SELF-SUPERVISED SPEAKER RECOGNITION WITH LOSS-GATED LEARNING

Ruijie Tao[1], *Kong Aik Lee*[2], *Rohan Kumar Das*[3], *Ville Hautamäki*[1,4] *and Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore [3]Fortemedia Singapore, Singapore
[4]School of Computing, University of Eastern Finland, Finland

ABSTRACT

In self-supervised learning for speaker recognition, pseudo labels are useful as the supervision signals. It is a known fact that a speaker recognition model doesn't always benefit from pseudo labels due to their unreliability. In this work, we observe that a speaker recognition network tends to model the data with reliable labels faster than those with unreliable labels. This motivates us to study a loss-gated learning (LGL) strategy, which extracts the reliable labels through the fitting ability of the neural network during training. With the proposed LGL, our speaker recognition model obtains a 46.3% performance gain over the system without it. Further, the proposed self-supervised speaker recognition with LGL trained on the VoxCeleb2 dataset without any labels achieves an equal error rate of 1.66% on the VoxCeleb1 original test set. We plan to release the codes later for public use.

Index Terms— self-supervised speaker recognition, pseudo label selection, loss-gated learning

1. INTRODUCTION

Speaker recognition aims to recognize persons from their voices [1–3]. Over the last decade, speaker recognition models trained via supervised learning have achieved remarkable performance [4–7]. However, these methods usually require a large set of data with manually annotated speaker labels. The creation of such annotated data is not only immensely costly, but also laborious. Self-supervised learning doesn't require such speaker labels [8, 9], that opens up the opportunity to leverage the abundant unlabeled speech resources.

The state-of-the-art self-supervised speaker recognition system consists of two stages [10, 11]. In Stage I, we solve the contrastive pretraining task through SimCLR [10, 12] or MoCo [13, 14], then the speaker encoder can learn the meaningful speech representation. Meanwhile, various loss functions are proposed to set contrastive targets [9, 15–17]. However, the performance at the Stage I is limited due to the lack of speaker identity information. Recently, an iterative Stage II [10, 11] is proposed to address this issue, where a clustering algorithm is applied to generate the pseudo labels for each utterance based on the learnt representation. With the pseudo labels, the network is trained with a classifier in a supervised manner. This process is repeated several times to improve the speaker encoder.

The state-of-the-art studies take the pseudo labels for fully supervised classification. Therefore, the quality of classification in the Stage II decides the upper bound of self-supervised speaker recognition [10, 11]. Usually, the pseudo labels include massive unreliable labels [18]. Such unreliable pseudo labels will adversely affect the performance of the encoder, that highlights the importance of having an effective and reliable selection of pseudo labels [19].

In this work, we hypothesize that neural networks model the data with reliable labels faster than the those with unreliable labels. Specifically, talking one utterance at the time, if the forward pass gives a low loss value, we can consider its pseudo label to be reliable, whereas in the case of high loss value, its pseudo label is unreliable. We design a toy experiment to validate our hypothesis. To this end, we propose a *loss-gated learning* (LGL) method to effectively select the data with reliable labels. A threshold is involved to retrain the data with small loss. Only the filtered data are then used to update the network. We believe the proposed LGL is capable of selecting the reliable labels to contribute towards an improved performance. In summary, we make the following contributions:

- We confirm that neural networks fit the reliable pseudo labels faster than the unreliable ones in self-supervised speaker recognition.
- Based on our finding, a LGL strategy is proposed to effectively select the reliable pseudo speaker labels.
- We compare the proposed LGL strategy to the baseline and fully supervised method for speaker recognition.

2. TWO-STAGE ARCHITECTURE

We now describe our two-stage baseline architecture. As shown in Fig. 1, a contrastive pretraining task is employed to learn meaningful speaker embeddings. Subsequently, we obtain the pseudo labels by clustering and train a classification network iteratively.

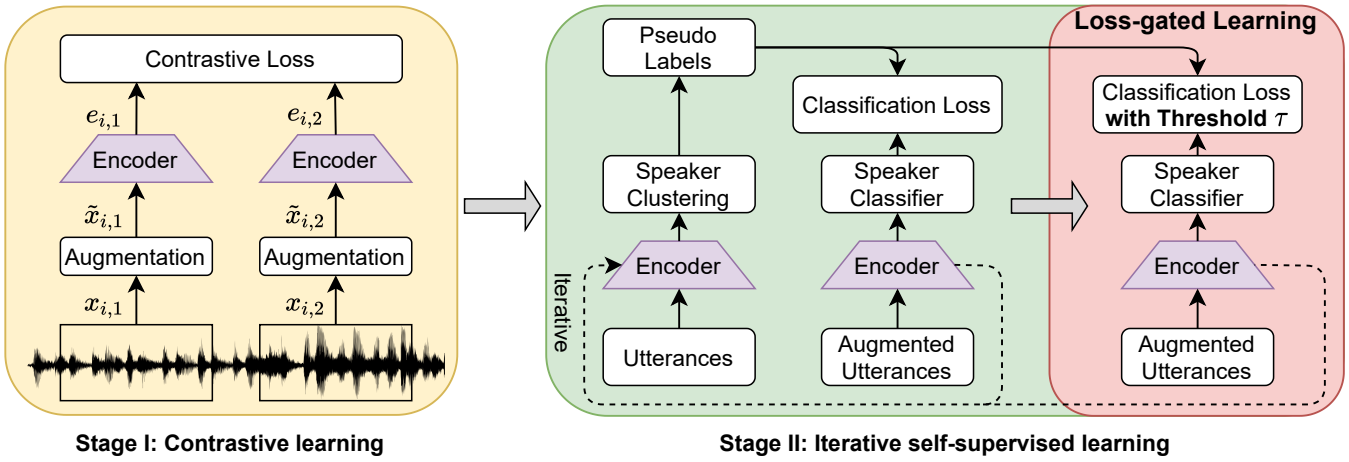

Fig. 1. Framework of self-supervised speaker recognition with loss-gated learning.

2.1. Stage I: Contrastive Learning

In Stage I, we design a self-supervised pretraining task from the *simple contrastive learning* (SCL) [10, 12]. For each mini-batch, we randomly select N unlabelled utterances $x_1, \cdots, x_N$. As shown in Fig. 1, we randomly consider two non-overlapping segments $x_{i,1}$ and $x_{i,2}$ with the same length for each utterance x_i. Then for each these segments, we apply the stochastic noise augmentation to get the augmented segments $\widetilde{x}_{i,1}$ and $\widetilde{x}_{i,2}$. These segments are fed into the speaker encoder $f(\cdot)$ to obtain the speaker embeddings $e_{i,j} = f(\widetilde{x}_{i,j})$, where $i \in \{1, \cdots, N\}$ and $j \in \{1, 2\}$. We assume that each utterance contains only one speaker. The segments drawn from the same utterance share the same speaker identity, and therefore form the positive pairs. On the other hand, we form negative pairs from the segments drawn from the different utterances. In order to attract the positive pairs and repel the negative pairs, we define the contrastive loss for each positive pair against all the negative pairs as [12]:

$$l_{i,j} = -\log \frac{\exp(\cos(e_{i,1}, e_{i,2}))}{\sum_{k=1}^{N}\sum_{l=1}^{2} \mathbb{1}_{\substack{k\neq i \\ j\neq l}} \exp(\cos(e_{i,j}, e_{k,l}))} \tag{1}$$

The loss function for each mini-batch is then given by:

$$L_{\text{scl}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2} l_{i,j} \tag{2}$$

Notice that the function $\cos(\cdot,\cdot)$ denotes the cosine similarity and we do not set the temperature parameter. By minimizing this loss function, the speaker encoder learns the utterance representations that discriminate positive pairs against negative pairs.

2.2. Stage II: Iterative Self-Supervised Learning

Stage II can be viewed from Fig. 1. First, we use the speaker encoder trained in Stage I as the initial model to extract the speaker embeddings for each utterance. Based on these embeddings, the k-means clustering [20] is performed to produce pseudo speaker labels. We interpret that the utterances in the same cluster share the same identity information. Next, we retrain the speaker encoder using these pseudo labels. The classification layer contains one fully connected layer. The *additive angular margin softmax* (AAM-softmax) loss [21] is used as the loss function. We repeat both these steps for several iterations until the system converges. It is noted that the encoder trained for speaker classification in each iteration will be used to generate embedding vectors for clustering in the next iteration.

3. LOSS-GATED LEARNING

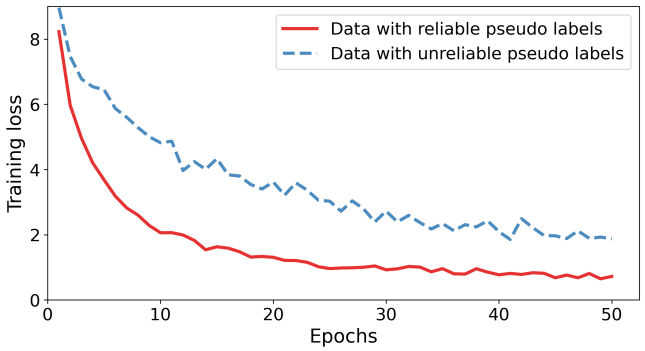

Fig. 2. The training loss on 1000 utterances trained by the pseudo labels.

Firstly, we make the definition of the reliable and unreliable pseudo labels. We can obtain an optimal one-to-one mapping between ground-truth speaker labels and clustered pseudo labels through the Hungarian algorithm [22]. If the ground-truth label of the utterance is the same as the mapped pseudo label, we define this data has the reliable pseudo label.

Otherwise, it has the unreliable pseudo label. It is noted that we do not know whether each data has the reliable or unreliable pseudo labels in our self-supervised speaker recognition experiments since we can not use the ground-truth labels.

Traditionally, the two-stage baseline performs clustering for the next iteration based on the current encoder until the evaluation performance stops improving. However, in this process, the erroneous information from those unreliable pseudo labels will propagate iteratively, which doesn't improve the model in practice [10, 19]. This prompts us to only use the data with reliable pseudo labels for training. In this way, we force the model to learn from data with accurate speaker identity.

Now the question is how to select the reliable pseudo labels effectively. The study in [23] shows that some classes were easier to learn and converge faster than other classes for image classification task. Similarly, the learning ability and convergence speed of data with reliable pseudo speaker labels should be different from those with unreliable labels. To validate this hypothesis, a toy experiment is conducted to compare the training performance of data with reliable and unreliable labels. We evenly select 1,000 utterances from 10 speakers and perform k-means clustering based on the speaker encoder (k=10). The network is then trained to distinguish these speaker labels. It is noted that only in this toy experiment, we use the ground-truth labels to distinguish the reliable and unreliable pseudo labels.

The average training loss curve is shown in Fig. 2. Although no reliability or unreliability information is provided to the network during training, we observe that the network converges much faster for reliable data than the unreliable ones automatically. In addition, there exists a distinct separation between the loss of the data with reliable and unreliable labels in the final converged stage.

From our toy experiment, the model tends to learn useful information from the reliable labels first and then extracts misleading information from unreliable labels. As shown in Fig. 1, after the pseudo label supervised learning in Stage II, we propose to select reliable data via continuing to train the network with LGL, instead of entering the next iteration directly. The LGL trains the encoder and classifier using the same set of pseudo labels, as in the original supervised learning, with a slight modification to the loss function. Specifically, we introduce a threshold τ into the loss function as follows:

$$L_{spk} = \sum_{i=1}^{N} l_i \mathbb{1}_{l_i<\tau} \tag{3}$$

where l_i is the classification loss for a data point. The assumption is that, data with a small loss is more likely to be more reliable compared to these with larger losses. Thus, LGL only uses the data with a small loss to update the parameters of the network. We train the encoder together with the classification layer until the system performs the best. Then we utilize the trained encoder to do the clustering for the next iteration.

4. EXPERIMENTS

4.1. Dataset

We use the development set of VoxCeleb2 [24] for training. Original, Extended and Hard VoxCeleb1 test sets (Vox1_O, Vox1_E and Vox1_H) [24, 25] are used for evaluation. No speaker label information is employed. The performance metric for the studies is the equal error rate (EER).

4.2. Network Architecture and Feature Extraction

In our experiments, the emphasized channel attention, propagation and aggregation in time-delay neural network (ECAPA-TDNN) in [6] is used as a speaker encoder. The channel size is set at 512. The input is the 80-dimensional log mel-spectrogram from the speech segments. On the other hand, the output is the 192-dimensional speaker embedding.

4.3. Implementation

During the training process, the mini-batch size is set at 256. The network parameters are optimized by Adam optimizer. The MUSAN [26] and RIRs [27] datasets are used for data augmentation. The MUSAN dataset includes ambient noise, music and babble noise, while the RIRs dataset contains the pre-computed room impulse responses. At the test stage, the cosine similarity score between the given utterance pairs is calculated.

4.3.1. Stage I: Contrastive Learning

In the contrastive learning stage, the discriminator training in [17] is added to build a robust encoder. The initial learning rate is 0.001 that decreases 5% in every 5 epochs. The duration of the input utterance is 1.8 seconds.

4.3.2. Stage II: Iterative Self-supervised Learning

For clustering, we employ the k-means algorithms without data augmentation by faiss library[1] [18]. We follow the previous work and set the number of the clusters as 6,000 [10]. For training the network with pseudo labels, we set the margin as 0.2 and the scale as 30 in the AAM-softmax loss and fix the learning rate as 0.001. In addition, the duration of the input utterance is 3 seconds. In LGL, we set the hyper-parameter τ as $\{1, 3, 3, 5, 6\}$ in the five iterations to guarantee that the number of selected data increases by at least 10% in each iteration. In the last iteration, we extend the channel size of the speaker encoder to 1024 to build a robust system.

[1] https://github.com/facebookresearch/faiss

Table 1. Performance of self-supervised speaker recognition with and without LGL in EER (%). A comparison to other existing works is shown as well.

Stage	Method	Vox1_O	Vox1_E	Vox1_H
-	Fully Supervised	1.21	1.42	2.64
I	Nagrani et al. [8]	22.09	-	-
	Chung et al. [9]	17.52	-	-
	Inoue et al. [28]	15.26	-	-
	Huh et al. [17]	8.65	-	-
	Zhang et al. [16]	8.28	-	-
	Xia et al. [14]	8.23	-	-
	Mun et al. [15]	8.01	-	-
	Ours	7.36	7.90	12.32
II	Cai et al. [10]	3.45	4.02	6.57
	Ours w/o LGL	3.09	3.81	6.32
	Thienpondt et al. [11]	2.10	-	-
	Ours with LGL	**1.66**	**2.18**	**3.76**

Table 2. Impact of LGL on performance in EER (%) and comparison to baseline on Vox1_O set.

Stage	Model	Ours w/o LGL	Ours with LGL
I	SCL	7.36	
II	Iteration-1	4.92	3.52
	Iteration-2	4.00	2.41
	Iteration-3	3.68	2.07
	Iteration-4	3.22	1.95
	Iteration-5	3.09	**1.66**

5. RESULTS

5.1. Proposed System Comparison to Existing Works

We now compare the proposed framework with the existing methods in Table 1. On Vox1_O set, we implement the two-stage architecture that achieves an EER of 7.36% and 3.09% in Stage I and Stage II, respectively. This proves the robustness of our baseline. Meanwhile, our proposed self-supervised speaker recognition system with LGL obtains an EER of 1.66% on Vox1_O set, which outperforms the best existing method [11] by 20.95%. It also outperforms the previous work [10] by 45.77% and 42.77% on Vox1_E and Vox1_H sets.

5.2. Impact of LGL: Iterative Analysis

We summarize the performance in each iteration with and without LGL on Vox1_O set in Table 2. From the table, it can be observed that the LGL can quickly promote the system performance in each iteration and finally brings 46.3% improvement compared with the baseline. In addition, the LGL is found to be more effective in the beginning since there are more unreliable data.

5.3. Robustness to Threshold of LGL

Then we study the robustness to threshold τ of LGL in the Iteration-1 in Table 3. Compared with the baseline method without threshold ($\tau = +\infty$), all thresholds setting from 1 to 5 can bring significant improvement on Vox1_O set. That proves LGL is relatively robust to this hyperparameter.

Table 3. Robustness to threshold in the Iteration-1 on Vox1_O set.

Threshold (τ)	1	2	3	4	5	$+\infty$
EER	**3.52**	3.77	3.74	4.10	4.14	4.92

5.4. Post-analysis of clustering

Now we evaluate the clustering performance by normalized mutual information (NMI). The higher NMI indicates the better clustering result. First, we compare the NMI in each iteration with and without LGL in the Fig. 3 (a), LGL improves NMI by a large margin and leads to a better clustering result. Then for self-supervised speaker recognition system with LGL, in the Fig. 3 (b), we compare the NMI of all data and the data selected by threshold only. The selected data has the very high NMI, proving that our LGL can successfully filter the reliable pseudo labels. It is noted that we do this analysis after we finish all the experiments and do not use it to guide the self-supervised training process.

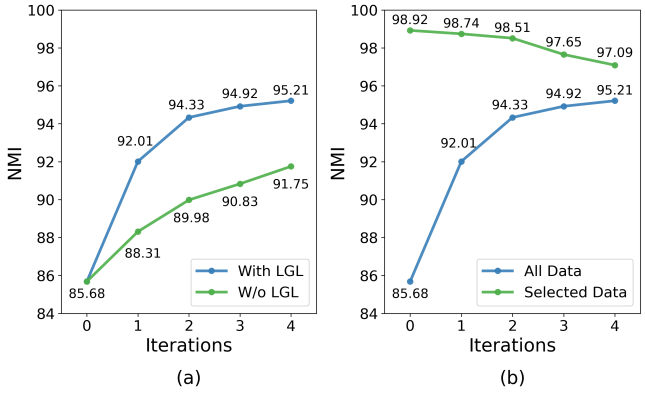

Fig. 3. (a) NMI for the system with and without LGL. (b) NMI of the selected data and all data for the system with LGL.

6. CONCLUSION

In this paper, we propose an effective LGL to select reliable pseudo labels for self-supervised speaker recognition. LGL utilizes neural network's propensity to extract reliable data automatically. The experiments on VoxCeleb datasets show that the proposed approach can improve the baseline model performance in EER by 46.3%. This showcases the importance of LGL to use abundant unlabelled data for speaker recognition studies.

7. REFERENCES

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

[2] Kong Aik Lee, Anthony Larcher, Helen Thai, Bin Ma, and Haizhou Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *Interspeech*, 2011, pp. 3317–3318.

[3] R. K. Das and S. R. M. Prasanna, "Investigating text-independent speaker verification from practically realizable system perspective," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1483–1487.

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

[5] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020, pp. 2977–2981.

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020, pp. 1–5.

[7] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

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

[9] Soo-Whan Chung, Hong-Goo Kang, and Joon Son Chung, "Seeing voices and hearing voices: Learning discriminative embeddings using cross-modal self-supervision," in *Interspeech*, 2020, pp. 3486–3490.

[10] Danwei Cai, Weiqing Wang, and Ming Li, "An iterative framework for self-supervised deep speaker representation learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6728–6732.

[11] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The IDLAB VoxCeleb Speaker Recognition Challenge 2020 system description," *arXiv preprint arXiv:2010.12468*, 2020.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning (ICML)*, 2020, pp. 1597–1607.

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

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

[15] 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.

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

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

[18] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, "Deep clustering for unsupervised learning of visual features," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[19] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[20] Stuart Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[21] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[22] Harold W Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[23] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 322–330.

[24] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

[26] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[27] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[28] Nakamasa Inoue and Keita Goto, "Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 1641–1646.