SPEAKER RECOGNITION WITH TWO-STEP MULTI-MODAL DEEP CLEANSING

Ruijie Tao[1], *Kong Aik Lee*[2], *Zhan Shi*[3] *and Haizhou Li*[1,3]

[1]College of Design and Engineering, National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Neural network-based speaker recognition has achieved significant improvement in recent years. A robust speaker representation learns meaningful knowledge from both hard and easy samples in the training set to achieve good performance. However, noisy samples (i.e., with wrong labels) in the training set induce confusion and cause the network to learn an incorrect representation. In this paper, we propose a two-step audio-visual deep cleansing framework to eliminate the effect of noisy labels in speaker representation learning. This framework contains a coarse-grained cleansing step to search for the peculiar samples, followed by a fine-grained cleansing step to filter out the noisy labels. Our study starts from an efficient audio-visual speaker recognition system, which achieves a close to perfect equal-error-rate (EER) of 0.16%, 0.23% and 0.42% on the Vox-O, E and H test sets. With the proposed multi-modal cleansing mechanism, four different speaker recognition networks achieve an average improvement of 5.9%. Code has been made available at: https://github.com/TaoRuijie/AVCleanse.

Index Terms— speaker recognition, noisy label, audio-visual, deep cleansing

1. INTRODUCTION

Automatic speaker recognition aims to distinguish a legitimate user from imposters based on their voices [1, 2]. Over the last decade, deep learning-based speaker representation, such as x-vector [3], xi-vector [4], and network architectures, such as ResNet [5] and ECAPA-TDNN [6], have achieved remarkable performance by training on the large-scale speech dataset. To further enhance the performance, previous works usually focus on the network structure [7], loss function [8] and back-end score calibration [9]. However, the existence of noisy labels in the training set has not been taken care of.

Large-scale dataset for speaker recognition typically consists of thousands of speakers and over millions of samples, whereby utterances from the same person share the same speaker labels. These datasets are typically collected from the Internet within automatic pipeline [10, 11]. However, it is not surprising to find that utterances assigned with the same label might actually come from different speakers, i.e., the noisy labels problem. From the study of noisy-label learning, training data can be divided into three categories according to the learning difficulty [12, 13]: easy, hard and noisy samples. Note that we use the term noisy samples referring to samples with wrong labels following the terminology in [10]. Due to the memorization effects [14], neural networks tend to fit the easy samples and converge rapidly in the early stage of training. In the later stage of training, the network has to distil correct but difficult knowledge from the peculiar samples. However, these peculiar samples contain hard and noisy samples, which deters the learning process.

Prior works dealing with noisy labels in speaker recognition include: training the speaker network and manually setting the threshold to filter out the unusual utterances [15]; dropping out the samples with large training loss as the noisy data in self-supervised learning [16]. However, there are two major problems with the existing approach. In terms of quality, it is challenging for the speaker network to recognize the difficult utterances [17]. In terms of logic, the speaker network has been trained on the entire dataset (including noisy ones) and push them to their class centre. Using the same network to decide the noisy ones is not a good choice.

In this paper, we propose a two-step multi-modal deep cleansing framework to solve the above mentioned problems. Firstly, we do a coarse-grained cleansing based on the speech modality only, which divides the training data into easy and peculiar samples. Secondly, we train a new speaker recognition network based on easy samples to do a fine-grained cleansing, which divides the peculiar samples into hard and noisy ones. Since the network does not train on the noisy samples, using it to do the cleansing is more reliable. On the other hand, from the biometric recognition study [18, 19], face image and speech utterance can provide complementary identity information. Motivated by this, we use the face recognition network in the second step to boost the system.

Our contribution can be summarized as follows. Firstly, we propose a robust audio-visual speaker recognition system which can achieve close-to-perfect verification. Secondly, a two-step audio-visual cleansing framework is designed to filter out the noisy data in the training set. Thirdly, four speaker recognition networks are trained on the original and cleansed dataset to show the impact of our method.

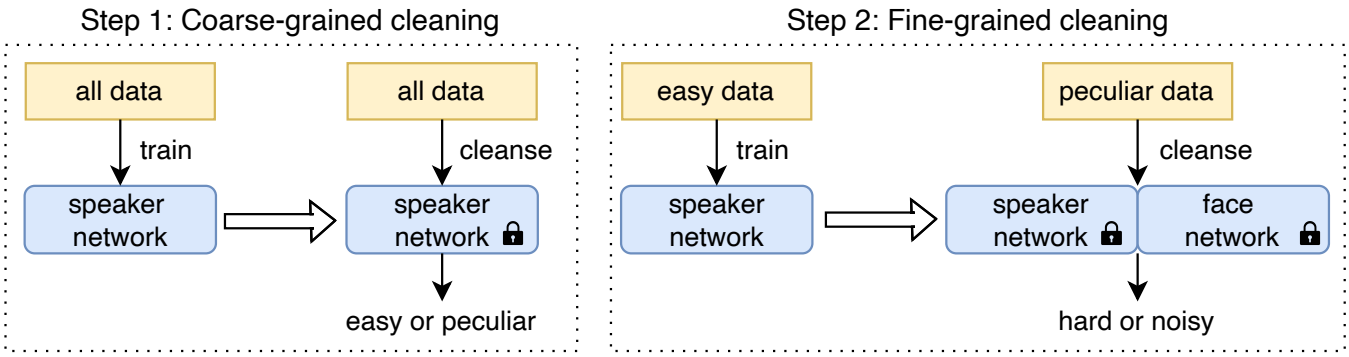

Fig. 1: The proposed two-step audio-visual deep cleansing framework. The lock represents that the network is frozen.

2. AUDIO-VISUAL SPEAKER RECOGNITION

In this section, we propose an efficient audio-visual speaker recognition system containing both speaker and face modalities since a reliable identity recognition system is the foundation of noisy sample cleansing.

For training, the speaker network is used to extract the speaker embedding from the input utterance. This embedding contains the characteristics of the speaker's voice. Then it follows a speaker classifier to distinguish the utterances from different speakers with an AAM-softmax loss [20]. Similarly, the face network takes one face image as the input and outputs the face embedding. This embedding is trained with a face classifier [21]. For testing, speaker and face embeddings are used together to enhance the verification performance.

Compared with previous works [22, 23], we take two strategies to simplify the system and improve the stability. Firstly, we align all the faces with the detected face landmarks during preprocessing [24] since the unaligned faces in the training set make recognition harder [25]. Secondly, existing approaches usually attempt to combine the speaker and face modality by the attention mechanism [18, 26]. However, we argue that training two modalities separately and directly concatenating two embeddings for testing is an effective and convenient solution. [19, 23]

3. AUDIO-VISUAL DEEP CLEANSING

We consider a training set with N video clips. Each video clip consists of synchronized speech utterance and face frames. And, there exists a few video clips with wrong labels, i.e., noisy samples. Our proposed audio-visual two-step deep cleansing framework is shown in Fig 1 to discover these noisy samples.

3.1. Step 1: Coarse-grained cleansing

Firstly, we design coarse-grained cleansing to divide the training set into easy and peculiar samples to narrow down the selection of noisy data. Our target is successfully dividing all the noisy samples into the peculiar class. It is noted that this step uses speech modality only since face modality can only assist in deciding the correctness of samples.

Firstly, we train a speaker network with all the utterances in the training set. Then we fix this network's parameters and extract the speaker embeddings from the entire training set. These embeddings are represented as $s_1, \cdots, s_N$. The corresponding speaker labels are annotated as $c_1, \cdots, c_N, c_i \in \{1, 2, \cdots, K\}$, where K is the number of speakers in the training set. Then we compute the average cosine similarity between the speaker embedding of each sample and all samples from the same speaker. The score x_i of the embedding s_i can be represented as:

$$x_i = \frac{1}{M_k} \sum_{j=1}^{N} \mathbb{1}_{c_i=c_j} cos(s_i, s_j) \tag{1}$$

Here M_k is the number of samples in the kth class. $\mathbb{1}$ is an indicator function evaluating 1 when $c_i = c_j$. For the computed scores, we set a threshold τ. Samples with scores smaller than τ are the easy samples; otherwise, they are deemed to be peculiar samples.

3.2. Step 2: Fine-grained cleansing

As mentioned earlier, the network without training on the noisy samples can provide an objective and accurate representation. Motivated by that, we train the network without noisy samples in fine-grained cleansing to filter out noisy samples.

3.2.1. Face modality

Firstly, we train a new speaker network with the easy samples found from the coarse-grained cleansing. Then considering sufficient annotated face datasets, the pre-trained face recognition is applied to our multi-modal cleansing proposal. It is noted that training a face network with the images of the found easy samples is also a reasonable solution.

3.2.2. Decision boundary

Then we train a classifier in the two-dimensional score space to separate noisy samples from the hard samples. The binary classifier is trained on a validation set consisting of target trials [10,11], i.e., two video clips from the same speakers, and the imposter trials, i.e., two video clips are from different speakers. We can compute the speaker and face cosine similarity for each trial using our multi-modal system. Then an SVM [27] is learnt on these two-dimensional scores and using the ground-truth labels in the validation. The SVM decision boundary can efficiently distinguish trials into the target and the imposter ones. Target and imposter trials are associated with the hard and noisy samples, respectively.

3.2.3. Deep cleansing

In this stage, we freeze the networks to extract the speaker and face embeddings of all the samples in the training set. Speaker and face embeddings are represented as $s_1, \cdots, s_N$ and $f_1, \cdots, f_N$, respectively. We compute the average cosine similarity between the embedding of each sample and all the samples from the same class. The speaker scores x_i of the embedding s_i is similar to that in (1). The face score y_i of the embedding f_i is computed as:

$$y_i = \frac{1}{M_k} \sum_{j=1}^{N} \mathbb{1}_{c_i=c_j} cos(f_i, f_j) \tag{2}$$

Finally, we apply the learnt SVM to predict the correctness of each training sample using x_i and y_i. The samples predicted as the target trials are defined as clean data; otherwise, they are noisy data.

4. EXPERIMENTAL SETUP

Our training set is the VoxCeleb2 [11], an audio-visual dataset derived from YouTube interviews. It contains 1,091,724 video clips from 5,994 speakers, and each video clip has only one synchronized and visible talking face.

For audio-visual speaker recognition, the original VoxCeleb2 is used for training. Here we also report the performance of the face network pre-trained on the large face dataset Glint360K [28] to show the effectiveness of face modality. The speaker network is the ECAPA-TDNN with a large channel size equal to 1024 (refer to ECAPA-L) [6]. The face network is the ResNet18 (training on VoxCeleb2) and ResNet50 (training on Glint360K) [20, 29]. We select Vox1-O set for validation, Vox1-E and Vox1-H for testing [10]. All training and test samples contain synchronized audio and visual data.

For audio-visual deep cleansing, the speaker network is an ECAPA-L, and the face network is a ResNet50 network. For fine-grained cleansing, we set the threshold τ to divide 92% of training data as the easy samples. Here we only need to ensure that no noisy samples are involved [15]. For coarse-grained cleansing, the face network is trained on Glint360K [28]. Only the cleansed samples are used to decide the class centre here, so we compute each sample's similarity score and find clean samples with five rounds.

Table 1: The EER (%) of audio-visual speaker recognition. '-Vox2' denotes training on VoxCeleb2 dataset, '-Glint' denotes training on Glint360K dataset.

Modality	System	Vox1-O	Vox1-E	Vox1-H
Speech	Sari et al. [23]	2.20	-	-
	Qian et al. [18]	1.62	1.75	3.16
	Chen et al. [26]	2.31	2.23	3.78
	(1) Ours-Vox2	**1.02**	**1.23**	**2.36**
Face	Sari et al. [23]	3.90	-	-
	Qian et al. [18]	3.04	2.18	4.23
	Chen et al. [26]	2.26	1.54	2.37
	(2) Ours-Vox2	**0.97**	**0.81**	**1.16**
	(3) Ours-Glint	**0.03**	**0.07**	**0.09**
Fusion	Sari et al. [23]	0.90	-	-
	Qian et.al. [18]	0.71	0.48	0.85
	Chen et al. [26]	0.59	0.43	0.74
	(1) + (2)	**0.16**	**0.23**	**0.42**
	(1) + (3)	**0.01**	**0.07**	**0.13**

To show the impact of our audio-visual deep cleansing framework, we train four speaker networks with or without cleansing to compare the performance. The networks include x-vector [3], ResNet34 [5], ECAPA-TDNN [6] with a small channel size equal to 512 (refer to ECAPA-S) and the ECAPA-L. All the experiments are repeated three times with the same setting. Vox1-O, E and H are used for in-domain evaluation and CnCeleb-VoxSRC22 [30][1] is used for cross-domain evaluation.

During training, we apply data augmentation for speaker and face networks to boost performance [31, 32]. During the evaluation, all test sets provide a set number of trials for verification, each containing two samples. For single modality, the cosine similarity between the speaker embedding (from the entire utterance) or the face embedding (from five face frames) of the given trial is calculated. For multi-modal, the combined speaker and face embedding is applied. The performance metric is the equal error rate (EER).

5. RESULTS AND ANALYSIS

5.1. Audio-visual speaker recognition

First, we report the performance of our audio-visual speaker recognition system in Tab 1. Both speaker and face networks perform better than existing approaches when trained on the VoxCeleb2 dataset. Here the face recognition network can

[1] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/data_workshop_2022/Track3_validation_trials.txt

Table 2: In-domain evaluation EER(%) of the speaker networks trained on the original and cleansed VoxCeleb2.

Network	Method	Vox1-O	Vox1-E	Vox1-H
X-vector	w/o cleanse	2.20	2.32	4.06
	with cleanse	2.09	2.13	3.76
	Δ	**5.0%**	**8.2%**	**7.4%**
ResNet34	w/o cleanse	1.31	1.41	2.58
	with cleanse	1.24	1.28	2.52
	Δ	**5.3%**	**9.2%**	**2.3%**
ECAPA-S	w/o cleanse	1.24	1.34	2.49
	with cleanse	1.17	1.28	2.37
	Δ	**5.6%**	**4.5%**	**4.8%**
ECAPA-L	w/o cleanse	1.02	1.23	2.36
	with cleanse	0.93	1.18	2.22
	Δ	**8.8%**	**4.1%**	**5.9%**

achieve even 0.03% EER when trained on Glint360K. For multi-modal verification, we obtain 0.16% and 0.01% EER on Vox1-O when the face network is trained on VoxCeleb2 and Glint360K, respectively. So our audio-visual system can accurately define the person's identity to guarantee the cleansing process.

5.2. Audio-visual deep cleansing

In Tab 2, we compare the performance of the speaker network trained on the original VoxCeleb2 and the VoxCeleb2 with our deep cleansing. For in-domain evaluation on VoxCeleb1, our audio-visual deep cleansing can remove the side-effect of the noisy data and boost the speaker recognition system with an average improvement of 5.9%. In Tab 3, for cross-domain evaluation on CnCeleb-VoxSRC22, our approach achieve an average improvement of 3.2%, which proves the network obtains a stronger generality.

Table 3: Cross-domain evaluation EER(%) of the speaker networks trained on the original and cleansed VoxCeleb2.

Method	X-vector	ResNet34	ECAPA-S	ECAPA-L
w/o cleanse	17.00	14.90	18.15	19.86
with cleanse	16.34	14.64	17.39	19.26
Δ	**3.9%**	**1.7%**	**4.2%**	**3.0%**

5.3. Visualization of results

We visualize our results in Fig 2. The left panel (a) represents audio-visual speaker recognition on Vox1-O. Each point denotes a test trial. The X-axis and Y-axis represent the speaker and face similarity score between the two samples in each trial, respectively. Target trials have high audio-visual scores, and imposter trials are the opposite, so the decision boundary is clear and reliable. The right panel (b) represents audio-visual deep cleansing on VoxCeleb2 with the same boundary.

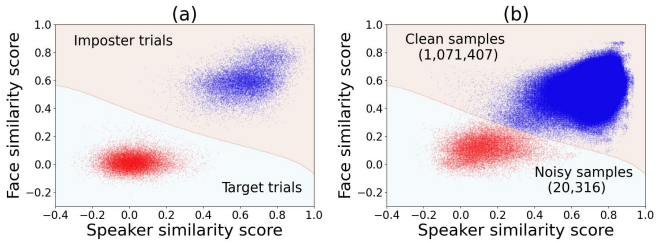

Fig. 2: Visualization of (a) audio-visual speaker recognition on Vox1-O; (b) audio-visual deep cleansing on VoxCeleb2.

Each point represents a training sample. The similarity score is between this point and the cleansed samples from the same speaker. Here most of the noisy samples have a very different representation to the mainstream samples in their class since they have a low audio-visual score. Our method finds 1.9% noisy samples on VoxCeleb2.

Fig. 3: Visualization of the samples with the speaker and face similarity score.

In Fig 3, each row contains four face images from the samples with the same speaker label, the left three are from the clean samples, and the right one is from the noisy sample. Both speaker and face similarity score has been marked below. Visual modality can assist speech modality in finding noisy samples.

6. CONCLUSION

In this paper, we design an audio-visual speaker recognition system that achieves close-to-perfect verification on the VoxCeleb1 test sets. A two-step audio-visual deep cleansing framework is proposed to automatically pick up noisy training sounds and strengthen the speaker recognition network. We observe that noisy samples (i.e., with wrong labels) are commonplace in large-scale datasets, and 5.9% of performance improvement could be achieved by simply removing a considerable percentage of noisy samples from the training set. In future work, we will study an end-to-end approach to combine the training and cleansing steps.

7. REFERENCES

[1] 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.

[2] Rohan Kumar Das and S. R. Mahadeva 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.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

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

[5] 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.

[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. 3830–3834.

[7] Tianchi Liu, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li, "MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7517–7521.

[8] 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.

[9] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Integrating frequency translational invariance in TDNNs and frequency positional information in 2D ResNets to enhance speaker verification," in *Interspeech*, 2021.

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

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

[12] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud, "Superloss: A generic loss for robust curriculum learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4308–4319, 2020.

[13] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3326–3334.

[14] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[15] Xiaoyi Qin, Na Li, Chao Weng, Dan Su, and Ming Li, "Simple attention module based speaker verification with iterative noisy label detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6722–6726.

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

[17] Lantian Li, Di Wang, and Dong Wang, "Pay attention to hard trials," *arXiv preprint arXiv:2209.04687*, 2022.

[18] Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, and Andrea Cavallaro, "Audio-visual tracking of concurrent speakers," *IEEE Transactions on Multimedia*, 2021.

[19] Rohan Kumar Das, Ruijie Tao, Jichen Yang, Wei Rao, Cheng Yu, and Haizhou Li, "Hlt-nus submission for 2019 nist multimedia speaker recognition evaluation," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 605–609.

[20] 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.

[21] 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.

[22] Suwon Shon, Tae-Hyun Oh, and James Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3995–3999.

[23] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf, "A multi-view approach to audio-visual speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6194–6198.

[24] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[25] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 88–97.

[26] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Multi-modality matters: A performance leap on voxceleb.," in *Interspeech*, 2020, pp. 2252–2256.

[27] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al., "Partial fc: Training 10 million identities on a single machine," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1445–1449.

[29] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang, "Learning for disparity estimation through feature constancy," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2811–2820.

[30] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

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

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