Braun and Habets *EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:34
DOI 10.1186/s13636-015-0077-2

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A multichannel diffuse power estimator for dereverberation in the presence of multiple sources

Sebastian Braun* and Emanuël A. P. Habets

Abstract

Using a recently proposed informed spatial filter, it is possible to effectively and robustly reduce reverberation from speech signals captured in noisy environments using multiple microphones. Late reverberation can be modeled by a diffuse sound field with a time-varying power spectral density (PSD). To attain reverberation reduction using this spatial filter, an accurate estimate of the diffuse sound PSD is required. In this work, a method is proposed to estimate the diffuse sound PSD from a set of reference signals by blocking the direct signal components. By considering multiple plane waves in the signal model to describe the direct sound, the method is suitable in the presence of multiple simultaneously active speakers. The proposed diffuse sound PSD estimator is analyzed and compared to existing estimators. In addition, the performance of the spatial filter computed with the diffuse sound PSD estimate is analyzed using simulated and measured room impulse responses in noisy environments with stationary noise and non-stationary babble noise.

Keywords: Dereverberation, Multichannel Wiener filter, Diffuse power estimation

1 Introduction

In speech communication scenarios, reverberation can degrade the speech quality and, in severe cases, the speech intelligibility [1]. State-of-the-art devices such as mobile phones, laptops, tablets, or smart TVs already feature multiple microphones to reduce reverberation and noise. Multichannel approaches are generally superior to single-channel approaches, since they are able to exploit the spatial diversity of the sound scene.

In general, there exist several very different classes of dereverberation algorithms. Algorithms of the first class identify the acoustic system and then equalize it (cf. [1] and the references therein). Given a perfect estimate of the acoustic system described by a finite impulse response, perfect dereverberation can be achieved by applying the multiple input/output inverse theorem [2] (i.e., by applying a multichannel equalizer). However, this approach is not robust against estimation errors of the acoustic impulse responses. As a consequence, this approach is also sensitive to changes in the room and to position changes of the microphones and sources. For a single source, more robust equalizers were recently developed in [3, 4]. Additive noise is usually not taken into account. It should be noted that many multi-source dereverberation algorithms also separate the speech signals of multiple speakers [5], which might not be necessary in some applications.

Algorithms of the second class are proposed, e. g., in [6–9], where the acoustic system was described using an auto-regressive model. The approach proposed in [6] estimates the clean speech for a single source based on multichannel linear prediction by enhancing the linear prediction residual of the clean speech. In [7–9], the received signal is expressed using an autoregressive model and the regression coefficients are estimated from the observations. The clean speech is then estimated using the regression coefficients. While in [8, 9] multi-source models were employed, the algorithm in [8] is evaluated only for a single-talk scenario. Linear prediction-based dereverberation algorithms are typically computationally complex and sensitive to noise. It is, for example, shown in [9] that the complexity and convergence time greatly increases with the number of sources.

*Correspondence: sebastian.braun@audiolabs-erlangen.de
International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany

Algorithms of the third class are used to compute spectral and spatial filters that can also be combined. Exclusively spectral filters are typically single-channel approaches. While early reflections add spectral coloration and can even improve the speech intelligibility, late reverberation mainly deteriorates the speech intelligibility due to overlap-masking [10]. The majority of single-channel dereverberation approaches aim at suppressing only late reverberation using spectral enhancement techniques as proposed in [11, 12] or more recently in [13, 14]. The late reverberant power spectral density (PSD) can be estimated using a statistical model of the room impulse response [15, 16]. The model parameters consist of the reverberation time and in some cases also the direct-to-reverberation ratio (DRR) and need to be known or estimated.

In the multichannel case, spatial or spectro-spatial filters can achieve joint noise reduction and dereverberation, typically in a higher quality than single-channel filters. Recently, an informed spatial minimum mean square error (MMSE) filter based on a multi-source sound field model was proposed in [17]. The reverberation is modeled by a diffuse sound field with a highly time-varying PSD and known spatial coherence. The filter is expressed in terms of the model parameters which include time- and frequency-dependent direction of arrivals (DOAs) and the diffuse sound PSD. As these parameters can be estimated online almost instantaneously, the filter can quickly adapt to changes in the sound field. This spatial filter provides an optimal tradeoff between dereverberation and noise reduction and provides a predefined spatial response for multiple simultaneously active sources. The dereverberation performance is determined by the estimation accuracy of the diffuse sound PSD which is a challenging task because the direct sound and reverberation cannot be observed separately.

There exist already some techniques to estimate the late reverberant or diffuse sound PSD or the signal-to-diffuse ratio (SDR), such as the single-channel method based on Polack's model, that requires prior knowledge about the reverberation time [11] or additionally the DRR [16]. Further suitable methods are the coherence-based SDR estimator proposed in [18] or a linearly constrained minimum variance (LCMV) beamformer placing nulls in the direction of direct sound sources while extracting the ambient sound [19]. In [20], we proposed a method to estimate the diffuse sound PSD using multiple reference signals, while we assumed at most one active source at a known position. In [21], a direct maximum likelihood estimate of the diffuse sound PSD given the observed signals was derived by assuming a noise-free signal model and using prior knowledge of the source position and the diffuse coherence. As the estimator presented in [21] considers only one sound source and no additive noise, we do not consider the estimator in the present work.

In this paper, the aim is to dereverberate multiple simultaneously active sources in the presence of noise without prior knowledge of the position of the sources. The processing is done in the short-time Fourier transform (STFT) domain using the informed spatial filter presented in [17]. In this work, we derive a diffuse sound PSD estimator similar to the one presented in [20] but extended for multiple simultaneously active sources and analyze it in detail. In addition, the influence of the blocking matrix used to create the reference signals is investigated. The PSD estimator depends only on the narrowband DOAs and the noise PSD matrix that can be estimated in advance using existing techniques [22–25]. While we investigate the influence of estimation errors of the DOAs and the noise PSD, these estimators are beyond the scope of this paper. The proposed dereverberation and noise reduction solution is suitable for online processing as the estimators and filters use only current and past observations and the introduced latency depends only on the STFT parameters.

The paper is structured as follows. In Section 2, the signal model is introduced, the spatial filter is derived, and the problem is formulated. Section 3 reviews some existing estimators for the diffuse sound PSD for comparison and derives the proposed estimator. The diffuse sound PSD estimators and the dereverberation system are evaluated in Section 4, and conclusions are drawn in Section 5.

2 Problem formulation

2.1 Signal model

We assume a general scenario with multiple sources in a reverberant and noisy environment. The sound field is captured using an array of M microphones with an arbitrary geometry. In the STFT domain, the microphone signals $Y_m(k,n)$, $m \in \{1,\dots,M\}$ are written into the vector $\mathbf{y}(k,n) = [Y_1(k,n),\dots,Y_M(k,n)]^{\mathrm{T}}$, where k denotes the STFT frequency index and n the time frame index. We describe the sound field using the model proposed in [19], which assumes $L < M$ plane waves propagating in a time-varying diffuse sound field with additive stationary noise, such as sensor noise and ambient noise. The microphone signals are described by

$$\mathbf{y}(k,n) = \sum_{l=1}^{L} \mathbf{a}_l(k,n) X_l(k,n) + \mathbf{d}(k,n) + \mathbf{v}(k,n) \quad (1a)$$

$$= \mathbf{A}(k,n)\,\mathbf{x}(k,n) + \mathbf{d}(k,n) + \mathbf{v}(k,n) \quad (1b)$$

where $X_l(k,n)$ denotes the lth plane wave as received by a reference microphone, $\mathbf{a}_l(k,n)$ is the relative propagation vector of the lth plane wave from the reference microphone to all M microphones, $\mathbf{d}(k,n)$ is the diffuse sound, and $\mathbf{v}(k,n)$ is the additive noise. The sum over l in (1a)

can be expressed as matrix-vector multiplication of the $M \times L$ matrix $\mathbf{A}(k,n) = [\mathbf{a}_1(k,n), \ldots, \mathbf{a}_L(k,n)]$ and the plane wave vector $\mathbf{x}(k,n) = [X_1(k,n), \ldots, X_L(k,n)]^{\mathrm{T}}$. The relative propagation vector of a plane wave for a linear microphone array with omnidirectional sensors is given by

$$\mathbf{a}_l(k,n) = \left[e^{j\lambda(k)r_1 \sin\theta_l(k,n)}, \ldots, e^{j\lambda(k)r_M \sin\theta_l(k,n)}\right]^{\mathrm{T}}, \quad (2)$$

where $\theta_l(k,n)$ is the DOA of the lth plane wave, $r_m = \|\mathbf{r}_m\|_2 - \|\mathbf{r}_{\mathrm{ref}}\|_2$ is the signed distance between the microphone at position $\mathbf{r}_m$ and the reference microphone at position $\mathbf{r}_{\mathrm{ref}}$, both given in cartesian coordinates, and $\lambda(k) = 2\pi\frac{kf_s}{Nc}$ is the spatial frequency with N, f_s, and c being the STFT length, the sampling frequency, and the speed of sound, respectively.

Each of the L plane waves models a directional sound component, which are mutually uncorrelated. Due to the spectral sparsity of speech signals and the modeling of the plane waves independently per time-frequency instant, the number of modeled plane waves L does not have to match the number of physical broadband sound sources exactly. The reverberation is modeled by the diffuse sound component $\mathbf{d}(k,n)$. In principle, $\mathbf{d}(k,n)$ can contain also other non-stationary diffuse noise components such as babble speech that can be observed for example in a cafeteria. The signal component $\mathbf{v}(k,n)$ models stationary or slowly time-varying additive components such as sensor noise and ambient noise.

Assuming that the components in (1) are mutually uncorrelated, the PSD matrix of the microphone signals is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) &= E\left\{\mathbf{y}(k,n)\,\mathbf{y}^{\mathrm{H}}(k,n)\right\} \\ &= \mathbf{A}(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{A}^{\mathrm{H}}(k,n) + \mathbf{\Phi}_{\mathbf{d}}(k,n) \\ &\quad + \mathbf{\Phi}_{\mathbf{v}}(k,n),\end{aligned} \quad (3)$$

where $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ is the PSD matrix of the plane wave signals, $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ is the PSD matrix of the diffuse sound, and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ denotes the noise PSD matrix. Since the L plane waves originate from uncorrelated plane waves, $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ is a diagonal matrix with the PSDs $\phi_l(k,n) = E\left\{|X_l(k,n)|^2\right\}$ on its main diagonal. Note that $\phi_l(k,n)$ is the PSD, at the reference microphone, of the lth plane wave arriving from $\theta_l(k,n)$.

Modeling reverberation as a scaled diffuse sound field holds statistically for the late reverberation tail and a finite time-frequency resolution [26, 27]. The diffuse sound PSD matrix can be expressed in terms of the scaled diffuse coherence matrix

$$\mathbf{\Phi}_{\mathbf{d}}(k,n) = \phi_{\mathrm{d}}(k,n)\,\mathbf{\Gamma}_{\mathrm{diff}}(k), \quad (4)$$

where $\phi_{\mathrm{d}}(k,n)$ is the PSD of the diffuse sound. The form given by (4) holds due to the spatial homogeneity of a diffuse sound field. The ideal diffuse coherence matrix $\mathbf{\Gamma}_{\mathrm{diff}}(k)$ can be calculated for various array configurations and diffuse fields. For a spherical isotropic diffuse sound field captured by omnidirectional microphones, the element with index $p,q \in \{1,\ldots,M\}$ of the matrix $\mathbf{\Gamma}_{\mathrm{diff}}(k)$ is given by [28]

$$\Gamma_{\mathrm{diff}}^{p,q}(k) = \mathrm{sinc}\left(\lambda(k)\,|r_p - r_q|\right), \quad (5)$$

where $\mathrm{sinc}(x) = \frac{\sin(x)}{x}$ for $x \neq 0$ and $\mathrm{sinc}(x) = 1$ for $x = 0$.

Since our goal is to jointly reduce reverberation and noise, we define the interference matrix

$$\mathbf{\Phi}_{\mathbf{u}}(k,n) = \mathbf{\Phi}_{\mathbf{d}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n). \quad (6)$$

In this work, the desired signal, denoted by $Z(k,n)$, is given by the sum of the L plane waves, i.e.,

$$Z(k,n) = \mathbf{1}^{\mathrm{T}}\mathbf{x}(k,n), \quad (7)$$

where $1 = [1,1,\ldots 1]^{\mathrm{T}}$ is a vector of ones with size $L \times 1$. In the following section, we derive a spatial filter that is applied to $\mathbf{y}(k,n)$ to obtain an estimate of $Z(k,n)$.

2.2 Spatial filter design

To estimate the desired signal given by (7), a spatial filter is applied to the microphone signals such that

$$\hat{Z}(k,n) = \mathbf{h}^{\mathrm{H}}(k,n)\,\mathbf{y}(k,n). \quad (8)$$

An estimate of the desired signal $Z(k,n)$ can be obtained using the multichannel Wiener filter (MWF) proposed in [17]. The filter minimizes the interference while preserving all directional components. The MWF is obtained by minimizing the cost function

$$J_{\mathrm{MWF}}(\mathbf{h}) = E\left\{\left|\mathbf{h}^{\mathrm{H}}(k,n)\mathbf{y}(k,n) - \mathbf{1}^{\mathrm{T}}\mathbf{x}(k,n)\right|^2\right\}. \quad (9)$$

The solution is the MWF for multiple plane waves and is given by

$$\mathbf{h}_{\mathrm{MWF}} = \left[\mathbf{A}\mathbf{\Phi}_{\mathbf{x}}\mathbf{A}^{\mathrm{H}} + \mathbf{\Phi}_{\mathbf{u}}\right]^{-1}\mathbf{A}\mathbf{\Phi}_{\mathbf{x}}\,\mathbf{1}. \quad (10)$$

The frequency and time indices k and n are omitted wherever necessary to shorten the notation. For each time-frequency bin, the L columns of the propagation matrix $\mathbf{A}(k,n)$ can be computed using (2) and L narrowband DOAs estimates. In the following, we assume that a suitable narrowband DOA estimator is available (for more information regarding the DOA estimation, we refer the reader to [29, 30]). Given an estimate of $\mathbf{\Phi}_{\mathbf{u}}(k,n)$, the PSD matrix of the plane waves at the microphones can be computed by

$$\widehat{\mathbf{\Phi}}_{\mathbf{Ax}}(k,n) = \mathbf{\Phi}_{\mathbf{y}}(k,n) - \mathbf{\Phi}_{\mathbf{u}}(k,n). \quad (11)$$

If we define the vector containing the plane wave PSDs at the reference microphone $\mathbf{q} = \mathrm{diag}\{\mathbf{\Phi}_{\mathbf{x}}\} = [\phi_1,\ldots,\phi_L]^T$, a least squares estimate of the plane wave PSDs can be obtained using [17]

$$\hat{\mathbf{q}} = (\mathbf{C}^{\mathrm{H}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{H}}\,\mathrm{vec}\{\widehat{\mathbf{\Phi}}_{\mathbf{Ax}}\}, \quad (12)$$

where vec{·} are the columns of a matrix stacked into a column vector and the $L^2 \times L$ matrix $\mathbf{C} = \left[\text{vec}\left\{\mathbf{a}_1\mathbf{a}_1^{\text{H}}\right\}, \ldots, \text{vec}\left\{\mathbf{a}_L\mathbf{a}_L^{\text{H}}\right\}\right]$. The $L \times 1$ vector obtained by (12) contains the estimated plane wave PSDs that are on the main diagonal of the matrix $\mathbf{\Phi}_{\mathbf{x}}(k,n)$, and all off-diagonal elements are zero since we assume uncorrelated plane waves.

The remaining challenge is to estimate the interference PSD matrix $\mathbf{\Phi}_{\mathbf{u}}(k,n)$. The stationary or slowly time-varying noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ is observable when the speakers are inactive and can be estimated using, e. g., [22–25]. In contrast, the diffuse sound PSD matrix $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ that originates from reverberation cannot be observed separately from the desired speech. Assuming that we know the spatial coherence of the diffuse sound field, our aim is to estimate the diffuse sound PSD $\phi_{\text{d}}(k,n)$. Given $\phi_{\text{d}}(k,n)$ and $\mathbf{\Gamma}_{\text{diff}}(k)$, we can then calculate $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ using (4).

3 Estimation of the diffuse sound PSD

In this section, we first review some estimators that can be used to obtain an estimate of the PSD of diffuse or reverberant sound and then derive a novel estimator that takes the presence of multiple plane waves as given by the signal model (1) into account.

3.1 Existing estimators

3.1.1 Based on a statistical reverberation model

The first estimator is based on a single-channel late reverberant PSD estimator proposed in [16]. This estimator is derived using a statistical reverberation model that depends on the (in general frequency dependent) room reverberation time $T_{60}(k)$ and the DRR $\kappa(k)$, which varies with the source-microphone distance. Let us first define $\phi_{\text{xd}}^m(k,n)$ as the reverberant signal PSD at the mth microphone of which an estimate is given by the mth element on the diagonal of the matrix $\mathbf{\Phi}_{\mathbf{y}}(k,n) - \mathbf{\Phi}_{\mathbf{v}}(k,n)$. The late reverberant PSD at the mth microphone $\phi_{\text{d}}^m(k,n)$ is estimated by [16]

$$\begin{aligned}\hat{\phi}_{\text{d}}^m(k,n) = [1-\kappa(k)]\ e^{-2\alpha(k)RN_L}\ \hat{\phi}_{\text{d}}^m(k,n-N_{\text{L}}) \\ + \kappa(k)\ e^{-2\alpha(k)RN_L}\ \phi_{\text{xd}}^m(k,n-N_{\text{L}}),\end{aligned} \tag{13}$$

where N_{L} corresponds to the number of frames between the direct sound and the start of the late reverberation, $\alpha(k) = 3\ln(10)/(T_{60}(k)f_{\text{s}})$ is the reverberation decay constant, and R is the hop size. As the diffuse sound field is assumed to be spatially homogeneous, the estimate of the diffuse sound PSD $\phi_{\text{d}}(k,n)$ can be obtained by spatially averaging $\hat{\phi}_{\text{d}}^1(k,n)\ldots\hat{\phi}_{\text{d}}^M(k,n)$ as [31]

$$\hat{\phi}_{\text{d}}^{\text{LRSV}}(k,n) = \frac{1}{M}\sum_{m=1}^{M}\hat{\phi}_{\text{d}}^m(k,n). \tag{14}$$

3.1.2 Based on the spatial coherence

The second estimator is the *coherence-based signal-to-diffuse ratio estimator* (CSDRE) [32]; a similar estimator is also presented in [33]. It calculates the SDR in mixed sound fields by exploiting the spatial coherence of a single directional component and the diffuse sound field. The diffuse PSD can then be extracted from the noise-free PSD and the SDR estimate. Let us denote $\Phi^{p,q}$ as the element p,q of any PSD matrix. The coherence of the mixed sound field between the microphones p and q is calculated from the input signal coherence and taking into account the additive noise as

$$\gamma_{\text{s}}^{p,q} = \frac{\Phi_y^{p,q}}{\sqrt{\Phi_y^{p,p}-\Phi_v^{p,p}}\sqrt{\Phi_y^{q,q}-\Phi_v^{q,q}}}. \tag{15}$$

As shown in [32], the SDR estimator can be calculated with (15), a DOA estimate, and the diffuse coherence between the microphones p,q given in (5). The SDR estimate is first calculated for each possible microphone pair, which results in $M!/((M-2)!\cdot 2)$ estimates and is then averaged over all microphone pair combinations assuming that the direct sound PSD is equal at all microphones according to (2). Finally, the diffuse PSD can be obtained by

$$\hat{\phi}_{\text{d}}^{\text{CSDRE}}(k,n) = \frac{\frac{1}{M}\sum_{m=1}^{M}\phi_{\text{xd}}^m(k,n)}{\text{SDR}(k,n)+1}. \tag{16}$$

3.1.3 Based on an ambient beamformer

A third diffuse sound PSD estimator was proposed in [19]. An ambient beamformer (ABF) is derived that is intended to capture the ambient sound, which is assumed to correlate well with the diffuse sound. This is achieved by minimizing the noise $\mathbf{v}(k,n)$ while placing nulls to the DOAs of the directional sound components and placing a unit response to the direction that has the maximum angular distance to all L DOAs. The ambient beamformer $\mathbf{h}_{\text{ABF}}$ is derived by solving

$$\mathbf{h}_{\text{ABF}}(k,n) = \arg\min_{\mathbf{h}} \mathbf{h}^{\text{H}}\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h} \tag{17a}$$

subject to

$$\mathbf{h}^{\text{H}}\mathbf{A}(k,n) = \mathbf{0}_{1\times L} \tag{17b}$$

$$\mathbf{h}^{\text{H}}\mathbf{a}_0(k,n) = 1, \tag{17c}$$

where $\mathbf{a}_0$ is a propagation vector corresponding to the DOA with maximum angular distance to all L DOAs. For further details, the reader is referred to [19]. The diffuse sound PSD estimate is then obtained by

$$\hat{\phi}_{\text{d}}^{\text{ABF}} = \frac{\mathbf{h}_{\text{ABF}}^{\text{H}}\mathbf{\Phi}_{\mathbf{y}}\mathbf{h}_{\text{ABF}} - \mathbf{h}_{\text{ABF}}^{\text{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{h}_{\text{ABF}}}{\mathbf{h}_{\text{ABF}}^{\text{H}}\mathbf{\Gamma}_{\text{diff}}\mathbf{h}_{\text{ABF}}}. \tag{18}$$

3.2 Discussion of the existing estimators

The following observations can be made regarding the existing estimators discussed in the previous section:

- The estimator presented in Section 3.1.1 requires prior information about the frequency-dependent reverberation time and DRR. In [34], it is shown that existing T_{60} estimators are strongly biased at low signal-to-noise ratios (SNRs). Furthermore, T_{60} estimators typically require a few seconds of data and therefore cannot adapt quickly to changes in the reverberation time.
- The single-source model as assumed in the approach presented in Section 3.1.2 has been shown to be inaccurate in multi-talk scenarios in [35].
- The single- and dual-channel approaches presented in Section 3.1.1 and 3.1.2 do not directly take all microphones into account.
- The estimator presented in Section 3.1.3 is suboptimal as it aims not directly to estimate the diffuse sound PSD. Furthermore, it requires a specific look direction.

To reduce the amount of required prior knowledge and to relax the assumptions for the diffuse PSD estimator, we propose a new estimator in the following section that

1. is able to respond immediately to changes in the sound field and is independent of the reverberation time and DRR,
2. is based on the multi-wave signal model (1), and
3. directly estimates the diffuse sound PSD using all microphones.

3.3 Maximum likelihood estimator using reference signals

In this section, we derive an estimator for the diffuse sound PSD $\phi_{\mathrm{d}}(k,n)$ based on multiple reference signals. In Section 3.3.1, the computation of the reference signals is described. In Section 3.3.2, a maximum likelihood estimator (MLE) for the diffuse sound PSD is derived based on the computed reference signals.

3.3.1 Generating the reference signals

The reference signal vector $\widetilde{\mathbf{u}}(k,n)$ is obtained as the output of a blocking matrix (BM) $\mathbf{B}(k,n) \in \mathbb{C}_{M\times K}$

$$\widetilde{\mathbf{u}}(k,n) = \mathbf{B}^{\mathrm{H}}(k,n)\,\mathbf{y}(k,n), \tag{19}$$

which creates a set of K reference signals which contain no direct signal components. Therefore, the blocking matrix has to fulfill the constraint

$$\mathbf{B}^{\mathrm{H}}(k,n)\,\mathbf{A}(k,n) = \mathbf{0}_{K\times L}. \tag{20}$$

In general, there is no unique solution for (20). Two common approaches are reviewed here: the *eigenspace*-based BM [36] and the *sparse* BM [37]. A blocking matrix for M microphones with L directional constraints consists of up to $K = M-L$ linearly independent columns. The *eigenspace BM* [36] is constructed as

$$\mathbf{B}_{\mathrm{e}} = \left[\mathbf{I}_{M\times M} - \mathbf{A}\,(\mathbf{A}^{\mathrm{H}}\mathbf{A})^{-1}\,\mathbf{A}^{\mathrm{H}}\right]\;\mathbf{I}_{M\times K}, \tag{21}$$

where $\mathbf{I}_{M\times K}$ is a truncated identity matrix, that selects the first K columns of the expression in square brackets. Using the eigenspace BM, each output signal of the BM is a linear combination of all microphone signals, where all coefficients of $\mathbf{B}_{\mathrm{e}}$ are non-zero. In contrast, the *sparse BM* [38] forms each output depending only on $L+1$ adjacent channels. Let $\mathbf{A}_{1:L,1:L}$ denote a matrix containing the first L rows and columns of $\mathbf{A}$, $\mathbf{A}_{m,:}$ denotes the mth row of $\mathbf{A}$, and $\boldsymbol{\beta}_m = \left(\mathbf{A}_{1:L,1:L}^{-1}\right)^{\mathrm{H}}\,\mathbf{A}_{m,:}^{\mathrm{H}}$. Then, the sparse BM is calculated as [37]

$$\mathbf{B}_{\mathrm{s}} = \begin{bmatrix} -\boldsymbol{\beta}_{L+1} & \cdots & -\boldsymbol{\beta}_{M} \\ \multicolumn{3}{c}{\mathbf{I}_{(M-L)\times(M-L)}} \end{bmatrix}. \tag{22}$$

Using (3), (4), and (19), it follows that the PSD matrix $\widetilde{\boldsymbol{\Phi}}_{\mathbf{u}}(k,n)$ of the blocking matrix output signal (19) depends only on the residual diffuse and residual noise PSD matrices, i.e.,

$$\begin{aligned}\widetilde{\boldsymbol{\Phi}}_{\mathbf{u}} &= \mathbf{B}^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{y}}\mathbf{B} \\ &= \underbrace{\mathbf{B}^{\mathrm{H}}\mathbf{A}\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{A}^{\mathrm{H}}\mathbf{B}}_{\mathbf{0}_{K\times K}} + \phi_{\mathrm{d}}\,\underbrace{\mathbf{B}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{diff}}\,\mathbf{B}}_{\widetilde{\boldsymbol{\Gamma}}_{\mathrm{diff}}} + \underbrace{\mathbf{B}^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{B}}_{\widetilde{\boldsymbol{\Phi}}_{\mathbf{v}}}\end{aligned} \tag{23}$$

where the matrices $\widetilde{\boldsymbol{\Gamma}}_{\mathrm{diff}}(k,n)$ and $\widetilde{\boldsymbol{\Phi}}_{\mathbf{v}}(k,n)$ denote the diffuse coherence matrix and the noise PSD matrix at the output of the blocking matrix, respectively. The direct sound PSD is zero due to (20).

3.3.2 Derivation of the maximum likelihood estimator

As proposed in [39], we introduce the error matrix that models the estimation errors of $\widetilde{\boldsymbol{\Phi}}_{\mathbf{u}}$ and $\widetilde{\boldsymbol{\Phi}}_{\mathbf{v}}$ as

$$\boldsymbol{\Phi}_{\mathbf{e}} \;=\; \underbrace{\widetilde{\boldsymbol{\Phi}}_{\mathbf{u}} - \widetilde{\boldsymbol{\Phi}}_{\mathbf{v}}}_{\widetilde{\boldsymbol{\Phi}}_{\mathbf{d}}} \;\; - \phi_{\mathrm{d}}\,\widetilde{\boldsymbol{\Gamma}}_{\mathrm{diff}}. \tag{24}$$

The matrix $\widetilde{\boldsymbol{\Phi}}_{\mathbf{d}}(k,n)$ can be estimated from the measured PSD matrix $\widetilde{\boldsymbol{\Phi}}_{\mathbf{u}}(k,n) = E\{\widetilde{\mathbf{u}}(k,n)\,\widetilde{\mathbf{u}}^{\mathrm{H}}(k,n)\}$ with (19) and the residual noise PSD matrix $\widetilde{\boldsymbol{\Phi}}_{\mathbf{v}}(k,n)$. As in prior work [20, 39], we assume the real and imaginary elements of $\boldsymbol{\Phi}_{\mathbf{e}}(k,n)$ to be independent zero-mean Gaussian distributions with equal variance. This is however not the case for the diagonal elements which are strictly real valued. Therefore, we define an operator $\mathcal{V}$ that creates a vector containing all real elements and all off-diagonal imaginary elements of a complex matrix $\boldsymbol{\Phi}$ of size $K\times K$ as

$$\begin{aligned}\mathcal{V}\{\boldsymbol{\Phi}\} = \big[&\Re\{\widetilde{\Phi}^{1,1}\},\,\Re\{\widetilde{\Phi}^{p,q}\},\,\ldots,\\ &\Im\{\widetilde{\Phi}^{1,2}\},\,\Im\{\widetilde{\Phi}^{i,j}\},\,\ldots\big]^{\mathrm{T}},\end{aligned} \tag{25}$$

where $p, q \in \mathbb{N}\{1,\ldots,K\}$ and $i, j \in \mathbb{N}\{1,\ldots,K\}$ with $i \neq j$. The column vector $\mathcal{V}\{\mathbf{\Phi}\}$ is of length $2K^2 - K$. Using this operator, we define the error vector $\mathcal{V}\{\mathbf{\Phi}_\mathbf{e}(k,n)\}$. The probability density function of this error vector can be modelled as a multivariate Gaussian distribution with zero mean and covariance $\sigma^2\mathbf{I}$ as

$$f\left(\mathcal{V}\{\mathbf{\Phi}_\mathbf{e}(k,n)\}\right) = \frac{1}{(\sqrt{2\pi}\sigma)^{2K^2-K}} \times \exp\left(-\frac{(\mathbf{m}-\phi_\mathrm{d}\,\mathbf{n})^\mathrm{T}(\mathbf{m}-\phi_\mathrm{d}\,\mathbf{n})}{2\sigma^2}\right) \tag{26}$$

where $\mathbf{m} = \mathcal{V}\left\{\widetilde{\mathbf{\Phi}}_\mathbf{d}\right\}$ and $\mathbf{n} = \mathcal{V}\{\widetilde{\mathbf{\Gamma}}_\mathrm{diff}\}$. By maximizing the log-likelihood function $\log(f)$, we obtain the least squares solution for $\mathbf{n}^\mathrm{T}\mathbf{n} \neq 0$

$$\hat{\phi}_\mathrm{d} = \max\left\{0,\ \left(\mathbf{n}^\mathrm{T}\mathbf{n}\right)^{-1}\mathbf{n}^\mathrm{T}\mathbf{m}\right\}, \tag{27}$$

where the $\max\{\cdot\}$ operation is included to ensure that the estimated PSD is positive also in the presence of estimation errors. Although we excluded the imaginary diagonal elements, it can be shown that the result is mathematically equivalent to the solution obtained in [20].

3.4 Dereverberation system overview

The system can be summarized as follows. Firstly, a microphone array captures the sound components. From the observed signals, the DOAs are estimated, which are used to construct the blocking matrix and the spatial filter. From the K blocking matrix outputs, the diffuse PSD is estimated and the interference matrix is constructed together with the noise PSD matrix that can be observed during speech pauses. Figure 1 shows the entire proposed system. Note that the proposed diffuse sound PSD estimator utilizes the DOAs and the noise PSD matrix that are also required to compute the spatial filter and hence can be implemented without significantly increasing the computational complexity of the entire dereverberation system.

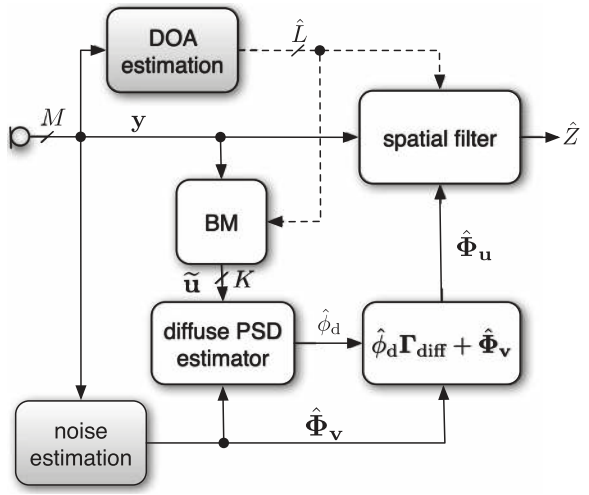

Fig. 1 Complete dereverberation system. Proposed dereverberation system for L sources and M microphones using a spatial filter. The late reverberant PSD is estimated from K reference signals using a maximum likelihood estimator. The estimators denoted by the *grey blocks* are beyond the scope of this paper

4 Performance evaluation

For all simulations, the following parameters were used: a sampling frequency of $f_\mathrm{s} = 16$ kHz, a hamming window of length of $N_\mathrm{win} = 32$ ms, a FFT length of $N = 2N_\mathrm{win}$, a hop size of $N_\mathrm{hop} = 0.25\,N_\mathrm{win}$ and recursive averaging for the online estimated PSD matrices with a time constant of 70 ms. The stationary noise PSD matrix was calculated in advance during periods of speech absence.

4.1 Analysis of the blocking matrices

A detailed evaluation of the eigenspace and sparse BM is given in [38]. There it is shown that for accurately estimated propagation vectors, the blocking ability of both BMs is in theory equal, but if the estimation accuracy is low, the blocking ability of the sparse BM is slightly lower compared to the eigenspace BM.

Figure 2 shows the beampatterns of the two blocking matrices for the DOAs $\{-73°, -51°, 10°, 21°\}$ using a uniform linear array (ULA) of $M = 8$ microphones with 2 cm spacing, where 0° is the broadside direction. Since the beampattern of $\mathbf{B}$ at each beamformer output (i.e.,

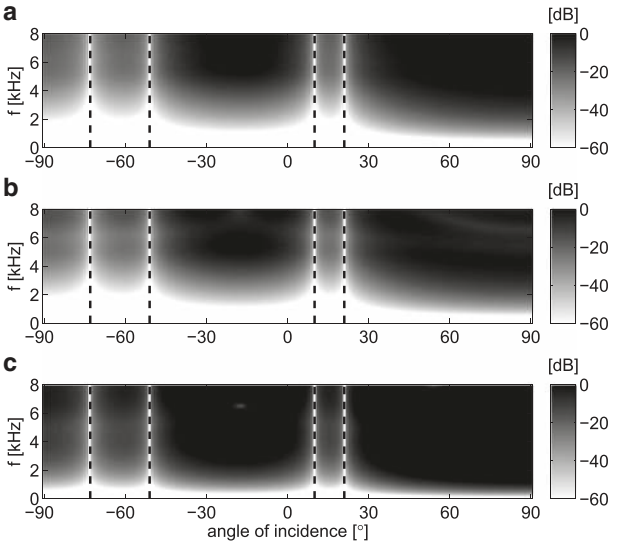

Fig. 2 Blocking matrix beampatterns. Beampatterns of the eigenspace and sparse blocking matrices for $L = 4$ broadband sources. **a** Eigenspace BM, last column. **b** Sparse BM, first column. **c** Sparse BM, last column. The DOAs are marked as *dashed lines*

number of its columns K) of the eigenspace BM is very similar, it is only shown for the last column. In contrast, the beampatterns of the sparse BM vary clearly. The low frequency performance of the sparse BM increases for each output element, due to the increasing spacing between the employed microphone pairs. In Fig. 2, it can be observed that the sparse BM attenuates low frequencies of ambient directions less or equal than the eigenspace BM, depending on the output element.

The average output gain of the blocking matrix $\mathbf{B}$ to a sound field with the coherence matrix $\mathbf{\Gamma}(k)$ is given by

$$G_{\Gamma}(k) = \mathrm{tr}\left\{\mathbf{B}^{\mathrm{H}}(k)\mathbf{\Gamma}(k)\mathbf{B}(k)\right\}, \tag{28}$$

where $\mathbf{\Gamma}(k)$ is either the ideal diffuse coherence matrix (5) or the identity matrix for spatially white noise fields. The power of diffuse and spatially white noise fields at the BM output is shown in Fig. 3. We can observe in Fig. 3 that the sparse BM attenuates diffuse sound less than the eigenspace BM, which might be an advantage for our application. On the other hand, spatially white noise is highly amplified by the sparse BM whereas the eigenspace BM slightly suppresses the noise.

4.2 Estimation considering multiple waves

We now analyze the performance of the proposed diffuse PSD estimator while varying the number of estimated simultaneous arriving plane waves $\hat{L}$ that might differ in practice from the actual number of directional sources L. For this experiment, four directional sound components are simulated. All source signals consist of independent white Gaussian noise, and the sources are randomly distributed around the array on the horizontal half plane with a random distance in the farfield of the array. The diffuse sound signals $\mathbf{d}(k,n)$ are generated using independent and identically distributed (i. i. d.) noise signals using the method proposed in [40]. The spatial coherence between the signals $\mathbf{d}(k,n)$ is chosen as the coherence of an ideal diffuse field (5) and are added with an SDR of 10 dB. The additive noise signals $\mathbf{v}(k,n)$ are simulated as well as i. i. d. processes with an SNR of 50 dB.

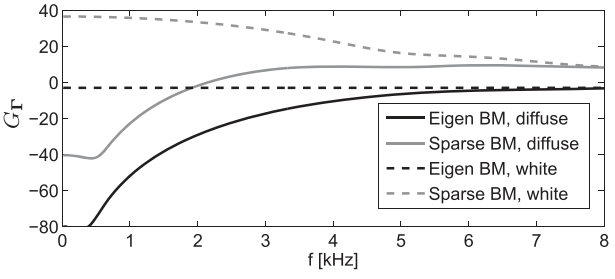

Fig. 3 Blocking matrix output gain. Average blocking matrix output gain for diffuse and spatially white noise fields

The soundfield is captured by a ULA of $M = 8$ microphones with an inter-microphone spacing of 2 cm. In this experiment, the DOAs of the L directional sound sources are known and are successively taken into account plus one extra DOA to investigate the effect of overestimation of L, i.e., $\hat{L} \in \{1,\ldots,L+1\}$. At the position of the extra DOA, no source is active. Note that the number of reference signals K, i.e., the length of vector $\widetilde{\mathbf{u}}(k,n)$, decreases with an increasing number of plane waves $\hat{L}$ taken into account.

Figure 4 shows the logarithmic estimation error $\mathrm{LE}(\hat{\phi}_{\mathrm{d}}) = \mathrm{LE}_{\mathrm{o}}(\hat{\phi}_{\mathrm{d}}) + \mathrm{LE}_{\mathrm{u}}(\hat{\phi}_{\mathrm{d}})$ of the diffuse PSD estimates, decomposed into overestimation $\mathrm{LE}_{\mathrm{o}}(\hat{\phi}_{\mathrm{d}})$ and underestimation $\mathrm{LE}_{\mathrm{u}}(\hat{\phi}_{\mathrm{d}})$ as computed by

$$\mathrm{LE}_{\mathrm{o}}(\hat{\phi}_{\mathrm{d}}) = \frac{1}{|\mathcal{T}|}\sum_{k,n}\left|\min\left\{0,\ 10\log_{10}\frac{\phi_{\mathrm{d}}(k,n)}{\hat{\phi}_{\mathrm{d}}(k,n)}\right\}\right| \tag{29a}$$

$$\mathrm{LE}_{\mathrm{u}}(\hat{\phi}_{\mathrm{d}}) = \frac{1}{|\mathcal{T}|}\sum_{k,n}\left|\max\left\{0,\ 10\log_{10}\frac{\phi_{\mathrm{d}}(k,n)}{\hat{\phi}_{\mathrm{d}}(k,n)}\right\}\right|, \tag{29b}$$

where the ideal diffuse PSD is obtained as the spatial average of the instantaneous diffuse sound power over all microphones, i. e., $\phi_{\mathrm{d}}(k,n) = \mathbf{d}^{\mathrm{H}}(k,n)\mathbf{d}(k,n)/M$, and $(n,k) \in \mathcal{T}$ is the set of time-frequency points, where the ideal diffuse PSD is above a certain threshold. The errors $\mathrm{LE}_{\mathrm{o}}(\hat{\phi}_{\mathrm{d}})$ and $\mathrm{LE}_{\mathrm{u}}(\hat{\phi}_{\mathrm{d}})$ are plotted on top of each other, such that the total bar height shows the total error $\mathrm{LE}(\hat{\phi}_{\mathrm{d}})$.

The estimation accuracy increases by increasing the number of directional constraints $\hat{L}$ for the BM. When the number of DOAs exceeds the actual number of plane waves ($\hat{L} > 4$), we observe no significant performance degradation. The eigenspace BM is slightly more suited for $L = 1$, whereas the sparse BM performs slightly better for $L > 1$. However, for unknown L, there is no significant performance difference between both tested BMs. In the

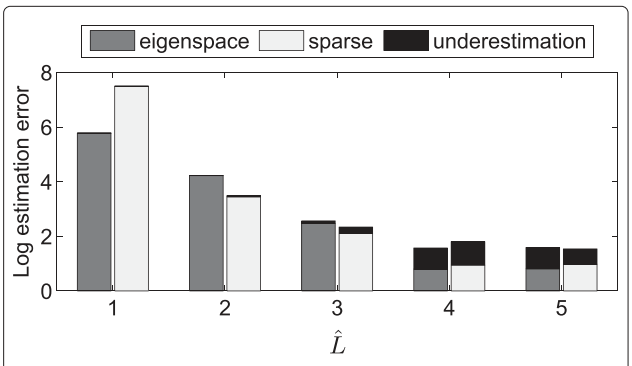

Fig. 4 Log error for different numbers of directional constraints. Accuracy improvement of the proposed diffuse PSD estimator for different blocking matrices for an increasing number of directional constraints

remainder of this work, we use the eigenspace BM which has been found to be more robust against DOA estimation errors [38].

4.3 Robustness against estimation errors

The accuracy of the proposed estimator depends basically on two parameters. The estimated DOAs and the estimated noise PSD matrix. The performance of the DOA estimation is mainly degraded by strong reverberation and noise. The robustness in the presence of estimation errors is analyzed using two experiments.

In the first experiment, we investigate the influence of DOA estimation errors. For this experiment, a scenario with a single speaker was simulated. The direct sound of the speaker was captured by a 4-microphone ULA with 4 cm microphone spacing. The diffuse noise is created as a noise field with the spatial coherence $\mathbf{\Gamma}_{\text{diff}}(k)$ and the noise amplitude was modulated by the smooth temporal envelope of the speech to simulate reverberation. These diffuse signals were added with a long-term SDR of 10 dB. Additional stationary white Gaussian noise was added with an SNR of 80 dB. To model the DOA estimation errors, a zero-mean Gaussian process with standard deviation σ_{DOA} is added to the known DOA θ_1 as

$$\hat{\theta}_1(k,n) = \theta_1 + \theta_{\text{e}}(k,n), \tag{30}$$

where $\theta_{\text{e}}(k,n)$ is the DOA error and $\sigma^2_{\text{DOA}} = E\left\{\theta^2_{\text{e}}(k,n)\right\}$ is the error variance. The evaluation is carried out over utterances by six different speakers. The logarithmic error with over- and underestimation (29) of the proposed estimator for different error variances is shown in Fig. 5. A DOA variance below 5° shows no significant influence on the estimation accuracy of the diffuse sound PSD. Large DOA estimation errors lead mainly to overestimation of the diffuse PSD due to leakage of the direct signal through the BM.

In the second experiment, we evaluated the influence of noise PSD estimation errors depending on the diffuse-to-noise ratio (DNR). We assumed spatially uncorrelated homogenous noise, i.e., $\mathbf{\Phi}_{\mathbf{v}} = \phi_v\,\mathbf{I}$, and the DNR is given by ϕ_d/ϕ_v. The noise estimation error was modeled by an over/underestimation factor c_v of the true noise PSD matrix, i. e., the estimated noise PSD matrix is modeled by $\hat{\mathbf{\Phi}}_{\mathbf{v}} = c_v\,\phi_v\,\mathbf{I}$. In Fig. 6, the relative diffuse PSD estimation error defined as $\Delta_{\text{d}} = \hat{\phi}_{\text{d}}/\phi_{\text{d}}$ is shown, where $\hat{\phi}_{\text{d}}$ is estimated using $\tilde{\mathbf{\Phi}}_{\mathbf{d}} = \mathbf{B}^{\text{H}}(\phi_{\text{d}}\,\mathbf{\Gamma}_{\text{diff}} + \mathbf{\Phi}_{\mathbf{v}})\mathbf{B} - \hat{\mathbf{\Phi}}_{\mathbf{v}}$ with (24) and finally applying (27). A relative estimation error Δ_{d} of 0 dB indicates a perfect estimation, whereas positive values indicate overestimation and negative value underestimation. For high DNRs, underestimation of the noise has only a very small effect on the relative estimation error Δ_{d}. When the noise is so much overestimated that the power of $\widetilde{\mathbf{\Phi}}_{\mathbf{d}}$ in (24) is basically zero, the estimated diffuse power is consequently zero, which results in maximum underestimation as can be seen as the large white area. When the noise is underestimated at low DNRs, the diffuse PSD is overestimated rather proportionally. For positive DNRs, the diffuse estimation error is always very small. However, if the DNR is low, the emphasis lies on noise reduction and diffuse PSD estimation errors do not have a severe negative effect on the spatial filter given by (10).

4.4 Performance in time-varying diffuse noise fields

We now analyze the estimator's performance in a time-varying diffuse sound field. In this experiment, a noise field with an ideal diffuse coherence was simulated in the same manner as in Sections 4.2 and 4.3. Two sources were simultaneously active at positions (−15°, 1.4 m) and (59°, 2.7 m), where the distance is measured from the center of the array. Only the direct path of the two sources was simulated, whereas the reverberation was simulated as a diffuse noise field that was shaped by the temporal envelope of the sum of both speech sources and added with an SDR of 10 dB. Spatially and temporally white noise was added with an SNR of 50 dB. Figure 7 shows the broadband ideal diffuse PSD and two settings for two ULAs of 4 and 8 microphones with 2 cm spacing. The narrowband

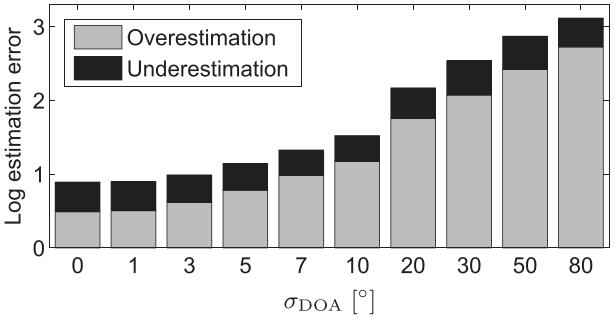

Fig. 5 Influence of DOA estimation errors on the diffuse PSD estimation accuracy

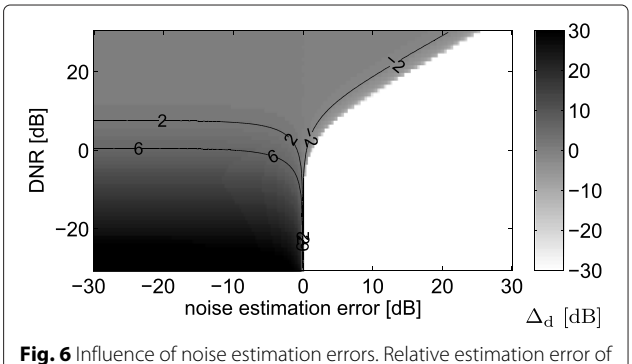

Fig. 6 Influence of noise estimation errors. Relative estimation error of the diffuse PSD Δ_{d} as a function of the noise estimation error and the DNR

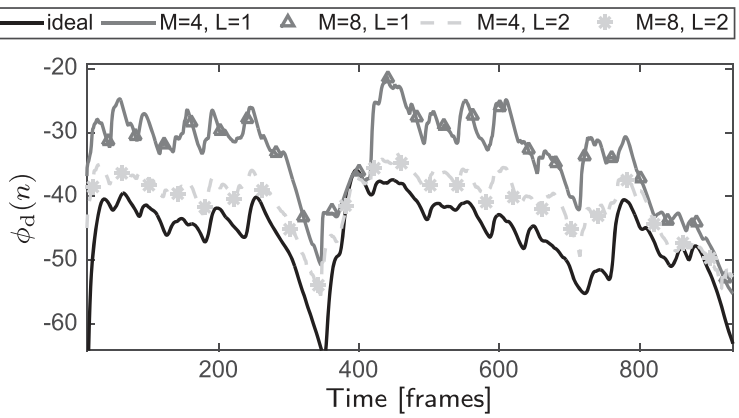

Fig. 7 Tracking of a time-varying diffuse noise field in the presence of two direct source signals. The lines for $M = 8$ are omitted since they are almost identical to the corresponding case with $M = 4$

DOAs are estimated online using TLS-ESPRIT [41] either estimating $\hat{L} = 1$ or $\hat{L} = 2$ DOAs per time-frequency bin. The true broadband diffuse PSD is drawn in black. We observe that by simultaneously blocking two instead of one plane waves in the reference signals $\widetilde{\mathbf{u}}(k, n)$, the accuracy of the estimator can be increased, while increasing the number of microphones has almost no effect on the estimation accuracy. Furthermore, it can be seen that the estimator is able to track the temporal changes.

4.5 Comparison to existing diffuse PSD estimators

In this section, we evaluate the performance of the proposed diffuse PSD estimator and the three estimators described in Sections 3.1.1–3.1.3, denoted by LRSV, CSDRE, and ABF, respectively. A ULA of $M = 8$ microphones with 2 cm spacing was simulated in a reverberant room of size $6 \times 5 \times 4$ m with a $T_{60} = 500$ ms using the well-known image method [42]. Two speech sources are located at 20° and −45° from the broadside direction of the array at distances of 2.7 and 1.9 m, respectively. White noise was added with different levels, described by the iSNR.

The logarithmic estimation error (29), where the diffuse signal component $\mathbf{d}(k, n)$ is the reverberant speech signal component 40 ms after the direct sound, is shown in Fig. 8. The ABF and the proposed MLE were computed by assuming either $\hat{L} = 1$ or $\hat{L} = 2$ simultaneous arriving plane waves, estimated via TLS-ESPRIT. The CSDRE is using the TLS-ESPRIT DOA estimator with $\hat{L} = 1$. The LRSV estimator is computed using the ideal parameters for the simulated reverberation time and DRR. Figure 8a shows the results obtained using a single active speech source; Fig. 8b shows the results for two continuously active speech sources. It can be observed that the ABF approach is very sensitive to noise and has a decreasing performance for decreasing iSNR. All other estimators are quite robust against noise and show only a significantly increasing error for very low iSNRs. The CSDRE has the highest overestimation in all situations, which is more critical than underestimation since it causes distortion of the desired signal. The LRSV estimator performs best with very low overestimation. The proposed MLE performs slightly worse than the LRSV with ideal parameters but better than the other estimators. The use of $\hat{L} = 2$ yields a lower overestimation in most situations for MLE and ABF, which is advantageous in terms of audible artifacts caused by overestimation.

The LRSV requires in addition to the noise PSD an estimate of the typically frequency-dependent reverberation time (which is here almost frequency independent due to the simulated impulse responses), the DRR, and the start time of the late reverberation, which are here assumed to be known. Especially at low iSNRs, online estimates of these parameters are strongly biased and hard to obtain [34], which is not reflected in the evaluation in Fig. 8. Note that the DOA-dependent approaches in this scenario use estimated DOAs without prior information and therefore contain estimation errors.

Since the performance of the LRSV estimator depends on the T_{60} parameter, we analyzed the performance as a function of this parameter. In the following experiment, the DRR was fixed and corresponds to the ideal value. The scenario is identical to the above two speaker scenario but the iSNR was set to 30 dB. Although the true reverberation time was $T_{60} = 500$ ms, the parameter $\hat{T}_{60}$ influencing (13) was varied between 100 and 1200 ms, which can be the case in the presence of T_{60} estimation inaccuracies. The logarithmic error depending on the $\hat{T}_{60}$ parameter is shown in Fig. 9. The proposed method, which is independent of the T_{60}, is shown as dashed lines. It can be observed that the LRSV estimator is only superior to the proposed method (i. e., has a smaller total error), where the estimated $\hat{T}_{60}$ is close to the true T_{60}.

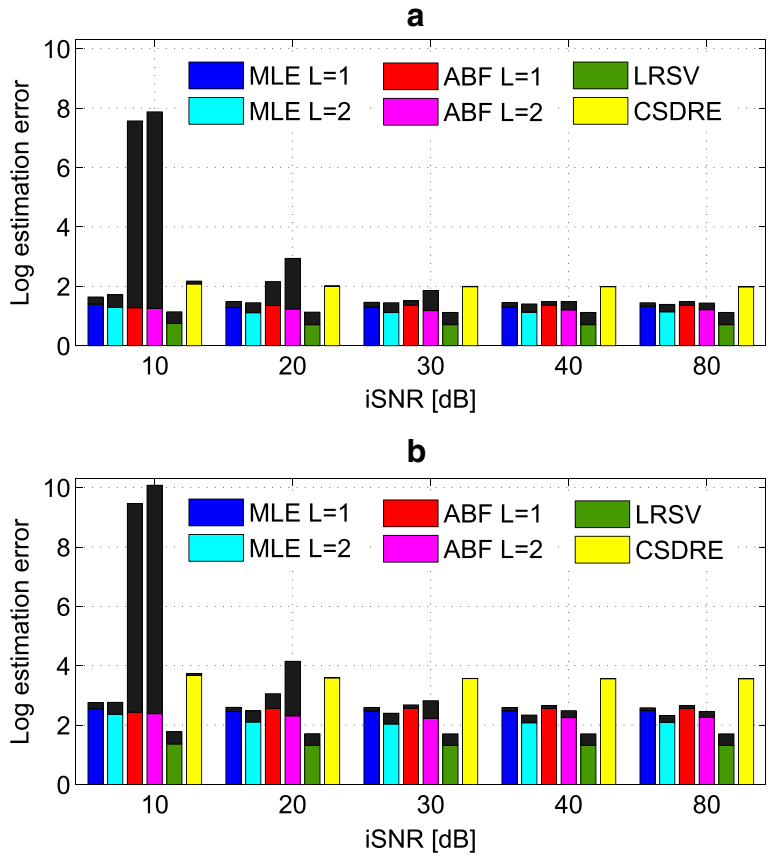

Fig. 8 Log error of diffuse PSD estimators. The coloured bars show the overestimation for different estimators. The underestimation is *black* on top of each bar. **a** Single active source. **b** Two active sources

4.6 Performance of the overall system

In this section, we evaluate the performance of the complete dereverberation system described by (10) for different acoustic scenarios.

In the first experiment, one, two, or three speakers were active simultaneously. The first speech signal was obtained by concatenating 6 speech signals of about 20 s (3 male, 3 female) from the EBU SQAM database [43], and the second and third signals were obtained by permutation of the speakers. The sources were positioned at $\theta = \{5°, -68°, 54°\}$ at distances of {2.7 m, 1.9 m, 2.3 m} from the broadside direction of a ULA with $M = 8$ and microphone spacing 1.75 cm. The room was again simulated by the image method with a $T_{60} = 500$ ms. Uncorrelated white Gaussian noise was added with iSNR $= 40$ dB. Either $\hat{L} = 1$ or $\hat{L} = 2$ DOAs were estimated per time-frequency instant using TLS-ESPRIT.

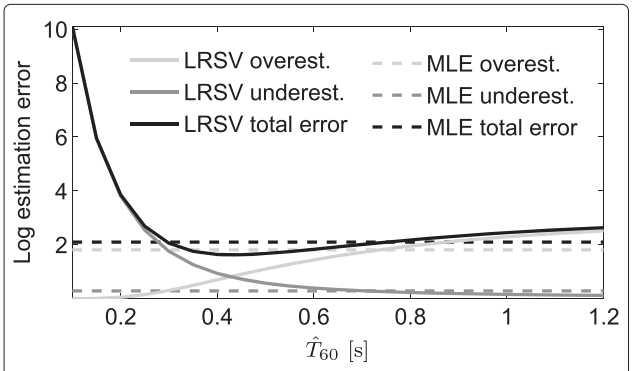

Fig. 9 Log error of the LRSV estimator (*solid lines*) depending on the T_{60} parameter compared to the proposed estimator with $\hat{L} = 2$ (*dashed lines*) at iSNR $= 30$ dB

The performance is evaluated using four objective measures, namely, the *perceptual evaluation of speech quality* (PESQ) [44], the *cepstral distance* (CD) [45], the *speech-to-reverberation modulation ratio* (SRMR) [46, 47], and the *segmental signal-to-interference ratio* enhancement (ΔsegSIR) given in decibels. The desired reference signal for the objective measures is the sum of the direct signal components (7) plus early reflections up to 40 ms after the direct sound; the interference is calculated as the sum of stationary noise and the late reverberation after 40 ms.

Figure 10 shows spectrograms of an excerpt of the signals for the described scenario. Figure 10a shows the spectrogram of the desired signal, which is the sum of

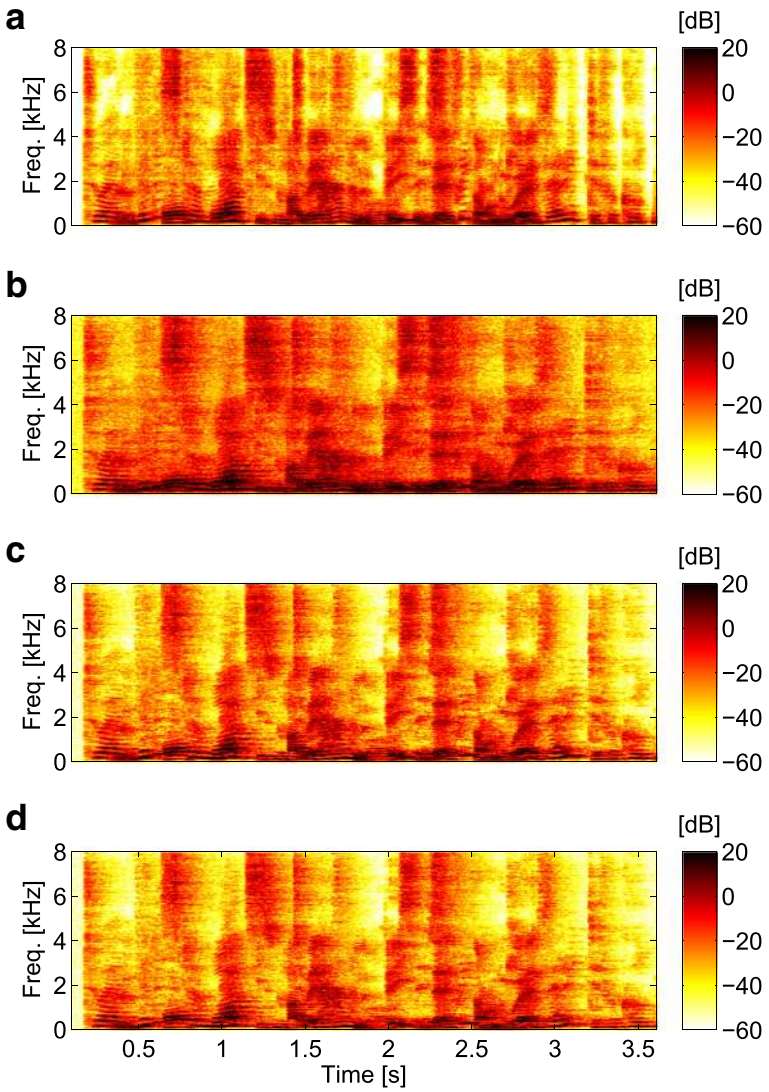

Fig. 10 Spectrograms of desired direct signal, reverberant and noisy microphone signal and processed signals obtained with the two proposed filters using $\hat{L} = 2$. **a** Direct signal. **b** Reverberant input signal. **c** MWF using LRSV. **d** MWF using MLE

the two direct signal components. Below is the reverberant and noisy input signal as captured by the reference microphone. Figures 10c,d show the spectrograms of the processed signals with the MWF using the LRSV and the proposed diffuse PSD estimator. It can be clearly observed that the stationary noise and the reverberation is reduced by the MWF, while the direct signals are preserved.

Tables 1, 2, and 3 show the results of the objective measures for one, two, and three simultaneous active speakers. The first column indicates the processing method and the method used to estimate the diffuse PSD ϕ_{d}. The second column shows the number of simultaneous DOAs that were estimated per time-frequency bin and were used to compute the diffuse PSD MLE and the spatial filter. We can observe that the measures improve over an unprocessed reference microphone signal for all methods. The approach using $\hat{L} = 1$ typically achieves the highest segmental SIR improvement but yields a higher CD for multiple sources. Although the higher overestimation of the diffuse PSD with $\hat{L} = 1$ increases the ΔsegSIR, it can be observed that the CD increases.

In terms of most performance measures, the LRSV slightly outperforms the MLE in Tables 1, 2, and 3. It should however be noted that the LRSV was computed using prior knowledge of the reverberation time and DRR.

In the second experiment, the system was evaluated in a realistic environment with measured impulse responses and recorded babble noise. We measured impulse responses of two common rooms, i.e., a meeting room (M) and a large presentation room (P). The

Table 1 Objective measures for simulated rooms, 1 active source

Method	$\hat{L}$	PESQ	CD	SRMR	ΔsegSIR [dB]
Unprocessed	-	2.08	4.54	2.26	–
MWF MLE	1	2.27	4.00	2.91	2.04
	2	2.22	4.06	2.96	1.88
MWF LRSV	1	2.38	3.85	2.90	2.10
	2	2.35	3.85	2.95	1.99

meeting room with a size of $6.7 \times 4.8 \times 2.8$ m and a $T_{60} \approx 700$ ms is not acoustically treated and has some strong early reflections caused by a large conference table and large windows. The presentation room with size of $10.4 \times 12.6 \times 3$ m and a $T_{60} \approx 650$ ms is acoustically treated but was almost empty besides some chairs. We used a similar array setup as in the simulations, i. e., a ULA with $M = 8$ and inter-microphone spacing 1.75 cm. We measured 3 positions in the meeting room and 6 positions in the presentation room, all located between $\pm 75°$ of the broadside array direction and at $1.5 \ldots 5$ m distance from the array. The test signals were created by convolving the impulse responses of two positions with different anechoic speech signals. Therefore, the scenario is constant double-talk from two different positions. Uncorrelated white Gaussian sensor noise was added with an iSNR of 50 dB, and diffuse cafeteria babble speech was added with an SDR of 15 dB. The stationary noise PSD matrix is estimated in advance by an arithmetic average over a period of 20 s during which the speakers were inactive. Due to the non-stationary nature of the babble speech, only the stationary part of the noise is captured in the time-invariant noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}$. The non-stationary diffuse components (babble speech and reverberation) are captured by the diffuse PSD estimate. For the evaluation, the direct desired signal component was generated by using windowed impulse responses $\mathbf{c}_{\mathrm{dir}}(t)$, where only the direct peak and early reflections are inside the window

$$\mathbf{c}_{\mathrm{dir}}(t) = w(t)\,\mathbf{c}(t), \tag{31}$$

where $\mathbf{c}(t)$ is a $M \times 1$ vector containing the measured impulse response, $w(t)$ is the window function and t is the discrete time index. The window function $w(t)$ is chosen as a crossfade between direct sound and late reverberation

Table 2 Objective measures for simulated rooms, 2 active sources

Method	$\hat{L}$	PESQ	CD	SRMR	ΔsegSIR [dB]
Unprocessed	-	2.06	3.72	1.88	–
MWF MLE	1	2.28	3.54	2.41	2.34
	2	2.25	3.39	2.46	2.20
MWF LRSV	1	2.37	3.46	2.36	2.20
	2	2.34	3.22	2.43	2.16

Table 3 Objective measures for simulated rooms, 3 active sources

Method	$\hat{L}$	PESQ	CD	SRMR	ΔsegSIR [dB]
Unprocessed	-	2.05	3.47	1.73	–
MWF MLE	1	2.17	3.40	2.22	2.18
	2	2.15	3.26	2.26	1.98
MWF LRSV	1	2.33	3.36	2.13	2.12
	2	2.31	3.07	2.19	2.05

that ensures that the direct sound peaks are weighted with 1 and fades to zero until 40 ms after the direct sound peak. The late reverberant impulse responses are obtained by

$$\mathbf{c}_{\mathrm{d}}(t) = \mathbf{c}(t) - \mathbf{c}_{\mathrm{dir}}(t). \tag{32}$$

Table 4 shows the objective measures for the measured test data set. We used TLS-ESPRIT to estimate $\hat{L} = 2$ DOAs, and the filter was computed using the proposed diffuse PSD estimator and the LRSV estimator. Due to the challenging scenario, the improvements are smaller than in the simulated scenarios. Nevertheless, an improvement of all measures is achieved compared to the unprocessed signals. The improvement for PESQ in Table 4 is sometimes very small. The reason is that PESQ is mainly a quality measure that does not quantify the amount of reverberation. However, informal listening tests confirmed that a significant dereverberation effect can be perceived, which is well represented by ΔsegSIR and SRMR.

5 Conclusions

We proposed a system for joint dereverberation and noise reduction for multiple simultaneously active desired direct sound plane waves. The system consists of an informed spatial filter that is computed using multiple DOAs per time-frequency bin and the PSD matrices of the diffuse sound and the noise. An estimator for the diffuse PSD was developed that uses a set of reference signals that are created by simultaneously blocking multiple active plane waves. The proposed estimator was

Table 4 Objective measures for measured rooms with 2 active sources and babble noise using $\hat{L} = 2$

Room	Method	PESQ	CD	SRMR	ΔsegSIR [dB]
M	Unprocessed	2.08	4.05	1.55	–
	MWF MLE	2.09	3.69	2.27	3.24
	MWF LRSV	2.27	3.51	2.27	2.98
P	Unprocessed	2.22	3.54	1.59	–
	MWF MLE	2.28	3.13	2.11	4.57
	MWF LRSV	2.41	3.13	2.11	3.51

compared to three existing estimators. The proposed estimator shows comparable or slightly more robust performance compared to all estimators under test except the well-established single-channel LRSV estimator. However, the LRSV estimator was computed with prior knowledge of the reverberation time and DRR, which might be difficult to estimate in noisy environments and in scenarios where the source positions and the room characteristics change over time. The objective measures of the dereverberation system show a comparable performance by using the proposed estimator or the LRSV estimator.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was partly funded by the German-Israeli Foundation for Scientific Research and Development (GIF).

Received: 17 August 2015 Accepted: 19 November 2015
Published online: 04 December 2015

References

1. PA Naylor, ND Gaubitch (eds.), *Speech Dereverberation* (Springer, London, UK, 2010)
2. M Miyoshi, Y Kaneda, Inverse filtering of room acoustics. IEEE Trans. Speech Audio Process. **36**(2), 145–152 (1988)
3. I Kodrasi, S Doclo, in *ICASSP*. Robust partial multichannel equalization techniques for speech dereverberation (Kyoto, Japan, 2012)
4. F Lim, PA Naylor, in *ICASSP*. Robust low-complexity multichannel equalization for dereverberation, (2013), pp. 689–693
5. Y Huang, J Benesty, J Chen, A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. IEEE Trans. Speech Audio Process. **13**(5), 882–895 (2005)
6. M Delcroix, T Hikichi, M Miyoshi, Dereverberation and denoising using multichannel linear prediction. Audio Speech Lang. Process. IEEE Trans. **15**(6), 1791–1801 (2007)
7. T Nakatani, T Yoshioka, K Kinoshita, M Miyoshi, J Biing-Hwang, Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Trans. Audio Speech Lang. Process. **18**(7), 1717–1731 (2010)
8. T Yoshioka, T Nakatani, Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Audio, Speech, Lang. Process. **20**(10), 2707–2720 (2012)
9. M Togami, Y Kawaguchi, R Takeda, Y Obuchi, N Nukaga, Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function. IEEE Trans. Audio, Speech, Lang. Process. **21**(7), 1369–1380 (2013)
10. K Kokkinakis, PC Loizou, The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners (L). J. Acoust. Soc. Am. **130**(3), 1099–1102 (2011)
11. K Lebart, JM Boucher, PN Denbigh, A new method based on spectral subtraction for speech de-reverberation. Acta Acoustica. **87**, 359–366 (2001)
12. EAP Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*. (PhD thesis, Technische Universiteit Eindhoven, 2007). http://alexandria.tue.nl/extra2/200710970.pdf
13. X Bao, J Zhu, An improved method for late-reverberant suppression based on statistical models. Speech Commun. **55**(9), 932–940 (2013)
14. S Mosayyebpour, M Esmaeili, TA Gulliver, Single-microphone early and late reverberation suppression in noisy speech. IEEE Trans. Audio Speech Lang. Process. **21**(2), 322–335 (2013)
15. JD Polack, *La transmission de l'énergie sonore dans les salles*. (PhD thesis, Université du Maine, Le Mans, France, 1988)
16. EAP Habets, S Gannot, I Cohen, Late reverberant spectral variance estimation based on a statistical model. IEEE Signal Process. Lett. **16**(9), 770–774 (2009)
17. O Thiergart, M Taseska, EAP Habets, *An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates*, (Marrakesh, Morocco, 2013)
18. O Thiergart, O Del Galdo, EAP Habets, in *ICASSP*. Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones, (2012)
19. O Thiergart, EAP Habets, in *ICASSP*. An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates, (2013)
20. S Braun, EAP Habets, in *EUSIPCO*. Dereverberation in noisy environments using reference signals and a maximum likelihood estimator (IEEE, 2013)
21. A Kuklasinski, S Doclo, SH Jensen, J Jensen, in *EUSIPCO*. Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids (Lisbon, Portugal, 2014), pp. 61–65
22. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001)
23. T Gerkmann, RC Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio, Speech, Lang. Process. **20**(4), 1383–1393 (2012)
24. M Souden, J Chen, J Benesty, S Affes, An integrated solution for online multichannel noise tracking and reduction. IEEE Trans. Audio, Speech, Lang. Process. **19**(7), 2159–2169 (2011)
25. M Taseska, EAP Habets, in *IWAENC*. MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator, (2012)
26. F Jacobsen, T Roisin, The coherence of reverberant sound fields. J. Acoust. Soc. Am. **108**, 204–210 (2000)
27. S Gergen, C Borss, N Madhu, R Martin, in *Proc. IEEE Intl. Conf. on Signal Processing, Communication and Computing (ICSPCC)*. An optimized parametric model for the simulation of reverberant microphone signals (IEEE, Hong Kong, 2012), pp. 154–157
28. MS Brandstein, DB Ward (eds.), *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, Berlin, Germany, 2001)
29. Z Chen, GK Gokeda, Y Yu, *Introduction to Direction-of-Arrival Estimation*. (Artech House, London, UK, 2010)
30. TE Tuncer, B Friedlander (eds.), *Classical and Modern Direction-of-Arrival Estimation* (Academic Press, Burlington, USA, 2009)
31. EAP Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*. (Ph.D. Thesis, Technische Universiteit Eindhoven, 2007)
32. O Thiergart, G Del Galdo, EAP Habets, On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation. J. Acoust. Soc. Am. **132**(4), 2337–2346 (2012)
33. M Jeub, CM Nelke, C Beaugeant, P Vary, in *EUSIPCO*. Blind Estimation of the Coherent-to-Diffuse Energy Ratio From Noisy Speech Signals (Barcelona, Spain, 2011)
34. ND Gaubitch, HW Löllmann, M Jeub, TH Falk, PA Naylor, P Vary, M Brookes, in *IWAENC*. Performance Comparison of Algorithms for Blind Reverberation Time Estimation from Speech (Aachen, Germany, 2012)
35. O Thiergart, EAP Habets, in *IWAENC*. Sound field model violations in parametric spatial sound processing, (2012)
36. S Markovich, S Gannot, I Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Trans. Audio, Speech, Lang. Process. **17**(6), 1071–1086 (2009)
37. S Markovich-Golan, S Gannot, I Cohen, in *IEEEI*. A weighted multichannel Wiener filter for multiple sources scenario, (2012)
38. Markovich-Golan, S, S Gannot, I Cohen, *A sparse blocking matrix for multiple constraints GSC beamformer*. (IEEE, Kyoto, Japan, 2012), pp. 197–200
39. HQ Dam, S Nordholm, HH Dam, SY Low, in *Asia-Pacific Conference on Communications*. Maximum likelihood estimation and cramer-rao lower bounds for the multichannel spectral evaluation in hands-free communication (IEEE, Perth, Australia, 2005)
40. EAP Habets, I Cohen, S Gannot, Generating nonstationary multisensor signals under a spatial coherence constraint. J. Acoust. Soc. Am. **124**(5), 2911–2917 (2008)
41. R Roy, T Kailath, ESPRIT - estimation of signal parameters via rotational invariance techniques. IEEE Trans. Acoust., Speech, Signal Process. **37**, 984–995 (1989)

42. JB Allen, DA Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
43. EB Union, Sound Quality Assessment Material Recordings for Subjective Tests. http://tech.ebu.ch/publications/sqamcd
44. ITU-T, Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs. International Telecommunications Union (ITU-T), 2001
45. N Kitawaki, H Nagabuchi, K Itoh, Objective quality evaluation for low bit-rate speech coding systems. IEEE J. Sel. Areas Commun. **6**(2), 262–273 (1988)
46. T Falk, C Zheng, W-Y Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio, Speech, Lang. Process. **18**(7), 1766–1774 (2010)
47. JF Santos, M Senoussaoui, TH Falk, in *IWAENC*. An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation (Antibes, France, 2014)