INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

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

Recent work has shown large performance gains using denoising DNNs for speech processing tasks under challenging acoustic conditions. However, training these DNNs requires large amounts of parallel multichannel speech data which can be impractical or expensive to collect. The effective use of synthetic parallel data as an alternative has been demonstrated for several speech technologies including automatic speech recognition and speaker recognition (SR). This paper demonstrates that denoising DNNs trained with real Mixer 2 multichannel data perform only slightly better than DNNs trained with synthetic multichannel data for microphone SR on Mixer 6. Large reductions in pooled error rates of 50% EER and 30% min DCF are achieved using DNNs trained on real Mixer 2 data. Nearly the same performance gains are achieved using synthetic data generated with a limited number of room impulse responses (RIRs) and noise sources derived from Mixer 2. Using RIRs from three publicly available sources used in the Kaldi ASpIRE recipe yields somewhat lower pooled gains of 34% EER and 25% min DCF. These results confirm the effective use of synthetic parallel data for DNN channel compensation even when the RIRs used for synthesizing the data are not particularly well matched to the task.

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 (SR) [1, 8], but unfortunately there is limited publicly available real microphone data appropriate for evaluating SR 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 SR performance in the NIST 2004, 2005, 2006, 2008 and 2010 SR evaluations [9, 10, 11, 12, 13, 14]. The complete set of wide-bandwidth Mixer 1 and 2 microphone recordings were used in this work and will be available from the LDC in a future release. The LDC has already released the Mixer 6 wide-bandwidth recordings [15] 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 SR 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. In this work we try to address the question of whether using real parallel multi-microphone data for developing channel robust SR 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 room impulse response (RIRs) and noise sources estimated using parallel microphone segments extracted from a small subset of the Mixer 2 data, and the second synthetic corpora uses RIRs drawn from three publicly available databases used in the Kaldi ASpIRE evaluation system [16]. 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 [17].

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 and/or reverberated version 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 multichannel 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 with the hope that the DNN will be optimized to improve the microphone data while leaving the telephone data unaltered. 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.

Denoising DNNs have been used to extract features that are beneficial for a range of different speech technologies and ap-

This work was sponsored by the Department of Defense under Air Force contract F19628-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

http://dx.doi.org/10.21437/Interspeech.2016-544

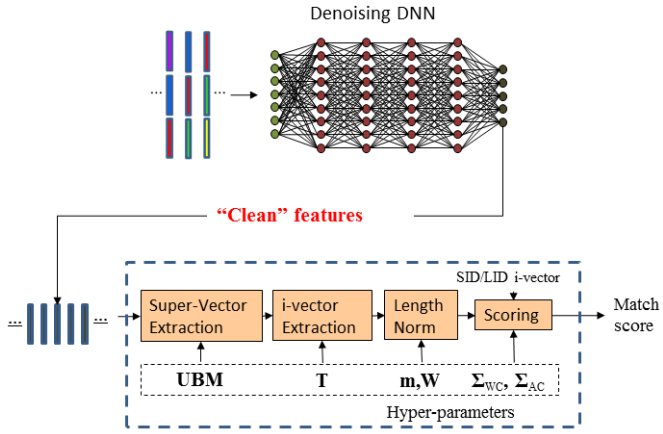

Figure 1: Hybrid denoising DNN i-vector system

plications. 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 SR. 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 is often referred to as the universal background model (UBM) to extract zeroth and first order statistics from the input feature vector sequence. A super vector created by stacking the first order statistics is transformed down to a lower dimensional sub-space using a linear transformation that depends on the zeroth order statistics (see [18] 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 [18] or PPCA [19].

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 [20]. Then full rank within class ($\mathbf{\Sigma}_{\text{wc}}$) and across class ($\mathbf{\Sigma}_{\text{ac}}$) covariance matrices are estimated using speaker labeled multi-session data and the "2 covariance model" described in [21] 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 SR. 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 [11, 10, 14] 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 channel 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 speakers (280 female and 266 male) over 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 [13, 15] for more details).

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 Pearlcorder
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

The six microphones selected for this work, based on their distance from the speaker, appear in bold in Table 2. We chose to evaluate target and non-target trials only on the same microphone and 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). Given the relatively small amount of data available, all sessions were used for evaluating microphone SR performance. During the initial course of our investigations we found that the high vocal effort speech significantly degraded SR performance on the telephone channel data compared to the performance observed over the microphone channels. Further analysis of high scoring false alarms revealed a significant degree of distortion in the telephone handset for the high vocal effort sessions. Therefore we have chosen to use the standard NIST 2010 speaker recognition task for measuring telephone SR performance instead of using the Mixer 6 telephone channel data.

A test set was created from the Mixer 6 data for evaluating microphone SR performance with 1,230 target and 224,897 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 potion of the SRE10 test set was used for evaluating SR 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 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 via filtering the original telephone speech with the estimated RIRs and adding noise. Specifically, 60 sec segments were extracted from eight Mixer 2 sessions 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 then applied to the waveform using an overlap-add synthesis of randomized windows of the noise estimate while maintaining the original SNR levels.

The Kaldi ASpIRE approach described in [16] 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 derivative 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 (i.e. 225ms 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 architecture in all cases consists of 5 layers with 1024 nodes per 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, $\mathbf{T}$, $\mathbf{m}$, $\mathbf{W}$ $\mathbf{\Sigma}_{\rm wc}$, $\mathbf{\Sigma}_{\rm ac}$ 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.

6. Experiments

In the following section, "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.

Performance for the baseline and DNN systems is presented in Table 3 (EER) and Table 4 (min DCF). In the tables, "AVG" is the average EER across microphones and "POOL" is the pooled

DNN Training	AVG (imp)	POOL (imp)
None (baseline)	11.5% (-)	21.2% (-)
Real Mixer 2	**7.23% (37%)**	**10.6% (50%)**
Mixer 2 RIRs	7.25% (37%)	11.1% (48%)
Kaldi/ASpIRE RIRs	9.66% (16%)	13.9% (34%)

Table 3: EER performance for real and synthetic parallel data (improvement relative to the baseline is in parentheses)

DNN Training	AVG (imp)	POOL (imp)
None (baseline)	0.728 (-)	0.978 (-)
Real Mixer 2	**0.581 (20%)**	**0.687 (30%)**
Mixer 2 RIRs	0.592 (19%)	0.730 (25%)
Kaldi/ASpIRE RIRs	0.632 (13%)	0.729 (25%)

Table 4: Min DCF Performance for real and synthetic parallel data (improvement relative to the baseline is in parentheses)

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% / 50% for the AVG / POOL EERs and 20% / 30% for the AVG / 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% / 34% relative improvement in AVG / POOL EER and 13% / 25% relative improvement in AVG / POOL min DCFs). The AIR, RWCP and REVERB 2014 databases provide RIRs from a broader range of acoustic environments than the offices used in Mixer 2 and Mixer 6 collections which may explain the degraded performance using the Kaldi/ASpIRE RIRs corpus.

DET plots for the four systems are shown in Figure 2. The apparent correlation of performance across microphones with the microphone distances listed in Table 2 is confirmed by an analysis similar to the one presented in [26]. Distance attenuation of the Mixer 6 microphones and system performance show a Spearman correlation of 0.793 for the baseline system and 0.650 for Real Mixer 2 DNN system, confirming that channel compensation helped mitigate the effect of distance from the microphone on system performance.

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 task. The results of this experiment are given in Table 5. 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

Collecting parallel multi-channel data from different environments over a range of microphones and microphone positions can be prohibitively expensive and impractical. In this work we have compared the use of real parallel multi-microphone speech

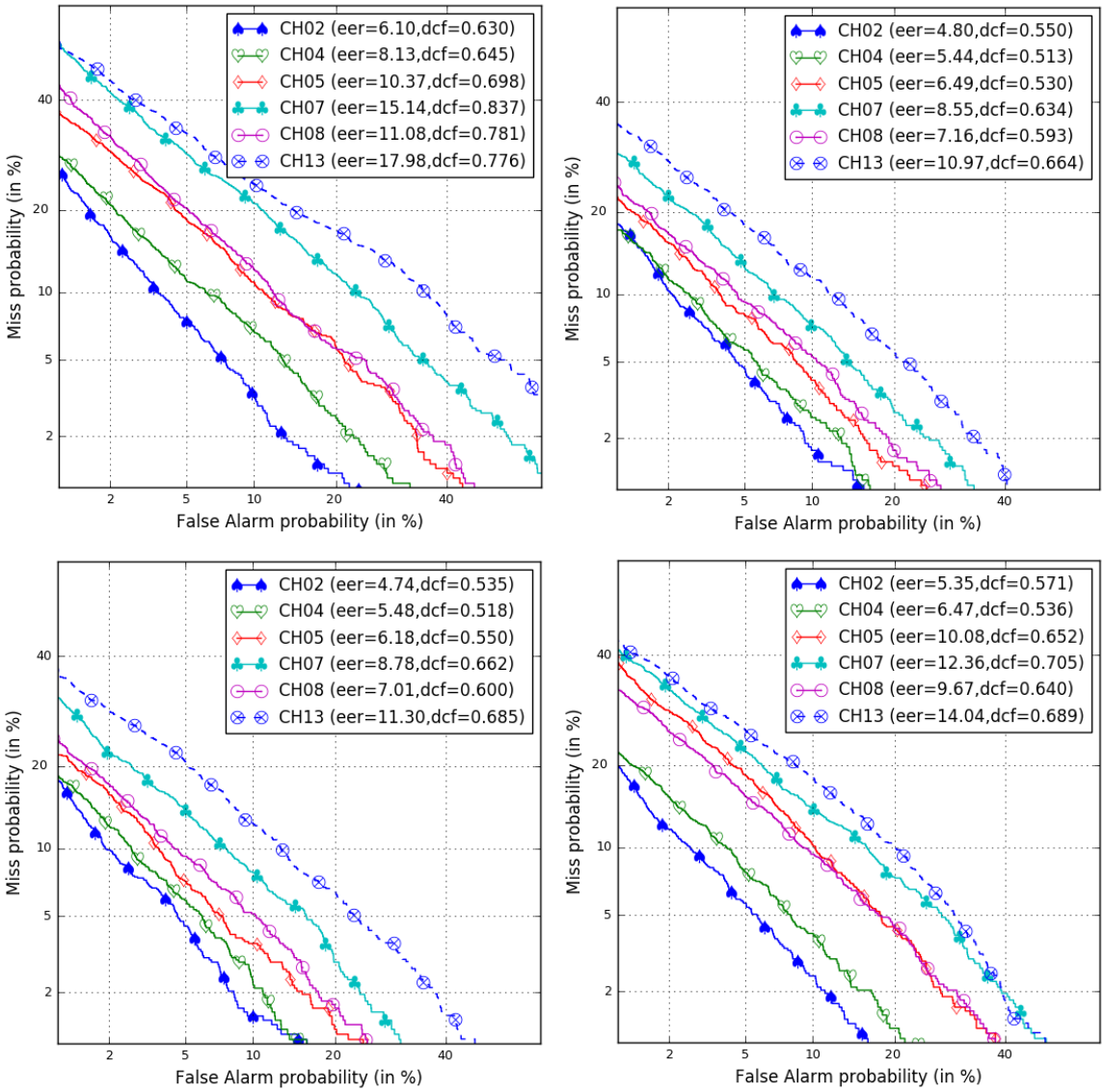

Figure 2: DET curves from baseline (upper left), real Mixer 2 DNN (upper right), Mixer 2 RIRs DNN (lower left) and Kaldi/ASpIRE RIRs DNN (lower right)

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 5: Performance on SRE10 telephone data

data and synthetic multi channel speech data for training denoising DNNs for channel compensation. DNNs from both the real Mixer 2 parallel data and a synthetic parallel corpus created using RIRs from a small subset of Mixer 2 perform comparably well on the Mixer 6 same-channel multi-microphone task yielding large relative performance improvements. Significant but lower performance gains were realized using data generated with RIRs drawn from three publicly available databases used in the Kaldi ASpIRE recipe. Importantly, all three denoising DNN systems did not adversely impact telephone SR performance as measured on the SRE10 telephone task implying that the DNN channel compensation can be applied universally to both telephone and microphone data. These results suggest that the substantial performance improvements demonstrated using DNN channel compensation for the SR task can be achieved with far smaller (though diverse) collections of parallel microphone data than has been acquired (at great expense) in the past.

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] F. Richardson, B. Nemsick, and D. Reynolds, "Channel compensation for speaker recognition using MAP adapted PLDA and denoising DNNs," *to appear in Odyssey*, 2016.

[9] 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.

[10] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki, and K. Walker, "The mixer and transcript reading corpora: Resources for multilingual, cross-channel speaker recognition research," in *Proc. of LREC*, 2006.

[11] 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.

[12] L. Brandschain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker recognition: Building the mixer 4 and 5 corpora," in *LREC*, 2008.

[13] 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.

[14] J. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. Martin, and M. Przybocki, "The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation," in *Proc. of IEEE Odyssey*, 2004.

[15] Linguistic Data Consortium, "Mixer 6 corpus specification v4.1," 2013.

[16] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASpIRE system : Robust LVCSR with TDNNS, ivector adaptation and RNN-LMS," in *Proc. of IEEE ASRU*, 2015.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Proc. of IEEE Odyssey*, 2010.

[22] M. Jeub, M. Schafer, and P. Vary, "A binaurl 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.

[26] 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.