τ to 1. The speaker embeddings were L2-normalized.

Method	Encoder	Batch Size	EER(%)	minDCF
Baseline [1]	ECAPA-TDNN	256	8.67	0.50
Nagrani et al. [34]	VGG-M	900	22.09	-
Keoage at al. [21]	Thin ResNet-34	256	13.46	0.85
Huh et al. [24]	Fast ResNet-34	200	8.65	0.45
Zhang et al. [18]	Thin ResNet-34	256	8.28	0.61
Xia et al. [19]	X-vector	4096	8.23	0.59
Mun et al. [20]	Fast ResNet-34	200	8.01	-
Tao et al. [25]	ECAPA-TDNN	256	7.36	-
W-ACSG (ours)	ECAPA-TDNN	256	**7.02**	**0.39**

Table 1. Performance comparison of existing and proposed methods on the Voxceleb1 evaluation set. All of the encoders do not have a classification head, and they were trained without using speaker labels.

To simulate different environments and reduce the effect of non-speaker information, we augmented two non-overlapping segments by adding noise, music, and reverberation randomly sampled from the MUSAN [32] and RIR [33] datasets. The noise types included ambient noise, television, and babble noise, with SNR ranging from 5 to 20 dB. For the reverberation, we performed a convolution operation with simulated room impulse responses. We removed the discriminator training and followed the same setting as in [25] to obtain the baseline result. The initial learning rate was 0.001 and it decreased 5% for every 5 epochs. The batch size was set to 256.

4. RESULTS AND ANALYSIS

4.1. Comparison with Existing Works

We obtained the baseline performance by directly feeding two speaker embeddings into the vanilla contrastive loss (Eq. 2). This approach obtained an EER of 8.67% and a minDCF of 0.50. We compared the performance of the proposed loss function with other state-of-the-art SSL methods in Table 1. Notably, the proposed method, W-ACSG, outperforms many previous approaches. For instance, [34] employed a cross-modal to provide self-supervision signals for speaker verification. Using a VGG-based network, they achieved promising results by adopting a large batch size. In the case of SimCLR, a larger batch size can potentially yield better performance, due to the presence of numerous negative pairs. As highlighted in [17], negative pairs played an important role in preventing model collapse; as a result, increasing their number helps contrastive learning.

Previous works by Huh *et al*. [24] and Tao *et al*. [25] incorporated a discriminator alongside a speaker encoder to suppress the channel effects. However, tuning a discriminator is difficult and not always stable. On the other hand, the MoCo-based approach [21] utilizes momentum training and does not require a large batch size. Our proposed method surpasses these classical approaches by introducing clean segments and a weighting mechanism.

[1]Followed the setting of Tao *et al.* [25] but removing the discriminator.

Method	Encoder	EER(%)	minDCF
W-ACSG (ours)	ECAPA-TDNN	7.02	0.39
w/o Clean segments		7.96	0.47
w/o Weighting		8.04	0.45
w/o both		8.67	0.50

Table 2. The effect of clean segments and weighting mechanism on SV performance.

Among the related works, [18] is the most similar to ours, as they also incorporated clean segments into a channel-invariant loss. However, their method minimized the Euclidean distance between clean and noisy segments, which was suboptimal in the embedding space when the speaker embeddings were pre-normalized. In our experiments, we only introduced one clean segment for each utterance but achieved better performance.

4.2. Ablation Study

We investigated the importance of clean segments and hard negative pairs individually. Table 2 shows that both methods contribute to the performance gains. When we included clean segments in the training pipeline, the EER was reduced from 8.67% to 7.96%. Removing the weighting mechanism increases the EER from 7.06% to 8.04%. Combining both methods results in the best performance.

Batch Size	EER(%)	minDCF
64	8.06	0.47
128	7.57	0.42
256	7.02	0.39

Table 3. Ablation study on batch size.

Table 3 shows the effect of batch size on performance. The results show that our system with a mini-batch of 256 achieves the best performance, which closely aligns with the findings in the contrastive learning literature. Our methods surpass many previous works, even with a batch size of 128.

5. CONCLUSIONS

This paper has explored the impact of clean samples in contrastive learning for speaker verification. We introduced a weighted contrastive loss within an asymmetric self-supervised learning framework for robust speaker representation learning. The proposed method incorporated clean segments to mitigate speaker-dependent information loss. To produce robust speaker embeddings, we optimized the encoder with a weighted loss function. The experimental results on Voxceleb1 demonstrated the merit of the framework.

6. REFERENCES

[1] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12–40, 2010.

[2] Youzhi Tu, Weiwei Lin, and Man-Wai Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99038–99049, 2022.

[3] Man-Wai Mak and Jen-Tzung Chien, *Machine learning for speaker recognition*, Cambridge University Press, 2020.

[4] Sefik Emre Eskimez, Peter Soufleris, Zhiyao Duan, and Wendi Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication*, vol. 99, pp. 101–113, 2018.

[5] Joseph P. Campbell, Wade Shen, William M. Campbell, Reva Schwartz, Jean-Francois Bonastre, and Driss Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[6] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[7] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[8] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[9] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. IEEE Spoken Language Technology Workshop*, 2016, pp. 165–170.

[10] 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.

[11] Joon Son Chung, Jaesung Huh, and Seongkyu Mun, "Delving into voxceleb: environment invariant speaker recognition," *arXiv preprint arXiv:1910.11238*, 2019.

[12] 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*, 2018, pp. 4879–4883.

[13] Yi Liu, Liang He, and Jia Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.

[14] Zili Huang, Shuai Wang, and Kai Yu, "Angular softmax for short-duration text-independent speaker verification.," in *Proc. Interspeech*, 2018, pp. 3623–3627.

[15] Chunlei Zhang and Kazuhito Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances.," in *Proc. Interspeech*, 2017, pp. 1487–1491.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Sung Hwan Mun, Woo Hyun Kang, Min Hyun Han, and Nam Soo Kim, "Unsupervised representation learning for speaker recognition via contrastive equilibrium learning," *arXiv preprint arXiv:2010.11433*, 2020.

[21] Théo Lepage and Réda Dehak, "Label-efficient self-supervised speaker verification with information maximization and contrastive learning," in *Proc. Interspeech*, 2022, pp. 4018–4022.

[22] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola, "What makes for good views for contrastive learning?," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6827–6839, 2020.

[23] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell, "What should not be contrastive in contrastive learning," in *Proc. International Conference on Learning Representations*, 2021.

[24] Jaesung Huh, Hee Soo Heo, Jingu Kang, Shinji Watanabe, and Joon Son Chung, "Augmentation adversarial training for self-supervised speaker recognition," *arXiv preprint arXiv:2007.12085*, 2020.

[25] Ruijie Tao, Kong Aik Lee, Rohan Kumar Das, Ville Hautamäki, and Haizhou Li, "Self-supervised speaker recognition with loss-gated learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6142–6146.

[26] Longxin Li, Man-Wai Mak, and Jen-Tzung Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2236–2245, 2022.

[27] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka, "Contrastive learning with hard negative samples," in *Proc. International Conference on Learning Representations*, 2021.

[28] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus, "Hard negative mixing for contrastive learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 21798–21809.

[29] Zhe Li, Man-Wai Mak, and Helen Mei-Ling Meng, "Discriminative speaker representation via contrastive learning with class-aware attention in angular space," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

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

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

[32] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[33] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5220–5224.

[34] Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6829–6833.