SINGLE-CHANNEL SPEAKER DIARIZATION BASED ON SPATIAL FEATURES

Mathieu Hu[1*], *Pablo Peso Parada*[2], *Dushyant Sharma*[2], *Simon Doclo*[3], *Toon van Waterschoot*[4], *Mike Brookes*[1], *Patrick A. Naylor*[1]

[1] Department of Electrical and Electronic Engineering, Imperial College London, UK
[2] Voicemail-To-Text Research, Nuance Communications Inc., Marlow, UK
[3] Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, Oldenburg, Germany,
[4] Dept. of Electrical Engineering (ESAT-STADIUS/ETC), KU Leuven, Belgium
mathieu.hu12@imperial.ac.uk, pablo.peso@nuance.com

ABSTRACT

Speaker diarization has gained much importance over the past five years in helping overcome key challenges faced by automatic meeting transcription systems. Current state-of-the-art algorithms can only utilize spatial information when multi-microphone recordings are available. In this paper, we propose the novel use of reverberation as a source of spatial information obtained from single-channel recordings to perform speaker diarization. The proposed system is shown to reduce speaker classification errors by 34% when compared with current MFCC based single-channel systems.

Index Terms— Speaker diarization, direct-to-reverberant ratio, spatial acoustic features

1. INTRODUCTION

Speaker diarization aims at answering the question 'which speaker spoke when?'. This information can be used to improve the performance of Automatic Speech Recognition (ASR) systems by allowing effective speaker adaptation or enable efficient tagging of large transcripts of meetings so that a user can search by speaker name, for example.

Previous algorithms for single-channel speaker diarization discriminated different speakers using speech dependent features such as Mel-frequency Cepstral Coefficients (MFCC) or Perceptual Linear Predictive (PLP) coefficients [1] preferably extracted from data captured by a close talking microphone [2]. In meeting scenarios, however, distant microphones are often preferred to headsets or lapel microphones. The use of distant microphones can cause the adverse effects of reverberation to become more prominent and reduce the performance of MFCC or PLP based systems due to temporal smearing of the received signal. The combination of reverberation and other adverse phenomena typical of meeting scenarios, e.g. poor Signal-to-Noise Ratio (SNR) or variable speech levels, further reduces the performance of such diarization systems [3].

When multi-channel signals are available, the use of beamforming as a preprocessor to enhance the signal has been explored in [4], while in [5], a feature based on the Time-Difference-of-Arrival (TDOA) between each pair of microphones has been proposed. In [6], a framework in which MFCCs and TDOAs are combined has also been proposed.

In many meeting scenarios, such as teleconferencing, there is only a single microphone signal available. In such scenarios, the current state-of-the-art speaker diarization systems are unable to benefit from any spatial information as beamforming preprocessing or TDOA features can only be used when there are at least 2 microphones available. Whereas in previous single-channel speaker diarization methods, reverberation would be seen as a hindrance, we propose in this paper to turn this hindrance into an advantage. The Direct-to-Reverberant Ratio (DRR) [7] is known to be strongly correlated with the distance between a microphone and the sound source [8]. We estimate the DRR from the single-channel signal using a non-intrusive algorithm [9] that does not require prior knowledge of the Room Impulse Response (RIR) or the source signal and use it as an additional feature that characterizes the acoustic channel from each speaker to the microphone.

The remainder of this paper is organized as follows: in Section 2.2, the diarization system is described. In Section 3, the simulated meeting data together with the evaluation metric are shown. The performance of the proposed method is then presented in Section 4 and conclusions are drawn in Section 5.

2. THE DIARIZATION SYSTEM

Throughout this paper, we assume that the speakers are stationary and that the received signal at the microphone $y(n)$ at time instant n is given by

$$x_i(n) = h_i * s_i(n), \tag{1}$$

$$y(n) = \sum_{i=1}^{P} x_i(n) + \nu(n) \tag{2}$$

where i represents a speaker index, $h_i = [h_i(0), h_i(1), \ldots, h_i(L-1)]$ the L sample time-invariant RIR relating the i^{th} speaker to the microphone, P the total number of speakers and $\nu(n)$ is the additive noise at the microphone.

2.1. Baseline diarization system

The baseline, selected for comparison in this paper, is the system proposed in [10]. It discriminates different speakers based on

*The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° ITN-GA-2012-316969.

978-1-4799-7450-4/15/$31.00 ©2015 IEEE

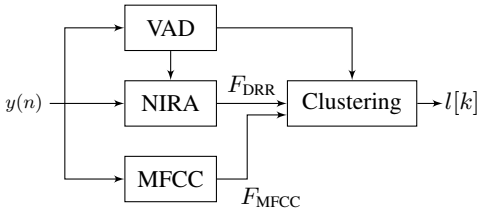

Figure 1: Block diagram of the proposed speaker diarization system.

MFCC features and a diarization system relying on the Information Bottleneck (IB) principle [11].

IB based diarization. The IB based diarization system clusters a given uniform linear segmentation A of the recorded signal $y(n)$ into a set C of clusters, which compresses the input variable while preserving the mutual information about a set B of relevance variables. This is achieved by the minimization of the following objective function using the agglomerative IB [12]:

$$J = I(A, C) - \beta I(C, B) \tag{3}$$

where β is a trade-off parameter and I is the mutual information. The relevance variables correspond the components of a background Gaussian Mixture Model (GMM) trained on the entire recording.

At each step of the agglomerative IB, two clusters are merged so that the loss of mutual information about the relevance variables is minimum. The optimal number of clusters is determined by a threshold on the normalized mutual information $\frac{I(B,C)}{I(A,B)}$. Further details can be found in [6].

The Short Time Fourier Transform (STFT) of the recorded signal $y(n)$ is computed after application of a 20 ms Hamming window with 50% overlap. From each frame, 19 MFCCs are extracted.

At the output of the agglomerative IB algorithm, the clusters are aligned with the boundaries of the initial segments. Those boundaries are realigned by computing the sequence of clusters that minimizes the cost function based on the posterior distribution of the relevance variables.

2.2. Proposed diarization system

The proposed diarization system, a block diagram of which is shown in Fig. 1, is built on top of the system described in [10]. From the received signal $y(n)$, 2 streams of features, namely the MFCC and DRR features, are extracted independently before being combined and clustered so that a label $l[k]$ is assigned to the k^{th} frame.

We propose to use DRR as a spatial feature for speaker diarization, defined for speaker i as follows [13]

$$\text{DRR}_i = 10\log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n=n_d+n_0} h_i^2(n)}{\sum_{n=0}^{n=n_d-n_0} h_i^2(n) + \sum_{n=n_d+n_0}^{\infty} h_i^2(n)}\right) \text{dB}, \tag{4}$$

where n_0 is the number of samples in a rectangular window of 8 ms and n_d is the time index (in sample) of the direct path arrival in the RIR, $h_i(n)$.

To evaluate this measure of reverberation, the RIR needs to be known or estimated. However, in this work DRR is estimated non-intrusively from the reverberant signal, i.e. without information on the RIR or the source signal. The non-intrusive room acoustic parameter estimator (NIRA) [14] employed is a data-driven approach which computes 106 features per frame from active speech segments in the input signal using the Voice Activity Detector (VAD) presented in Section 2.4. These features are derived from pitch period, importance weighted signal to noise ratio, zero-crossing rate, Hilbert transformation, power spectrum of long term deviation, MFCCs, line spectrum frequency and modulation representation. A bidirectional long short-term memory model [15] is trained with these features computed from speech segments to obtain DRR estimates every 10 ms. The estimation of DRR is made level independent by normalizing the power of the input utterance.

2.3. Feature combination

A background GMM, M_j, is now estimated for each stream of features, F_j. The VAD described in Section 2.4 is applied to exclude features extracted from estimated pauses in the training of the GMM. The combined distribution, which is required in the IB based diarization system, is then calculated as:

$$p(b|a) = \sum_{j\in\{\text{MFCC, DRR}\}} p(b|M_j, a)P_j$$

where $b \in B$ is a relevance variable, $a \in A$ is an input feature and the stream weights, P_j, satisfy $P_{\text{MFCC}} + P_{\text{DRR}} = 1$. These weights are empirically estimated from a development set to maximize performance.

2.4. VAD

A VAD based on the P.56 method [16, 17] is applied to extract segments with active speech from the input signal. The time boundaries of detected speech are also given to the clustering block to discard speech features that were extracted from detected silence frames.

3. EXPERIMENTAL SETUP

This section describes the experimental setup used to evaluate the diarization systems. NIRA was trained on the training data of the REVERB Challenge. The measured RIRs used in the training and the simulated meeting data were captured in different rooms [18].

3.1. Simulated meeting data

The simulated meetings were generated by convolving clean speech with measured RIRs taken from the evaluation set of the REVERB Challenge database [18]. Recorded fan noise was added at 20 dB SNR.

Speech data. The nearly-anechoic speech data consisted of utterances taken from the WSJCAM0 corpus [19] which were recorded by a close-talking noise-cancelling head-mounted microphone. There were, in total, 14 speakers: 7 male and 7 female. Each utterance was only used once across the simulated meetings.

RIRs. The RIRs of the REVERB Challenge were captured in 3 rooms of different size by a circular microphone array. For each room, the RIR that was recorded at a distance of 0.5 m is labeled as *near* and the RIR that was recorded at a distance of 2 m is labeled

	Room 1	Room 2	Room 3
T_{60} (s)	0.25	0.5	0.7
Near DRR (dB)	16	11	11
far DRR (dB)	6	0	2

Table 1: T_{60} in s and DRR in dB for the *near* and *far* positions in each of the three rooms.

as *far*. The ground truth DRRs together with the reverberation time T_{60} are given in Table 1.

In each simulated meeting, the reverberant signal consisted of 10 to 14 utterances spoken by 2 different speakers speaking in turn. Spatial differences between the 2 speakers are simulated by convolving each speakers' signal with different RIRs, i.e. the utterances spoken by one speaker are convolved with the *near* RIR while the utterances spoken by the other speaker are convolved with the *far* RIR.

In total, 42 simulated meeting audio streams were generated, 14 streams per room. Among the 42 audio streams, 18 streams contained speakers of different gender, 12 streams contained only male speakers and 12 streams contained only female speakers.

Noise. Fan noise signals, recorded in each room, were added to the simulated meeting data at a mean SNR of 20 dB, defined as [18]

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{N}\sum_{i=1}^{P}\tilde{x}_i^2(n)}{\sum_{n=1}^{N}\nu^2(n)}\right), \quad (5)$$

$$\tilde{x}_i(n) = \tilde{h}_i * s_i(n) \quad (6)$$

where N is the total number of samples in the considered recording, $\tilde{h}_i = [h_i(0), h_i(1), \ldots, h_i(n_{50}-1)]$ is the truncated RIR containing the n_{50} taps within the first 50 ms and $\tilde{x}_i(n)$ is the direct sound together with the early reflections up to 50 ms.

Development and Evaluation sets. The simulated data were broken down into two non-overlapping subsets: a development and an evaluation set.

The development set was used to determine the optimum weights P_{MFCC} and $P_{\text{DRR}} = 1 - P_{\text{MFCC}}$ to use in the IB based diarization system to minimize the Diarization Error Rate (DER) defined below. The optimum weights were then used on the evaluation set.

3.2. Evaluation

The performance of the diarization system was evaluated in terms of DER [20]. The DER corresponds to the fraction of time that is not attributed the correct speaker. That score can be broken down into the missed speaker time, false alarm time and speaker error time. The first two solely depend on the VAD as they respectively correspond to the fraction of time of estimated silence when speech was present and the fraction of time of estimated speech when there was a silence.

The DER is computed as follows [20]:

$$\text{DER} = \frac{\sum_{\sigma\in S}\tau(\sigma)\left\{\max\left(N_{\text{ref}}(\sigma),\, N_{\text{sys}}(\sigma)\right) - N_{\text{correct}}(\sigma)\right\}}{\sum_{\sigma\in S}\tau(\sigma)N_{\text{ref}}(\sigma)} \quad (7)$$

where S is the set of speech segments taken between 2 consecutive speaker change points after merging the estimated and the reference diarization files, σ is a segment and $\tau(\sigma)$ is the duration of σ. The symbols $N_{\text{ref}}(\sigma)$ and $N_{\text{sys}}(\sigma)$ respectively denote the number of reference and system speakers in σ. The number of reference speakers speaking in σ for whom their mapped system speakers are also speaking in σ is denoted by $N_{\text{correct}}(\sigma)$.

4. RESULTS

4.1. Development set

Figure 2 shows the variation of DER with P_{DRR} on the development set. It can be observed that the use of DRR features alone in the IB based diarization system does not achieve low DER. However, by combining them with MFCCs, the performance of the diarization system can be improved to outperform the similar system solely relying on MFCCs.

The lowest DER in the development set is achieved for $P_{\text{DRR}} = 0.18$. This weight reduces the DER from 1.95% to 1.2%, achieving a relative improvement of 38%.

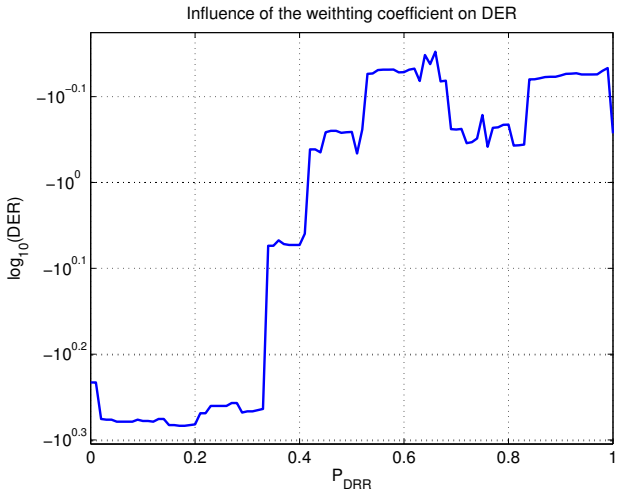

Figure 2: DER of the development set as a function of DRR weight ($P_{\text{DRR}} = 1 - P_{\text{MFCC}}$).

As shown in Table 2, the inclusion of DRR features in the diarization system reduces the DER on average. Although the baseline and the proposed systems perform similarly in Room 1, significant improvements are observed in Room 2 and Room 3, i.e. rooms with higher amount of reverberation. A relative improvement of 56% and 46% are respectively seen in the 2 latter rooms.

	Overall	Room 1	Room 2	Room 3
Baseline	1.95%	0.95%	2.80%	2.10%
Proposed	1.20%	0.94%	1.22%	1.44%

Table 2: Mean DER of the baseline and the proposed method for the development set.

Analyzing the breakdown of the DER in terms of gender, the proposed system achieves lower DER for simulated meeting data containing mixed genders or 2 female speakers. In these cases, the mean DERs were respectively 1.09% and 1.79% for the proposed system against 2.28% and 3.06% for the baseline.

4.2. Evaluation set

The weights $P_{\text{MFCC}} = 0.82$, $P_{\text{DRR}} = 0.18$ determined from the development set are now applied to evaluate the performance of the diarization system on the evaluation set.

Figure 3 shows that the inclusion of DRR features improves by 23% the performance of the diarization system on average. The DER as well as the speaker error time in Room 2 and Room 3 are decreased while the DER in Room 1 is increased by 9% on average. This degradation in performance is due to the limitations of the NIRA DRR estimator, which is reported to have high estimation errors for environments with low reverberation [9]. Analyzing the Root-Mean-Square Deviation (RMSD) of the estimated DRR in each room, the RMSD indeed decreased as T_{60} increased. The RMSD values were respectively 2.7 dB, 2.5 dB and 2.2 dB for Room 1, Room 2 and Room 3.

Similar observations can be made about the speaker error time, which decreases from 2.35% and 1.34% down to 0.85% and 1.03% for Room 2 and Room 3 respectively. The speaker error time increases from 1.07% to 1.24% in Room 1. The variance of the DER is also significantly reduced in Room 2 and Room 3 and slightly increased in Room 1 as shown in Table 3.

	Room 1	**Room 2**	**Room 3**	**Overall**
Baseline	0.68%	4.03%	1.06%	2.50%
Proposed	0.75%	1.16%	0.69%	0.91%
RMSD (dB)	2.7	2.5	2.2	2.5

Table 3: Standard deviation in speaker error time of the baseline and proposed systems for the evaluation set together with the RMSD of the estimated DRR.

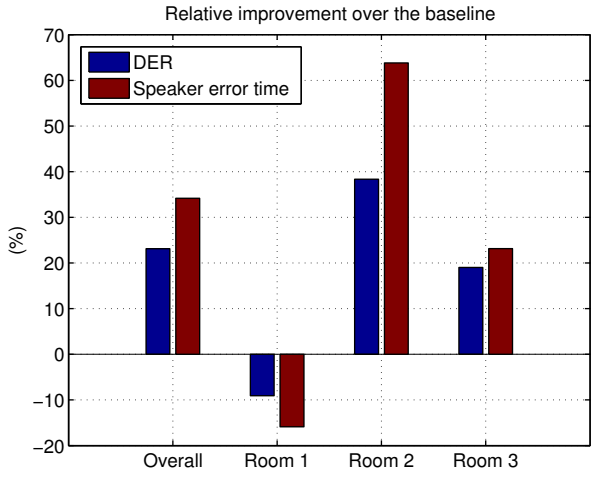

Figure 3: Relative improvement in DER and speaker time error by inclusion of DRR features.

Figure 4 shows that the estimated DRRs mostly take values around 10 dB and 2 dB, depending on the identity of the active speaker. This figure corresponds to the simulated meeting which had the highest DER and speaker error time, which were respectively 16.73% and 13.7%. The presence of the DRR features respectively decreased the DER and speaker error time to 7.13% and 4.1%, achieving a relative improvement of 57% and 70% respectively.

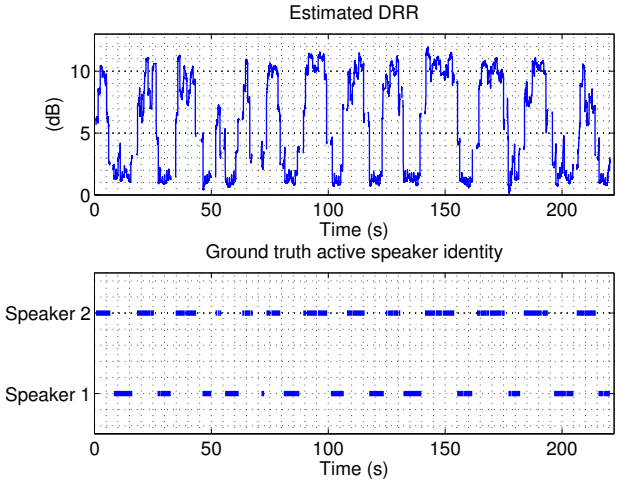

Figure 4: Estimated DRR along with the ground truth speaker identity.

Table 4 shows that the proposed system decreases the mean speaker error time when the speakers have the same gender while the mean speaker error time slightly increases when the speakers have different gender. This shows that the DRR feature is beneficial when the MFCCs have low discrimination capabilities.

	Male - Male	**Mixed**	**Female - Female**
Baseline	1.07%	0.76%	3.20%
Proposed	0.74%	0.77%	1.68%

Table 4: Mean speaker error time broken down by gender for the evaluation set.

5. CONCLUSION

In this work, we proposed to take advantage of the presence of the speaker-dependent variation of reverberation in single-channel recorded meetings for speaker diarization problems. The DRR has been shown to be a feature that can discriminate different speakers located in a room at different positions relative to the microphone.

By combining non-intrusive estimates of the DRRs with MFCC features, the proposed diarization system was evaluated on simulated meeting data created using recorded RIRs and noise signals and was shown to give a relative improvement of 34% in average in terms of speaker error time. The standard deviation of the speaker error time was also reduced in the 2 rooms with higher T_{60}.

We might suggest that a good way forward would be selective combination, such that if the T_{60} indicates an environment with low reverberation, spatial features are not used, otherwise if the T_{60} indicates a moderately or highly reverberant environment, we use the spatial features as proposed in this paper. We note that the NIRA algorithm can also be used to estimate the T_{60} in a non-intrusive manner.

6. REFERENCES

[1] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sept. 2005, pp. 2437–2440.

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1557–1565, 2006.

[3] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: a review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[4] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 2011–2022, 2007.

[5] N. Evans, C. Fredouille, and J.-F. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 4061 – 4064.

[6] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic combination of mfcc and tdoa features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 431–438, Feb. 2011.

[7] P. A. Naylor, E. A. P. Habets, J. Y. C. Wen, and N. D. Gaubitch, "Models, measurement and evaluation," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010, ch. 2.

[8] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1793–1805, Sept. 2010.

[9] P. Peso Parada, D. Sharma, and P. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 4718–4722.

[10] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization beyond two acoustic feature streams," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4950–4953.

[11] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *The 37th annual Allerton Conference on Communication, Control and Computing*, Allerton, IL, USA, Sept. 1999, pp. 368–377.

[12] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen, and K. Müller, Eds. MIT Press, 2000, pp. 617–623.

[13] J. Eaton, A. H. Moore, N. D. Gaubitch, and P. A. Naylor, "The ACE challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2015.

[14] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, P. A. Naylor, and T. van Waterschoot, "A single-channel non-intrusive C50 estimator with application to reverberant speech recognition," *submitted to IEEE Trans. Audio, Speech, Lang. Process.*, 2015.

[15] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT–the Munich open-source CUDA recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 15, 2014.

[16] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[17] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2013.

[18] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2013.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, MI, USA, May 1995, pp. 81–84.

[20] NIST, "Spring 2006 (rt-06s) rich transcription meeting recognition evaluation plan," http://www.itl.nist.gov/iad/mig//tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf, Feb. 2006.