Far-Field End-to-End Text-Dependent Speaker Verification based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation

Xiaoyi Qin[1,2], *Danwei Cai*[1], *Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China

ming.li369@dukekunshan.edu.cn

Abstract

In this paper, we focus on the far-field end-to-end text-dependent speaker verification task with a small-scale far-field text dependent dataset and a large scale close-talking text independent database for training. First, we show that simulating far-field text independent data from the existing large-scale clean database for data augmentation can reduce the mismatch. Second, using a small far-field text dependent data set to fine-tune the deep speaker embedding model pre-trained from the simulated far-field as well as original clean text independent data can significantly improve the system performance. Third, in special applications when using the close-talking clean utterances for enrollment and employing the real far-field noisy utterances for testing, adding reverberant noises on the clean enrollment data can further enhance the system performance. We evaluate our methods on AISHELL ASR0009 and AISHELL 2019B-eval databases and achieve an equal error rate (EER) of 5.75% for far-field text-dependent speaker verification under noisy environments.

Index Terms: speaker verification, text-dependent, far-field, transform learning, data augmentation

1. Introduction

In the past decade, the performance of automatic speaker verification (ASV) has improved dramatically. The i-vector based method [1] and the deep neural network (DNN) based methods [2, 3] have been widely used in telephone channel and close-talking scenarios.

Recently, smartphones and virtual assistants become very popular. People use pre-defined words to wake up the system. To enhance the security level and be able to provide preconized service, the wake-up words based text-dependent speaker verification is adapted to determine whether the wake-up speech is indeed uttered by the claimed speaker [4, 5, 6]. However, in many Internet of Things (IoT) applications, e.g., smart speakers and smart home devices, text-dependent speaker verification under far-field and complex environmental settings are still challenging due to the effects of room reverberation and various kinds of noises and distortions. To reduce the effects of room reverberation and environmental noise, various approaches with single channel microphone or multi-channel microphone array, have been proposed at different levels of the text independent ASV system. At the signal level, linear prediction inverse modulation transfer function [7] and weighted prediction error (WPE) [8, 9] methods have been used for dereverberation. DNN based denoising methods for single-channel speech enhancement [10, 11, 12, 13] and beamforming for multi-channel speech enhancement [8, 14, 15] have also been explored for ASV system under complex environments. At the feature level, sub-band Hilbert envelopes based features [16, 17, 18], warped minimum variance distortionless response (MVDR) cepstral coefficients [19], blind spectral weighting (BSW) based features [17], power-normalized cepstral coefficients (PNCC) [20] and DNN bottleneck features [21] have been applied to ASV system to suppress the adverse impacts of reverberation and noise. At the model level, reverberation matching with multi-condition training models have also been successfully employed within the universal background model (UBM) or ivector based front-end systems [22, 23]. Multi-channel i-vector combination for far-field speaker recognition is also explored in [24]. In back-end modeling, multi-condition training of probabilistic linear discriminant analysis (PLDA) models was employed in i-vector system [25]. The robustness of deep speaker embeddings for far-field text-independent speech has also been investigated in [26, 27]. Finally, at the score level, score normalization [22] and multi-channel score fusion [28] have been applied in far-field ASV system to improve the robustness.

In this work, we focus on the far-field end-to-end text-dependent speaker verification task at the model level. Previous studies [4, 5, 6] on end-to-end deep neural network based text-dependent speaker verification directly use large-scale text dependent database to train the systems. However, in real-world applications, people may want to use customized wake-up words for speaker verification, and different smart home devices may have different wake-up words even for products from the same company. Hence collecting a large-scale far-field text-dependent speech database for each new or customized wake-up words may not be possible. This motivates us to explore the transfer learning concept and use a small far-field text-dependent speech dataset to fine-tune the existing deep speaker embedding network trained from large-scale text independent speech databases, like NIST SRE databases or voxceleb [29, 30].

Furthermore, we propose a new research topic on far-field text-dependent speaker verification, which is to use the close-talking clean data for enrollment and employ the real far-field noisy utterances for testing. This scenario corresponds to the case that only one clean utterance recorded by cell phone is used to enroll the speaker for the smart home devices. In this work, we investigate an enrollment data augmentation scheme to reduce the mismatch and improve the ASV performance.

2. Corpora

2.1. Text-dependent corpora

The AISHELL-2019B-eval dataset[1] is an open source wake-up words speech database which includes 86 speakers.

A wake-up word of four Chinese characters “ni hao, mi ya” (“Hello, Mia” in English) is employed in the dataset. The av-

[1]https://www.aishelltech.com/aishell_2019B_eval

Table 1: *The details of the training, enrollment and testing data used in this study*

Data	AISHELL catalogue	Type	Quality	Speakers	Utterances	Hours
Train-TI-clean	ASR0009	text independent	clean close-talking	1997	984907	1000
Train-TD-mixed	subset of 2019B-eval	text dependent	clean close-talking + clean & noisy far-field	67	13400	37.22
Enroll-TD-far-field	subset of 2019B-eval	text dependent	clean far-field + noisy far-field	19	760	–
Enroll-TD-clean	subset of 2019B-eval	text dependent	clean close-talking	19	380	–
Test-TD-far-field	subset of 2019B-eval	text dependent	noisy far-field	19	1520	–
Test-TD-clean	subset of 2019B-eval	text dependent	clean close-talking	19	380	–

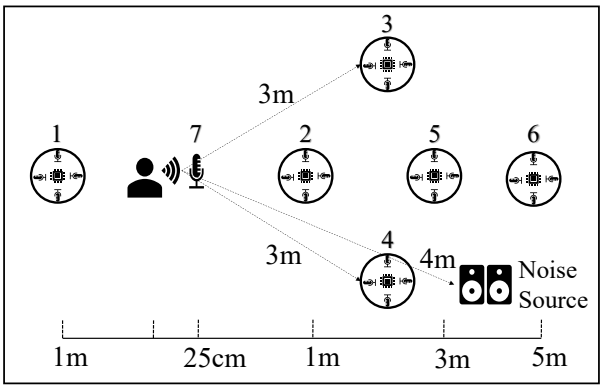

Figure 1: *The setup of the AISHELL-2019B-EVAL database*

erage duration for each wake-up utterance is around 1 second. During the recording process, seven recording positions were set in the real smart home environment and each speaker records 80 utterances, with the first 60 utterances in a quiet environment and the remaining 20 utterances in the noisy environment. The recording devices included six 16-channel circular microphone arrays (*mic 1* $\sim$ *6*) and one close-talking microphone (*mic 7*) for high-quality clean speech recording. TV and music inference sources were used to simulate the noisy condition. The details of the recording environment are shown in figure 1.

We split the 86 speakers in the dataset into subset A and B with 67 and 19 speakers, respectively. The subset A with 67 speakers is used for training, and subset B with 19 speakers is used for enrollment and testing.

For the training data with 67 speakers, we selected recordings from four channels (*channel 0, 4, 8, 12*) in each microphone array (*mic 1* $\sim$ *6*). Recordings from the high-quality microphone (*mic 7*) are also selected for the training data.

The remaining subset B from 19 speakers are used for enrollment and testing. Usually, the speech at a distance of more than 3-5 meters from the microphone is considered as far-field speech. Therefore, the collected speech of *mic 5* and *6* is regarded as far-field speech. Since we only focus on the single channel based far-field recognition in this work, we selected *channel 0* from *mic 5 and 6* as our enrollment and testing data.

In terms of the enrollment data, we randomly selected 20 out of the first 120 utterances in the quiet environment from the *mic 7* as the clean enrollment data (Enroll-TD-clean). Furthermore, with *channel 0* from *mic 5 and 6*, ten utterances in the quiet condition and ten utterances in the noisy condition were randomly selected as the far-field enrollment data (Enroll-TD-far-field).

Another 20 utterances different from the Enroll-TD-clean in *mic 7* are randomly selected as clean testing data (Test-TD-clean). The remaining utterances in the noisy condition of *channel 0, 4, 8, 12* in *mic 5 and 6* are used as the far-field testing data (Test-TD-far-field).

2.2. Text-independent corpora

The AISHELL-ASR0009[2] is a Chinese Mandarin speech recognition dataset. In this study, we use the high-quality microphone channel of the dataset, which contains 1,997 speakers with 984,907 close-talking utterances and around 1000 hours in total. The average duration of the utterance is 3.54s.

The details of the training, enrollment, and testing data used in this study are summarized in Table 1.

3. Methods

In this section, we describe the methods used for text-dependent speaker recognition. Firstly, front-end speech enhancement and dereverberation for far-field speech are employed. Then we describe the deep speaker embedding DNN for text-independent ASV. Moreover, we investigate the transfer learning strategy to adapt a text-independent DNN model to a text-dependent DNN model. Finally, to reduce the mismatch between enrollment and testing speech, we introduce enrollment data augmentation.

3.1. Front-end signal enhancement for far-field speech

To reduce the mismatch between the far-field testing data and clean close-talk training data, speech enhancement, and speech dereverberation are widely used to enhance the speech quality. We used DNN-based speech enhancement and weighted prediction error (WPE) dereverberation for enrollment and testing speech.

3.1.1. Speech enhancement

In this paper, DNN-based speech enhancement is used for speech denoising. Taking the network configuration in [31], the text-dependent DNN-based speech enhancement (SE) model is trained to estimate the ideal binary mask (IBM) for noisy speech. The clean channel (*mic 7*) and the noisy channel (*mic 3* $\sim$ *6*) of AISHELL-2019B-eval dataset is used to train the DNN SE model.

3.1.2. Deverberation

The weighted prediction error (WPE) algorithm is a successful algorithm to reduce late reverberation [32]. During the enrolling and testing, we use the single-channel WPE to dereverberate the sound with a dereverberation filter of 10 coefficients. The WPE codes are from `http://www.kecl.ntt.co.jp/icl/signal/wpe`.

[2] `http://www.aishelltech.com/jcsjnewls`

3.2. Deep speaker embedding system

3.2.1. Model architecture

Thanks to the fast development in the deep neural network, the superiority of deep speaker embedding systems have been shown in text-independent speaker recognition for closed talking [2, 3] and far-field scenarios [26, 27, 33]. In this paper, we adopt a deep speaker embedding system which is initially designed for the far-field text-independent speaker verification in our previous work, and details of the model architecture can be found in [33].

Our network structure contains three main components: a front-end pattern extractor, an encoding layer, and a back-end classifier. We build the front-end pattern extractor on the well known ResNet-18 architecture [34], which learns three-dimensional high-level descriptions for the given 64-dimensional raw Mel-filterbank energies. After the front-end ResNet, the output is a temporal representation of the input feature. To get the single utterance-level representation, we adopt a global average pooling (GAP) layer, which accumulates mean statistics along with the time-frequency axis. 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 in an end-to-end manner with a softmax classifier based cross-entropy loss.

After training, the utterance-level speaker embedding is extracted after the penultimate layer of the neural network for the given utterance. Cosine similarity serves as a back-end scoring method when testing.

3.2.2. Training data augmentation for far-field ASV

Typically, training data augmentation is often used to improve the robustness of the deep speaker embedding model. In this study, we augment the training data with reverberation and different kinds of noise to simulate the real-world far-field speech and reduce the mismatch between the training and testing data.

We use *pyroomacoustics* [35] to simulate the room acoustic based on room impulse response (RIR) generator using Image Source Model (ISM) algorithm. The width and length of the room size are randomly set to 4 to 12 meters with a height of 3 meters. A single microphone is randomly generated and randomly placed at the center, corner, or middle front of the room. Then the foreground speech source is positioned at 0.5, 1, 3, 5 or 8 meters from the microphone.

To simulate the noisy environment, we place the interference noise source at 0.5, 2, 4 meters from the microphone array with the signal-to-noise ratio (SNR) between 0 to 20 dB. There are four types of noise: ambient noise, music, television, and babble noise. Specifically, the ambient and the music noise are selected from the MUSAN dataset [36]. The television noise is generated with one music file and one speech file from MUSAN. The babble noise is constructed by mixing three speech files into one, which results in three overlapping voices simultaneously with the foreground speech.

3.3. Mixed training data with transfer learning

Collecting a text-dependent corpus with sufficient speakers for DNN speaker embedding system training is expensive. However, a text-dependent deep speaker embedding model trained with a small number of speakers is not able to learn discriminant speaker information and is very likely to overfit on the training data with few speakers.

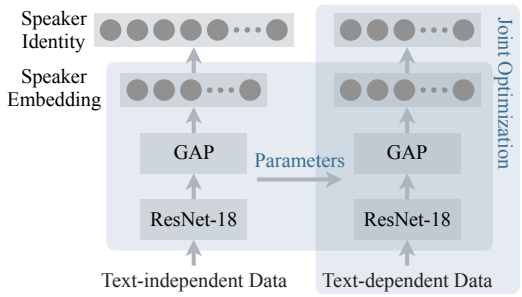

Figure 2: *Transfer the text-independent deep speaker embedding model to text-dependent model.*

Therefore, we investigate the transfer learning strategy to adopt a text-independent deep speaker embedding model to a text-dependent model. With transfer learning, the adapted text-dependent DNN model takes the advantages of the pre-trained model with a large number of speakers, without training the whole network from scratch. Specifically, after the text-independent deep speaker model is trained, transfer learning adapts the front-end local pattern extractor, the encoding layer, and the embedding extraction layer to the text-dependent task. The fine-tuning steps are as follows:

- Train the text-independent deep speaker model with a large amount of text-independent data with sufficient speakers;
- Retain all parameters of the model except for the output speaker classification layer; Replace the original output layer with a speaker classifier which classifies the speakers in the target text-dependent training data;
- Train the new model with the text-dependent data until it converges. The parameters of the front-end ResNet and the back-end classifier are jointly optimized.

Figure 2 shows the transfer learning process of the text-dependent deep speaker embedding model.

3.4. Enrollment data augmentation

In far-field speaker recognition, the enrollment speech and testing speech are usually in different environmental settings, which leads to the mismatch between enrollment and testing utterance. In this paper, we investigate the scenario corresponds to the case that only one clean utterance recorded by cell phone is used to enroll the speaker for all the smart home devices. To reduce the mismatch, we augment the enrollment speech with diverse environmental settings. Embedding level average fusion is adopted to enhance the enrollment template. The augmented enrollment speeches may cover different environmental settings of the testing speech.

We use the *RIR-generator* toolkit to generate far-field audio. The toolkit is based on image method, which is one of the most commonly used methods in generating synthetic room impulse response [37]. The codes are from `https://github.com/ehabets/RIR-Generator`.

4. Experiments

4.1. Baseline system and fine-tuned model

We train four deep speaker embedding systems with different training data and one with fine-tune training strategy. Table 2 shows the details of our experimental setup and results.

Table 2: *EER of different speaker embedding systems with three enrollment-testing scenarios. (Ori is the original far-field speech.)*

Model	Training data	Enrollment data	Testing data	Ori	+ SE	+ WPE
Text-independent	Train-TI-clean	Enroll-TD-clean	Test-TD-clean	3.38%	-	-
		Enroll-TD-clean	Test-TD-far-field	24.41%	18.45%	22.40%
		Enroll-TD-far-field	Test-TD-far-field	22.12%	16.55%	16.17%
Text-independent	Train-TI-clean + far-field simulation	Enroll-TD-clean	Test-TD-clean	3.38%	-	-
		Enroll-TD-clean	Test-TD-far-field	15.68%	16.48%	22.80%
		Enroll-TD-far-field	Test-TD-far-field	10.34%	12.53%	14.42%
Text-independent	Train-TI-clean + far-field simulation + Train-TD-mixed	Enroll-TD-clean	Test-TD-clean	4.39%	-	-
		Enroll-TD-clean	Test-TD-far-field	**14.33%**	11.92%	11.74%
		Enroll-TD-far-field	Test-TD-far-field	**9.17%**	11.32%	20.91%
Text-dependent	Train-TD-mixed	Enroll-TD-clean	Test-TD-clean	3.35%	-	-
		Enroll-TD-clean	Test-TD-far-field	16.33%	14.16%	26.21%
		Enroll-TD-far-field	Test-TD-far-field	12.86%	11.89%	14.58%
Text-dependent Fine-tuned	Train-TI-clean + far-field simulation + Train-TD-mixed	Enroll-TD-clean	Test-TD-clean	4.25%	-	-
		Enroll-TD-clean	Test-TD-far-field	**7.86%**	9.97%	15.01%
		Enroll-TD-far-field	Test-TD-far-field	**5.79%**	6.66%	6.67%

Several observations from the results are discussed in the following. Firstly, the speech enhancement and dereverberation algorithms can improve system performance when mismatches occur at training data, enrollment data, and testing data. But when using far-field speech for training and testing, the speech enhancement and dereverberation degrade the system performance, partly due to the mismatch between the training data (far-field speech) and the enhanced speech data. Secondly, simulating far-field text independent data from the existing large-scale clean database for data augmentation can increase the robustness of the deep speaker embedding model and improve the system performance. The system trained with text-independent data (clean + simulated data) outperforms the model trained within-domain text-dependent far-field data. Finally, the fine-tuned model achieves the best performance among all the systems. Comparing to the text-independent model directly trained with the same mixed training data, the fine-tuned model achieves 36.9% relative improvement in terms of equal error rate (EER) at far-field enrollment far-field testing. Comparing to the text-dependent model trained with text-dependent data, our fine-tuned model obtains 55.0% relative improvement in terms of EER.

4.2. Enrollment data augmentation

From table 2, the clean enrollment far-field testing scenarios always have a worse performance comparing to the far-field enrollment far-field testing scenarios, even when speech enhancement and dereverberation are applied. The main reason is the mismatch between the enrollment utterance and the testing utterance. We thus investigate enrollment data augmentation to compensate the mismatch between the enrollment utterance and the testing utterance. Using the *RIR-Generator* toolkit, we simulate far-field speech with different settings of t60 (time required to reduce the sound pressure level by 60dB is the reverberation time after the sound source stops sounding) and source-microphone distance. We augment the original enrollment utterance with different numbers of simulated far-field utterances. The simulated far-field enrollment utterances with the original enrollment utterance are averaged at embedding level. The experimental results on the fine-tuned text-dependent model are shown in table 3. The results show that the enrollment data augmentation can reduce the gap between the far-field enrollment far-field testing and the clean enrollment far-field testing.

Table 3: *EER of enrollment data augmentation*

Enrollment condition	EER
Clean utterance	7.86%
Real far-field utterance	5.79%
Clean utterance + 1 simulated far-field utterance	6.83%
Clean utterance + 5 simulated far-field utterances	6.66%
Clean utterance + 10 simulated far-field utterances	6.64%
Clean utterance + 20 simulated far-field utterances	6.60%

5. Conclusions

In this paper, we focus on far-field end-to-end text-dependent speaker verification. Firstly, we employ the transfer learning concept and use a small far-field text-dependent speech dataset to fine-tune the existing deep speaker embedding network trained from large-scale text in-dependent speech database. Also, in special applications when using the close-talking clean utterances for enrollment and employing the real far-field noisy utterances for testing, we augment the enrollment speech with simulated noisy far-field speech to reduce the mismatch between the enrollment and testing utterance. Further work includes extending the signal-channel far-field ASV to multi-channel microphone array far-field ASV.

6. Acknowledgement

This research was funded in part by the National Natural Science Foundation of China (61773413), Natural Science Foundation of Guangzhou City (201707010363), Six Talent Peaks project in Jiangsu Province (JY-074), Science and Technology Program of Guangzhou City (201903010040). We thank Weixiang Hu, Yu Lu, Zexin Liu, and Lei Miao from Huawei Digital Technologies Co., Ltd, China who provided insight and expertise that greatly assisted this research.

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

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

[3] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Odyssey*, 2018, pp. 74–81.

[4] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep Neural Networks for Small Footprint Text-dependent Speaker Verification," in *ICASSP*, 2014, pp. 4052–4056.

[5] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification," in *ICASSP*, 2016, pp. 5115–5119.

[6] S. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-End Attention based Text-dependent Speaker Verification," in *SLT*, 2016, pp. 171–178.

[7] B. J. Borgstrom and A. McCree, "The Linear Prediction Inverse Modulation Transfer Function (lP-IMTF) Filter for Spectral Enhancement, with Applications to Speaker Recognition," in *ICASSP*, 2012, pp. 4065–4068.

[8] L. Mosner, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and Beamforming in Far-Field Speaker Recognition," in *ICASSP*, 2018, pp. 5254–5258.

[9] 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.

[10] X. Zhao, Y. Wang, and D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[11] M. Kolboek, Z.-H. Tan, and J. Jensen, "Speech Enhancement Using Long Short-Term Memory based Recurrent Neural Networks for Noise Robust Speaker Verification," in *SLT*, 2016, pp. 305–311.

[12] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification," in *Interspeech*, 2016, pp. 2204–2208.

[13] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-End Speech Enhancement for Commercial Speaker Verification Systems," *Speech Communication*, vol. 99, pp. 101–113, 2018.

[14] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *ICASSP*, 2016, pp. 196–200.

[15] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[16] T. Falk and Wai-Yip Chan, "Modulation Spectral Features for Robust Far-Field Speaker Identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, 2010.

[17] S. O. Sadjadi and J. H. L. Hansen, "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 937–945, 2014.

[18] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature Normalization for Speaker Verification in Room Reverberation," in *ICASSP*, 2011, pp. 4836–4839.

[19] Q. Jin, R. Li, Q. Yang, K. Laskowski, and T. Schultz, "Speaker Identification with Distant Microphone Speech," in *ICASSP*, 2010, pp. 4518–4521.

[20] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coeffients (PNCC) for Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1315–1329, 2016.

[21] T. Yamada, L. Wang, and A. Kai, "Improvement of Distant-Talking Speaker Identification Using Bottleneck Features of DNN," in *Interspeech*, 2013, pp. 3661–2664.

[22] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation Matching for Speaker Recognition," in *ICASSP*, 2008, pp. 4829–4832.

[23] A. R. Avila, M. Sarria-Paja, F. J. Fraga, D. O'Shaughnessy, and T. H. Falk, "Improving the Performance of Far-Field Speaker Verification Using Multi-Condition Training: The Case of GMM-UBM and i-Vector Systems," in *Interspeech*, 2014, pp. 1096–1100.

[24] A. Brutti and A. Abad, "Multi-Channel i-vector Combination for Robust Speaker Verification in Multi-Room Domestic Environments," in *Odyssey*, 2016, pp. 252–258.

[25] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *ICASSP*, 2012, pp. 4257–4260.

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

[27] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, and M. Graciarena, "Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings," in *Interspeech*, 2018, pp. 1106–1110.

[28] M. Ji, S. Kim, H. Kim, and H.-S. Yoon, "Text-Independent Speaker Identification using Soft Channel Selection in Home Robot Environments," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 1, pp. 140–144, 2008.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A Large-Scale Speaker Identification Dataset," in *Interspeech*, 2017, pp. 2616–2620.

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

[31] Y. Xu, J. Du, L. Dai, and C. Lee, "An Experimental Study on Speech Enhancement based on Deep Neural Networks," *IEEE Signal Processing Letters*, pp. 65–68, 2014.

[32] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. 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.

[33] D. Cai, X. Qin, W. Cai, and M. Li, "Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment," in *Interspeech*, 2019.

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

[35] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *ICASSP*, 2018, pp. 351–355.

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

[37] J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *Journal of the Acoustical Society of America*, pp. 943–950.