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

The SJTU System for Short-duration Speaker Verification Challenge 2021

Bing Han, Zhengyang Chen, Zhikai Zhou, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{hanbing97, zhengyang.chen, zhikai.zhou, yanminqian}@sjtu.edu.cn

Abstract

This paper presents the SJTU system for both text-dependent and text-independent tasks in short-duration speaker verification (SdSV) challenge 2021. In this challenge, we explored different strong embedding extractors to extract robust speaker embedding. For text-independent task, language-dependent adaptive snorm is explored to improve the system performance under the cross-lingual verification condition. For text-dependent task, we mainly focus on the in-domain fine-tuning strategies based on the model pre-trained on large-scale out-of-domain data. In order to improve the distinction between different speakers uttering the same phrase, we proposed several novel phrase-aware fine-tuning strategies and phrase-aware neural PLDA. With such strategies, the system performance is further improved. Finally, we fused the scores of different systems, and our fusion systems achieved 0.0473 in Task1 (rank 3) and 0.0581 in Task2 (rank 8) on the primary evaluation metric.
Index Terms: speaker verification, phrase-aware fine-tuning, cross-lingual verification, SdSV challenge 2021

1. Introduction

Speaker verification system has gained great improvement with the development of deep learning. From the phone-channel [1] to in-the-wild condition [2], researchers have proposed different architectures [3, 4, 5, 6], different losses [7, 8, 9], and different training strategies [10, 11, 12] to improve the system performance under different conditions. However, there are still some challenges unresolved when the speaker verification system is applied in the real world, such as the short duration problem and cross-lingual problem.

In this paper, we introduce the SJTU system submitted to the short-duration speaker verification (SdSV) challenge 2021 [13]. The SdSV challenge 2021 includes two tasks. The task1 is the text-dependent task, where the speaker verification system should verify the test speaker identity and speaking phrase at the same time. Task2 is text-independent task, the system should only consider the speaker identity. Particularly, the SdSV challenge introduces a new challenging verification condition, the cross-lingual verification for task2, where one speaker may speak different languages at the enrollment and test stage.

The SdSV 2021 is the second challenge of SdSV series and many competitive systems have been proposed in the last challenge. For text-independent task in the last challenge, Jenthe et al. proposed a new data mining strategy HPM [14] and introduced adaptive snorm to improve system's cross language verification robustness. Peng et al. introduced a greedy fusion algorithm [15] to further improve the performance of the fusion system. Besides, teams mainly focused on the back-end optimization [16, 17] in text-dependent task.

In this challenge, we first explored different well-performed architectures and trained them on all the available data. Then, we focus on the in-domain data fine-tuning strategies to further improve the system performance. To solve the cross-language verification problem in text-independent task, we trained another language identification network to introduce the language information to the adaptive s-norm [18] procedure. For text-dependent task, we implemented different methods to increase the distinction between the target trial and different non-target trials. We used an ASR system to classify the speaker phrase during the test stage and filter out the phrase-mismatch (speaker utters a wrong pass-phrase) trials directly. To better distinguish different speakers uttering the same phrase, we proposed several novel phase-aware fine-tuning strategies and phrase-aware neural PLDA. Based on such strategies, the performance of our systems is further improved.

The rest of the paper is organized as follows: Section 2 introduces the dataset used in this challenge. Section 3 introduces our embedding extractor architectures and the proposed fine-tuning strategies. The experimental results and corresponding analysis are given in Section 4. Finally, we make the conclusion in Section 5.

2. Datasets

2.1. Training Data

SdSV challenge adapted a fixed training condition where the system should only be trained on the designed set. The main training and evaluation data for SdSV challenge is the DeepMine [19, 20] dataset which was recorded in realistic environments of Iran. And the collection protocol was designed to incorporate various kinds of noises during the recording. The main language is Persian while the most of the participants also participated in the English partition.

- Task 1 in-domain data: A dataset which is designed for building text-dependent speaker verification system. It consists of 101k utterances from 963 different speakers. The content of all utterances is limited to a fixed set including five Persian phrases and five English phrases.
- Task 2 in-domain data: A dataset which has no restrictions on utterance content. It contains 125k utterances collected from 588 speakers while some of them have only Persian phrases.

In addition to the in-domain training data, other opening datasets allowed to be used in the training process are described as follows.

[†] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-2136

Voxceleb [21, 22]: Voxceleb 1&2 contain more than one million utterances from 7245 celebrities, which are collected from videos uploaded to YouTube.

Librispeech [23]: A dataset which comprises 281k utterances from 2338 speakers. It's sourced from audio books and the majority of speech is US English.

Common Voice Farsi [24]: The Common Voice corpus is a massively-multilingual collection of transcribed speech. And only Persian part of it is used in this challenge.

2.2. Evaluation

The evaluation data for task 1&2 are both part of the DeepMind in-domain data.

- In Task 1, each trial consists of a test segment along with a model identifier which indicates three enrollment utterances and a phrase ID that uttered in the utterances. These trials can be classified into four basic types including TC, TW, IC and IW. The text-dependent speaker verification system should accept the TC trials and reject the other three types as imposture trials.
- In Task 2, the enrollment data consists of one to several variable-length utterances from the Persian language while the test utterances might from the different language (English). For this task, systems should accept the trials if enroll and test utterances are both from the same speaker without considering language mismatch.

The main metric adopted by SdSV challenge is normalized minimum detection cost function (minDCF), which is defined as a weighted sum of false alarm and miss error probabilities.

3. Methods

In this section, we will introduce the embedding extractors, fine-tuning strategies and several post-processing methods used in our system. In our experiment, the embedding extractors are firstly trained on all the available data for both task1 and task2 in a text-independent mode. Then, we fine-tune the pre-trained models using in-domain data. Finally, post-processing methods are used to further improve the system performance.

3.1. Speaker Embedding Extractors

To build a robust speaker verification system for SdSV challenge, all datasets including Voxceleb, Common Voice Farsi, Librispeech, and DeepMind in-domain data are combined for training the speaker embedding extractors. In order to reduce the duration mismatch between the training and test data, all the utterances are randomly chunked into segments of 2 seconds during the training stage. The acoustic feature we used is 40 dimensional Fbank with 25 ms frame length and 10 ms shift. To increase the quantity and diversity of the training data, we apply online data augmentation during the training process. The additive noises from MUSAN corpus [25] and the impulse responses from RIR [26] are used for augmentation.

In our system, we mainly adopt three different speaker verification architectures, including the ResNet34 [27], ECAPA-TDNN [5] and DPN68 [6, 28].

ResNet34: ResNet has achieved superior performance in speaker verification for its high-efficiency modeling complex data structure. We use the ResNet34 introduced in [27] as our resnet based architecture. In this architecture, input features are processed by the initial convolution layer and 4 residual blocks, then a following statistic pooling layer aggregates the frame-level features into segment-level representation. Finally, a 256-dimensional fully connected layer transforms it into a fixed vector to represent the speaker.

ECAPA-TDNN: The ECAPA-TDNN [5] has achieved great success in speaker verification system and has been used in the VoxSRC2020 [29] winning system. We set the channel number for ECAPA-TDNN to 1024 in our experiment. Channel attention with and without global context are both applied and we denote the corresponding architectures as Ecapa and Ecapa-Glob respectively.

DPN68: The DPN (Dual Path Network) [6] is firstly applied in speaker verification task in [28], which leverages the advantage of ResNet and DenseNet [30] at the same time. Here, we use the DPN68 architecture as one of our embedding extractors in our systems.

Additive angular margin softmax loss (AAM) [7] is used to optimize all the embedding extractors. The scale parameter and the margin of AAM loss are set to 32 and 0.2 respectively. We train each model for 165 epochs and the learning rate exponentially decreases from 0.1 to 1e-5 during the training process.

3.2. In-domain Fine-tuning

In this section, we will introduce our fine-tuning strategies based on the pre-trained model presented in the last section to further improve the system performance on the in-domain evaluation set.

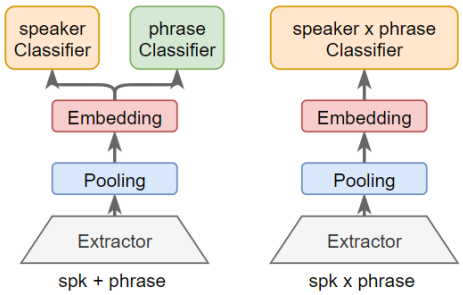

Figure 1: *Text-dependent Mode Fine-tuning for Task 1*

3.2.1. Text-Dependent Mode Fine-tuning for Task 1

To encode the phrase information into speaker embedding and further enlarge the distance of different speakers uttering the same phrase, we fine-tune the embedding extractors in text-dependent mode for task1. Strategies are introduced as follows:

spk + phrase: As shown in Figure. 1 (left), in the fine-tuning stage, there are two separate heads for speaker and phrase classification and we fine-tune the embedding extractors in a multi-task way.

spk × phrase: Here, utterances in different phrases spoken by the same speaker are considered as the different classes. As shown in Figure.1 (right), there is only one classification head, but both speaker and phrase information are considered.

Since the classification of the phrase requires all the information of a sentence, the inputs are not chunked during the training process and variable-length inputs in the same batch are zero-padded to the same length.

3.2.2. Text-Independent Mode Fine-tuning for Task1 and Task2

In our experiment, the phrase mismatch trials (IW and TW) in task1 can be filtered out by ASR system introduced in sec-

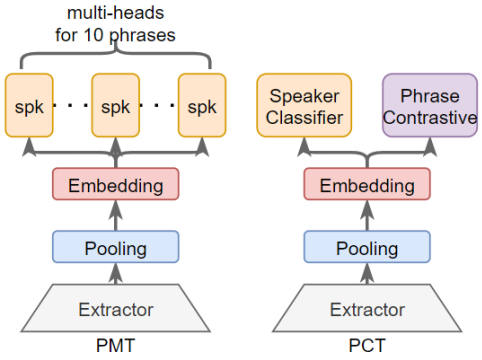

Figure 2: *Text-independent Mode Fine-tuning for Task 1*

tion 3.3.3. Therefore, the model only needs to verify utterances in the same phrase, and task1 can also be regarded as text-independent task. For usual fine-tune methods of task1 and task2, we use AAM softmax to optimize the pre-trained models on in-domain data.

Especially for task1, to strengthen model's ability to discriminate speakers uttering the same phrase, we proposed two phrase-aware text-independent fine-tuning strategies, including phrase-aware multi-head training (PMT) and phrase-aware contrastive training (PCT).

PMT: For task 1, all the phrases are drawn from a fixed set of ten phrases consisting of five Persian and five English phrases. As shown in Figure 2 (left), different speaker classification heads are used for the utterances in different phrases. With such a training strategy, the distance between different speakers within the same phrase can be enlarged.

PCT: As shown in Figure 2 (right), in this fine-tuning strategy, we introduce a contrastive learning loss that can be jointly optimized with the AAM softmax loss. In our experiment, generalized end-to-end loss [9] is adopted to calculate the contrastive loss and we sample two utterances for each speaker in the training batch. To improve the distinction between different speakers uttering the same phrase, we constrain that all the utterances in the same batch are from the same phrase.

3.3. Post Processing

3.3.1. Language-dependent AS-Norm

The most difficulty of task 2 is cross-language trials. To minimize the language mismatch between the different utterances, we introduce the language information to the adaptive symmetric score normalization (AS-Norm), which is defined in Equation 1. The cohort set for each enroll model, $\mathcal{E}_{e,lan}^{top}$, is decided by the enroll model and the language of test utterance, where the language of cohort set is the same as test utterance. To detect the language of the test utterances, a TDNN based language identification is trained on task 2 in-domain data.

$$s(e,t) = \frac{s(e,t) - \mu(S_t(\mathcal{E}_t^{top}))}{\sigma(S_t(\mathcal{E}_t^{top}))} + \frac{s(e,t) - \mu(S_e(\mathcal{E}_{e,lan}^{top}))}{\sigma(S_e(\mathcal{E}_{e,lan}^{top}))}$$

3.3.2. Phrase-aware Neural PLDA

The neural PLDA (NPLDA) [31] was successfully applied in the NICT system [32] of SdSV challenge 2020. To enhance the NPLDA for text-dependent task, we constrain the input pair for NPLDA from the same phrase to improve the different speakers' distinction ability within the same phrase. Our NPLDA was initialized by phrase-dependent PLDA model trained on task2 in-domain data and the parameters were set the same as [31], except learning rate = 5e-5 and epochs = 5.

3.3.3. ASR System

For text-dependent task1, an ASR system is trained to filter the trials that enroll and test utterances are from different phrases. We adopt the conformer [33] based joint CTC-attention automatic speech recognition model (ASR) from the ESPnet [34] Librispeech recipe. The ASR system is first trained on the Librispeech dataset and then fine-tuned on the task1 in-domain data. During evaluation, we use the ASR system to recognize the phrases and classify each utterance based on its Levenshtein edit distance with the references. According to the phrase label generated by ASR, we directly filter out IW and TW trials by setting the score to a very low value. It is noted that all the results for task1 provided in the experiments are revised based on this ASR system.

4. Experiments

4.1. Task 1: Text-Dependent Speaker Verification

4.1.1. Pre-trained Model

All the embedding extractors introduced in section 3.1 are firstly pre-trained on all the available datasets and the corresponding results are listed in Table 1. From the results, we can see that ResNet34 using cosine similarity measurement obtains the best performance. In comparison with Ecapa-TDNN based models, DPN68 also exhibits a better performance but worse than ResNet34. In addition, the cosine scoring method outperformed the PLDA in most cases, and we will only provide the cosine result in the following sections for analysis.

Table 1: *Results comparison of Pre-trained Models on Task 1*

Model Type	Cosine		PLDA	
	EER	minDCF	EER	minDCF
ResNet34	**2.680**	**0.0811**	2.680	0.1032
Ecapa	3.093	0.0994	3.093	0.0984
Ecapa-Glob	2.887	0.1008	3.093	0.0943
DPN68	2.680	0.0882	3.299	0.0984

4.1.2. In-domain Fine-tuning

Text-Dependent Mode: Table 2 illustrates the comparison of different text-dependent mode fine-tuning strategies introduced in 3.2.1. Limited by GPU memory and time consumption, large models with whole utterances as input are difficult to train. Here, experiments are only based on Resnet34. From the result, we can see that fine-tuning will bring improvement to the text-dependent task. Especially, "speaker + phrase" obtained the best performance among these results.

Text-Independent Mode: Our proposed text-independent mode phrase-aware fine-tuning strategies are investigated in

Table 2: *Text-dependent Mode Fine-tuning of Task 1*

Text-dependent Finetune	EER	minDCF
-	2.680	0.0811
speaker $\times$ phrase	2.474	0.0740
speaker $+$ phrase	**2.268**	**0.0718**

Table 3: *Main Results of Task 2. Fine-tune is applied on Task 2 in-domain data with AAM softmax*

Model	ResNet34		Ecapa		Ecapa-Glob		DPN68	
	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
Pre-trained	2.657	0.1036	2.793	0.1252	2.929	0.1281	2.452	0.0975
+ Fine-tune	2.316	0.0933	2.793	0.1173	2.589	0.1219	2.316	0.0832
++ A-snorm	**1.981**	0.0872	2.520	0.0991	2.316	0.1045	2.112	**0.0752**

this section and the results are shown in Table 4. It's obvious that all fine-tuning systems perform better than the pre-trained model. In this table, we also list the fine-tuning results with PCT (phrase-aware contrastive training) and PMT (phrase-aware multi-head training) for all models. Compared with the usual fine-tuning method on the in-domain data, PCT and PMT both achieve excellent performance improvement on both EER and minDCF. In addition, ResNet34 with PCT strategy performs the best within all the models.

Table 4: *Text-independent Phrase-aware Fine-tune for Task 1. The usual Fine-tune without PCT and PMT is also done on Task 1 for the comparison where the in-domain data is trained with AAM softmax in text-independent mode.*

Model	Fine-tune	PCT	PMT	EER	minDCF
ResNet34				2.680	0.0811
	✓			2.680	0.0734
	✓	✓		**2.260**	**0.0642**
	✓		✓	2.680	0.0708
Ecapa				3.093	0.0994
	✓			3.093	0.0992
	✓	✓		2.680	0.0863
	✓		✓	2.887	0.0852
Ecapa Glob				2.887	0.1008
	✓			2.680	0.0863
	✓	✓		2.680	0.0855
	✓		✓	3.093	0.0772
DPN68				2.680	0.0882
	✓			2.887	0.0866
	✓	✓		2.680	0.0759
	✓		✓	2.680	0.0759

4.1.3. Phrase-aware Neural PLDA

As mentioned in Table 1, PLDA cannot provide a satisfactory performance compared to cosine similarity. To further enhance the fusion system performance, we also trained the phrase-aware NPLDA introduced in section 3.3.2 to improve the result of PLDA which can be used in the final fusion system. We conduct this investigation based on the models fine-tuned with PCT strategy which obtain the best result in section 4.1.2 and the corresponding NPLDA result can be found in Table 5. Compared with traditional PLDA, NPLDA shows its effectiveness and brings an excellent improvement on minDCF which is comparable with cosine back-end results.

4.2. Task 2: Text-Independent Speaker Verification

The main results of task2 are listed in the Table 3. In the task2, all the embedding extractors are also first pre-trained on all the available datasets and then fine-tuned on the in-domain data. According to the results, we can see that DPN68 achieves the best result within all the models. Besides, compared with the pre-trained models, fine-tuning with in-domain data delivers a significant performance improvement. Besides, we also conduct an investigation of language-dependent asnorm. DPN68 with asnorm outperforms others with the lowest minDCF 0.0752.

Table 5: *NPLDA Results of Task 1*

Model	PLDA		NPLDA	
	EER	minDCF	EER	minDCF
ResNet34	3.093	0.1060	2.887	0.0734
Ecapa	2.887	0.1106	**2.680**	0.0844
Ecapa-Glob	2.887	0.1021	2.887	0.0738
DPN68	3.505	0.1009	2.887	**0.0656**

4.3. Fusion Result

Table 6: *Fusion Results on Dev and Eval Set*

Task	Dev set		Eval set		Rank
	EER	minDCF	EER	minDCF	
Task1	2.268	0.0493	1.44	0.0473	3
Task2	1.703	0.0579	1.23	0.0581	8

Finally, the scores from all the systems including different models, back-ends and fine-tuning strategies are weighted summed to get the fusion system and we use the development set for fusion weights tuning. The results of fusion systems on the development set and evaluation set are shown in Table 6. From the table, we can see that a fusion system could further improve the performance. Our primary submission is produced by fusion systems and achieved 0.0473 in Task1 (rank 3) and 0.0581 in Task2 (rank 8) on minDCF.

5. Conclusion

In this paper, we give a detailed description of our submission to Task 1 & 2 of SdSV Challenge 2021. Several strong embedding extractors are explored in our experiment. For text-independent task, another language identifier is used to introduce language information to the adaptive snorm. For text-dependent task, an ASR system is used to filter the IW and TW trials. We proposed several phrase-aware fine-tuning and post-processing methods to strengthen model's ability to verify speakers within the same phrases. Based on these strong systems, our final fusion system achieved 3^{rd} and 8^{th} place in task1 and task2 respectively.

6. Acknowledgements

This work was supported by the China NSFC projects (No. 62071288 and No. U1736202). Experiments have been carried out on the PI super- computer at Shanghai Jiao Tong University. And the author would like to thank the support of Houjun Huang and Xu Xiang from AISpeech Ltd.

7. References

[1] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The nist speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[2] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[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)*. IEEE, 2018, pp. 5329–5333.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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," *arXiv preprint arXiv:2005.07143*, 2020.

[6] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," *arXiv preprint arXiv:1707.01629*, 2017.

[7] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," *arXiv preprint arXiv:1904.03479*, 2019.

[8] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Interspeech*, 2018, pp. 3623–3627.

[9] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[10] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

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

[12] Z. Chen, S. Wang, and Y. Qian, "Adversarial domain adaptation for speaker verification using partially shared network," *Proc. Interspeech 2020*, pp. 3017–3021, 2020.

[13] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (sdsv) challenge 2021: the challenge evaluation plan." arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[14] 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*, 2020, pp. 756–760.

[15] P. Shen, X. Lu, and H. Kawai, "Investigation of NICT Submission for Short-Duration Speaker Verification Challenge 2020," in *Proc. Interspeech 2020*, 2020, pp. 751–755.

[16] Z. Chen and Y. Lin, "Improving X-Vector and PLDA for Text-Dependent Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 726–730.

[17] A. Lozano-Diez, A. Silnova, B. Pulugundla, J. Rohdin, K. Veselỳ, L. Burget, O. Plchot, O. Glembek, O. Novotnỳ, and P. Matejka, "But text-dependent speaker verification system for sdsv challenge 2020," *Proc. Interspeech 2020*, pp. 761–765, 2020.

[18] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Proc. Interspeech 2017*, 2017, pp. 1567–1571.

[19] 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 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[20] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.

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

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, 2018, pp. 1086–1090.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] 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," *arXiv preprint arXiv:1912.06670*, 2019.

[25] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[26] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[27] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[28] X. Xiang, "The xx205 system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2011.00200*, 2020.

[29] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] S. Ramoji, P. Krishnan, and S. Ganapathy, "NPLDA: A Deep Neural PLDA Model for Speaker Verification," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 202–209.

[32] P. Shen, X. Lu, and H. Kawai, "Investigation of nict submission for short-duration speaker verification challenge 2020," *Proc. Interspeech 2020*, pp. 751–755, 2020.

[33] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition."

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.