INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Our Learned Lessons from Cross-Lingual Speaker Verification: The CRMI-DKU System Description for the Short-duration Speaker Verification Challenge 2021

Xiaoyi Qin[1,3], *Chao Wang*[2], *Yong Ma*[2], *Min Liu*[2], *Shilei Zhang*[2], *Ming Li*[1,3*]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] China Mobile Research Institute, Beijing, China
[3] Data Science Research Center, Duke Kunshan University, Kunshan, China

xiaoyi.qin@dukekunshan.edu.cn, ming.li369@dukekunshan.edu.cn

Abstract

In this paper, we present our CRMI-DKU system description for the Short-duration Speaker Verification Challenge (SdSVC) 2021. We introduce the whole pipeline of our cross-lingual speaker verification system, including data preprocessing, training strategy, utterance-level speaker embedding extractor, domain-adaptation, and score calibration. We also propose methods to learn language-invariant features and perform domain adaptation to reduce the cross-lingual mismatch. In addition, we explore a semi-supervised method to utilize the unlabeled training data. The final submitted score level fusion system achieves 0.0476 minDCF and 0.98% EER on the evaluation set.

Index Terms: speaker verification, short duration, cross-lingual, domain adaptation

1. Introduction

In recent years, the X-Vector[1] based deep learning system has achieved great success in the field of automatic speaker verification (ASV). The existence of large-scale datasets (such as VoxCeleb1&2[2, 3]) and powerful modeling framework (such as ResNet[4], TDNN and ECAPA-TDNN[5] based backbone module) improve the performance of speaker verification systems significantly. The application of the Softmax Cross-Entropy loss function and its angular margin variants[6, 7] further reduce the variability of intra-class and enlarge the inter-class distance to obtain more discriminative features. However, speaker verification system is sensitive to the cross-domain mismatch. The issues of cross-dataset, far-field[8, 9] or cross-lingual[10] scenarios will degrade the ASV system performance. Although Probability Linear Discriminant Analysis(PLDA)[11], adversarial learning, and domain adaptation methods[12, 13] are proposed to address these issues, it is still a very challenging task and needs further investigation.

The SdSV Challenge[14, 15] proposes two tasks based on the DeepMine[16] dataset to explore these challenges. In this paper, we focus on task2: Text-Independent Speaker Verification. There are two partitions in this task. The first partition consists of typical text-independent trials where the enrollment and test utterances are from the same language (Persian). The second partition consists of text-independent cross-language trials where the enrollment utterances are in Persian, and the test utterances are in English[15]. The training data is fixed and consists of the VoxCeleb 1&2[2, 3], LibriSpeech[17], Mozilla Common Voice Farsi[18], and DeepMine (Task 2 Train Partition). It should be noted that most of the aforementioned datasets are in English except DeepMine. Moreover, although the language of the Common Voice dataset is Persian, it has no speaker labels. For solving these problems, we propose methods to learn language-invariant features and perform domain adaptation to cope with the cross-lingual trial file. We also train a language recognition system to divide the test trials into the same-lingual trial set and the cross-lingual trial set. In order to utilize the unlabeled in-domain data (Persian data), we also introduce a semi-supervised method to create the pseudo-label for the Common Voices data and include those data into training data. In addition, we also apply gradient reversal layer and domain adversarial training methods to make the speaker verification system less sensitive to the language mismatch.

The rest of the paper is organized as follows: Section2 describes the usage of databases. Section 3 will show the details of our speaker verification system, include data processing and utterance-level speaker embedding extractor. The training strategies and proposed robust methods are shown in Section 4. Section 5 and 6 provide the experimental results and conclusion, respectively.

2. Training dataset

Same as SdSVC 2020, the predefined fixed training set of SdSV 2021 consists of the following databases: VoxCeleb 1&2, LibriSpeech, Mozilla Common Voice Farsi, DeepMine Task 2 Train Partition.

2.1. VoxCeleb1&2 and LibriSpeech

The well-known VoxCeleb1&2 and LibriSpeech are mainly employed as the pre-train data. The majority utterances of these two databases are in English.

2.2. DeepMine

The DeepMine dataset is the in-domain data, which is from the same resource as the evaluation data. There are 588 speakers in Task 2 Train Partition of DeepMine, among them, 498 speakers have both Persian and English utterances. The in-domain data are utilized in several ways to reduce the domain mismatch:

- Mixed into pre-train data to fine-tune the model;
- Fine-tuning the pre-train model only with the in-domain data.

In addition, the in-domain data are also used in the score

* Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2021-398

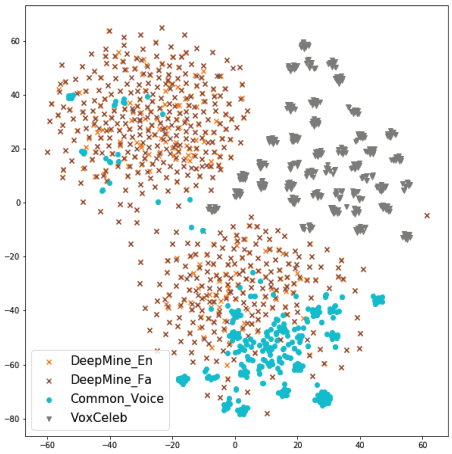

Figure 1: *The t-SNE plot of speaker embeddings in different datasets .*

normalization, domain-adaptation, and language-invariant feature learning.

2.3. Mozilla Common Voice Farsi

The Mozilla Common Voice Farsi dataset in total has 285867 utterances without the speaker label. We adopt the pre-train model to extract the speaker embeddings from those 285k+ utterances. We use the K-Means clustering to estimate the optimal K as the estimated speaker member. As shown in Fig.1, although the language of the Common Voice dataset is also Persian, it is not the in-domain data completely. However, we also try to add the Common Voice data into the model training and score normalization.

3. Speaker Verification System

In this section, We will introduce our baseline speaker verification system pipeline, including the acoustic feature extraction, data augmentation, utterance-level speaker embedding extractor, language recognition system, etc.

3.1. Data processing

Acoustic Feature. The acoustic features are 80-dimensional log Mel-filterbank energies with a frame length of 25ms and hop size of 10ms. The extracted features are mean-normalized before feeding into the deep speaker network.

Data Augmentation. We perform online data augmentation[19] using the MUSAN dataset[20]. Furthermore, we adopt the speed perturbation based on the SoX speed function to augment the speaker labels. The strategy also has a successful application in speech and speaker recognition tasks[21, 22]. The size of training data is tripled since two new versions of the original signal are created with speed factors of 0.9 and 1.1. The newly generated pitch-shifted data are considered as from new speakers. Therefore, we have in total 21795 virtual speakers in the speaker verification pre-training model.

3.2. Utterance-level Speaker Embedding Extractor

3.2.1. ResNet34 and ResNet34SE with Statistic Pooling

For the ResNet34 module, we adopt the same structure as in[23]. The network structure contains three main components: a front-end pattern extractor, an encoder layer, and a back-end classifier. The ResNet34[4] structure is employed as the front-end pattern extractor, the 128-dimensional fully connected layer following the encoder layer based on global statistic pooling (GSP) is adopted as the speaker embedding layer. The ArcFace[7] (s=32,m=0.2) is used as a classifier. The detailed configuration of the neural network is the same as in[24].

In addition, we also adopted the Squeeze-and-Excitation Module(SE)[25] which has been popular application in speaker verification system. Different from the ResNet34 system, we increase the widths from {32, 64, 128, 256} to {64, 128, 256, 512}. The output dim of the bottleneck layer is 256. The pooling layer and the classifier is the same as ResNet34 System.

3.2.2. ECAPA-TDNN with ASP

The ECAPA-TDNN Network[5] achieves great success in recent speaker verification tasks and provides the start-of-the-art performance. For this model, 1024 feature channels are used to scale up the network. The dimension of the bottleneck in the SE-Block is set to 256. The front-end feature extractor is followed by an attentive statistics pooling (ASP) layer[26] that calculates the mean and standard deviations of the final frame-level features. The classifier is the same as the ResNet system in Section 3.3.1.

3.3. Language Recognition

Considering there are two different trial cases, we train a language recognizer to separate them and further enhance the system performance. We employ the ECAPA-TDNN backbone as a front-end feature extractor. English and Persian are the output of binary language classification. We only adopt the Softmax Cross-Entropy as loss function in this task. The accuracy of language recognition reaches 100% in the dev dataset.

4. Training Strategy

In this section, we will introduce our training strategy, including mix-training, fine-tuning, domain adaptation, learning language-invariant features, semi-supervised learning, and score calibration.

4.1. Mix-training and Fine-tune

Based on our previous studies[8, 24, 27], the mix-training and fine-tune strategy are effective methods to improve the system performance.

Mix-training. Although the data are in different domains, mix-training is a direct and effective method to improve the system performance with reasonable amount of cross-domain data. We mix the DeepMine data into the VoxCeleb1&2 and LibriSpeech data to train a cross-lingual model; the pre-train model has been trained by VoxCeleb1&2 and LibriSpeech for 20 epochs before adding DeepMine data.

Fine-tune. The fine-tune strategy achieves good results in 2020 far-field speaker verification challenge[8, 28]. Given limited target domain data, the pre-train model can learn speaker information and the discriminative speaker embeddings, since

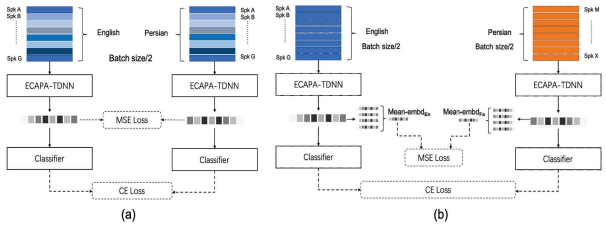

Figure 2: *Learning language-invariant features. The Fig.2 (a) stands for the training pipeline of Learning language-invariant features with Within Sample Method, and the Fig.2 (b) stands for the training pipeline of Learning language-invariant features with Batch Language Sample Method.*

Table 1: *The performance of different training strategy in SdSV dev set. Mix-training is mix the DeepMine data into the VoxCeleb1&2 and LibriSpeech data to train a cross-lingual model.*

Model ECAPA-TDNN	**SdSV21 Dev**	
	EER[%]	**mDCF**$_{0.01}$
Pre-training (Vox1&2 + Librispeech)	4.497	0.165
Mix-training	1.839	0.076
+ Common Voice (pseudo labels)	1.769	0.077
Fine-tune	1.909	0.079
+ GRL	11.3027	0.3622
+ Lang-invariant (With-in sample)	1.8996	0.099
+ Lang-invariant (Batch Language sample)	**1.6384**	0.079
+ Mean-Sub	**1.699**	**0.074**

we feed the target-domain data into the pre-train model with a small learning rate to make the speaker verification model also perform well on the target domain. Different from mixed training, fine-tuning is an effective and efficient method when only limited in-domain data are available.

4.2. Domain adaptation

Due to the existence of the domain gap (cross-lingual mismatch), we adopt multiple domain adaptation methods. Although there are many adversarial learning methods to make domain adaptation, these methods do not significantly improve or even degrade the system performance in our experiments for this task. In model training, for the methods of gradient reversal layer (GRL) and domain adversarial layer, although the accuracy increases, EER and minDCF becomes worse. As shown in Table.1, the proposed language-invariant feature learning method has a slight and stable improvement. However, Mean Subtraction is more effective in this task.

4.2.1. Learning language-invariant features

We design the two methods for learning language-invariant features.

Within Sample Method. Inspired by [29], we apply the within-sample language-invariant loss for cross-lingual speaker verification. The DeepMine dataset in the task2 part has 498 speakers with both Persian and English utterances. Therefore, we set the size of one batch as $B = N \times 2$. B stands for batch size with a number of N speakers in one batch. Each speaker random chose one English utterance and one Persian utterance

Table 2: *The performance of various speaker verification systems in the SdSV dev set.*

Model	**SN**	**Mean-Sub**	**SdSV21 Dev**	
			EER[%]	**mDCF**$_{0.01}$
ECAPA-TDNN				
1 + Mix-training	-	-	1.839	0.076
2 + Mix-training	✓	-	1.909	0.070
3 + Mix-training	✓	✓	**1.578**	**0.067**
4 + Fine-tune	-	-	1.909	0.079
5 + Fine-tune	✓	-	1.699	0.070
6 + Fine-tune	✓	✓	1.699	0.074
7 + Lang-invariant	-	-	1.830	0.079
8 + Lang-invariant	✓	-	1.717	0.072
9 + Lang-invariant	✓	✓	1.647	0.069
ResNet34-SE				
10 + Mix-training	-	-	1.569	0.059
11 + Mix-training	✓	-	1.359	0.055
12 + Mix-training	✓	✓	**1.307**	**0.056**
13 + Fine-tune	-	-	1.4989	0.075
ResNet34				
14 + Mix-training	-	-	1.438	0.065
15 + Mix-training	✓	-	1.499	0.059
16 + Mix-training	✓	✓	**1.359**	**0.049**
Fusion				
3+6+9+12+16(Average)			1.020	0.042
3+6+9+12+16 (Bosaris)			1.020	0.043
1+2+...+16 (Average)			1.020	0.041
1+2+...+16 (Bosaris)			**0.950**	**0.040**

in each batch. The loss function consists of the following formula:

$$L_{loss} = L_{CE}(\hat{Y}_{pre}, Y_{label}) + \frac{1}{N}\sum_{i=0}^{N} L_{MSE}(X^i_{En}, X^i_{Fa}) \quad (1)$$

The L_{CE} stands for the Cross Entropy Loss and the L_{MSE} denotes the Mean Squared Error between the English embedding and the Persian embedding for per speaker. $X^i_{En/Fa}$ denotes the English or Persian speaker embedding for the i_{th} speaker.

Batch Language Sample Method. This method does not depend on the selection of speakers. One batch samples are formed by half English and half Persian utterances. One batch is formed by $B = 2 \times N$ samples. N is the utterances number of English/Persian. The English and Persian samples are randomly chosen. The loss function consists of the following formula:

$$L_{loss} = L_{CE}(\hat{Y}_{pre}, Y_{label}) + L_{MSE}(\frac{1}{N}\sum_{i=0}^{N} X^i_{En}, \frac{1}{N}\sum_{j=0}^{N} X^j_{Fa}) \quad (2)$$

For the Batch Language Sample Method, the batch size should be big enough to smooth out the variability of small samples. In this experiment, we set the batch size as 512/1024. The details of the model training are shown in Fig.2.

4.2.2. Domain-adaption

Since there is a domain mismatch between cross-domain embeddings, to further reduce to the domain gap, we propose to adopt a direct method: Mean subtraction(Mean-Sub). We assume that the domain gap of speaker embeddings is the $X_{gap} = X_{Mean_{Fa}} - X_{Mean_{En}}$. For the cross-lingual trial, we transfer the test data to target domain of enrollment data through

Table 3: *The performance of various single speaker verification systems under different trial cases.*

Model	Dev Fari		Dev En	
	EER[%]	$\text{mDCF}_{0.01}$	EER[%]	$\text{mDCF}_{0.01}$
ECAPA-TDNN	0.984	0.049	2.537	0.120
+ Lang invariant	1.208	0.070	3.082	0.100
ResNet34SE	1.105	0.046	1.992	0.076
ResNet34	0.880	0.040	2.349	0.087

$X_{test} += X_{gap}$. The results of various domain adaptation methods are shown in Table.1

4.3. Semi-supervised learning

Although the Common Voice data does not have speaker labels, the t-SNE visualization results show that the data distribution of Common Voice and DeepMine is similar. Therefore, we adopt a semi-supervised method to utilize the data. The procedure is shown in the following:

- Step 1. Using the current speaker embedding network to extract speaker embeddings for all utterances from the Common Voice dataset.
- Step 2. Running a clustering algorithm (K-means) to determine the optimal K.
- Step 3. Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of K, and choose the K for which WSS first starts to diminish.

According to our calculation following steps 1, 2 and 3, there are about 30 clusters. In order to avoid inaccurate classification, we just chose 300 utterances which are close enough to each cluster centroid. The result is shown in Table.1. However, we believe that there are more than 30 speakers in the Common Voices Farsi dataset.

4.4. Score Calibration

Based on[30], we set the imposter cohort of the adaptive score normalization to contain in-domain Farsi and English data. Here, the score between the enrollment utterance e and the test utterance t is denoted as $s(e,t)$. We adopt the adaptive S-norm2[30], which is defined as

$$s(e,t)_{as-norm2} = \frac{1}{2}\left(\frac{s(e,t)-\mu(S_e(\varepsilon_t^{top}))}{\sigma(S_e(\varepsilon_t^{top}))} + \frac{s(e,t)-\mu(S_t(\varepsilon_e^{top}))}{\sigma(S_t(\varepsilon_e^{top}))}\right) \quad (3)$$

where $\mu(S_t(\cdot))$ and $\sigma(S_t(\cdot))$are mean and standard deviation of S_t. The adaptive cohort can be selected to as X closest (most positive scores) files to either the enrollment file ε_e^{top} or the test file ε_t^{top}[30].

5. Experimental results

5.1. Single System

In this experiment, we found that mix-training is better than fine-tuning. We consider there might be two reasons: 1) in-domain data is sufficient; 2) the speaker embedding extractor framework we adopted have powerful modeling capability, fine-tuning only with in-domain data may cause overfitting.

Table 4: *Evaluation of submitted system on the SdSVC eval set.*

Model	SdSV21 Eval	
	EER[%]	$\text{mDCF}_{0.01}$
3+6+9+12+16 (Average)	-	0.049373
model id 1-16 (Average)	0.98	0.047553
3+6+9+12+16 (Bosaris)	-	0.053411
model id 1-16 (Bosaris)	-	0.048552

Comparing with the other two speaker verification systems in Table.2, the ResNet34-SE achieves the best performance in the SdSV dev set without any trick. The Mean-Sub method in all systems shows to beneficial and enhances performance by 5% to 15% relatively in terms of EER and mDCF, respectively. From Table.3, we can observe that the most challenging part of this task is still the cross-lingual trial. Even though we tried to train language-invariant feature to reduced the gap on cross-language trials, the performance of the same-lingual trials is affected.

5.2. Fusion System

The average weight and weight based on the Bosaris toolkit are also employed in the score level fusion and calibration. From Table.4, we can find that the fusion on the score level by taking a weighted average over the calibrated scores of each individual system achieves the best result on the test set. Fusion of all systems leads to a relative improvement over the ResNet34-SE System by 20% in EER and MinDCF on the SdSVC dev set. This shows that multiple network frameworks can prove sufficient to learn complementary speaker information. Finallyour submitted system achieves the 0.950% EER and 0.040 $mDCF_{0.01}$ on the development set and 0.98% EER and 0.0476 $mDCF_{0.01}$ on the evaluation set.

6. Conclusions

In this paper, we present our baseline system and robust training methods to reduce the cross-lingual mismatch. Furthermore, the unlabeled in-domain Common Voices data could be further utilized to enhance the performance. A fusion of three systems based on ECAPA-TDNN and ResNet architecture in conjunction with the proposed strategy results in a good performance on Task 2 of the 2021 SdSVC with an EER of 0.98% and a MinDCF of 0.0476.

7. Acknowledgements

This research is funded in part by the National Natural Science Foundation of China (61773413), the Fundamental Research Funds for the Central Universities (2042021kf0039), Key Research and Development Program of Jiangsu Province (BE2019054), Science and Technology Program of Guangzhou City (201903010040, 202007030011) and Six Talent Peaks Project in Jiangsu Province (JY074)

8. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "x-vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A Large-

Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc.CVPR*, 2016, pp. 770–778.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep Hypersphere Embedding for Face Recognition," in *Proc. CVPR*, 2017, pp. 212–220.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.

[8] X. Qin, D. Cai, and M. Li, "Far-Field End-to-End Text-Dependent Speaker Verification based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation," in *Proc. Interpseech 2019*.

[9] L. Zhang, J. Wu, and L. Xie, "NPU Speaker Verification System for INTERSPEECH 2020 Far-Field Speaker Verification Challenge," in *Proc. Interspeech*, 2020, pp. 3471–3475.

[10] J. Thienpondt, B. Desplanques, and K. Demuynck, "Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization," in *Proc. Interspeech*, 2020, pp. 756–760.

[11] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. ECCV 2006*, pp. 531–542.

[12] Y. Tu, M. Mak, and J. Chien, "Variational Domain Adversarial Learning With Mutual Information Maximization for Speaker Verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.

[13] L. Li, M. W. Mak, and J. T. Chien, "Contrastive Adversarial Domain Adaptation Networks for Speaker Recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2020.

[14] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV Challenge 2020: Large-Scale Evaluation of Short-Duration Speaker Verification," in *Proc. Interspeech*, 2020, pp. 731–735.

[15] ——, "Short-duration Speaker Verification (SdSV) Challenge 2021: the Challenge Evaluation Plan," in *arXiv:1912.06311*.

[16] H. Zeinali, L. Burget, and J. Cernocky, "A Multi Purpose and Large Scale Speech Corpus in Persian and English for Speaker and Speech Recognition: The Deepmine Database," in *Proc. ASRU*, 2019, pp. 397–402.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *arXiv:1912.06670*.

[19] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

[20] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484*.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[22] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker Augmentation and Bandwidth Extension for Deep Speaker Embedding," in *Proc. Interspeech*, 2019, pp. 406–410.

[23] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. Odyssey*, 2018, pp. 74–Sd81.

[24] W. Wang, D. Cai, X. Qin, and M. Li, "The DKU-DukeECE Systems for VoxCeleb Speaker Recognition Challenge 2020," *arXiv.2010.12731*.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. CVPR*, 2018.

[26] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," *in Proc. Interspeech*, 2018.

[27] D. Cai, X. Qin, W. Cai, and M. Li, "The DKU-SMIIP System for the Speaker Recognition Task of the VOiCES from a Distance Challenge," in *Proc. Interspeech*, 2019.

[28] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. Narayanan, and H. Li, "The INTERSPEECH 2020 Far-Field Speaker Verification Challenge," in *Proc. Interspeech*, 2020, pp. 3456–3460.

[29] D. Cai, W. Cai, and M. Li, "Within-Sample Variability-Invariant Loss for Robust Speaker Recognition Under Noisy Environments," in *Proc. ICASSP*, 2020, pp. 6469–6473.

[30] P. Matjka, O. Novotn, O. Plchot, L. Burget, M. D. Snchez, and J. ernock, "Analysis of Score Normalization in Multilingual Speaker Recognition," in *Proc. Interspeech*, 2017.