The INTERSPEECH 2020 Far-Field Speaker Verification Challenge

Xiaoyi Qin[1], Ming Li[1,5], Hui Bu[4], Wei Rao[2], Rohan Kumar Das[2],
Shrikanth Narayanan[3], Haizhou Li[2]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]Department of Electrical & Computer Engineering, National University of Singapore, Singapore
[3]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, USA
[4]AI Shell Foundation, Beijing, China
[5]School of Computer Science, Wuhan University, Wuhan, China
ming.li369@dukekunshan.edu.cn

Abstract

The INTERSPEECH 2020 Far-Field Speaker Verification Challenge (FFSVC 2020) addresses three different research problems under well-defined conditions: far-field text-dependent speaker verification from single microphone array, far-field text-independent speaker verification from single microphone array, and far-field text-dependent speaker verification from distributed microphone arrays. All three tasks pose a cross-channel challenge to the participants. To simulate the real-life scenario, the enrollment utterances are recorded from close-talk cellphone, while the test utterances are recorded from the far-field microphone arrays. In this paper, we describe the database, the challenge, and the baseline system, which is based on a ResNet-based deep speaker network with cosine similarity scoring. For a given utterance, the speaker embeddings of different channels are equally averaged as the final embedding. The baseline system achieves minDCFs of 0.62, 0.66, and 0.64 and EERs of 6.27%, 6.55%, and 7.18% for task 1, task 2, and task 3, respectively.

Index Terms: Speaker verification, Far-field, Cross channel matching, Distributed microphone array, Enrollment augmentation

1. Introduction

Automatic speaker verification (ASV) is an enabling technology in speech processing and biometric authentication. It has been deployed in many real-life applications, such as access control, and law enforcement. With the advent of deep learning, speaker recognition performance has improved remarkably in both close-talk and far-field settings. However, it is still far from perfect, for example, speaker recognition under noisy and far-field conditions remains a challenging task. The INTER-SPEECH 2020 Far-Field Speaker Verification Challenge aims at providing a common platform for the research community to advance the state-of-the-art.

A typical deep speaker network firstly learns frame-level speaker representation with the local pattern extractor, which is usually a time-delayed neural network (TDNN) [1] or a convolutional neural network (CNN) [2]. The learnt frame-level feature sequence is then converted into a fixed-dimension representation by different pooling mechanisms such as statistics pooling [1], attentive pooling [3], and learnable dictionary encoding [2]. Since speaker verification in the open set settings is essentially a metric learning problem, several discriminative classification losses such as A-softmax [4] and AM-softmax [5] are employed to enhance the recognition performance.

To compensate for the adverse impacts of reverberation and noise in the far-field scenario, various approaches have been proposed for ASV systems. At signal level, weighted prediction error [6, 7] is employed for dereverberation. DNN-based denoising [8, 9, 10] and beamforming [11, 12] are investigated for single-channel and multi-channel speech enhancement respectively. At the modeling level, data augmentation [13, 14, 15] and transfer learning [16] are proven to be effective with limited target domain data. To learn a noise-invariant speaker embedding, adversarial training [17, 18] and variability-invariant loss [19] are investigated. Also, joint training of speech enhancement network and speaker embedding network can improve the ASV performance under noisy conditions [20, 21, 22]. For deep speaker modeling with microphone array, a multi-channel training framework is proposed for speaker embedding extraction [23]. Moreover, in the testing phase, enrollment data augmentation is proposed to reduce the mismatch between the enrollment and testing utterances [16].

Recently, far-field speaker recognition attracts more and more attention from the research community. The Voices Obscured in Complex Environmental Settings (VOiCES) Challenge launched in 2019 aims to benchmark state-of-the-art speech processing methods in far-field and noisy conditions [24]. The wake-up word dataset *Hi Mia* has also been released to facilitate the studies in far-field text-dependent speaker verification [25]. However, some research questions still require further exploration for speaker verification in the far-field and complex environments. Those open challenges including but not limited to,

1. Far-field text-dependent speaker verification for wake up control

2. Far-field text-independent speaker verification with complex environments

3. Far-field speaker verification with cross-channel enrollment and test

4. Far-field speaker verification with single multi-channel microphone array

5. Far-field speaker verification with multiple distributed microphone arrays

6. Far-field speaker verification with front-end speech enhancement methods

7. Far-field speaker verification with end-to-end modeling using data augmentation

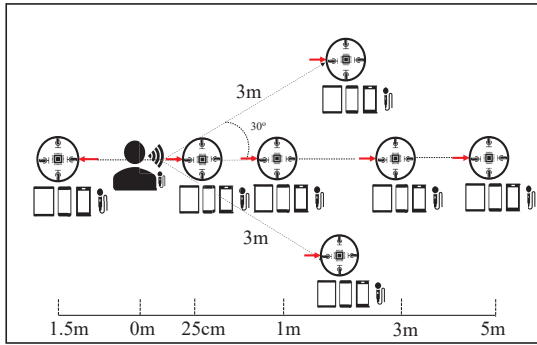8. Far-field speaker verification with front-end and back-end joint modeling

Figure 1: *The setup of the recording environment*

9. Far-field speaker verification with transfer learning and domain adaptation

To this end, we collect a large scale far-field speaker verification dataset with real speakers in multiple scenarios, which include text-dependent, text-independent, cross channel enrollment and test, distributed microphone array, etc. We also launch the Far-Field Speaker Verification Challenge 2020 (FFSVC 2020) based on this dataset. It focuses on far-field distributed microphone arrays under noisy conditions in real scenarios. The objectives of this challenge are to: 1) benchmark the current speech verification technology under this challenging condition, 2) promote the development of new ideas and techniques in speaker verification, 3) provide an open, free, and large scale speech dataset to the community that exhibits the far-field characteristics in real scenes.

The challenge consists of three tasks with different setups,

- Task 1: far-field text-dependent speaker verification from single microphone array

- Task 2: far-field text-independent speaker verification from single microphone array

- Task 3: far-field text-dependent speaker verification from distributed microphone arrays

All three tasks pose a cross-channel challenge, that is to have enrollment speech from the close-talk cellphone, and to have test speech from far-field microphone array(s). This paper provides the description of the challenge, the dataset, and the reference baseline.

2. Challenge Dataset

2.1. DMASH Dataset

The Distributed Microphone Arrays in Smart Home (DMASH) dataset is recorded in real smart home scenarios with two different rooms. Figure 1 shows the recording setup of DMASH dataset. The recording devices include one close-talk microphone and seven groups of devices at seven different positions of the room. A group of recording devices include one iPhone, one Android phone, one iPad, one microphone, and one circular microphone array with a radius of 5cm. The red arrow in figure 1 points to channel 0 of microphone arrays.

During data collection, each speaker visits three times with a gap of 7-15 days. In the first visit (F), the noise sources include an electric fan and the TV broadcast or the office ambient noise. The recording environment of the second visit (S) is quiet. In the third visit (T), the electric fan is the only noise.

Table 1: *The summary of the FFSVC20 challenge data*

Utt ID	Content	Noise
001-030	*ni hao mi ya* (text-dependent)	F: TV/Office + electric fan T: electric fan
091-	text independent	S: quiet

In each visit, more than 300 utterances for each speaker are recorded. The first 30 utterances are of fixed content: *'ni hao mi ya'* in Mandarin Chinese. The next 60 utterances are defined as semi-text-dependent, in which the text content is *'ni hao mi ya'* followed by some random text. The remaining utterances are text-independent. All speech recordings are in Mandarin Chinese.

2.2. FFSVC 2020 Challenge Dataset

The FFSVC 2020 challenge dataset is part of the DMASH dataset. It includes the recordings from the close-talk microphone, the iPhone at 25cm distance, and three randomly selected circular microphone arrays. For the circular microphone arrays, only four recording channels are used. Under this data selection protocol, each utterance have 14 (1 + 1 + 4 × 3) recording channels.

In FFSVC 2020, the training partition includes 120 speakers and the development partition includes 35 speakers. Additionally, any publicly open and freely accessible dataset shared on openslr before Feb. 1st, 2020 can be used for training[1].

Table 1 shows the details of the challenge data. More information about the dataset can be found in [26].

For each task, the evaluation data includes 80 speakers. There is no overlapping among the speakers in the training, development, and evaluation sets for each task. Moreover, there is no overlapping among the evaluation data of the three tasks. Recordings from the iPhone at 25cm distance are selected for enrollment. For testing, one microphone array is used in task 1 and task 2; 2-4 microphone arrays are randomly selected in task 3. For each true trial, the enrollment and the testing utterance are from different visits of the same speaker.

3. The Baseline System

3.1. Data Processing

3.1.1. Data augmentation

To improve the robustness and generalization of the deep speaker network, we used pyroomacoustics toolkit [27] to simulate the room acoustic and generate far-field training data. The room width was randomly set between six to eight meters, and the locations of the speaker, noise, and microphones were also randomly distributed. The noise sources were from MU-SAN dataset [28], and the signal-to-noise-ratio (SNR) was between 0 to 20 dB.

3.1.2. Voice activity detection

Two voice activity detection (VAD) methods were explored in the systems. The first one was the energy-based VAD. The second method was the gradient boosting algorithm-based voice activity detection (GVAD) [29] for the far-field speeches. GVAD is a classifier that separates the speech segments from

---

[1]Dataset published on openslr before SLR85, including SLR85.

Table 2: *The ResNet-34 network architecture,* **C***(kernal size, stride) denotes the convolutional layer,* **S***(kernal size, stride) denotes the shortcut convolutional layer,* [·] *denotes the residual block.*

Layer	Output Size	Structure
Conv1	$32 \times 64 \times L$	$\mathbf{C}(3 \times 3, 1)$
Residual Layer 1	$32 \times 64 \times L$	$\begin{bmatrix} \mathbf{C}(3 \times 3, 1) \\ \mathbf{C}(3 \times 3, 1) \end{bmatrix} \times 3$
Residual Layer 2	$64 \times 32 \times \frac{L}{2}$	$\begin{bmatrix} \mathbf{C}(3 \times 3, 2) \\ \mathbf{C}(3 \times 3, 1) \\ \mathbf{S}(1 \times 1, 2) \end{bmatrix} \begin{bmatrix} \mathbf{C}(3 \times 3, 1) \\ \mathbf{C}(3 \times 3, 1) \end{bmatrix} \times 3$
Residual Layer 3	$128 \times 16 \times \frac{L}{4}$	$\begin{bmatrix} \mathbf{C}(3 \times 3, 2) \\ \mathbf{C}(3 \times 3, 1) \\ \mathbf{S}(1 \times 1, 2) \end{bmatrix} \begin{bmatrix} \mathbf{C}(3 \times 3, 1) \\ \mathbf{C}(3 \times 3, 1) \end{bmatrix} \times 5$
Residual Layer 4	$256 \times 8 \times \frac{L}{8}$	$\begin{bmatrix} \mathbf{C}(3 \times 3, 2) \\ \mathbf{C}(3 \times 3, 1) \\ \mathbf{S}(1 \times 1, 2) \end{bmatrix} \begin{bmatrix} \mathbf{C}(3 \times 3, 1) \\ \mathbf{C}(3 \times 3, 1) \end{bmatrix} \times 2$
Encoding	512	Global Statistics Pooling
Embedding	128	Fully Connected Layer
Classifier	10544	Fully Connected Layer

Table 3: *The ResNet-50 network architecture. The meaning of parameters in this table can be referred to the caption of Table 2.*

Layer	Output Size	Structure
Conv1	$64 \times 128 \times L$	$\mathbf{C}(7 \times 7, 2)$
Residual Layer 1	$256 \times 64 \times \frac{L}{2}$	$\begin{bmatrix} \mathbf{C}(1 \times 1, 1) \\ \mathbf{C}(3 \times 3, 2) \\ \mathbf{S}(1 \times 1, 1) \end{bmatrix} \times 3$
Residual Layer 2	$512 \times 32 \times \frac{L}{4}$	$\begin{bmatrix} \mathbf{C}(1 \times 1, 1) \\ \mathbf{C}(3 \times 3, 2) \\ \mathbf{S}(1 \times 1, 1) \end{bmatrix} \times 4$
Residual Layer 3	$1024 \times 16 \times \frac{L}{8}$	$\begin{bmatrix} \mathbf{C}(1 \times 1, 1) \\ \mathbf{C}(3 \times 3, 2) \\ \mathbf{S}(1 \times 1, 1) \end{bmatrix} \times 6$
Residual Layer 4	$2048 \times 8 \times \frac{L}{16}$	$\begin{bmatrix} \mathbf{C}(1 \times 1, 1) \\ \mathbf{C}(3 \times 3, 2) \\ \mathbf{S}(1 \times 1, 1) \end{bmatrix} \times 5$
Encoding	2048	Global Statistics Pooling
Embedding	1024	Fully Connected Layer
Classifier	2447	Fully Connected Layer

the non-speech segments. The training data of GVAD was the simulated far-field speeches from the AISHELL-1 dataset (SLR33) [30], as described in Section 3.1.1. The 'speech' and 'non-speech' labels are generated with an energy-based VAD on the original clean data of SLR33. All the far-field speeches of FFSVC 2020 dataset are processed with the trained GVAD before testing.

3.2. Acoustic Feature Extraction

Audios were resampled to 16,000 Hz and pre-emphasized before feature extraction. Two acoustic features were used: (1) 64-dimensional log Mel-filterbank energies with a frame length of 25ms and hop size of 10ms and (2) 30-dim MFCCs. The former features were used for ResNet-34 and the latter features were used for ResNet-50. The extracted features were mean-normalized before feeding into the deep speaker network.

3.3. Deep Speaker Embedding

ResNet based networks were applied for FFSVC. Two different ResNet architectures were investigated: (1) ResNet-34 and (2) ResNet-50.

3.3.1. ResNet-34

The network structure contains three main components: a front-end pattern extractor, an encoding layer, and a back-end classifier. The ResNet-34 structure [31] is adopted as the front-end pattern extractor. It learns a frame-level representation from the input spectral features. The global statistics pooling (GSP) layer is then used as the encoder layer to compute the mean and standard deviation of the input frame-level feature sequence. The GSP layer outputs an utterance-level representation with speaker information. A fully-connected layer with a classification output layer then processes the utterance-level representation. Each unit in the output layer is represented as a target speaker identity. All the components in the pipeline are jointly learned with cross-entropy loss. The detailed configuration of the neural network is in Table 2.

We pre-trained the deep speaker network with larger scale text-independent mix-dataset (close-talk and its simulation data). The pre-training data contained 10,554 speakers, including SLR33, SLR38, SLR47, SLR49, SLR62, and SLR68 from openslr.org. In the pre-training stage, the model was trained for 50 epochs with an initial learning rate of 0.1. The learning rate was divided by ten every 20 epochs. The network was optimized by stochastic gradient descent.

According to previous works, fine-tuning was an effective transfer learning approach for far-field ASV [16]. In task 1 and task 3, the fine-tuning data is SLR85 dataset and the first 30 utterances of FFSVC 2020 training dataset. The remaining FFSVC 2020 training dataset is used to fine-tune the model for task 2. To prevent overfitting during fine-tuning, data augmentation is also employed to simulate the far-field data for the clean close-talk channel. The real and simulated far-field data jointly fine-tune the pre-trained model. The learning rate is set to 0.001 when fine-tuning.

3.3.2. ResNet-50

The network structure of ResNet-50 is similar as ResNet-34 and also composed of three main components described in Section 3.3.1. The details of configuration is shown in Table 3.

SLR33, SLR38, SLR62, SLR82, SLR85 from openslr.org and FFSVC training set were used for training. The training data contained 2,447 speakers. The model was trained for 25 epochs with an initial learning rate of 0.1. The learning rate was divided by ten every 5 epochs. The network was optimized by stochastic gradient descent.

3.4. Back-end Scoring

The cosine distance scoring (CDS) and probabilistic linear discriminant analysis (PLDA) [32, 33] served as the back-end scoring methods.

Table 4: *Performance of the speaker verification systems. "Model" represents the types of deep speaker embedding network; "Scoring" represents the back-end scoring; "DA" represents whether performing data augmentation; "Enrollment" represents whether some methods were applied for enrollment utterances; "Test" represents the types of utterances selected as test for scoring; "Single" represents that utterance from one channel of microphone array(s) is selected as test for scoring; "Multi" represents that the utterances of all channels from the microphone array(s) are used as test for scoring.*

| | | | | | | Development Set | | | | | | Evaluation Set | | | | | |
| | | | | | | Task1 | | Task2 | | Task3 | | Task1 | | Task2 | | Task3 | |
ID	Model	Scoring	DA	Enrollment	Test	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER
1	ResNet50	CDS	N	–	Multi	0.86	8.04%	0.97	10.96%	0.88	7.22%	0.87	9.93%	0.95	11.87%	0.86	10.68%
2	ResNet34	CDS	Y	–	Single	0.64	6.30%	0.65	6.23%	0.64	5.82%	0.71	7.02%	0.72	6.93%	0.68	7.78%
3	ResNet34 (Baseline System)	CDS	Y	–	Multi	**0.57**	6.01%	**0.58**	5.83%	**0.59**	5.42%	**0.62**	6.37%	**0.66**	6.55%	**0.64**	7.18%
4	ResNet34	PLDA	Y	–	Multi	0.58	5.92%	0.60	5.69%	0.61	5.36%	0.63	6.28%	0.67	6.48%	0.67	7.10%
5	ResNet34	CDS	Y	EDA	Multi	0.60	5.87%	0.61	5.61%	0.60	5.33%	0.64	6.23%	0.68	6.36%	0.71	7.03%

3.5. Enrollment Data Augmentation

In far-field speaker verification, the mismatch between enrollment and testing utterances generally exists due to the different recording environments. Data augmentation on enrollment utterances is proven to be effective in reducing this mismatch [25]. In this paper, instead of randomly simulating the far-field enrollment data, we used the background noise of the testing utterance to perform enrollment augmentation. Specifically, a GVAD was adopted to detect the non-speech parts of the testing utterance for each trial. These non-speech parts were used as the background noise to get a simulated enrollment utterance. The speaker embeddings from the simulated and original utterance were equally weighted to get the final enrollment embedding.

4. Experiment Results

Table 4 shows the performance of the speaker verification systems under development and evaluation condition, respectively. The performance metrics for FFSVC 2020 are equal error rate (EER) and minimum detection cost function (minDCF) with $P_{target} = 0.01$. The minDCF is used as primary metric.

Experimental results of System ID 1 and ID 3 in Table 4 show that larger scale of training data and data augmentation could significantly improve the performance while ResNet34 is shallower than ResNet50.

During testing, different channels from the microphone array(s) are equally weighted at the embedding level before scoring(ID 3 in table 4). One channel of the microphone array(s) is selected as single-channel testing for comparison (ID 2 in table 4). The results of enrollment data augmentation (EDA) are also given. Additional attention should be paid to our results in metric: we observed that enrollment augmentation with PLDA can improve the EER, but not the minDCF. Finally, the embedding level averaging model(ID 3) which achieves the best single-system results on evaluation and development datasets is selected as baseline system for FFSVC 2020 challenge.

5. Conclusions

The primary purpose of the FFSVC 2020 is to investigate how well the speaker verification technology processes the real-world audio data, especially for the far-field distributed microphone arrays. The challenge data will be released as a large scale speech database after the competition. This paper also provides the description of the baseline system. We believe that this challenge and the published corpus will promote the advancement of research and technology development in far-field speaker recognition.

6. Acknowledgements

7. 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] 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. 74–81.

[3] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. INTERSPEECH*, 2018, pp. 3573–3577.

[4] 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. 6738–6746.

[5] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[6] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[9] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[10] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[11] L. Mosner, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and beamforming in far-field speaker recognition," in *Proc. ICASSP*, 2018, pp. 5254–5258.

[12] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[13] D. Cai, X. Qin, W. Cai, and M. Li, "The DKU System for the Speaker Recognition Task of the 2019 VOiCES from a Distance Challenge," in *Proc. INTERSPEECH*, 2019, pp. 2493–2497.

[14] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, G. Lavrentyeva, V. Volokhov, and A. Kozlov, "STC speaker recognition systems for the voices from a distance challenge," *arXiv preprint arXiv:1904.06093*, 2019.

[15] P. Matejka, O. Plchot, H. Zeinali, L. Mosner, A. Silnova, L. Burget, O. Novotny, and O. Glembek, "Analysis of BUT submission in far-field scenarios of voices 2019 challenge," in *Proc. INTERSPEECH*, 2019, pp. 2448–2452.

[16] 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. INTERSPEECH*, 2019, pp. 4045–4049.

[17] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training Multi-Task Adversarial Network for Extracting Noise-Robust Speaker Embedding," in *Proc. ICASSP*, 2019, pp. 6196–6200.

[18] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial Speaker Verification," in *Proc. ICASSP*, 2019, pp. 6216–6220.

[19] 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.

[20] Y. Shi, Q. Huang, and T. Hain, "Robust speaker recognition using speech enhancement and attention model," *arXiv preprint arXiv:2001.05031*, 2020.

[21] S. Shon, H. Tang, and J. Glass, "VoiceID Loss: Speech Enhancement for Speaker Verification," in *Proc. INTERSPEECH*, 2019, pp. 2888–2892.

[22] F. Zhao, H. Li, and X. Zhang, "A Robust Text-independent Speaker Verification Method Based on Speech Separation and Deep Speaker," in *Proc. ICASSP*, 2019, pp. 6101–6105.

[23] D. Cai, X. Qin, and M. Li, "Multi-Channel Training for End-to-End Speaker Recognition Under Reverberant and Noisy Environment," in *Proc. INTERSPEECH*, 2019, pp. 4365–4369.

[24] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. R. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (VOICES) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[25] X. Qin, H. Bu, and M. Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *Proc. ICASSP*, 2020, pp. 7609–7613.

[26] X. Qin, M. Li, H. Bu, R. K. Das, W. Rao, S. Narayanan, and H. Li, "The FFSVC 2020 Evaluation Plan," *arXiv preprint arXiv:2002.00387*, 2020.

[27] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 6638–6648.

[30] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and A speech recognition baseline," *arXiv preprint arXiv:1709.05522*, 2017.

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

[32] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. of Odyssey: Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010.

[33] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.