ROBUST COHERENCE-BASED SPECTRAL ENHANCEMENT FOR DISTANT SPEECH RECOGNITION

Hendrik Barfuss, Christian Huemmer, Andreas Schwarz, and Walter Kellermann

Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg
Cauerstr. 7, 91058 Erlangen, Germany
{barfuss,huemmer,schwarz,wk}@int.de

ABSTRACT

In this contribution to the 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) we extend the acoustic front-end of the CHiME-3 baseline speech recognition system by a coherence-based Wiener filter which is applied to the output signal of the baseline beamformer. To compute the time- and frequency-dependent postfilter gains the ratio between direct and diffuse signal components at the output of the baseline beamformer is estimated and used as approximation of the short-time signal-to-noise ratio. The proposed spectral enhancement technique is evaluated with respect to word error rates of the CHiME-3 challenge baseline speech recognition system using real speech recorded in public environments. Results confirm the effectiveness of the coherence-based postfilter when integrated into the front-end signal enhancement.

Index Terms— Robust automatic speech recognition, Postfiltering, Spectral enhancement, Coherence-to-diffuse power ratio, Wiener filter

1. INTRODUCTION

For a satisfying user experience of human-machine interfaces it is crucial to ensure a high accuracy in automatically recognizing the user’s speech. As soon as no close-talking microphone is used, the recognition accuracy suffers from reverberation as well as background noise and active interfering speakers picked up by the microphones in addition to the desired speech signal [1] [2]. Signal processing techniques for robust speech recognition in noisy environments can be categorized into two major categories, namely front-end (e.g., speech enhancement [3] [4] [5]) and back-end (e.g., acoustic-model adaptation [6] [7] [8]) processing techniques.

The 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) [9] targets the performance of state-of-the-art Automatic Speech Recognition (ASR) systems in real-world scenarios. In this year’s challenge, the primary goal is to improve the ASR performance of real recorded speech of a person talking to a tablet device in realistic noisy environments by employing front-end and/or back-end signal processing techniques.

In this contribution to the CHiME-3 challenge, we focus on front-end speech enhancement and extend the CHiME-3 baseline front-end signal processing, consisting of a Minimum Variance Distortionless Response (MVDR) beamformer, by a coherence-based postfilter. The postfilter is realized as a Wiener filter, where an estimate of the ratio between direct and diffuse signal components at the output of the baseline MVDR beamformer is used as an approximation of the short-time Signal-to-Noise Ratio (SNR) to compute the time- and frequency-dependent postfilter gains. The employed postfilter is Direction-of-Arrival (DoA)-independent and has a low computational complexity.

An overview of the overall signal processing pipeline is given in Fig. 1. Whereas the purpose of the beamformer is to reduce the signal components from interfering point sources by spatial filtering, the postfilter shall remove diffuse interference components, e.g., reverberation, from the beamformer output signal. The output of the front-end signal enhancement (consisting of MVDR beamformer and postfilter) is further processed by feature extraction/transformation and acoustic modeling following the CHiME-3 baseline ASR system, which provides a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM)-based as well as an HMM-Deep Neural Network (DNN)-based speech recognizer [9].

The remainder of this article is structured as follows: In Section 2 the proposed front-end signal enhancement is introduced in detail, followed by a brief review of the employed ASR system in Section 3. The performance of the front-end speech enhancement is evaluated with respect to word error rates (WERs) of the baseline ASR system, which are presented in Section 4. A conclusion and an outlook to future work is given in Section 5.
2. FRONT-END ENHANCEMENT TECHNIQUES

The front-end speech enhancement considered in this article consists of an MVDR beamformer (provided by the CHiME-3 baseline) and a single-channel coherence-based postfilter. In the following, the baseline MVDR beamformer is briefly reviewed, followed by a detailed presentation of the proposed postfilter.

2.1. Signal model

For a consistent presentation of the front-end speech enhancement considered in this work, we first introduce a signal model which will be used throughout this article.

The $N$ microphone signals of the microphone array in the short-time Fourier transform (STFT) domain at frame $l$ and frequency $f$ are given as:

$$\mathbf{x}(l, f) = \mathbf{h}(l, f)S(l, f) + \mathbf{n}(l, f), \quad (1)$$

where vector

$$\mathbf{x}(l, f) = [X_0(l, f), X_1(l, f), \ldots, X_{N-1}(l, f)]^T \quad (2)$$

contains the microphone signals, $S(l, f)$ denotes the clean source signal, and $\mathbf{n}(l, f)$ includes sensor noise as well as diffuse background noise components and is defined analogously to $\mathbf{x}(l, f)$ in (2). Assuming free-field propagation of sound waves, $\mathbf{h}(l, f)$ represents the steering vector modeling the sound propagation between the desired source located at direction $(\phi_d, \theta_d)$ and all $N$ microphones:

$$\mathbf{h}(l, f) = [e^{-jk_d^T \mathbf{p}_0}, e^{-jk_d^T \mathbf{p}_1}, \ldots, e^{-jk_d^T \mathbf{p}_{N-1}}]^T, \quad (3)$$

where wavevector $\mathbf{k}_d$ is defined as [10]:

$$\mathbf{k}_d = -\frac{2\pi f}{c} \left[\sin(\theta_d) \cos(\phi_d), \sin(\theta_d) \sin(\phi_d), \cos(\theta_d)\right]^T, \quad (4)$$

with speed of sound $c$ and operator $(\cdot)^T$ denoting the transpose of a vector or matrix. $\phi$ and $\theta$ denote azimuth and elevation angle, respectively, and are defined as in [10] with $(\phi, \theta) = (90^\circ, 90^\circ)$ denoting broadside. Furthermore, the $n$-th microphone position in Cartesian coordinates is captured by the three-dimensional vector $\mathbf{p}_n$, $n \in \{0, \ldots, N-1\}$.

The beamformer output $Y_{\text{BF}}(l, f)$ is obtained by multiplying each microphone signal with a complex-valued filter weight $W_n(l, f)$, followed by a summation over all microphone channels:

$$Y_{\text{BF}}(l, f) = \mathbf{w}^H(l, f)\mathbf{x}(l, f), \quad (5)$$

where

$$\mathbf{w}(l, f) = [W_0(l, f), \ldots, W_{N-1}(l, f)]^T \quad (6)$$

contains the beamformer filter coefficients $W_n(l, f)$. Subsequently, the postfilter is applied to the beamformer output signal, yielding the overall output signal

$$Y(l, f) = G(l, f)Y_{\text{BF}}(l, f), \quad (7)$$

where $G(l, f)$ describes the postfilter gains. After front-end signal enhancement, $Y(l, f)$ is fed into the CHiME-3 baseline acoustic back-end system [9].

2.2. Minimum variance distortionless response beamformer

The filter weights of the MVDR beamformer are determined such that the power of the noise components at the output of the beamformer is minimized, subject to a distortionless constraint in target look direction. Thus, the constrained optimization problem of the MVDR beamformer is given as [10]

$$\mathbf{w}_{\text{MVDR}}(l, f) = \arg\min_{\mathbf{w}(l, f)} \mathbf{w}^H(l, f)\mathbf{S}_{\text{nn}}(l, f)\mathbf{w}(l, f) \quad (8)$$

subject to

$$\mathbf{w}^H(l, f)\mathbf{d}(f) = 1, \quad (9)$$

where $\mathbf{S}_{\text{nn}}(l, f)$ is the multichannel spatio-spectral covariance matrix of the noise components at the input of the beamformer, and vector $\mathbf{d}(f)$ in (9) represents the steering vector corresponding to the beamformer’s desired look direction $(\phi_d, \theta_d)$, defined as

$$\mathbf{d}(f) = [e^{-jk_d^T \mathbf{p}_0}, \ldots, e^{-jk_d^T \mathbf{p}_{N-1}}]^T = \mathbf{h}(l, f). \quad (10)$$

Eq. (8) represents the minimization of the noise variance at the output of the beamformer, whereas (9) contains the distortionless constraint which ensures that a plane wave coming
As illustrated in Fig. 1, we apply a postfilter to remove diffuse noise components from the output of the MVDR beamformer. The postfilter gain \( G(l, f) \) at frame \( l \) and frequency \( f \) is given as (11):

\[
G(l, f) = \max \left\{ 1 - \mu \frac{1}{1 + \text{SNR}(l, f)} \cdot G_{\text{min}} \right\}, \quad (12)
\]

with overestimation factor \( \mu \), and gain floor \( G_{\text{min}} \). The postfilter in (12) is a Wiener filter using the short-time SNR to compute the filter gains \( G(l, f) \). In this work, we approximate the short-time SNR in (12) by the estimated Coherent-to-Diffuse Power Ratio (CDR), which is the ratio between direct and diffuse signal components. From (12) it can be seen that a low CDR value, which corresponds to strong diffuse signal components being present at the input of the system, leads to low filter gains and vice versa.

The CDR between two omnidirectional microphones is defined as (12):

\[
\text{CDR}(l, f) = \frac{\Gamma_n(l, f) - \Gamma_s(l, f)}{\Gamma_s(l, f) - \Gamma_n(l, f)}, \quad (13)
\]

where \( \Gamma_s(l, f) \) is the spatial coherence function of both microphone signals. Moreover, the spatial coherence functions for the direct and diffuse sound components are given as

\[
\Gamma_n(l, f) = e^{2\pi f \Delta t}, \quad (14)
\]

\[
\Gamma_d(l, f) = \frac{\text{sinc}(2\pi f \frac{d}{c})}{\text{sinc}(2\pi f \frac{d}{c})}, \quad (15)
\]

respectively, with Time Difference of Arrival (TDOA) \( \Delta t \) and microphone spacing \( d \).

Many different CDR estimators have been proposed in the literature, see, e.g., [13, 14, 15]. The CDR estimator we use in this work was proposed in [12] and is given by (16), where \( \text{Re}\{\cdot\} \) and \( | \cdot | \) represent the real part and magnitude of (\cdot), respectively. Moreover, \( \hat{\Gamma}_s(l, f) \) and \( \text{CDR}(l, f) \) are the estimated coherence and CDR of the two microphone signals, respectively. Note that \( l \) and \( f \) have been omitted in (16) for brevity. As can be seen from (16), the employed estimator does not require the DoA of the speech source, since \( \Gamma_s(l, f) \) is not required for calculating \( \text{CDR}(l, f) \). In [12] it was shown that the employed estimator is unbiased and robust in the sense that deviations of the coherence estimate \( \hat{\Gamma}_s(l, f) \) from the assumed model do not lead to large deviations of the CDR estimate. A more detailed investigation of the employed CDR estimator is available in [12] [16].

When applying the coherence-based postfilter to the output of a beamformer, two aspects need to be considered: First, since the microphone array of the CHiME-3 challenge consists of five forward-facing microphones, the CDR estimator (initially designed for a pair of microphones) has to be adapted to exploit all available microphone signals. To do so, we apply the CDR estimator to every pair of non-failing microphones, i.e., ten pairs for five microphones, to obtain the CDR estimate of each microphone pair. From each of these estimates, we calculate the respective diffuseness values as (16) [17]:

\[
D(l, f) = \frac{1}{1 + \text{CDR}(l, f)}. \quad (17)
\]

Subsequently, we take the arithmetic average of all microphone pair-specific diffuseness values, and calculate the final CDR estimate as

\[
\hat{\text{CDR}}_{\text{In}}(l, f) = \frac{1 - \overline{D}(l, f)}{\overline{D}(l, f)}, \quad (18)
\]

where \( \hat{\text{CDR}}_{\text{In}}(l, f) \) describes the final CDR estimate at the input of the system, and \( \overline{D}(l, f) \) denotes the average diffuseness obtained by calculating the mean of all microphone pair-specific diffuseness values. Second, note that the obtained CDR estimate \( \hat{\text{CDR}}_{\text{In}}(l, f) \) is an estimate of the CDR at the input of the signal enhancement system, i.e., the beamformer. However, what we actually need is the CDR at the output of the beamformer. This can be obtained by applying a correction factor \( A_{\Gamma}(l, f) \) to \( \hat{\text{CDR}}_{\text{In}}(l, f) \). Thus, the CDR estimate at the output of the beamformer \( \hat{\text{CDR}}_{\text{BF}}(l, f) \) is defined as

\[
\hat{\text{CDR}}_{\text{BF}}(l, f) = \frac{\hat{\text{CDR}}_{\text{In}}(l, f)}{A_{\Gamma}(l, f)}, \quad (19)
\]
The HMM-GMM system is designed to provide WERs at relatively low computational costs. In addition, an HMM-DNN system, consisting of beamformer and coherence-based postfilter which is applied to the beamformer output.

where \( A_T(l, f) \) is given as \( \text{[18]} \)

\[
A_T(l, f) = w^H(l, f) J_{\text{diff}}(f) w(l, f),
\]

where \( J_{\text{diff}}(f) \) is the spatial coherence matrix of a diffuse noise field.

Fig. 2 shows the block-diagram of the employed front-end enhancement system, consisting of beamformer and coherence-based postfilter.

3. BACK-END ACOUSTIC MODELING

As indicated in Fig. 1, we employ the acoustic back-end system provided by the CHiME-3 baseline ASR system. It provides an HMM-GMM system, consisting of 2500 tied triphone HMM states which are modeled by 15000 Gaussians. The HMM-GMM system is designed to provide WERs at relatively low computational costs. In addition, an HMM-DNN ASR system providing state-of-the-art ASR performance is contained in the CHiME-3 baseline. It employs a seven-layer DNN with 2048 neurons per hidden layer and is based on the Kaldi toolkit \( \text{[19]} \). The DNN training process includes pre-training using restricted Boltzmann machines, cross-entropy training, and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion. For a more detailed presentation of the baseline ASR systems, see \( \text{[9]} \).

4. EXPERIMENTAL RESULTS

In the following, we investigate the impact of our proposed front-end enhancement on the STFT spectra of a noisy speech utterance, and evaluate the speech recognition accuracy of the front-end with respect to WERs using the CHiME-3 baseline ASR systems.

4.1. Setup and parameters

For all experiments, we use half-overlapping sine windows of 1024 samples to obtain the complex-valued STFT representation of the signals, which is equal to the baseline processing presented in \( \text{[9]} \). The signals were processed at a sampling rate of 16 kHz. The DoA of the desired source, which is required for the MVDR beamformer design, was obtained using the baseline localization algorithm \( \text{[9]} \). For realizing the coherence-based postfilter, we chose gain floor \( G_{\text{min}} = 0.1 \) and overestimation factor \( \mu = 1.3 \). The short-time coherence estimates \( \hat{\Gamma}_x(l, f) \) were obtained by recursive averaging of the auto- and cross-power spectra with forgetting factor \( \lambda = 0.68 \), as in \( \text{[12]} \).

The ASR task included sets of real and simulated noisy utterances in four different environments: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). For each environment, a training set, a development set, and an evaluation set consisting of real and simulated data was provided \( \text{[9]} \).

4.2. Illustration of front-end impact in the STFT domain

In Fig. 3 we illustrate the impact of the MVDR beamformer and the coherence-based postfilter on the STFT spectra of a noisy utterance, with the number of frames \( l \) and frequency \( f \) on the horizontal and vertical axis, respectively. Note that the coarse temporal resolution of the STFT spectra is due to the baseline block-processing. As a reference, the spectrum of the close-talking microphone (channel 0) is shown in Fig. 3(a). It contains the desired utterance plus little background noise. The recorded desired signal is a male speaker saying “Our guess is no” in the café environment. The spectrum of microphone 1 is illustrated in Fig. 3(b). As can be seen, low- as well as high-frequency noise is acquired by the microphone, whereas most of the noise is present in the frequency range of speech. Applying the baseline MVDR beamformer leads to a reduction of the interfering components, as illustrated in Fig. 3(c). A comparison of Fig. 3(c) with Fig. 3(d) shows that applying the coherence-based postfilter to the MVDR beamformer output yields a significant reduction of interference across the entire frequency range, but it also removes low-frequency components of the desired signal. The estimated diffuseness \( D_{\text{BF}}(l, f) \) at the beamformer output is illustrated in Fig. 3(e). Comparing Figs. 3(c) and 3(d) shows that the estimated diffuseness \( D_{\text{BF}}(l, f) \) is very low whenever the desired source is active, which is to be expected, since the CDR will be high whenever the desired source is active. A final comparison of Figs. 3(a) and 3(d) reveals the similarity between the front-end output signal \( Y(l, f) \) and the close-talking microphone signal \( S(l, f) \), which indicates the effectiveness of the proposed front-end signal enhancement technique.

4.3. Evaluation of estimation accuracy

Table I summarizes the average WERs (in %) of the baseline (MVDR) and the extended (MVDR+PF) front-end enhancement obtained for the CHiME-3 baseline HMM-GMM and HMM-DNN ASR (termed HMM-DNN+sMBR in the tables.
Table 1. Average WERs (in %) obtained with the baseline (MVDR) and extended (MVDR+PF) front-end signal enhancement for the baseline HMM-GMM and HMM-DNN ASR systems.

| Acoustic model | Test data | Training data | Development set | Evaluation set |
|----------------|-----------|---------------|-----------------|-----------------|
|                |           |               | Real data       | Sim. data       |
|                |           |               | 18.67           | 18.07           |
|                |           |               | 16.70           | 14.38           |
| HMM-GMM        | Noisy     | Noisy         | 20.87           | 9.67            |
| HMM-DNN+sMBR   | MVDR      | MVDR          | 17.70           | 8.22            |
|                | MVDR+PF   | MVDR+PF       | 16.13           | 11.55           |
| HMM-GMM        | MVDR+PF   | MVDR+PF       | 14.97           | 10.17           |

Table 2. WERs (in %) obtained with the extended front-end signal enhancement for the baseline HMM-DNN ASR system in each scenario.

| Environment | Development set | Evaluation set |
|-------------|-----------------|----------------|
|             | Real data       | Sim. data       |
|             | Real data       | Sim. data       |
| BUS         | 17.63           | 8.94           |
| CAF         | 14.65           | 12.23          |
| PED         | 12.97           | 8.42           |
| STR         | 14.64           | 11.11          |

to be consistent with [9]) systems. The WERs were averaged over all four acoustic environments. In the first column the employed acoustic model is specified. The test and training data sets are indicated in the second and third column, whereas the respective results for the development and evaluation data set are given in the fourth and fifth column. The ASR systems have always been trained on the output signals of the applied front-end enhancement. As a reference, the first row in Table 1 contains the WERs obtained for the noisy unprocessed microphone signals. Note that the results in the case of no front-end enhancement (Noisy) and for the baseline MVDR beamformer (second row in Table 1) only differ slightly from the presented results in [9]. The slight deviations are due to random initialisation and machine-specific issues.

When comparing the results of the HMM-GMM ASR system in the first and second row, one can observe that the baseline front-end enhancement only improves the WERs for simulated data. In the case of real data, the recognition accuracy of the baseline front-end processing is significantly worse than without front-end signal processing. For the HMM-DNN-based recognizer, significant WER improvements can be observed for simulated data, whereas for real data there is no clear advantage of the baseline front-end processing compared to no front-end processing.

A comparison of the results for the HMM-GMM ASR system in the second and third row shows that applying the coherence-based postfilter to the MVDR beamformer output signal drastically decreases the average WER for real data with an improvement of 4.74 and 9.89 percentage points for the development and evaluation data set, respectively. It can also be seen that the WERs of the extended front-end are slightly increased for simulated data. The reason for this may be that the employed postfilter parameters $\mu$ and $G_{\min}$ are suboptimal for the simulated data set. The results for the baseline (MVDR) and the proposed front-end (MVDR+PF) obtained with HMM-DNN ASR system in the second and third row show the same tendencies. Our proposed front-end enhancement yields significantly lower WERs for real data and a worse recognition accuracy for simulated data. In the case of real data, the WERs were decreased by 2.73 and 5.2 percentage points for the development and evaluation data set, respectively, by applying the coherence-based postfilter.

It is interesting to note that for our proposed front-end, the HMM-DNN ASR system only yields a better recognition performance than the HMM-GMM system for the development data, whereas for the real evaluation data the HMM-GMM ASR system achieves lower WERs. Especially for the simulated evaluation data, the HMM-GMM ASR is superior to the HMM-DNN-based recognizer. One explanation for this phenomenon might be a suboptimal architecture of the DNN which we did not optimize as part of this contribution. Finally, we can observe that only applying the postfilter to the MVDR output signal yields significantly lower WERs with both baseline ASR systems for real data, compared to the unprocessed signal, which confirms the effectiveness of our pro-
posed postfilter.

In Table 2 the scenario-specific WERs of our proposed front-end enhancement obtained with the baseline HMM-DNN ASR system are provided. Judging from the obtained WERs, the BUS environments seems to be the most challenging scenario for real data, whereas the highest WER for simulated data was obtained for the café scenario.

5. CONCLUSION

In this contribution to the CHiME-3 challenge, we proposed an extension of the baseline front-end speech enhancement by a coherence-based postfilter. The postfilter is realized as a Wiener filter, where an estimate of the ratio between direct and diffuse signal components at the output of the baseline MVDR beamformer is used as an approximation of the short-time SNR to compute the filter gains. To estimate the ratio between direct and diffuse signal components, we used a DoA-independent estimator, which can be efficiently realized since it only requires an estimate of the auto- and cross-power spectra at the microphone signals. As a consequence, the proposed postfilter has a very low computational complexity as well. Both the baseline and the extended front-end speech enhancement have been evaluated on real and simulated data with respect to WERs using the baseline HMM-GMM and HMM-DNN ASR systems. The results confirmed that the proposed coherence-based postfilter significantly improves the recognition accuracy of the enhanced speech compared to the MVDR beamformer when applied to real data. The improved recognition accuracy in addition to the low computational complexity makes the proposed postfilter very suitable for real-time robust distant speech recognition. Future work includes the analysis of the performance of DoA-dependent CDR estimators for the CHiME-3 data. Also combining DoA-dependent and DoA-independent CDR estimators in different frequency ranges will be investigated. Moreover, using spatial diffuse-ness features as an additional input to a DNN-based acoustic model, as proposed in [20], is another avenue for future work.

6. ACKNOWLEDGEMENT

We would like to thank Stefan Meier and Christian Hofmann for their continuous support and fruitful discussions.

Fig. 3. Illustration of impact of front-end signal processing on the recorded noisy microphone signal, with recorded close-talking desired signal $S(l, f)$ in (a), microphone signal $X_1(l, f)$ in (b), baseline beamformer output signal $Y_{BF}(l, f)$ in (c), and postfilter output signal $Y(l, f)$ in (d). Fig. (e) shows the diffuseness $D_{BF}(l, f)$ which was estimated from the beamformer output signal in (c), and which has been used to compute the postfilter gains.
7. REFERENCES

[1] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, “Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?,” in INTERSPEECH, 2013, pp. 2992–2996.

[2] T. Yoshioka and M.J.F. Gales, “Environmentally robust ASR front-end for deep neural network acoustic models,” Computer Speech and Language (CSL), vol. 31, no. 1, pp. 65–86, 2015.

[3] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” IEEE Trans. Speech Audio Process. (SAP), vol. 11, no. 5, pp. 466–475, Sep. 2003.

[4] A. Krueger and R. Haeub-Umbach, “Model-based feature enhancement for reverberant speech recognition,” IEEE Trans. Audio, Speech and Lang. Process. (ASLP), vol. 18, no. 7, pp. 1692–1707, Sep. 2010.

[5] M.J.F. Gales and Y.-Q. Wang, “Model-based approaches to handling additive noise in reverberant environments,” in Proc. Joint Workshop Hands-free Speech Comm. Microphone Arrays (HSCMA). 2011, pp. 121–126, IEEE.

[6] X. Li and J. Bilmes, “Regularized adaptation of discriminative classifiers,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP). 2006, vol. 1, pp. 237–240, IEEE.

[7] H. Liao, “Speaker adaptation of context dependent deep neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP). 2013, pp. 7947–7951, IEEE.

[8] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP). 2013, pp. 7893–7897, IEEE.

[9] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines.” in Submitted to IEEE Workshop Automat. Speech Recog., Understanding (ASRU). Dec. 2015, IEEE.

[10] H.L. Van Trees, Detection, Estimation, and Modulation Theory, Optimum Array Processing, Detection, Estimation, and Modulation Theory. Wiley, 2004.

[11] E. Haensler and G. Schmidt, Acoustic Echo and Noise Control: A Practical Approach, Wiley-Interscience, 2004.

[12] A. Schwarz and W. Kellermann, “Unbiased coherent-to-diffuse ratio estimation for dereverberation,” in Proc. Int. Workshop Acoustic Echo, Noise Control (IWAENC). Sep. 2014, IEEE.

[13] M. Jeub, C. M. Nelke, C. Beaugeant, and P. Vary, “Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals,” in Proc. 19th European Signal Processing Conference (EUSIPCO), Aug. 2011, pp. 1347–1351.

[14] O. Thiergart, G. Del Galdo, and E.A.P. Habets, “Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP). Mar. 2012, IEEE.

[15] O. Thiergart, G. Del Galdo, and E.A.P. Habets, “On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation,” J. Acoust. Soc. Am. (JASA), vol. 132, pp. 2337, Oct. 2012.

[16] A. Schwarz and W. Kellermann, “Coherent-to-diffuse power ratio estimation for dereverberation,” IEEE Trans. Audio, Speech and Lang. Process. (ASLP), Apr. 2015.

[17] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, “The diffuse sound field in energetic analysis,” J. Acoust. Soc. Am. (JASA), vol. 131, no. 3, pp. 2141–2151, Mar. 2012.

[18] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in Microphone Arrays, Digital Signal Processing, pp. 39–60. Springer Berlin Heidelberg, Jan. 2001.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and others, “The Kaldi speech recognition toolkit,” 2011.

[20] A. Schwarz, C. Hmmer, R. Maas, and W. Kellermann, “Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP). Apr. 2015, IEEE.