Multi-Channel Training for End-to-End Speaker Recognition under Reverberant and Noisy Environment

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

Despite the significant improvements in speaker recognition enabled by deep neural networks, unsatisfactory performance persists under far-field scenarios due to the effects of the long range fading, room reverberation, and environmental noises. In this study, we focus on far-field speaker recognition with a microphone array. We propose a multi-channel training framework for the deep speaker embedding neural network on noisy and reverberant data. The proposed multi-channel training framework simultaneously processes the time-, frequency- and channel-information to learn a robust deep speaker embedding. Based on the 2-dimensional or 3-dimensional convolution layer, we investigate different multi-channel training schemes. Experiments on the simulated multi-channel reverberant and noisy data show that the proposed method obtains significant improvements over the single-channel trained deep speaker embedding system with front end speech enhancement or multi-channel embedding fusion.

Index Terms: speaker recognition, far-eld microphone array, multi-channel training, deep embeddings

1. Introduction

Automatic speaker verification (ASV) refers to accept or reject a claimed speaker by analyzing the speech. It is widely used in many real-world biometric authentication applications such as call center, mobile payment systems, personalized services of smart speakers and so on.[1].

In the past decade, the performance of speaker recognition has improved significantly. The i-vector based method [2] and the deep neural network (DNN) based methods [3, 4] have promoted the development of speaker recognition technology in telephone channel and closed talking scenarios. However, speaker recognition under far-field and complex environmental settings is still challenging due to the effects of the long range fading, room reverberation, and complex environmental noises. Speech signal propagating in long range suffers from fading, absorption and reflection by various objects, which change the pressure level at different frequencies and degrade the signal quality [5]. Reverberation includes early reverberation and late reverberation. Early reverberation (i.e., reflections within 50 to 100 ms after the direct wave arrives at the microphone) can improve the received speech quality, while late reverberation will damage speech. The adverse effects of reverberation on speech signal includes smearing spectro-temporal structures, amplifying the low-frequency energy, and flattening the formant transitions, etc. [6]. Also, the complex environmental noises "fill in" regions with low speech energy in the time-frequency plane and blur the spectral details [5]. These effects result in the loss of speech intelligibility and speech quality, imposing great challenges in far-field speaker recognition and far-field speech recognition.

To compensate the adverse impacts of room reverberation and environmental noise, various approaches, based on single-channel microphone or multi-channel microphone array, have been proposed at different stages of the 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 [9, 14, 15] have also been investigated for ASV under complex environment. At 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 [20], power-normalized cepstral coefficients (PNCC) [21, 22] and DNN bottleneck features [23] 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 been successfully employed within the universal background model (UBM) or i-vector based front-end systems [24, 25]. Multi-channel i-vector combination for far-field speaker recognition is also explored in [26]. In back-end modeling, multi-condition training of probabilistic linear discriminant analysis (PLDA) models was employed in i-vector system [27]. The robustness of deep speaker embeddings for far-field speech has also been investigated in [22, 28]. Finally, at the score level, score normalization [24] and multi-channel score fusion [29, 30] have been applied in far-field ASV system to improve the robustness.

In this study, we focus on far-field speaker recognition at the model level. A multi-channel training framework based on the state-of-the-art deep speaker embedding network is used for far-field speaker recognition under the reverberant and noisy environment with a multi-channel microphone array. The multi-channel training framework utilizes the information carried out by multiple speech observations at different spatial locations and simultaneously processes the time-, frequency- and channel-information to learn a robust deep speaker embedding. Based on 2-dimensional (2D) or 3-dimensional (3D) convolution layer, we investigate three different multi-channel training schemes: 2D convolution with multi-channel 2D input features, 3D convolution with 3D input features, and incorporating 3D convolution with 2D convolution. To the best of our knowl-

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 also thank Weixiang Hu, Yu Lu, Zexin Liu and Lei Miao from Huawei Digital Technologies Co., Ltd, China.

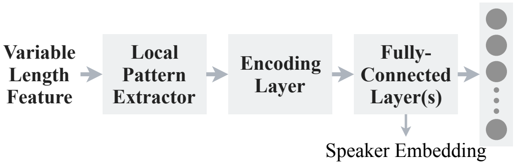


Figure 1: *Deep speaker embedding framework*

edge, this study is the first work to investigate the DNN speaker embedding system for reverberant and noisy speech from a microphone array with multi-channel training framework.

2. Deep Speaker Embedding

In this section, we describe the deep speaker embedding framework. As demonstrated in figure 1, it consists of a local pattern extraction network, an utterance-level encoding layer, and fully-connected layers for speaker embedding and speaker classification.

To simulate the real-world test utterance with variable length in the training stage, the network takes variable length feature sequence as input and produce utterance level result. Given the input feature sequence, the local pattern extractor learns high-level representations at the frame-level. Typically, the local pattern extractor can be a convolutional neural network (CNN) [4] or a time-delayed neural network (TDNN) [3]. After the front-end local pattern extractor, the output is still a temporal representation of the input feature. An encoding layer is then applied on top of the frame level representations to aggregate them into an utterance level representation. Several encoding methods has been investigated under the deep speaker embedding framework. The most common one is the average pooling layer, which aggregates the statistics (i.e., mean, or mean and standard deviation) over the whole utterance [3, 4]. Self-attentive pooling layer [31], learnable dictionary encoding (LDE) layer [32], long-short-term memory (LSTM) layer [33], dictionary-based NetVLAD layer [34, 35] also have been proposed to serve the encoding layers. The utterance level representation after the encoding layer is further processed through a fully connected layer followed by a speaker classifier.

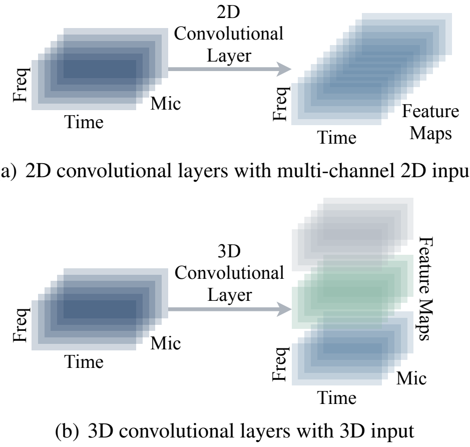
After training, the utterance level speaker embedding can be extracted after the penultimate layer of the neural network for the given variable-length feature sequence.

In this method, we adopt a residual convolutional neural network (ResNet) [36] as the local pattern extractor. For a given feature sequence, the ResNets learned descriptions are a three-dimensional tensor block of shape $C \times H \times W$, where C denotes the number of channels, H and W denotes the height and width of the feature maps. 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. Given feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, the output of GAP is a fixed-dimensional utterance-level representation $\mathbf{V} = [v_1, v_2, \cdots, v_C]$, where v_c is

$$v_c = \frac{1}{H \times W} \sum_{i=1}^{i=H} \sum_{j=1}^{j=W} \mathbf{F}_{c,i,j} \tag{1}$$

3. Multi-Channel Training

Given the microphone array with M channels, the spectro-temporal feature for recording channel m can be represented as $\mathbf{X}_m \in \mathbb{R}^{F\times T}$, where F is the feature dimension, and T is the number of time frame. The features of multi-channel microphone array utterance can be seen as either a multi-channel 2D features or a 3D feature representation $\mathbf{X} \in \mathbb{R}^{M\times F\times T}$.



(a) 2D convolutional layers with multi-channel 2D input

(b) 3D convolutional layers with 3D input

Figure 2: *2D and 3D convolutional layers for multi-channel training (Freq denotes frequency, Mic denotes microphone array channels)*

The feature representation is then fed into the DNN speaker embedding network with multi-channel input. In this paper, we explore three multi-channel training schemes based on the deep speaker embedding network.

3.1. 2D CNN with Multi-Channel 2D Features

Given the multi-channel 2D features, the convolutional layer with 2-dimensional kernel takes $\mathbf{X}$ as M 2D feature planes and produces the output feature maps of $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of output feature planes, H and W denotes the height and width of the feature maps. Formally, the c^{th} feature map of $\mathbf{F}$ can be describe as

$$\mathbf{F}_c = \sum_{m=0}^{M} \mathbf{K}(c,m) \star \mathbf{X}_m \tag{2}$$

where $\mathbf{K}(c,m)$ is the 2D filter weights for input channel m and output channel c, and $\star$ is the valid 2D cross-correlation operator.

In this study, the first convolutional layer of the ResNet is designed to receive multiple channel features. With 2D convolution, the way how multi-channel training work is the same as processing three color channel picture in computer vision.

3.2. 3D CNN with 3D Features

The second scheme for multi-channel training is the use of 3D convolutional layers. 3D CNN has been applied for far-field multi-channel speech recognition in [37]. The 3D convolutional layer receives 4-dimensional input feature maps with size $C \times D \times H \times W$, where C is the number of feature maps, D, H, W are the depth, height and width of the feature map respectively. The output $\mathbf{F}_{\text{out}}$ can be defined as

$$\mathbf{F}_{\text{out},c} = \sum_{k=0}^{C_{\text{in}}} \mathbf{K}(c,k) \star \mathbf{F}_{\text{in},k} \tag{3}$$

where $\mathbf{K}(c,k)$ is the 3D filter weights for input channel k and output channel c, $\star$ is the valid 3D cross-correlation operator, $\mathbf{F}_{\text{out},c}$ is the c^{th} feature map of the output feature and $\mathbf{F}_{\text{in},k}$ is the k^{th} feature map of the input feature.

Figure 2 shows the difference between the 2D and 3D convolutional layer. The 3D convolution retains the channel axis and keeps the channel information in the whole CNN, while the 2D convolution aggregates the channel information together into the 2D feature maps after the first convolutional layer.

In this method, all the 2D convolutional layers in the original ResNet are replaced by the 3D convolutional layers. In this way, the 3D ResNets learned descriptions are a 4-dimensional feature maps $\mathbf{F} \in \mathbb{R}^{C\times D\times H\times W}$. We thus modify the GAP layer to aggregate the mean statistics along the time-, frequency- and channel-axis, the output $\mathbf{V} = [v_1, v_2, \cdots, v_C]$ can be represented as

$$v_c = \frac{1}{D\times H\times W}\sum_{i=1}^{i=D}\sum_{j=1}^{j=H}\sum_{k=1}^{k=W}\mathbf{F}_{c,i,j,k} \tag{4}$$

3.3. Incorporate 3D CNN with 2D CNN

As stated before, the 3D convolution retains the channel axis in the whole CNN, while the 2D convolution drops the channel axis after the first convolutional layer. However, using 3D convolutional layers in ResNet may greatly increase the model size. This motivates us to incorporate the 3D CNN with 2D CNN. To match the dimension between the 3D convolution feature maps (4D tensor) and 2D convolution feature maps (3D tensor), a 3D convolution layer with kernel size of $D_{\text{in}} \times 1 \times 1$ is adapted to covert the 4-dimensional feature maps into a 3-dimensional feature maps, where D_{in} is designed to match the channel size of the input feature maps. In this way, the channel axis of the 4-dimensional feature maps output has length 1, and it is then reshaped to 3-dimensional feature maps and fed into the 2D CNN.

4. Corpora and Data Simulation

4.1. Corpora

The AISHELL-ASR0009-[ZH-CN][1] is a Chinese Mandarin speech recognition dataset. In this study, we use the high-quality channel of the dataset, which contains 1,997 speakers with 984,907 close-talk utterances, for training and testing. The average duration of the utterance is 3.54s. We split the dataset into two parts, with 1947 speakers for training and 50 speakers for testing. In the testing set, 20 utterances from each speaker are selected for enrollment. The 1,000 (20×50) enrolling utterances with 24,001 testing utterances form the final trials, which contains 23,520,980 non-target trials and 480,020 target trials. For each trial, we only use one utterance for enrollment, and one utterance for testing.

4.2. Data Simulation

We use *pyroomacoustics* [38] to simulate the room acoustic based on 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 6-channel circular microphone array with a radius between 5 to 15 cm is randomly generated and randomly placed at the center, corner or middle front of the room. Then the foreground speech source is placed at 0.5, 1, 3, 5 or 8 meters from the microphone array.

Table 1: *The network architecture,* $\mathbf{C}$*(kernal size, stride) denotes the convulutional layer,* $[\cdot]$ *denotes the residual block.*

Layer	**Output Size**	**Structure**
Conv1	$16\times 64\times L$	$\mathbf{C}(3\times 3, 1)$
Residual Layer 1	$16\times 64\times L$	$\begin{bmatrix}\mathbf{C}(3\times 3,1)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}\times 2$
Residual Layer 2	$32\times 32\times \frac{L}{2}$	$\begin{bmatrix}\mathbf{C}(3\times 3,2)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}\begin{bmatrix}\mathbf{C}(3\times 3,1)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}$
Residual Layer 3	$64\times 16\times \frac{L}{4}$	$\begin{bmatrix}\mathbf{C}(3\times 3,2)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}\begin{bmatrix}\mathbf{C}(3\times 3,1)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}$
Residual Layer 4	$128\times 8\times \frac{L}{8}$	$\begin{bmatrix}\mathbf{C}(3\times 3,2)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}\begin{bmatrix}\mathbf{C}(3\times 3,1)\\ \mathbf{C}(3\times 3,1)\end{bmatrix}$
Encoding	128	Global Average Pooling
Embedding	256	Fully Connected
Output	1947	Fully Connected

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 [39]. 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. This results in three overlapping voices simultaneously with the foreground speech. To simulate the real training-test condition, we split the whole MUSAN dataset equally into two subsets. We use half of the noise to generate the training data, and the whole set to generate the test data.

5. Experimental Results

5.1. Single-Channel Training Results

For the i-vector system, the 20-dimensional MFCCs with their first and second derivatives are computed for training a 1024 component Gaussian Mixture Model-Universal Background Model (GMM-UBM) with full covariance. A single factor analysis is employed to extract 600-dimensional i-vectors [2]. Then, Gaussian probabilistic linear discriminant analysis (PLDA) with full rank is used for modeling and scoring [40]. The training data includes the whole clean training set. We also use simulated reverberant and noisy data (**ivector-AUG**) with clean data together to train PLDA.

For the deep speaker embedding systems, each audio is converted to 64-dimensional Mel-filterbank energies. The front-end local pattern extractor is based on the well known ResNet-18 architecture [36]. The detailed architecture is described in table 1. After training, the speaker embedding adopts cosine similarity for scoring. In the deep speaker embedding system with ResNet + GAP setting, a cosine similarity backend is sufficient to achieve good performance [4, 22]. For training data, the original clean speech is used to train the deep speaker embedding system. We also use simulated reverberant and noisy data (**DNN-AUG**) with clean data together to train models and to reduce the mismatch between training and testing.

In table 2, we report the equal error rate (EER) for single channel training conditions. The performance of the **best** and

[1] The dataset is available at `http://www.aishelltech.com/jcsjnewls`

Table 2: *EER for single channel training systems*

Testing Condition		**ivector**	**ivector-AUG**	**DNN**	**DNN-AUG**
Clean speech		1.77%	2.02%	1.39%	1.55%
Far-field speech	best	29.28%	17.21%	28.47%	5.89%
	worst	29.49%	17.54%	28.91%	5.97%
	fusion	27.74%	14.42%	27.01%	4.95%
+ WPE	best	24.51%	15.12%	24.83%	7.42%
	worst	24.87%	15.31%	25.11%	7.71%
	fusion	22.94%	12.94%	23.12%	6.17%
+ NN-GEV		21.61%	14.92%	26.18%	7.82%
+ BeamformIt		28.91%	17.28%	30.13%	6.07%
+ WPE + NN-GEV		17.77%	13.59%	21.36%	9.45%

worst channel as well as the **embedding level fusion** result for the multi-channel speech are reported. We use weighted prediction error (WPE) [8] for dereverberation, Generalized eigenvalue beamformer with DNN estimated power-spectral density masks (NN-GEV) [14] and BeamformIt tool with weighted delay-and-sum [41] for signal enhancement. We used the NN-GEV model trained by the authors of [8].

Although the performance of the i-vector and DNN embedding system trained with clean speech degrade severely for reverberant and noisy testing data, training the deep embedding system with clean and simulated far-field speech together (DNN-AUG) can significantly reduce the mismatch between training and testing, resulting in 79% reduction in terms of EER. Moreover, the WPE dereverberation, beamforming techniques, and the combination of them can improve the performance of the clean data trained systems. However, for the DNN-AUG system, the speech enhancement algorithms result in worse performance, partly due to the mismatch between the training data (clean and corrupted data) and the enhanced speech data.

5.2. Multi-Channel Training Results

Table 3 shows the EER for multi-channel training system. Speeches from six channels of the circular microphone array are jointly fed into the multi-channel DNN. Four kinds of multi-channel training framework are adopted:

- ResNet-18 2D: use 2D convolutional layers, the *Conv1* layer in table 1 has 6 input channels.
- ResNet-18 3D: all 2D convolutional layers in ResNet-18 are replaced by 3D convolutional layers.
- 3D conv + ResNet-18 2D: apply a 3D convolutional layer on top of the ResNet-18 2D architecture.

From the results, all the multi-channel training deep speaker embedding systems outperform the single-channel training DNN system with embedding level fusion as well as the i-vector system with dereverberation and denoising. Comparing to the single-channel DNN system, the performance of the ResNet-18 2D multi-channel training system achieves 18.18% reduction in terms of EER, and the ResNet-18 3D system obtains 28.08% reduction.

5.3. Model Size and System Performance

To investigate the relationship between the model size and the system performance, we train models with different parameters, and the results are presented in table 4. To be specific, the notations we use in table 4 are as follows:

Table 3: *EER for multi-channel training systems*

Training Condition	**Model & Testing Condition**	**EER**
Single-Channel	ResNet-18 embedding fusion	4.95%
Multi-Channel	ResNet-18 2D	4.05%
	ResNet-18 3D	3.56%
	3D conv + ResNet-18 2D	3.79%

Table 4: *Comparison of model size, real-time (RT) factor for embedding extraction, and system performance. Increment of model parameters and reduction of EER for each model comparing to single-channel trained ResNet-18 are also provided.*

System	**#Parameters**	**RT**	**EER**
Single-channel ResNet-18	1233k (-)	0.016×6	4.95% (-)
Single-channel ResNet-54	2804k (127%)	0.042×6	4.60% (7.1%)
ResNet-18 2D	1234k (.8‰)	0.016	4.05% (18%)
ResNet-54 2D	**2805k (127%)**	**0.043**	**3.46% (30%)**
ResNet-18 3D	2607k (111%)	0.543	3.56% (28%)
3D (16) + ResNet-18 2D	1236k (2.4‰)	0.016	4.07% (18%)
3D (32) + ResNet-18 2D	1240k (5.6‰)	0.017	3.93% (21%)
3D (64) + ResNet-18 2D	1246k (1.1%)	0.020	3.92% (21%)
3D (128) + ResNet-18 2D	1259k (2.1%)	0.024	4.00% (19%)
3D (256) + ResNet-18 2D	**1285k (4.2%)**	**0.038**	**3.79% (23%)**

- ResNet-54: 6 residual blocks for each residual layer is adopted instead of 2 residual blocks for each residual layer in ResNet-18.
- 3D (k) + ResNet-18 2D: apply a 3D convolutional layer with k output channels on top of the ResNet-18.

From table 4, we can see that with very little increase in model parameters, the multi-channel training framework can significantly improve the system performance comparing to the single-channel model. The multi-channel trained ResNet-18 2D model achieves 18% reduction in EER with only 0.8‰ increments in model parameters. With almost the same number of parameters, the single-channel trained ResNet-54 model obtains only 7.1% reduction in EER, while the ResNet-18 3D model obtains 28% performance gain.

Another benefit of using the multi-channel training framework is the extraction time for deep speaker embedding. The real-time factor (tested on one CPU core with 2.20GHz) for embedding extraction is given in table 4. The multi-channel training schemes have a lower computation time for it receives multi-channel input to extract one speaker embedding without extracting multi-channel embedding independently.

6. Conclusions

In this paper, we propose a multi-channel training framework within the deep speaker embedding network for speaker recognition under reverberant and noisy environment. The method receives the time-, frequency-, and spatial-information from the multi-channel input to learn a robust speaker embedding. With very little increase in model parameters, the proposed method significantly outperforms the i-vector system with front-end signal enhancement as well as the single-channel deep speaker embedding system. Future works include testing the multi-channel framework in real-world far-field data and exploring the multi-channel training with different input channels.

7. References

[1] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] 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.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "x-vectors: Robust DNN Embeddings for Speaker Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[4] 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: The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[5] M. Wolfel and J. McDonough, *Distant Speech Recognition*. John Wiley & Sons, Incorporated, 2009.

[6] P. Assmann and Q. Summerfield, "The Perception of Speech Under Adverse Conditions," in *Speech Processing in the Auditory System*, 2004, pp. 231–308.

[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 *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4065–4068.

[8] 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.

[9] L. Mosner, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and Beamforming in Far-Field Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5254–5258.

[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 *IEEE Spoken Language Technology Workshop*, 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 *Proceedings of the Annual Conference of the International Speech Communication Association*, 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 *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 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. Hansen, "Hilbert Envelope Based Features for Robust Speaker Identification Under Reverberant Mismatched Conditions," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5448–5451.

[18] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature Normalization for Speaker Verification in Room Reverberation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4836–4839.

[19] Q. Jin, R. Li, Q. Yang, K. Laskowski, and T. Schultz, "Speaker Identification with Distant Microphone Speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4518–4521.

[20] 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.

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

[22] 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 *Proceedings of the Annual Conference of the International Speech Communication Association*, 2019.

[23] T. Yamada, L. Wang, and A. Kai, "Improvement of Distant-Talking Speaker Identification Using Bottleneck Features of DNN," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2013, pp. 3661–2664.

[24] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation Matching for Speaker Recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4829–4832.

[25] 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 *Proceedings of the Annual Conference of the International Speech Communication Association*, 2014, pp. 1096–1100.

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

[27] 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 *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4257–4260.

[28] 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 *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 1106–1110.

[29] Q. Jin, T. Schultz, and A. Waibel, "Far-Field Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2023–2032, 2007.

[30] Mikyong Ji, Sungtak Kim, Hoirin Kim, and Ho-Sub 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.

[31] G. Bhattacharya, J. Alam, and P. Kenny, "Deep Speaker Embeddings for Short-Duration Speaker Verification," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 1517–1521.

[32] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A Novel Learnable Dictionary Encoding Layer for End-to-End Language Identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5189–5193.

[33] W. Cai, D. Cai, S. Huang, and M. Li, "Utterance-level end-to-end language identification using attention-based CNN-BLSTM," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[34] J. Chen, W. Cai, D. Cai, Z. Cai, H. Zhong, and M. Li, "End-to-end Language Identification using NetFV and NetVLAD," in *The 11th International Symposium on Chinese Spoken Language Processing*, 2018.

[35] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level Aggregation For Speaker Recognition In The Wild," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[37] S. Ganapathy and V. Peddinti, "3-D CNN Models for Far-Field Multi-Channel Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5499–5503.

[38] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 351–355.

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

[40] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2011, pp. 249–252.

[41] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.