ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096925

SELF-SUPERVISED AUDIO-VISUAL SPEAKER REPRESENTATION WITH CO-META LEARNING

Hui Chen[1,2], *Hanyi Zhang*[2], *Longbiao Wang*[2,*], *Kong Aik Lee*[3,*], *Meng Liu*[2], *Jianwu Dang*[2]

[1]Tianjin International Engineering Institute, Tianjin University, Tianjin, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

In self-supervised speaker verification, the quality of pseudo labels determines the upper bound of its performance and it is not uncommon to end up with massive amount of unreliable pseudo labels. We observe that the complementary information in different modalities ensures a robust supervisory signal for audio and visual representation learning. This motivates us to propose an audio-visual self-supervised learning framework named Co-Meta Learning. Inspired by the *Co-teaching+*, we design a strategy that allows the information of two modalities to be coordinated through the *Update by Disagreement*. Moreover, we use the idea of *model-agnostic meta learning* (MAML) to update the network parameters, which makes the hard samples of two modalities to be better resolved by the other modality through gradient regularization. Compared to the baseline, our proposed method achieves a 29.8%, 11.7% and 12.9% relative improvement on Vox-O, Vox-E and Vox-H trials of Voxceleb1 evaluation dataset respectively.

***Index Terms*—** self-supervised learning, speaker verification, audio-visual data, co-teaching+, meta-learning

1. INTRODUCTION

Deep-learning methods have been broadly applied for speaker verification (SV) task and obtained excellent performance [1–5]. However, all these methods usually require large amounts of data with speaker labels, while the creation of data set with annotated speaker labels is very difficult and expensive.

In self-supervised learning, the many previous efforts [6–9] have been made to obtain great performance with these large amounts of unlabeled data. The iterative framework [10, 11], current state-of-the-art self-supervised speaker verification paradigm includes two stages. In stage I, a speaker encoder is pretrained by contrastive learning, such as SimCLR [12]. In stage II, pseudo labels are estimated using the pre-trained model and then a new model is trained based on the estimated pseudo labels. Stage II is typically repeated for multiple iterations to continuously improve the performance.

This two-stage iterative training framework has achieved excellent performance. Though there still exist some problems, such as the pseudo labels. One problem with that is label noise which will confuse and degrade the generalization performance of the model [13].

While the evaluation phase only allows audio data, multi-modal audio-visual data can be used for training. The use of another modality could be beneficial for representation learning since different modalities contain complementary information. Cai *et al*. [14] applied the self-supervised learning framework to multi-modal audio-visual data. Although this approach brings further improvement, there is still room for improvement. The mutual information is only used when labels are obtained at the clustering stage in this approach, which does not fully utilize the multi-modal mutual information.

To tackle these problems and to improve the existing methods, in this work, we present a self-supervised learning framework with Co-Meta learning. More specifically, inspired by Co-teaching+ [15] which was proposed to deal with noisy label in supervised learning, we use the divergence strategy so that the networks can select samples to learn the complementary information during the training stage. Further, in order to avoid the models being inclined to focus on hard samples that are inconsistent between the two modalities while ignoring the samples which perform poorly on both modalities, we introduce the gradient regularization inspired by meta-learning [16] which is the core idea of this paper. The two models can make full use of the difficult samples with the other modality which contribute towards an improved performance.

2. TWO-STAGE ARCHITECTURE

We now describe our two-stage baseline architecture, which is similar to the framework in [14]. A contrastive pretraining task is employed to learn meaningful speaker embeddings

* Corresponding author

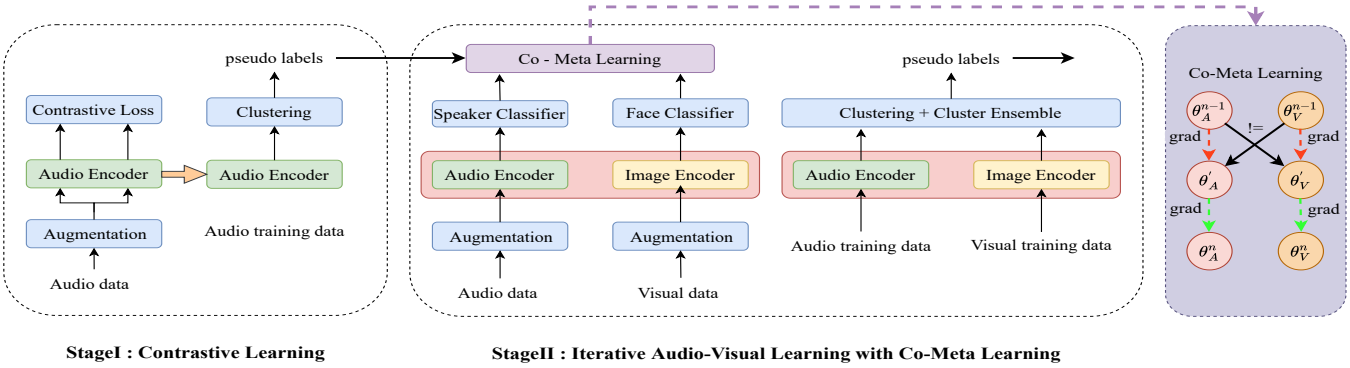

Fig. 1. Architecture of the proposed audio-visual self-supervised network with Co-Meta learning.

with only audio data. Subsequently, pseudo labels of all training utterances are found by clustering the embedding vectors of these utterances. In the iterative stage, the audio and visual representation encoders are employed to learn representations of the corresponding modality, and a cluster ensemble algorithm is then used to fuse the clustering results of the two modalities. Details of stage I and II are described below.

2.1. Stage I: Contrastive Learning

We design the contrastive self-supervised learning (CSL) framework similar to the *SimCLR* in [11, 12]. We randomly select N unlabelled data samples $x_1, x_2, ..., x_N$ for each mini-batch. We assume that each data sample defines its own speaker class and performs utterances discrimination. For each utterance x_i, we randomly select two non-overlapping segments with the same length before data augmentation. Then we apply the stochastic noise augmentation to get the correlated views $\tilde{x}_{i,1}$ and $\tilde{x}_{i,2}$. During training, a neural network encoder $f(\cdot)$ extracts the speaker embedding $e_{i,j} = f(\tilde{x}_{i,j})$, where $i \in \{1, 2, ..., N\}$ and $j \in \{1, 2\}$. The contrastive loss for each positive pair against all the negative pair is defined, as follows:

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

where $\mathbb{1}$ is an indicator function evaluating to 1 when $k \neq i$ and $l \neq j$, cos denotes the cosine similarity and we do not set the temperature, the total loss for each mini-batch is given by:

$$L_{CSL} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2} l_{i,j} \quad (2)$$

The encoder obtains the audio representations at the utterance level; thus, the learned representation is expected to contain considerable attributes related to the speaker identity.

2.2. Stage II: Iterative Audio-Visual Learning

First, we use the speaker encoder pre-trained with the CSL loss to extract the speaker embeddings for each utterance. Then, we employ a clustering algorithm to generate cluster assignments and pseduo labels. In this baseline, we use the k-means clustering [17] in view of its simplicity and capability to handle a large dataset.

Next, we train the audio and visual encoders using these pseudo labels. Through a multi-modal dataset $\mathcal{D}$ with audio-modality $\mathcal{D}_a = \{x_{a,1}, x_{a,2}, ..., x_{a,N}\}$ and visual-modality $\mathcal{D}_v = \{x_{v,1}, x_{v,2}, ..., x_{v,N}\}$, encoders are trained independent for each modality. For each sample x_i, its audio component $x_{a,i}$ and visual component $x_{v,i}$ share the same pseudo label. The audio encoder $f_a(\cdot)$ and the visual encoder $f_v(\cdot)$ are discriminatively trained with an audio classifier and a visual classifier using the *additive angular margin softmax* (AAM-softmax) loss [18]. Then, clustering on the audio representation (i.e., speaker embedding) and the visual representation (i.e., face embedding) gives audio pseudo labels and visual pseudo labels respectively. Since the complementary information from different modalities, we apply a clustering on the joint representations to obtain the more robust pseudo labels [14]. We choose the joint pseudo labels as the reference clustering output and calculate cluster correspondence for the audio and visual pseudo labels. The consistent label set is obtained after the re-labeling process. Majority voting [19, 20] is then employed to determine consensus of a pseudo label for each sample. We repeat both training and clustering steps for multiple iterations until the system converges.

3. CO-META LEARNING

The two-stage audio-visual baseline described in Section 2 only uses the complementary information in the clustering stage. Furthermore, the encoders for each modality are trained independent, which don't use the information to improve the models in the training stage. This prompts us to let the models transfer their knowledge between modalities in

the training stage. Now, the question is how to strengthen the connection of the two modalities. To this end, we adopt the Co-teaching+ [15] and propose a gradient correction method which forms the core idea this paper. The full algorithm is outlined in Algorithm 1, and illustrated further in Fig.1. Note that Co-teaching+ [15] was proposed to deal with noisy label problem in supervised learning context. Here, we use it to deal with noisy pseudo labels in self-supervised learning.

3.1. Co-teaching+: Training of multi-modal networks with noisy labels

To handle noisy labels and connect two modalities, the audio model and visual model feed forward and predict the audio component and visual component respectively of the same mini-batch. Then, they select prediction disagreement instances according to their predictions $\left\{\bar{y_1}^{(A)}, ..., \bar{y}_{\frac{|\bar{D}|}{N}}^{(A)}\right\}$ (predicted by audio model) and $\left\{\bar{y_1}^{(V)}, ..., \bar{y}_{\frac{|\bar{D}|}{N}}^{(V)}\right\}$ (predicted by visual model) as follow:

$$\bar{D}' = \left\{\left(x_i, y_i : \bar{y_i}^{(A)} \neq \bar{y_i}^{(V)}\right)\right\} \tag{3}$$

The operation corresponds to line 7 in Algorithm 1, which makes the two classifiers keep diverged. In lines 8 and 9, from the disagreement data subset $\bar{D}'$, each network selects its own small-loss data $\bar{D}'^{(A)}$ (resp. $\bar{D}'^{(V)}$), but back propagates the small loss data $\bar{D}'^{(A)}$ and $\bar{D}'^{(V)}$ corresponding to visual and audio data points to its peer network respectively and updates itself parameters after the gradient regularization of meta-learning (Section 3.2).

3.2. Gradient Regularization

The goal of meta-learning is to train a model on a variety of learning tasks, such that it can solve new learning tasks using only a small number of training samples [16]. In our approach, we introduce the meta-learning to learn the transferable meta-knowledge between audio data and visual data. Similar to MAML, we use all the samples in the current mini-batch as the support set, while the instances selected by divergence strategy as the query set. We use the support set to calculate the gradient of each parameter and complete the first gradient update. These operations correspond to lines 5 and 6 in Algorithm 1. Using the parameters obtained from this update, the query set completes the second gradient update. These correspond to lines 10 and 11 in Algorithm 1. Note that the gradient is taken with respect to the loss function with the updated parameters in lines 5 and 6. Then the obtained result is used to update its parameters. So that, the learning ability of the model on its difficult samples can be improved by another modality.

Algorithm 1: Framework with Co-Meta Learning.

Input: θ_A and θ_v, training set $\mathcal{D}$, mini-batches N, learning rate α,β, estimated noise rate τ, epoch E_k and E_{max};

1 **for** e $= 1, 2, ..., E_{max}$ **do**
2 **Shuffle** $\mathcal{D}$ into N mini-batches;
3 **for** $n = 1, ..., N$ **do**
4 **Fetch** n-th mini-batch $\bar{D}$ from $\mathcal{D}$;
5 **Get** $\theta_A^{'} = \theta_A^{n-1} - \alpha \bigtriangledown \ell\left(\bar{D}; \theta_A^{n-1}\right)$;
6 **Get** $\theta_V^{'} = \theta_V^{n-1} - \alpha \bigtriangledown \ell\left(\bar{D}; \theta_V^{n-1}\right)$;
7 **Select** prediction disagreement $\bar{D}'$ by Eq.(3);
8 $\bar{D}'^{(A)} = argmin_{D':\left|D'\right| \geq \lambda(e)\left|\bar{D}'\right|} \ell\left(D'^{(A)}; \theta_A^{n-1}\right)$;
9 $\bar{D}'^{(V)} = argmin_{D':\left|D'\right| \geq \lambda(e)\left|\bar{D}'\right|} \ell\left(D'^{(V)}; \theta_V^{n-1}\right)$;
10 **Update** $\theta_A^n = \theta_A^{n-1} - \beta \bigtriangledown_{\theta_A^{n-1}} \ell\left(\bar{D}'^{(V)}; \theta_A^{'}\right)$;
11 **Update** $\theta_V^n = \theta_V^{n-1} - \beta \bigtriangledown_{\theta_V^{n-1}} \ell\left(\bar{D}'^{(A)}; \theta_V^{'}\right)$;
12 **end**
13 **Update** $\lambda(e) = 1 - \min\left\{\frac{e}{E_k}\tau, \tau\right\}$;
14 **if** $e \bmod 10 = 0$ **then**
15 Update pseudo labels by clustering algorithm;
16 **end**
17 **end**

4. EXPERIMENTS

4.1. Dataset

We use the audio-visual Voxceleb2 [21] dataset for model training. Since not every audio file has a corresponding video file in the official version released online, we use a subset which contains both modalities. For each video file, we extract face image as visual data at one frame per second. Labels of two modalities are not used in the training stage.

For evalution, the experiments are conducted on the Original, Extended and Hard Voxceleb1 test sets [21, 22]. Visual verification evaluation is also using the test dataset. The face images extracted at one fps are downloaded from the official website[1].

4.2. Data Augmentation

Augmentation for audio data: We use the MUSAN [23] and RIRs [24] datasets, MUSAN dataset includes ambient noise, music and babble noise, and the RIRs dataset contains the pre-computed room impulse responses.

Augmentation for visual data: Similar to [14], we also do a random crop after resizing to 3 x 224 x 224, and then sequentially apply these augmentations including random horizontal flipping, random color distortions, random grey scaling, and random Gaussian blur for the images. Note that the augmentation is performed at a probability of 0.6.

[1]https:///www.robots.ox.ac.uk/vgg/research/ CMBiometrics

Table 1. *EER (%) comparison on Vox1_O test set for each iteration of the proposed Co-Meta learning and our baseline.*

Stages	Models	Baseline		with Co-teaching+		with Co-meta	
		Audio	Visual	Audio	Visual	Audio	Visual
Stage1	Contrastive Learning	7.16					
Stage2	Round 1	4.98	6.10	4.17	5.99	3.97	5.94
	Round 2	3.53	4.02	3.71	4.01	3.12	3.78
	Round 3	2.06	2.32	2.44	2.67	2.01	2.19
	Round 4	1.99	1.89	1.89	1.78	1.67	1.59
	Round 5	1.88	1.87	1.79	1.80	1.45	1.38
	Round 6	1.85	-	1.77	-	**1.27**	-

4.3. Network Architecture

The speaker encoder is an ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in Time-Delay Neural Network) network [3], with a channel size of 512. The output is the 192-dimensional speaker embedding. The setup for the audio encoder in Stage II are the same as Stage I. And the standard ResNet-34 in [25] is used as a visual encoder. A fully connected layer is added between the pooling layer and the final classification layer to produce a 192-dimensional embedding.

4.4. Implementation Details

4.4.1. Contrastive Learning

We add the discriminator training [6] in the CSL framework to improve the robustness. The duration, which is randomly generated for the utterance, is 1.8 seconds. 80-dimensional Log mel-spectrogram is used as the feature for the audio segments. The network parameters are optimized with the Adam optimizer [26] with an initial learning rate at 0.001.

4.4.2. Iterative Learning

We use the same clustering algorithm in [14], and set the number of clusters to 6000. When training the audio-visual network on pseudo labels with Co-Meta learning, we set the margin as 0.2 and the scale as 32. For Co-Meta learning, we set the estimated noise rate τ as 0.2, E_k as 10 and E_{max} as 200. And we set the initial learning rate α as 0.01 and learning rate β as 0.001 for the Adam optimizer.

5. RESULTS AND ANALYSIS

5.1. Impact of Co-Meta

We summarize the performance in each iteration of our baseline, and test Co-teaching+ and Co-Meta learning on the Vox1_O test set in Table 1. From the table, it can be observed that the Co-teaching+ strategy can slightly improve the performance. While the Co-teaching+ strategy may make the model more inclined to discover the hard samples of the two modalities, it ignores many samples which are equally bad on both modalities. With gradient regularization, the networks achieve a balance between these two issues. Finally, the system brings 31.4% improvement compared with our baseline.

5.2. Proposed Method Comparison to Existing Works

Table 2 shows a comparison between the proposed framework with the existing methods. The two-stage framework with Co-Meta learning achieves an EER of 1.27% as final result on Vox1_O set, which outperforms the best existing method [27] by 13.6%. It also outperforms the previous work [14] by 29.8%, 11.7% and 12.9% on Vox1_O, Vox1_E and Vox1_H sets.

Table 2. *EER(%) comparison for different self-supervised speaker verification systems on Vox1-O, Vox1-E and Vox1-H sets.*

Method	Clustering	Iter	Vox1-O	Vox1-E	Vox1-H
ID [11]	AHC(7500)	7	2.10	-	-
JHU [28]	AHC(7500)	5	1.89	-	-
SNU [29]	K-M(6000)	4	1.66	-	-
DKU [14]	AHC(7500)	5	1.81	2.06	3.80
LGL [9]	K-M(6000)	5	1.66	2.18	3.76
DLG-LC [27]	K-M(6000)	5	1.47	**1.78**	**3.19**
OURS	K-M(6000)	7	**1.27**	1.82	3.31

6. CONCLUSIONS

In this study, we proposed an effective Co-Meta learning approach to better utilize the audio-visual complementary information for self-supervised speaker verification. Co-Meta learning utilizes the *update by disagreement* and *gradient regularization* to strengthen the connection between the two modalities. Experiments results on Vox1-O set show that the proposed method can improve the performance in terms of EER by 31.4% compared to the baseline. Ours results on Vox1-E and Vox1-H are also comparable with the best results reported in the literature. In the future, we will explore the combination of meta learning and other methods to improve the performance.

7. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China under Grant 62176182. The work of Kong Aik Lee is supported by the Agency for Science, Technology and Research (A*STAR), Singapore, through its Council Research Fund (Project No. CR-2021-005).

8. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[2] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1385–1389, 2021.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] Pooyan Safari, Miquel India, and Javier Hernando, "Self-attention encoding and pooling for speaker recognition," *arXiv preprint arXiv:2008.01077*, 2020.

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

[6] 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.

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

[8] Themos Stafylakis, Johan Rohdin, Oldrich Plchot, Petr Mizera, and Lukas Burget, "Self-supervised speaker embeddings," *arXiv preprint arXiv:1904.03486*, 2019.

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

[10] Danwei Cai, Weiqing Wang, and Ming Li, "An iterative framework for self-supervised deep speaker representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 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*. PMLR, 2020, pp. 1597–1607.

[13] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[14] Danwei Cai, Weiqing Wang, and Ming Li, "Incorporating visual information in audio based self-supervised speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1422–1435, 2022.

[15] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama, "How does disagreement help generalization against label corruption?," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

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

[18] 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.

[19] Eric Bauer and Ron Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine learning*, vol. 36, no. 1, pp. 105–139, 1999.

[20] Louisa Lam and SY Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 27, no. 5, pp. 553–568, 1997.

[21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

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

[24] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] Bing Han, Zhengyang Chen, and Yanmin Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," *arXiv preprint arXiv:2208.01928*, 2022.

[28] Jejin Cho, Jesus Villalba, and Najim Dehak, "The jhu submission to voxsrc-21: Track 3," *arXiv preprint arXiv:2109.13425*, 2021.

[29] Sung Hwan Mun, Min Hyun Han, and Nam Soo Kim, "Snu-hil system for the voxceleb speaker recognition challenge 2021," .