INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The DKU System for the Speaker Recognition Task of the 2019 VOiCES from a Distance Challenge

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

In this paper, we present the DKU system for the speaker recognition task of the VOiCES from a distance challenge 2019. We investigate the whole system pipeline for the far-field speaker verification, including data pre-processing, short-term spectral feature representation, utterance-level speaker modeling, back-end scoring, and score normalization. Our best single system employs a residual neural network trained with angular softmax loss. Also, the weighted prediction error algorithms can further improve performance. It achieves 0.3668 minDCF and 5.58% EER on the evaluation set by using a simple cosine similarity scoring. Finally, the submitted primary system obtains 0.3532 minDCF and 4.96% EER on the evaluation set.

Index Terms: speaker recognition, far-field speech, deep ResNet, angular softmax, WPE

1. Introduction

In the past decade, the performance of speaker recognition has improved significantly. The i-vector based method [1] and the deep neural network (DNN) based methods [2, 3] 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 [4]. Reverberation includes eaarlay 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 degrade the speech quality. The adverse effects of reverberation on speech signal includes smearing spectro-temporal structures, amplifying the low-frequency energy, and flattening the formant transitions, etc. [5]. Also, the complex environmental noises “fill in” regions with low speech energy in the time-frequency plane and blur the spectral details [4]. 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 for the adverse impacts of room reverberation and environmental noise, various approaches have been proposed at different stages of the speaker recognition system. At the signal level, dereverberation [6], denoising [7, 8, 9, 10], and beamforming [11, 12] can be used for speech enhancement. At feature level, sub-band Hilbert envelopes based features [13, 14], warped minimum variance distortionless response (MVDR) cepstral coefficients [15], blind spectral weighting (BSW) based features [16] 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 has been successfully employed within the universal background model (UBM) or i-vector based front-end systems [17, 18]. In back-end modeling, multi-condition training of probabilistic linear discriminant analysis (PLDA) models were employed in i-vector system [19]. The robustness of deep speaker embeddings for far-field speech has also been investigated in [20]. Finally, at the score level, score normalization [17] and multi-channel score fusion [21, 22] have been applied in far-field ASV system to improve the robustness.

The “VOiCES from a Distance Challenge 2019” is designed to foster research in the area of speaker recognition and automatic speech recognition (ASR) with the special focus on single channel far-eld audio, under noisy conditions [23]. Our system pipeline consists of the following six main components, including data pre-processing, short-term spectral feature extraction, utterance-level speaker modeling, back-end scoring, score normalization, as well as fusion and calibration.

This paper is organized as follows: Section 2 describes the details of our submitted system. Section 3 clarifies the data usage, with experimental results and analysis. Conclusions are drawn in section 4.

2. System descriptions

2.1. Data pre-processing

2.1.1. Data augmentation

We adopt two kinds of data augmentation strategies. The first is the same as the x-vector system available at Kaldi Voxceleb recipe, which employs additive noises and reverberation. We also use *pyroomacoustics* [24] to simulate the room acoustic based on RIR generator using Image Source Model (ISM) algorithm. The microphones, distractors, and speech source are similar to the room settings presented in [25]. We use the music and noise part of the MUSAN dataset [26] to generate the television noise, and the ‘us-gov’ part to create babble noise.

For the systems described below, we use the Kaldi data augmentation strategy for the MFCC i-vector system and the TDNN x-vector system, and *pyroomacoustics* data augmentation strategy for the remaining systems.

2.1.2. Dereverberation

The weighted prediction error (WPE) algorithm is a successful algorithm to reduce late reverberation [6]. The method estimates the optimal dereverberation filter coefficients based on iterative optimization. During the enrolling and testing, we use the single-channel WPE to dereverberate the sound with a dere-

http://dx.doi.org/10.21437/Interspeech.2019-1435

verberation filter of 10 coefficients. The WPE codes are from `http://www.kecl.ntt.co.jp/icl/signal/wpe`.

2.2. Short-term spectral feature

Four features including Mel-frequency cepstral coefficient (MFCC), power-normalized cepstral coefficients (PNCC), Mel-filterbank energies (Mfbank) and gammatone-Filterbank energies (Gfbank) are adopted in our systems.

2.2.1. MFCC

Two kinds of MFCC features with a different number of cepstral filterbanks are adopted, which result in 20- and 30-dimensional MFCCs (MFCC-20 and MFCC-30). MFCC-20 is for the i-vector system, and MFCC-30 is for the TDNN x-vector system. Short-time cepstral mean subtraction (CMS) over a 3-second sliding window is applied. For the MFCC-20, their first and second derivatives are computed before applying the CMS.

2.2.2. PNCC

PNCC has proved to be more robust in various types of additive noise and reverberant environments compared to MFCC in ASR [27]. The major features of PNCC processing include the use of a power-law nonlinearity that replaces the traditional log nonlinearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppress background excitation, and a module that accomplishes temporal masking [27]. 20-dimensional PNCC are extracted using a 25 ms window with 10 ms shifts. First and second derivatives are computed before applying CMS.

2.2.3. Log Mel-filterbank energies

Each audio is converted to 64-dimensional log Mel-filterbank energies with cepstral filterbanks ranging from 20 to 7600 Hz (Mfbank-16k). We also downsample the audio to 8000 sample rate and use cepstral filter banks within the range of 20 to 3800 Hz to calculate Mfbank-8k features. A short-time cepstral mean subtraction is applied over a 3-second sliding window.

2.2.4. Gammatone-Filterbank Energies

Gammatone filters are approximations to the filtering system of human ear [28]. The Gammatone filterbanks are selected within the range of 50 to 8000 Hz to compute the 64-dimensional Gammatone-filterbank energies. Short-time CMS is then applied over a 3-second sliding window.

2.3. Utterance-level speaker modeling

We extract the utterance-level speaker embeddings from three state-of-the-art modelings, including the i-vector system [1], the TDNN x-vector system [2], and the deep ResNet system [3].

2.3.1. i-vector

We train two i-vector systems on the MFCC-20 and PNCC features respectively. The extracted 60-dimensional features are used to train a 2048 component Gaussian mixture model-universal background model (GMM-UBM) with full covariance matrices. Then zero-order and first-order Baum-Welch statistics are computed on the UBM for each recording's MFCC feature, and single factor analysis is employed to extract i-vectors with 600 dimensions [1].

2.3.2. TDNN x-vector

The x-vector system is developed by adapting the Kaldi Voxceleb recipe. For the x-vector extractor, a DNN is trained to discriminate speakers in the training set. The first five timed delayed layers operate at frame-level. Then a temporal statistics pooling layer is employed to compute the mean and standard deviation over all frames for an input segment. The resulted segment-level representation is then fed into two fully connected layers to classify the speakers in the training set. After training, speaker embeddings are extracted from the 512-dimensional affine component of the first fully connected layer.

2.3.3. Deep ResNet

We follow the deep ResNet system as described in [29, 3, 30], and we increase the widths (number of channels) of the residual blocks from $\{16, 32, 64, 128\}$ to $\{32, 64, 128, 256\}$. The network architecture contains three main components: a front-end ResNet, a pooling layer, and a feed-forward network. The front-end ResNet transforms the raw feature into a high-level abstract representation. The subsequent pooling layer outputs a single utterance-level representation. Specifically, means statistics are accumulated for each feature map, and finally 256-dimensional utterance-level representation is produced. 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 unified loss function. We adopt the typical softmax loss as well as the angular softmax loss (A-softmax) [31]. A-softmax learns angularly discriminative features by generating an angular classication margin between embeddings of different classes. The superiority of A-softmax has been shown in both face recognition [31], language recognition and speaker recognition [3].

After training, the 256-dimensional utterance-level speaker embedding is extracted after the penultimate layer of the neural network for the given utterance. In the testing stage, the full-length feature sequence is directly fed into the network, without any truncate or padding operation.

Based on the deep ResNet framework, we investigate multiple kinds of short-term spectral features and loss functions. Finally, we have four networks trained with different setups:

- Mfbank-8k + Softmax: ResNet system trained on Mfbank-8k features with softmax loss.
- Mfbank-16k + Softmax: ResNet system trained on Mfbank-16k features with softmax loss.
- Mfbank-16k + A-softmax: ResNet system trained on Mfbank-16k features with A-softmax loss.
- Gfbank + A-softmax. ResNet system trained on Gfbank-features with A-softmax loss.

2.4. Back-end modeling

In back-end modeling, we either use cosine similarity based scoring, or Probabilistic Linear Discriminant Analysis (PLDA) based scoring.

2.4.1. Cosine similarity

We use cosine similarity as a scoring method for the ResNet based systems. The scores of any given enrollment-test pair are calculated as the cosine similarity of the two embeddings.

Table 1: *Development subset results for the speaker recognition task of the VOiCES from a distance challenge (SN represents Score Normalization, devW represents whitening using development subset)*

Front-end	Back-end	WPE	SN	Development subset			Evaluation		
				minC	**actC**	**EER[%]**	**minC**	**actC**	**EER[%]**
MFCC i-vector	PLDA	-	√	0.4935	0.6747	6.33	0.8037	0.8294	12.92
	CORAL + devW + PLDA	√	√	0.4527	0.4703	6.12	0.6870	0.6891	11.89
PNCC i-vector	PLDA	-	√	0.5073	0.6745	6.12	0.6791	0.7803	10.18
	CORAL + devW + PLDA	√	-	0.4594	0.4697	5.29	0.6498	0.7152	10.09
x-vector	CORAL + PLDA	-	√	0.4018	0.4151	4.96	0.6377	0.6492	09.13
	CORAL + PLDA	√	-	0.3617	0.3688	4.52	0.5417	0.5544	07.54
Mfbank-8k	CORAL + devW + PLDA	-	-	0.4557	0.5246	5.41	0.6608	0.7128	10.92
ResNet + Softmax	CORAL + devW + PLDA	√	-	0.3934	0.4611	4.59	0.5929	0.6424	09.75
Mfbank-16k	cosine similarity	-	-	0.3608	1	3.81	0.6262	1	08.75
ResNet + Softmax	cosine similarity	√	-	0.3245	1	3.02	0.5507	1	07.91
Mfbank-16k	cosine similarity	-	-	0.2735	1	**2.73**	0.4156	1	**05.84**
ResNet + A-Softmax	cosine similarity	√	-	**0.2485**	1	**2.41**	**0.3668**	1	**05.58**
Gfbank	cosine similarity	-	-	0.3065	1	3.52	0.4411	1	06.78
ResNet + A-Softmax	cosine similarity	√	-	**0.2680**	1	3.14	**0.4056**	1	06.49

2.4.2. Gaussian PLDA

We use Correlation Alignment (CORAL) [32, 33] to align the distributions of out-of-domain and in-domain features in an unsupervised way by aligning second-order statistics, i.e., covariance. To minimize the distance between the covariance of the out-of-domain and in-domain features, a linear transformation $\mathbf{A}$ to the original source features and the Frobenius norm is used as matrix distance metric:

$$\min_{\mathbf{A}} \|\mathbf{C}_{\hat{S}} - \mathbf{C}_T\|_F^2 = \min_{\mathbf{A}} \|\mathbf{A}^T\mathbf{C}_S\mathbf{A} - \mathbf{C}_T\|_F^2 \tag{1}$$

where $\mathbf{C}_S$ and $\mathbf{C}_T$ are covariance matrix of the source-domain and target-domain features, $\mathbf{C}_{\hat{S}}$ is covariance of the transformed source features, and $\|\cdot\|_F^2$ denotes the matrix Frobenius norm.

The embeddings after domain adaptation are whitened and unit-length normalized. The whitening transforms is estimated with either the training set or the development subset.

The Gaussian PLDA model [34] with a full covariance residual noise term is trained on the speaker discriminant features. After the PLDA is trained, the scores of any given enrollment-test pair are calculated as the log-likelihood ratio on the PLDA model.

2.5. Score normalization

After scoring, results from all trials are subject to score normalization. We utilize Adaptive Symmetric Score Normalization (AS-Norm) in our systems [35]. The adaptive cohort for the enrollment file are selected to be X closest (most positive scores) files to the enrollment utterance e as $\mathcal{E}_e^{\text{top}}$. The cohort scores based on such selections for the enrollment utterance are then:

$$S_e(\mathcal{E}_e^{\text{top}}) = \{s(e, \varepsilon) | \forall \varepsilon \in \mathcal{E}_e^{top}\} \tag{2}$$

Then the AS-Norm is

$$\tilde{s}(e,t) = \frac{1}{2}\left(\frac{s(e,t) - \mu[S_e(\mathcal{E}_e^{\text{top}})]}{\sigma[S_e(\mathcal{E}_e^{\text{top}})]} + \frac{s(e,t) - \mu[S_t(\mathcal{E}_t^{\text{top}})]}{\sigma[S_t(\mathcal{E}_t^{\text{top}})]}\right) \tag{3}$$

2.6. System fusion and calibration

All the subsystems are fused and calibrated using the BOSARIS toolkit [36] which learn a scale and a bias for each subsystem. The final fusion is a score-level equal-weighted sum after applying the scale and the bias.

3. Experiments

3.1. Data usage

The training data includes VoxCeleb 1 [37] and VoxCeleb 2 [38]. The original distribution of VoxCeleb split each video into multiple short segments. During training, the segments from the same video are concatenated into a single sound wave, which results in 167897 utterances from 7245 speakers. No voice activity detection (VAD) is applied.

For the development data, we only use a subset of the development dataset provided by the VOiCES challenge. The total of 196 speakers in the original development dataset is split into two subgroups, each with 98 speakers. One subset is used as the new development set, and the other is used as the domain adaptation and score normalization corpus. In this way, we reduce the original 4,005,888 trials into 999,424 trials. Since a part of the development, data is used as the domain adaption and score normalization data, we can not provide the experimental results on the whole development data. So *all the experimental results on the development set presented in this paper use the new subtrials.*

3.2. System performance on single systems

In table 1, the systems of different front-end speaker discriminant features with the top one back-end are provided.

From the results in table 1, several observations are drawn as follows. First, the PNCC based i-vector system obtains a noticeable performance gain under strong reverberation and low SNR (signal to noise ratio) environments (evaluation set) compared to MFCC based i-vector system. For the development set with mild reverberation and higher SNR (about 20dB), the per-

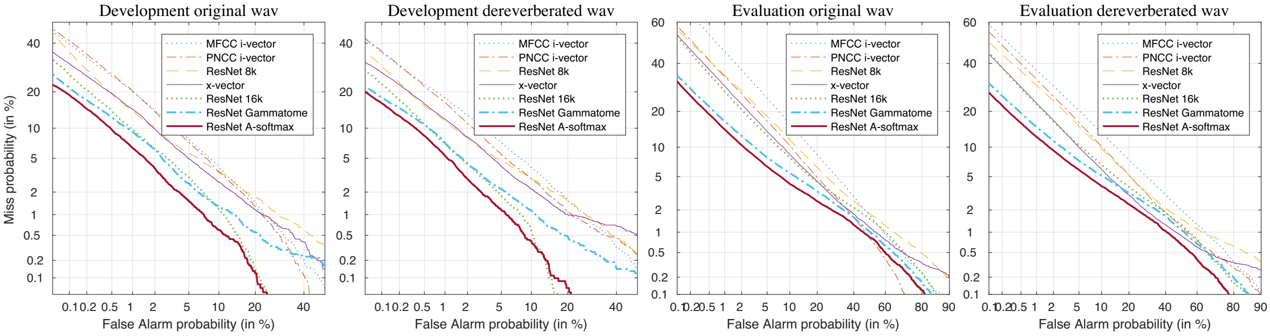

Figure 1: *DET plots for development and evaluation dataset with original or dereverberated sound wave*

Table 2: *System performance on different fusion system*

Fusion strategy	**Development subset**				**Evaluation**			
	minC	**actC**	**EER[%]**	**Cllr**	**minC**	**actC**	**EER[%]**	**Cllr**
Best single system (ResNet + A-softmax + WPE)	0.2485	1	2.41	0.8060	0.3668	1	5.58	0.8284
Each embedding with top 1 back-end	0.1831	0.1857	1.93	0.0808	**0.3205**	**0.3214**	**4.60**	**0.2335**
Each embedding with top 2 back-end	0.1644	0.1659	1.48	0.0710	0.3555	0.3578	4.79	0.2684
Each embedding with top 3 back-end (submission)	**0.1473**	**0.1484**	**1.21**	**0.0577**	0.3532	0.3609	4.96	0.2683

formance gain is not so obvious. Also, the WPE dereverberation algorithm results in 10% gain compared to the original wave for both i-vector and neural network based systems. Moreover, the ResNet + softmax system trained on 16k Mfbank achieves 17.5% relative performance gain in terms of minDCF compared to the 8k Mfbank. The last observation from the results is the performance of the system with A-softmax loss. Compared to the ResNet + softmax system, the ResNet + A-softmax system significantly improve the system performance by more than 20% on both development and evaluation sets.

The Detection Error Tradeoff (DET) curves in figure 1 provide a clear comparison among the subsystems we used in the VOiCES challenge.

The final best signal system is the ResNet + A-softmax network combined with cosine similarity scoring. Applying dereverberation to the enrollment and testing data can further improve the performance. On the development set, the final minDCF and EER are 0.2485 and 2.41% respectively. On the evaluation set, the final minDCF and EER are 0.3668 and 5.58%.

The performance degradation on the evaluation set can be observed from results. This performance degradation mainly due to the more challenging reverberation environments and much lower SNR in the evaluation data, which lead to the mismatch between development and evaluation data.

3.3. System performance on fused systems

For the seven kinds of front-end systems, the embeddings from the original audio and the de-reverberated audio are extracted respectively, resulting in 14 types of front-end speaker discriminant features. Then, different back-end modeling methods, including cosine scoring, a different set of PLDA modeling, and different setting of score normalization, are applied to these features. For each speaker embedding, the top three back-end methods with the best performance on the particular embedding are selected, and finally, we get 42 individual scores for the final fusion.

The final results on the development subset and the evaluation set are shown in table 2. Our final submission obtains minDCF of 0.1473 and 0.3532 on the development and evaluation set respectively.

After the evaluation, we investigate the system performance fused with different back-ends. It is interesting to find that although fusion with the top 3 back-ends for each front-end embeddings improves the performance by 20% relatively compared to fusion with top 1 back-ends, the results on the evaluation show the opposite: fusion with the top 3 back-ends for each front-ends degrades the performance by 10% compared to the fused system with top 1 back-ends. This is mainly because of the mismatch between the development and evaluation data.

4. Conclusions

We presented the components and analyzed the results of the DKU-SMIIP speaker recognition system for the VOiCES from a Distance Challenge 2019. We use different acoustic features, different front-end modeling methods, and various backend scoring methods. To further improve the performance, we use WPE to dereverberate the development and evaluation data. This enabled a series of incremental improvements, and the fusion showed that different subsystems are complementary to each other at score level.

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

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

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

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] S. O. Sadjadi and J. H. Hansen, "Hilbert Envelope Based Features for Robust Speaker Identification Under Reverberant Mismatched Conditions," in *ICASSP*, 2011, pp. 5448–5451.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] M. K. Nandwana, J. V. Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The VOiCES from a Distance Challenge 2019 Evaluation Plan," *arXiv:1902.10828 [eess.AS]*, 2019.

[24] 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.

[25] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices Obscured in Complex Environmental Settings (VOICES) corpus," in *Interspeech*, 2018, pp. 1566–1570.

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

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

[28] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex Sounds and Auditory Images," in *Auditory Physiology and Perception*. Oxford, UK: Y. Cazals, L. Demany, and K. Horner, (Eds), Pergamon Press, 1992, pp. 429–446.

[29] W. Cai, Z. Cai, W. Liu, X. Wang, and M. Li, "Insights into End-to-End Learning Scheme for Language Identification," in *ICASSP*, 2018, pp. 5209–5213.

[30] W. Cai, J. Chen, and M. Li, "Analysis of length normalization in end-to-end speaker verification system," in *Interspeech*, 2018, pp. 3618–3622.

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

[32] B. Sun, J. Feng, and K. Saenko, "Return of Frustratingly Easy Domain Adaptation," in *AAAI*, 2016, pp. 2058–2065.

[33] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker Verification in Mismatched Conditions with Frustratingly Easy Domain Adaptation," in *Odyssey*, 2018, pp. 176–180.

[34] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249–252.

[35] P. Matjka, O. Novotn, O. Plchot, L. Burget, M. D. Snchez, and J. ernock, "Analysis of Score Normalization in Multilingual Speaker Recognition," in *Interspeech*, 2017, pp. 1567–1571.

[36] N. Brümmer and E. De Villiers, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," *arXiv preprint arXiv:1304.2865*, 2013.

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

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