

Speaker Recognition Using Real vs Synthetic Parallel Data for DNN Channel Compensation

Fred Richardson, Michael Brandstein, Jennifer Melot and Douglas Reynolds

MIT Lincoln Laboratory

{frichard,msb,Jennifer.Melot,dar}@ll.mit.edu

Abstract

The effective use of synthetic multi-channel data for training denoising DNNs has been demonstrated for several speech technologies such as ASR and speaker recognition. This paper compares the use of real and synthetic data for training denoising DNNs for multi-microphone speaker recognition. Large reductions in error rates (37% and 50% for the AVG and POOL EERs and 20% and 30% for the AVG and POOL min DCFs) are attained on Mixer 6 microphone data using Mixer 1 and 2 multi-microphone data to train a denoising DNN. Nearly the same reduction in error rate is realized using room impulse response and noise estimates (RIRs) derived from the Mixer 1 and 2 data and applied to just the telephone channel. Applying RIRs from three publicly available databases used in the Kaldi Aspire evaluation system yields lower but significant reductions in error rate (16% and 34% relative improvement in AVG and POOL EER and 13% and 25% relative improvement in AVG and POOL min DCFs). In all cases, the telephone channel performance on SRE10 is improved by the denoising DNNs with the real Mixer 1 and 2 trained DNN reducing EER by 12% and min DCF by 8.9%.

Index Terms: denoising DNN, multi-condition training, channel compensation, speaker recognition

1. Introduction

Recently there has been a great deal of interest in using deep neural networks (DNNs) for channel compensation under reverberant or noisy channel conditions such as those found in microphone data [1, 2, 3, 4, 5, 6]. The 2015 Aspire challenge [7] evaluated automatic speech recognition (ASR) performance on conversational speech recorded over far-field microphones in different rooms. Details about the recording environments used for the Aspire evaluation data were not disclosed to performers prior to the evaluation and the performers were limited to using Fisher telephone data to train their systems. The top performing ASR systems in the Aspire challenge all used some form of denoising DNN trained on synthetic parallel microphone data generated from the Fisher telephone recordings [7].

The denoising DNN approach has also been shown to work well for speaker recognition[1], but unfortunately there is limited publicly available real microphone data appropriate for evaluating speaker recognition performance. The Mixer 1 and 2, Mixer 4 and 5, and Mixer 6 corpora collected by the Linguistic Data Consortium (LDC) include multi-session parallel microphone data that was used to measure cross-channel speaker recognition performance in the NIST 2004, 2005, 2006, 2008 and 2010 speaker recognition evaluations [8, 9, 10, 11, 12, 13]. The complete set of wide-bandwidth Mixer 1 and 2 microphone recordings were used in this work and have been made available

to the LDC for future public release. The LDC has already released the Mixer 6 wide-bandwidth recordings [14] which are also used in this work. For brevity the Mixer 1 and 2 corpora will be referred to simply as Mixer 2.

While future collections of real multi-microphone multi-session data may be essential for evaluating the performance of speaker recognition and other speech technologies under real and challenging channel conditions it may not be possible to collect enough data for performers to use for system development without disclosing channel characteristics of the evaluation data. In this work we try to address the question of whether using real parallel multi-microphone data for developing channel robust speaker recognition systems has advantages over using synthetic multi-channel data. For our analysis we use the Mixer 2 real parallel microphone corpora and two synthetic parallel channel corpora derived from the Mixer 2 telephone data. The first synthetic corpora uses RIRs and noise sources estimated using parallel microphone segments extracted from the Mixer 2 data and the second synthetic corpora uses RIRs drawn from three publicly available databases used in the Kaldi Aspire evaluation system [15]. For evaluation purposes we use the conversational portion of the Mixer 6 parallel microphone corpora where the target and non-target trials are all over the same microphone. For both Mixer 2 and Mixer 6, the wide bandwidth microphone recordings are down sampled to 8 KHz using the same technique described in [16].

2. DNN Channel Compensation

A denoising DNN is a neural network regression model trained to reconstruct data from a clean target channel given the same data from a different possibly noisy or reverberant channel or from the same channel as the target. The objective function for the denoising DNN is the minimum mean squared error between the output of the DNN and the target channel's data. The denoising DNNs output layer uses a linear activation function (instead of the softmax activation function used for a neural network classifier). For this work we use either the Mixer 2 multi-channel corpus or a synthetic parallel corpus for training the DNN with the telephone channel used as the target data. Both the microphone and the target telephone channels are used as input features to the DNN. A 5 layer 1024 node DNN architecture is used in all cases. The hidden layers of the DNN use the same number of nodes and the sigmoid activation function.

The denoising DNN has been used to extract features that are beneficial for a range of different speech technologies and applications. The focus of this work is to use features estimated by the denoising DNN as the input to an i-vector system for channel robust speaker recognition. A simplified block diagram of the hybrid i-vector/DNN system is shown in Figure 1. The i-vector system uses a Gaussian mixture model (GMM) which

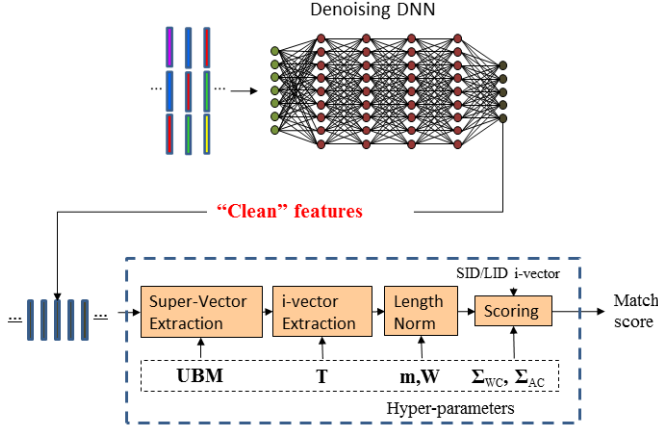


Figure 1: Hybrid denoising DNN i-vector system

is often referred to as the universal background model (UBM) to extract zero'th and first order statistics from the input feature vector sequence. A super vector created by stacking the first order statics is transformed down to a lower dimensional subspace using a linear transformation that depends on the zeroth order statistics (see [17] for more details). This transformation requires a total variability matrix $\mathbf{T}$ which is estimated from a large set of super-vectors using an EM-algorithm [17] or PPCA [18].

The i-vector is treated as a single low dimensional representation of a waveform that contains both speaker and channel information. Mean vector $\mathbf{m}$ and whitening matrix $\mathbf{W}$ are used to transform the i-vectors to have a unit normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ before applying length normalization [19]. Then full rank within class (Σ_{wc}) and across class (Σ_{ac}) covariance matrices are estimated using speaker labeled multi-session data and the "2 covariance model" described in [20] is used for PLDA scoring.

3. Microphone and Telephone Corpora

The Mixer 2 and Mixer 6 conversational microphone speech collections were used in this work for evaluating microphone channel compensation techniques for speaker recognition. For the Mixer 2 data there are 239 speakers (123 female and 116 male) with 1035 sessions (averaging 4.3 sessions/speaker). The sessions were recorded over 8 microphones (see Table 1) and a telephone channel in parallel at three different locations: ICSI, ISIP and LDC (see [10, 9, 13] for more details).

In order to train a denoising DNN on Mixer 2 data, a matched filter was used to time align the data from each microphone channels to the telephone channel. Audio files were rejected if the alignment process failed. At the end of the process a total of 873 sessions out of the 1035 available sessions had data for all channels.

The Mixer 6 microphone collection has data from 546 speaker (280 female and 266 male) with 1400 sessions. There are a maximum of 3 sessions per a speaker (the average is 2.5). The sessions were recorded over 14 microphones listed in Table 2 in two office rooms at the LDC (see [12, 14] for more details). Six microphones were selected for this work based on their distance from the speaker and appear in bold in Table 2 (microphones 02, 04, 05, 08, 07 and 13). We chose to evaluate target and non-target trials only on the same microphone and

Chan	Microphone
01	AT3035 (Audio Technica Studio Mic)
02	MX418S (Shure Gooseneck Mic)
03	Crown PZM Soundgrabber II
04	AT Pro45 (Audio Technica Hanging Mic)
05	Jabra Cellphone Earwrap Mic
06	Motorola Cellphone Earbud
07	Olympus Pearlcororder
08	Radio Shack Computer Desktop Mic

Table 1: Mixer 2 microphones

Chan	Microphone	Distance (inches)
02	Subject Lavalier	8
04	Podium Mic	17
10	R0DE NT6	21
05	PZM Mic	22
06	AT3035 Studio Mic	22
08	Panasonic Camcorder	28
11	Samson C01U	28
14	Lightspeed Headset On	34
07	AT Pro45 Hanging Mic	62
01	Interviewer Lavalier	77
03	Interviewer Headmic	77
12	AT815b Shotgun Mic	84
13	AcoustImagic Array	110
09	R0DE NT6	124

Table 2: Mixer 6 microphones

same room since all sessions from a given speaker in Mixer 6 were recorded in the same room.

Mixer 6 also includes sessions with varying vocal effort (high, low and normal). During the course of this work we found that the performance of the high vocal effort data was particularly poor on the telephone channel. The performance of our baseline system described in Section 5 on the NIST 2010 Speaker Recognition Evaluation (SRE10), Mixer 2 and Mixer 6 is summarized in Table 3. Our initial investigation revealed that at least some of the Mixer 6 data appears to have distortion on the telephone channel. Since the high vocal effort speech does not appear to adversely affect the other microphone channels and there are at most 3 microphone sessions per a speaker in Mixer 6, we chose to retain the high vocal effort data for the purpose of evaluating microphone speaker recognition performance. Following the approach described in [21], we analyzed the relationships among microphone distance attenuation, the ratio of speech plus noise power to noise power (SNRp), and the baseline system performance [21, 7]. Distance attenuation and system performance showed a Spearman correlation of 0.7928786 for the baseline system and 0.6499239 for the system with channel compensation, suggesting that channel compensation helped mitigate the effect of distance from the microphone on system performance. Our method of calculating SNRp used SAD marks rather than the reference transcripts used in [21]. SNRp and distance attenuation were not significantly correlated, with a Spearman rho of -0.1748497. Also, functions of SNRp were not significantly correlated with system performance, with rho values near 0.

A test set was created from the Mixer 6 data for evaluating microphone performance with 1,230 target and 224,897

Task	EER	DCF
SRE10 Tel	5.77	0.662
Mixer2 Tel	0.20	0.0352
Mixer 6 Tel	10.89	0.910

Table 3: Baseline system performance on telephone channel data

non-target trials for each of the 6 channels (7,371 target and 1,347,686 non-target trials pooled across all microphones). The telephone portion of SRE10 test set was used for evaluating speaker recognition performance on telephone data. The SRE10 test set consists of 7,094 target and 405,066 non-target trials.

4. Synthesized Corpora

The Mixer 2 telephone channel data was modified using room impulse responses and noise sources (RIRs) in two different ways. The first approach involved estimating the RIRs and additive noise from a very limited portion of Mixer 2 and then simulating the entire data set by generating synthetic microphone data by filtering the original telephone speech with the estimated RIRs and adding noise. Specifically 60 sec segments were extracted from eight Mixer 2 session across all eight parallel microphones. Each telephone microphone pair was time aligned and the channel impulse responses were estimated via Welch’s averaged periodogram over the speech segments while the additive noise was derived from the non-speech portions. Given the limited reverberant conditions of the original recording environment, the estimated impulse responses were truncated to a 100ms duration. Each Mixer 2 telephone recording was transformed for each microphone by randomly selecting one of the eight RIRs to create the synthetic multi-channel corpus. The additive noise was applied across the original telephone waveform by using an overlap-add synthesis of randomized windows of the noise estimate while maintaining the original SNR levels.

The Kaldi Aspire approach described in [15] was used to create a second synthetic corpus. RIRs were drawn from three different sources: the Aachen Impulse Response (AIR) database [22], the RWCP sound scene database [23] and the 2014 Reverb challenge database [24]. Both the Reverb Challenge and RWCP databases provided noise sources which were added at randomly selected SNR levels of 0, 5, 10, 15 or 20 dB. The RIRs were randomly selected eight times for each Mixer 2 telephone recording.

5. Experimental Setup

Denoising DNNs were trained using 40 Mel frequency cepstral coefficients (MFCCs) including 20 derivatives coefficients extracted from a 25ms window of speech every 10ms. The input to the DNN consist of the MFCCs feature vectors stacked in a 21 frame window with 10 frames before and after the center frame (210ms of speech) with the center frame corresponding to the target feature vector. The target data for the DNN is a single MFCC feature vector extracted from the telephone channel data. The MFCCs are normalized using a non-linear warping (see [25]) to fit a unit Gaussian distribution over a sliding 300 frame window for both the DNN input and output features. The DNNs are trained using stochastic gradient descent (SGD) with a mini-batch size of 256 and a learning rate of 0.1. In most cases SGD training is completed in fewer than 20 epochs. The DNN

DNN Training	AVG	POOL
None (baseline)	11.5%	21.2%
Real Mixer 2	7.23%	10.6%
Mixer 2 RIRs	7.25%	11.1%
Kaldi/Aspire RIRs	9.66%	13.9%

Table 4: Performance (EER) for real and synthetic parallel data

architecture in all cases consists of 5 layers with 1024 nodes per a layer and uses a sigmoid activation function.

The i-vector systems use a 2048 component Gaussian mixture model and 600 dimensional i-vector sub-space. The GMM, $\mathbf{T}$, $\mathbf{m}$, $\mathbf{W}$, Σ_{wc} , Σ_{ac} parameters are all estimated using the Switchboard 1 and 2 data sets. The baseline system uses 40 MFCC feature vectors with mean and variance normalization. For our experimental results we report both the equal error rate (EER) and minimum decision cost function (min DCF) for a target prior of 0.01.

In Section 6 “Real Mixer 2” refers the Mixer 2 parallel corpus, “Mixer 2 RIRs” refers to the synthetic corpus generated using the Mixer 2 derived RIRs and “Kaldi/Aspire RIRs” refers to the synthetic corpus generated using RIRs drawn from the AIR, RWCP or 2014 Reverb challenge databases.

6. Experiments

Performance for the baseline and DNN systems is presented in Table 4 (EER) and Table 5 (min DCF). In the tables, “AVG” is the average EER across microphones and “POOL” is the pooled performance for scoring all microphones together. The difference between the AVG and POOL results to some extent reflects the calibration of a given system.

In all cases, the DNN systems perform significantly better than the baseline system with the DNN trained on real Mixer 2 data giving the largest relative improvement of 37% and 50% for the AVG and POOL EERs and 20% and 30% for the AVG and POOL min DCFs. The DNN trained using the Mixer 2 RIRs corpus performs almost as well as the DNN trained on the Real Mixer 2 corpus except that the POOL min DCF is significantly worse. The DNN trained on the Kaldi/Aspire RIRs corpus does not perform as well as the other DNNs but is still significantly better than the baseline (16% and 34% relative improvement in AVG and POOL EER and 13% and 25% relative improvement in AVG and POOL min DCFs). The AIR, RWCP and Reverb 2014 databases may contain RIRs from a much broader range of acoustic environments than the offices used in Mixer 2 or Mixer 6 collections which could explain the degraded performance.

As noted earlier, it is important for the denoising DNNs to improve microphone performance without degrading performance on conversational telephone speech. To assess the performance impact of the denoising DNN on telephony data we evaluated the DNNs on the SRE10 telephone data set. The results of this experiment are given in Table 6. Note that there is actually a small gain in performance for the Real Mixer 2 denoising DNN on SRE10 (a 12% reduction in EER and 8.9% reduction in min DCF) and minor gains for the other two DNNs.

7. Conclusions

In this work we have compared the use of real parallel multi-microphone speech data and synthetic multi channel speech data for training denoising DNNs for speaker recognition. Both the real Mixer 2 parallel data and the synthetic data generated

DNN Training	AVG	POOL
None (baseline)	0.728	0.978
Real Mixer 2	0.581	0.687
Mixer 2 RIRs	0.592	0.730
Kaldi/Aspire RIRs	0.632	0.729

Table 5: Performance (min DCF) for real and synthetic parallel data

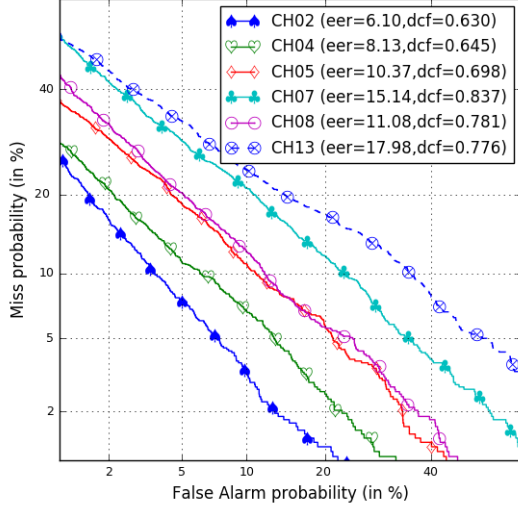


Figure 2: Baseline

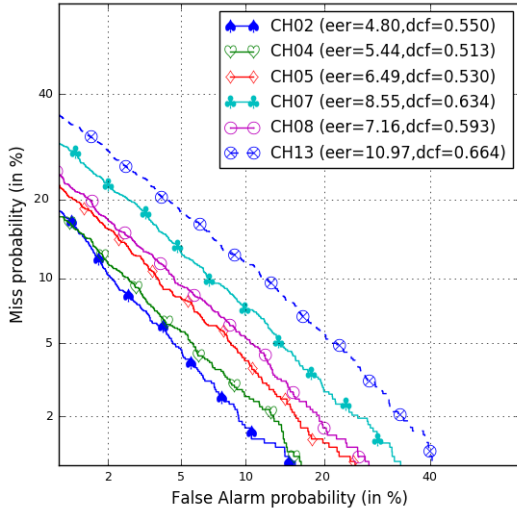


Figure 3: DNN trained using real Mixer 2

DNN Training	EER	DCF
None (baseline)	5.77	0.662
Real Mixer 2	5.05	0.603
Mixer 2 RIRs	5.24	0.632
Kaldi/Aspire RIRs	5.38	0.647

Table 6: Performance on SRE10 telephone data

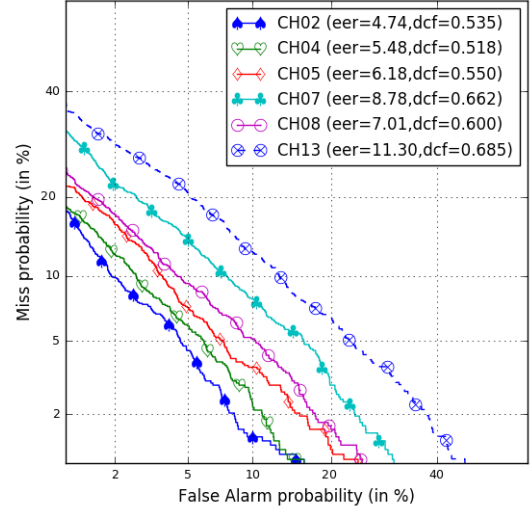


Figure 4: DNN trained using Mixer 2 RIRs

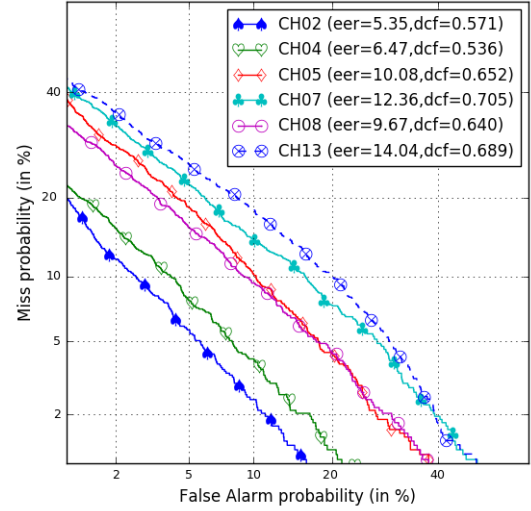


Figure 5: DNN trained using Kaldi/Aspire RIRs

using RIRs estimated from the Mixer 2 data perform very well on the Mixer 6 same-channel multi-microphone test data yielding large improvements in both EER and min DCF relative to the baseline system (relative improvements of 37% and 50% for the AVG and POOL EERs and 20% and 30% for the AVG and POOL min DCFs). Smaller but still significant performance gains are realized using data generated with RIRs and noise sources drawn from three publicly available databases. The RIRs and noise sources drawn from these databases may span a broader range of acoustic environments compared to those used in the Mixer 2 and Mixer 6 collections which could explain the lower reduction in error relative to the Mixer 2 derived RIRs and noise sources. While we have not attempted to address the question of whether or not synthetic multi-channel data should be used for evaluating speaker recognition performance, it appears that this data can be used effectively for developing channel robust speaker recognition systems through the use of denoising DNNs.

8. References

- [1] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi, "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2015.
- [2] M. Karafiat, F. Grezl, L. Burget, I. Szoke, and J. Cernocky, "Three ways to adapt a cts recognizer to unseen reverberated speech in but system for the aspire challenge," in *Proc. of Interspeech*, 2015.
- [3] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders," in *Reverb Challenge Workshop*, 2014.
- [4] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *Proc. of Interspeech*, 2015.
- [5] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *International Conference on Signal Processing*, 2014.
- [6] A. Nugraha, K. Yamamoto, and S. Nakagawa, "Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2014.
- [7] M. Harper, "The automatic speech recognition in reverberant environments (aspire) challenge," in *Proc. of IEEE ASRU*, 2015.
- [8] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The mixer 3, 4 and 5 corpora," 2007.
- [9] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybicki, and K. Walker, "The mixer and transcript reading corpora: Resources for multilingual, cross-channel speaker recognition research," in *Proc. of LREC*, 2006.
- [10] C. Cieri, J. P. Campbell, H. Nakasone, D. Miller, and K. Walker, "The mixer corpus of multilingual, multichannel speaker recognition data," in *Proc. of IEEE Odyssey*, 2004.
- [11] L. Brandschain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker recognition: Building the mixer 4 and 5 corpora," in *LREC*, 2008.
- [12] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. of LREC*, 2010.
- [13] J. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. Martin, and M. Przybicki, "The mmsr bilingual and crosschannel corpora for speaker recognition research and evaluation," in *Proc. of IEEE Odyssey*, 2004.
- [14] Linguistic Data Consortium, "Mixer 6 corpus specification v4.1," 2013.
- [15] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system : Robust lvsr with tdnn, ivector adaptation and rnn-lms," in *Proc. of IEEE ASRU*, 2015.

- [16] W. Campbell, D. Sturim, B. Borgstrom, R. Dunn, A. McCree, T. Quatieri, and D. Reynolds, "Exploring the impact of advanced front-end processing on nist speaker recognition microphone tasks," in *Proc. of IEEE Odyssey*, 2012.
- [17] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front end factor analysis for speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 19, no. 4, pp. 788–798, may 2011.
- [18] A. McCree, D. Sturim, and D. Reynolds, "A new perspective on gmm subspace compensation based on ppca and wiener filtering," in *Proc. of Interspeech*, 2011.
- [19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. of Interspeech*, 2011, pp. 249–252.
- [20] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Proc. of IEEE Odyssey*, 2010.
- [21] J. Melot, N. Malyska, J. Ray, and W. Shen, "Analysis of factors affecting system performance in the aspire challenge," in *Proc. of IEEE ASRU*, 2015.
- [22] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *IEEE Inter. Conf. on DSP*, 2009.
- [23] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *LREC*, 2000.
- [24] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2016).
- [25] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. of ICASSP*, 2002.