Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings

Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen Stauffer, Colleen Richey, Aaron Lawson, Martin Graciarena

Speech Technology and Research Laboratory, SRI International, Menlo Park, California, CA

mahesh.nandwana@sri.com

Abstract

This article focuses on speaker recognition using speech acquired using a single distant or far-field microphone in an indoors environment. This study differs from the majority of speaker recognition research, which focuses on speech acquisition over short distances, such as when using a telephone handset or mobile device or far-field microphone arrays, for which beamforming can enhance distant speech signals. We use two large-scale corpora collected by retransmitting speech data in reverberant environments with multiple microphones placed at different distances. We first characterize three different speaker recognition systems ranging from a traditional universal background model (UBM) i-vector system to a state-of-the-art deep neural network (DNN) speaker embedding system with a probabilistic linear discriminant analysis (PLDA) back-end. We then assess the impact of microphone distance and placement, background noise, and loudspeaker orientation on the performance of speaker recognition system for distant speech data. We observe that the recently introduced DNN speaker embedding based systems are far more robust compared to i-vector based systems, providing a significant relative improvement of upto 54% over the baseline UBM i-vector system, and 45.5% over prior DNN-based speaker recognition technology.

Index Terms: Speaker recognition, speaker embeddings, distant speech, reverberation

1. Introduction

Automatic speaker recognition systems measure the similarity between an unknown voice and voices previously enrolled in the system. The performance of automatic speaker recognition systems can degrade significantly when a mismatch exists between the system training, speaker enrollment, and test conditions. These mismatches are introduced by a number of factors, such as background noise; transmission channel; codecs; room reverberation; vocal effort; language; emotions; etc. [1, 2, 3]. This study focuses in particular on mismatches caused by distant or far-field speech acquired from a single microphone in the context of speaker recognition.

Automatic speaker recognition from distant speech is particularly challenging due to the effects of reverberation. Reverberation affects the spectro-temporal characteristics of the speech signal. In a reverberant environment, sound waves arrive at the microphone via a direct path, by multiple paths, and after reflecting off from surrounding walls and objects. Early reflections (i.e., reflections arriving within 50–80 ms after direct sound) tend to build up a level louder than the direct sound, which results in internal smearing known as the *self-masking effect*. The echoes arriving after early reflections are called late reflections. Late reflections tend to smear the direct sound over time and mask succeeding sounds. This phenomenon is referred to as the *overlap-masking effect* [4, 5]. These effects blur the details of the speech spectrum, which, in turn, hurts the performance of speaker recognition systems.

The field of automatic speech recognition has advanced greatly for distant or far-field speech conditions especially because of the CHiME speech separation challenge series [6], REVERB challenge [7], and IARPA Automatic Speech Recognition in Reverberant Environment (ASpIRE) challenge [8]. Moreover, the commercial success of digital personal assistants has also contributed to the growth of far-field ASR [9]. But research in the speaker recognition community has tended to focus on distant speech acquired in relatively clean conditions, such as in the NIST speaker recognition evaluation 2008 dataset, or artificially reverberated speech data. Lacking in the research is an analysis of speaker recognition using distant speech in realistic scenarios that include background noises such as a television, music, or other people talking in the background.

Past literature has reported several approaches to alleviate the impact of distant speech on speaker recognition systems. In [10, 11], using novel, robust acoustic features, such as modulation spectral features and mean Hilbert envelop coefficients (MHEC), was proposed. Multi-style training methods were used in [12, 13, 14] to minimize the effect of reverberation mismatch. Reverberation compensation at the score level was proposed in [2]. Multichannel signal-processing techniques (e.g., microphone arrays) were employed to improve the robustness of speaker identification (SID) systems by dereverberating the signal in [15, 16, 17, 18].

Two major shortcomings befall existing work on speaker recognition under reverberant conditions. First, the majority of studies use software simulations to generate data representing reverberant conditions [10], and those studies using actual data employ very few speakers [2], which results in limited significance in subsequent analysis. Second, multichannel or microphone array based solutions are inapplicable when only single-channel or prerecorded data is available [16].

The main contributions of this work are as follows. First, the distant speech datasets used in this work are collected in actual reverberant environments as opposed to software simulated data. Second, this study utilizes a large number of speakers using two different datasets in a variety of real-world conditions. Third, we perform speaker recognition using a single microphone instead of a microphone array. Fourth, we aim to provide a comparison of three different speaker recognition systems and assess the impact of microphone distance and placement, background noise, and loudspeaker orientation. Also, to the best of our knowledge, this study is the first work to show the robustness of a DNN speaker embeddings based speaker recognition system for distant speech data.

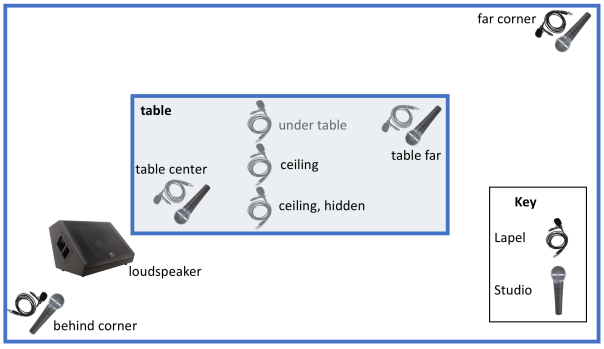

Figure 1: *Typical microphone layout in a room of the SRI Distant Speech Collect.*

2. Corpora

This section describes the distant speech corpora used in this work to evaluate the performance of speaker recognition systems. We use two different data sets: (i) SRI Distant Speech Collect and (ii) Voices Obscured in Complex Environments Settings (VOICES). These datasets were collected by taking audio files from existing corpora based on close-talking speech and then retransmitting the audio via a high-quality loudspeaker in a number of different rooms. This method not only serves as a close representation of real-world distant conditions compared to software simulations, but it is also cost-effective relative to recruiting speakers for original voice recordings.

2.1. SRI Distant Speech Collect Dataset

The first set of data, the SRI Distant Speech Collect dataset, was collected by playing audio out of a loudspeaker and recorded using 11 microphone channels at once, placed at different distances inside the room. Of these 11 microphones, 7 were omnidirectional lapel condenser microphones (AKG C417L), and the remaining 4 were cardioid dynamic studio microphones (SHURE SM58). Three pairs of lapel and studio microphones were placed in front of the speaker (two on the table and one in the far corner of the room) whereas one pair of lapel and studio microphones were placed behind the speaker (in the corner of the room). The remaining three microphones were positioned under the table, mounted on the ceiling, and placed inside a small box on the ceiling, respectively. The window blinds in the room were kept up to maximize reverberation while recording the data. The location of microphone placements in a room is shown in Figure 1.

This data was recorded in three different rooms with carpeted floor, and their dimensions and characteristics are summarized in Table 1. The data was collected for research in the area of speaker recognition, language recognition, and keyword spotting. This data is not publicly available as of now.

The retransmission data for speaker recognition was sourced from the Forensic Voice Comparison (FVC) dataset [19], which is comprised of Australian English speakers. A total of 197 speakers were sub-selected from a total of 552 speakers. The source data was collected using a close-talking, head-mounted microphone at a sampling frequency of 16 kHz.

Table 1: *Dimensions and characteristics of rooms for SRI Distant Speech Collect dataset.*

Room	Dimensions (in)	Room Characteristics
Room 1	164x135x107	Normal office furniture
Room 2	146x107x107	Reverberant walls and ceiling
Room 3	225x158x109	HVAC noise, long windows

2.2. Voices Obscured in Complex Environment Settings (VOICES) Dataset

Voices Obscured in Complex Environments Settings (VOICES) is a large dataset freely available to the public for research and was collected along the lines of SRI Distant Speech Collect but with several additional variations during speech acquisition. It has three different background noises and a speaker rotation mimicking human head movement. It was collected to foster research in the area of automatic speech recognition (ASR), speaker recognition, speech activity detection (SAD), and speech enhancement.

A comparison of the SRI Distant Speech Collect dataset and the VOICES dataset across a number of different parameters is summarized in Table 2. The details related to the data-collection protocols and availability of VOICES dataset can be found in [20]. Note that the VOICES dataset shares rooms 2 and 3 with SRI Distant Speech Collect.

Table 2: *Comparison of SRI Distant Speech Collect and VOICES datasets across different parameters.*

	SRI Distant	VOICES
Number of Speakers	197	300
Number of Mics	11	12
Number of Rooms	3	2
Source Dataset	FVC	LibriSpeech
Speech Type	Conversation	Read
Background Noise	No	Babble, Music, TV
Loudspeaker Orientation	No	0° to 180°
Freely available	No	Yes

3. Speaker Recognition Systems

In this section, we describe the speaker recognition systems developed for our experiments. We use three different speaker recognition systems, which include a traditional UBM i-vector based system, a hybrid alignment framework i-vector system based on DNN bottleneck features [21], and a state-of-the-art DNN speaker embedding based system [22, 23]. These systems use a probabilistic linear discriminant analysis (PLDA) back-end classifier to compute the speaker-similarity scores.

All three systems use DNN-based speech activity detection (SAD) with two hidden layers containing 500 and 1000 nodes, respectively. The SAD DNN is trained using 20-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) features, stacked with 31 frames. The MFCCs are mean and variance normalized over a 201-frame window before training the SAD DNN. The threshold for selecting the speech versus non-speech frames was 0.5 in training and evaluation, except during speaker embedding extractor DNN training, which used a threshold of -1.5 based on the findings of [23].

3.1. UBM I-Vector System

This is a traditional i-vector system [24], which uses 20-dimensional MFCCs with a frame length of 25 ms and a step size of 10 ms that are mean and variance normalized over a sliding window of three seconds. The MFCCs are contextualized with deltas and double deltas to create a 60-dimensional feature

vector. The universal background model (UBM) is a gender-independent, 2048-component, diagonal-covariance Gaussian mixture model (GMM). This system uses a 400-dimensional i-vector extractor. For training of UBM and i-vector extractor, we used the original PRISM training list, including degradations [25].

3.2. Hybrid Aligned Bottleneck I-Vector System

The hybrid alignment framework is based on DNN bottleneck features and was developed to improve calibration of DNN-based speaker recognition systems across varying conditions [21]. The framework uses two sets of features: a first set of features to determine the frame alignments (zero-order statistics) in the Baum-Welch statistics calculation and a second set of features for calculating first-order statistics. This process resulted in a more robustly calibrated DNN-based system compared to using concatenated MFCC and bottleneck features, by restricting the use of bottleneck features only to the alignment of standard acoustic features during i-vector extraction.

For this system, we use a DNN BN extractor trained from 20-dimensional power-normalized cepstral coefficients (PNCC) [26] contextualized with principal component analysis discrete cosine transform (pcaDCT) [27] with a window of 15 frames to create 90-dimensional inputs to the DNN that are then mean and variance normalized using a sliding window of three seconds. The DNN is trained to discriminate 1933 senones using Fisher and Switchboard telephone data and consists of five layers of 1200 nodes, except the fourth hidden layer, which has 80 nodes and forms the bottleneck extraction layer. The first-order features, aligned with the BN features and 2048-component, diagonal-covariance UBM, are MFCCs of 20 dimensions also contextualized with pcaDCT using a 15-frame window with an output of 60 dimensions. In all cases, the principal component analysis (PCA) transform for pcaDCT is learned using a subset of the DNN training data. For training the UBM and i-vector extractors, we used the original PRISM training list, including degradations [25]. This system also extracts 400-dimensional i-vectors.

3.3. DNN Speaker Embedding System

In recent years, speaker discriminative training of DNNs has been used to extract a low-dimensional representation of speaker characteristics from one of the DNN's hidden layers. This low-dimensional representation, rich in speaker information, is referred to as the speaker embedding. These speaker embeddings replace the i-vectors used in the above systems. DNN-based speaker embeddings have resulted in new state-of-the-art text-independent speaker recognition technology because of its ability to generalize to unseen conditions [28, 22].

For training the speaker embedding extractor, we used 52,456 audio files sourced from the non-degraded subset of PRISM training lists[25]. We then augmented this data with four copies of four different degradation types including a random selection of audio compression; a random selection of instrumental music at a 5 dB signal-to-noise (SNR) ratio; a random selection of noises at a 5 dB SNR; and a random selection of reverberated signals with low reverberation. This augmentation resulted in a total of 891,752 segments from 3,296 speakers for training the embeddings extractor. More details on this system are found in [23], where the system is denoted as raw+CNLRMx4.

3.4. Probabilistic Linear Discriminant Analysis (PLDA) Classifier

We use gender-independent probabilistic linear discriminant analysis (PLDA) [29] to compute the scores of speaker recognition systems. The fixed-dimensional speaker representation from each of these systems (either i-vector or speaker embedding) were further transformed using linear discriminant analysis (LDA) to 200 dimensions followed by length normalization and mean centering [30]. For training the PLDA model and LDA, we used the full PRISM training lists, which include noise and reverb degradations. Additional transcoded data was added to this PLDA training data [3].

One thing to note here is that the i-vector extractor (UBM/T) are trained to original PRISM lists as it doesn't respond well to augmentation [22] whereas DNN embedding extractor is trained on raw PRISM lists with 16x augmentation. Our assumption is that each of the i-vector systems have been developed over a long period of time with different types of training data and we use what is probably the most common training set collection for these according to literature from numerous research teams. This does not mean that it is the optimal training set, but rather a set that the community has settled on over years of i-vector research.

4. Experimental Evaluation

In this section, we benchmark each of the described speaker recognition systems on the SRI Distant Speech Collect and VOICES data sets. We also analyze the impact of microphone distance and placement, background noise, and loudspeaker orientation on speaker recognition system performance. We report our results in terms of equal error rates (EER) in percentage.

4.1. Evaluation Protocol

Audio files from the SRI Distant Speech Collect data set were cut into 20-second chunks based on SAD output and were then used in enrollment and verification. We performed enrollment on a single 20-second cut of audio from the source data and verified for different microphones placed at different positions on a single 20-second cut. For the VOICES dataset, the enrollment/test segments were 14-seconds long and speech dense.

4.2. Benchmarking Results

First, we present our benchmarking results for different rooms from the SRI Distant Speech Collect and VOICES data sets on the UBM-IV, Hybrid-IV, and speaker embeddings systems. We report a single measure of Equal-Error Rate (EER) per room

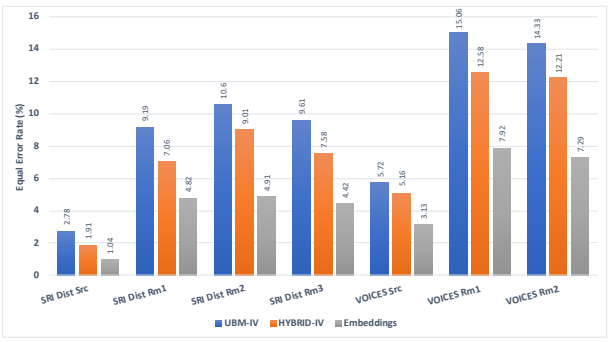

Figure 2: *Benchmarking of different speaker recognition systems on the SRI Distant Speech Collect and VOICES corpora.*

and per corpus pooled the trials from all microphones prior to calculating the EER. Enrollment was performed on the source data, and testing on the distant speech. While the VOICES data set contains various distractor sounds, in this section we only report on the subset of the data that contains no added background noise. These results are summarized in Fig. 2.

We observe that the speaker embedding-based system consistently outperformed the i-vector-based system by a considerable margin in reverberant conditions as well as for the source datasets. The relative gain over the baseline system ranged from 47% to 54% for SRI Distant Speech Collect for different rooms.

4.3. Impact of Microphone Distance and Placement

Next, we present the results on the impact of microphone distance and placement from the loudspeaker in Table 3. We observe that in this case, the speaker embeddings based system also outperformed the i-vector based system by a large margin. The equal error rates increased with distance. Noteworthy is the significant challenge that hidden microphones, such as under-the-table microphones, pose for speaker recognition systems.

Table 3: *Impact of microphone position on the performance of three different speaker recognition systems in terms of EER (%).*

Mics	Table-center	Hidden	Table-far	Ceiling	Behind	Far-corner
SRI Distant Room 1						
UBM-IV	5.8	10.9	6.2	9.5	11.3	9.6
Hybridiv	3.7	9.1	4.9	7.7	8.9	6.7
Embedding	**2.3**	**5.8**	**2.8**	**4.8**	**6.4**	**4.4**
SRI Distant Room 2						
UBM-IV	8.8	10.5	9.1	11.0	12.1	12.1
Hybridiv	7.3	9.1	7.3	9.3	10.1	10.5
Embedding	**4.0**	**5.4**	**3.7**	**5.0**	**5.5**	**5.2**
SRI Distant Room 3						
UBM-IV	5.9	11.1	8.7	9.4	11.2	10.6
Hybridiv	4.9	8.4	6.7	7.7	8.7	8.7
Embedding	**2.5**	**5.0**	**3.8**	**4.2**	**5.7**	**4.9**
VOICES Room 1						
UBM-IV	10.7	17.5	13.0	16.2	14.9	15.1
Hybridiv	8.6	15.1	11.5	13.1	12.4	12.3
Embedding	**5.0**	**9.9**	**6.7**	**8.5**	**8.2**	**7.2**
VOICES Room 2						
UBM-IV	10.9	16.3	13.1	13.2	14.7	16.6
Hybridiv	9.6	13.6	11.2	11.5	12.4	13.6
Embedding	**5.4**	**8.6**	**6.3**	**6.8**	**7.6**	**8.3**

4.4. Impact of Background Noise

In this section, we study the influence of background noises (i.e. "distractors") of different types on speaker recognition systems using the VOICES database. In order to mimic a realistic test case, we enroll speakers using a recording from the close Lapel microphone (Table, center) in room 1 with no distractor noise. The test segments originate from all microphones and were recorded with different types of background noise.

The results shown in Table 4 reflect the impact on the EER of various types of distractors on the three SID systems. We observe that all three systems struggle most with the Babble noise. Perhaps because it is very speech-like, but also perhaps because Babble was the only distractor played out of 3 separate sources while music and television were played out of a single source only. In terms of robustness to distractors, the embeddings system remains the clear winner in terms of Equal-Error Rate over the two other baseline systems. Interestingly though, the relative difference between the best and second-best system decreases from 42% with no distractors to 37% and 38% with Television and Music, to only 34% with Babble.

Table 4: *Impact of room distractor on the performance of three different speaker recognition systems in terms of EER (%). Each condition has above* $18k/2.8M$ *target/impostor trials.*

Distractor	None	Television	Music	Babble
UBM-IV	17.2	19.3	19.2	20.9
Hybridiv	14.9	16.7	17.2	18.1
Embedding	**8.6**	**10.5**	**10.5**	**11.9**

4.5. Impact of Loudspeaker Orientation

Because the speaker may not always be facing the microphone in distant recordings, it is important to assess the impact of speaker orientation on the performance of speaker identification systems. In this section, we study the influence of loudspeaker orientation using the VOICES database. We picked four Lapel microphones (table center, table far, corner far, under table) positioned roughly in a straight line from the default loudspeaker position (90 degrees). In order to isolate the "orientation" variable, we design trials to have enrollment and test samples from the same microphone and the same room (room 1), with no distractors sounds. We then design three enrollment conditions and three test conditions using three ranges of loudspeaker orientation: Left (0-20 degrees), Straight (80-100 degrees) and Right (160-180 degrees). One thing to note is that since different files are used for enrollment and testing in the experiments on the VOICES database, symmetrical enrollment/test conditions (e.g. left/right vs right/left) may give different results even though speaker verification is a symmetrical task.

Results shown in Table 5 show that a perpendicular loudspeaker orientation to the microphone of interest induces a degree of distortion that challenges even the most robust embeddings SID systems, with a pooled EER going from 8.5% to above 13% in some cases. Also, it is interesting to note that in this case obtaining a "matched" enrollment to the test sample, even from the same microphone in the same room doesn't help, e.g. it is better to enroll with "Straight" when testing on "Left" or "Right" even if the test orientation and microphone are known.

Table 5: *Impact of loudspeaker orientation on the performance of the embeddings SID system in terms of EER (%). Each condition has above* $1k/150k$ *target/impostor trials.*

Enroll \ **Test**	Left	Straight	Right
Left	13.3	10.3	13.0
Straight	10.1	**8.5**	10.9
Right	10.7	9.7	11.3

5. Conclusions

We investigated the impact of distant speech on the performance of speaker recognition systems. The corpora used in this work was collected in actual reverberant rooms rather created by software simulations. We benchmarked the performance of three speaker recognition systems on two different datasets. We observed that speaker embedding based speaker recognition systems gave very impressive gains over i-vector based systems.

6. 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] 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.

[3] M. McLaren, V. Abrash, M. Graciarena, Y. Lei, and J. Pesán, "Improving robustness to compressed speech in speaker recognition," *INTERSPEECH-2013*, pp. 3698–3702, 2013.

[4] A. K. Nábělek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap-and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, 1989.

[5] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," *Speech processing in the auditory system*, pp. 231–308, 2004.

[6] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015.

[7] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[8] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 547–554, 2015.

[9] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin *et al.*, "Acoustic modeling for Google home," *INTERSPEECH-2017*, pp. 399–403, 2017.

[10] T. H. Falk and W.-Y. 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.

[11] S. O. Sadjadi and J. H. L. Hansen, "Mean hilbert envelope coefficients (MHEC) for robust speaker and language identification," *speech communication*, vol. 72, pp. 138–148, 2015.

[12] 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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4257–4260, 2012.

[13] 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," *INTERSPEECH-2014*, pp. 1096–1100, 2014.

[14] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4253–4256, 2012.

[15] J. González-Rodríguez, J. Ortega-García, C. Martín, and L. Hernández, "Increasing robustness in GMMpeaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," *Fourth International Conference on Spoken Language Processing (ICSLP)*, vol. 3, pp. 1333–1336, 1996.

[16] L. Wang, K. Odani, and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 12, 2012.

[17] M. Ladislav, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and beamforming in far-field speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5254–5258, 2018.

[18] M. Ladislav, O. Plchot, P. Matejka, O. Novotny, and J. H. Cernocky, "Dereverberation and beamforming in robust far-field speaker recognition," *INTERSPEECH-2018*, 2018.

[19] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow, "Forensic database of voice recordings of 500+ Australian English speakers," *URL: http://databases. forensic-voice-comparison. net*, 2015.

[20] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, M. Graciarena, A. Lawson, M. K. Nandwana *et al.*, "Voices obscured in complex environmental settings (VOICES) corpus," *INTERSPEECH-2018*, 2018.

[21] M. McLaren, D. Castan, L. Ferrer, and A. Lawson, "On the issue of calibration in DNN-based speaker recognition systems," *INTERSPEECH-2016*, pp. 1825–1829, 2016.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[23] M. McLaren, D. Castan, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," *Speaker Odyssey*, 2018.

[24] 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.

[25] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," *Proceedings of NIST 2011 workshop*, 2011.

[26] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4101–4104, 2012.

[27] M. McLaren and Y. Lei, "Improved speaker recognition using DCT coefficients as features," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4430–4434, 2015.

[28] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *INTERSPEECH-2017*, pp. 999–1003, 2017.

[29] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[30] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," *INTERSPEECH 2011*, pp. 249–252, 2011.