Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators

Sebastian Braun, *Student Member, IEEE*, Adam Kuklasiński, Ofer Schwartz, *Student Member, IEEE*, Oliver Thiergart, Emanuël A. P. Habets, *Senior Member, IEEE*, Sharon Gannot, *Senior Member, IEEE*, Simon Doclo, *Senior Member, IEEE*, and Jesper Jensen

***Abstract*—Reduction of late reverberation can be achieved using spatio-spectral filters, such as the multichannel Wiener filter. To compute this filter, an estimate of the late reverberation power spectral density (PSD) is required. In recent years, a multitude of late reverberation PSD estimators have been proposed. In this paper, these estimators are categorized into several classes, their relations and differences are discussed, and a comprehensive experimental comparison is provided. To compare their performance, simulations in controlled as well as practical scenarios are conducted. It is shown that a common weakness of spatial coherence-based estimators is their performance in high direct-to-diffuse ratio conditions. To mitigate this problem, a correction method is proposed and evaluated. It is shown that the proposed correction method can decrease the speech distortion without significantly affecting the reverberation reduction.**

***Index Terms*—Dereverberation, speech enhancement, array processing, power spectral density estimation, diffuse sound.**

I. Introduction

STRONG room reverberation and interfering noise can impair the intelligibility of speech in communication scenarios such as mobile phones, conferencing systems, smart TVs, hearing aids, but also decrease the performance of automatic speech recognition systems [1], [2].

Manuscript received June 26, 2017; revised November 15, 2017 and January 22, 2018; accepted January 24, 2018. Date of publication February 8, 2018; date of current version April 11, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tuomas Virtanen. *(Corresponding author: Sebastian Braun.)*

S. Braun and E. A. P. Habets are with the International Audio Laboratories Erlangen (a joint institution between Friedrich-Alexander Universität Erlangen-Nürnberg and Fraunhofer Institute for Integrated Circuits), Erlangen 91058, Germany (e-mail: sebastian.braun@audiolabs-erlangen.de; emanuel.habets@audiolabs-erlangen.de).

A. Kuklasiński and J. Jensen are with the Oticon A/S, Smørum 2765, Denmark, and also with the Department of Electronic Systems and the Signal and Information Processing Section, Aalborg Univeristy, Aalborg 9220, Denmark (e-mail: adku@oticon.com; jesj@oticon.com).

O. Schwartz and S. Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel (e-mail: ofer.shwartz@live.biu.ac.il; sharon.gannot@biu.ac.il).

O. Thiergart is with the Fraunhofer Institute for Integrated Circuits, Erlangen 91058, Germany (e-mail: oliver.thiergart@iis.fraunhofer.de).

S. Doclo is with the Cluster of Excellence Hearing4all, Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg 26111, Germany (e-mail: simon.doclo@uni-oldenburg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2804172

Many methods for dereverberation exist, including blind channel identification [3]–[5] and inverse filtering [6]–[8], multichannel linear prediction [9]–[12], modification of the linear prediction residual [13], [14], spectral suppression [15]–[17], or spatio-spectral filtering [18]. The multichannel Wiener filter (MWF) and related beamformer-postfilter systems, which reside in the class of spatio-spectral filtering techniques, have been widely used for joint reverberation and noise reduction [19]–[24]. Due to its low complexity, high robustness in practice and direct integration into other speech enhancement systems, the MWF is very popular in practical systems. The MWF is typically derived in the short-time Fourier transform (STFT) domain assuming a narrowband signal model. The Wiener filter requires estimates of the second-order statistics of the desired and undesired signal components, where the accuracy of these estimates determines the performance of the MWF. When focusing on dereverberation, the late reverberation is often modeled in the STFT domain as an additive diffuse sound field with a time-varying power spectral density (PSD) and a time-invariant spatial coherence. In the following we refer to this model as the spatial coherence model. As the diffuse spatial coherence can be calculated analytically for known microphone array geometries, the remaining challenge is to obtain an accurate estimate of the late reverberation PSD, which directly affects the performance of the MWF and hence the quality of the dereverberated signal. As the late reverberation PSD is highly time-varying, it is challenging to obtain an accurate estimate.

To the best of our knowledge, the first multichannel methods to estimate the late reverberation PSD were proposed in [25], [26], whereas an explicit spatial coherence model was first used to estimate the coherent-to-diffuse ratio (CDR) in [27]–[29]. Although not in the context of dereverberation, methods to estimate the direct sound PSD in a diffuse noise field [30], or the diffuse sound PSD [31]–[34] have been proposed. In the past years, a multitude of estimators for the late reverberation or diffuse sound PSD have been developed assuming that the sound field can be described by the direct sound propagating as a plane wave in a time-varying diffuse field and additive stationary noise. Nevertheless, also temporal reverberation models can be exploited. Existing late reverberation PSD estimators can be divided into four classes, where the first three classes use the spatial coherence reverberation model. The first two classes are direct PSD estimators, whereas the third class comprises indirect PSD estimators, which require an additional step to obtain

the reverberation PSD. In contrast to the first three classes, the fourth class is based on temporal reverberation models.

Estimators in the first class model the reverberation as a diffuse sound field with known spatial coherence and block the direct sound utilizing direction of arrival (DOA) information. This simplifies the estimation procedure, since the resulting (in some methods multiple) signals after the blocking operation contain only filtered diffuse sound and noise. In [21], the error PSD matrix of the blocking output signals is minimized, whereas in [24] and [35], maximum likelihood (ML) estimators are derived given the blocked, or blocked and additionally filtered, signals. In [35] the solution is obtained using the Newton method, whereas in [24] the solution is obtained by a root finding procedure. Thiergart et. al. [36–37] developed several spatial filters to extract the diffuse sound while blocking the direct sound. In [36] a spatial filter is derived that maximizes the diffuse-to-noise ratio (DNR) at its output, while in [37] a linearly constrained minimum variance (LCMV) beamformer is proposed with a novel constraint set to block the direct sound and to extract the diffuse sound.

The spatial coherence-based estimators in the second class use no blocking of the direct sound, therefore the unknown PSDs of direct sound and reverberation have to be estimated jointly. In [38] a closed-form ML estimator for the direct sound and reverberation PSDs is presented without taking additive noise into consideration. The method presented in [39] obtains the ML estimator of the direct and reverberation PSDs using the Newton method. In [40], a batch expectation-maximization (EM) algorithm to estimate the direct and reverberation PSDs in the ML sense is presented, where unlike in all other methods considered in this paper, this method also estimates the spatial coherence matrix of the reverberation. In [41], the direct and reverberation PSDs are estimated jointly in the least-squares sense by minimizing the Frobenius norm of an error matrix.

Estimators in the third class are considered as indirect PSD estimators based on the spatial coherence model, assuming that the reverberation is diffuse: Rather than estimating the diffuse PSD directly, an estimate can be obtained by first estimating the CDR and then to estimate the diffuse PSD. To limit the number of algorithms under test, we constrain ourselves to the best performing CDR estimator reported in [42].

Estimators in the fourth class utilize temporal models to describe the reverberation, and make no assumption on the spatial coherence. In this class, the reverberation is described either using Polack's model [15], [16], or using a narrowband moving average model [43]. However, when the estimated late reverberation PSD is used in the MWF for dereverberation, an assumption on the spatial coherence of the late reverberation is required.

The reverberation PSD estimators in these four classes can be used equivalently in the MWF or similar beamformer-postfilter systems [22] for dereverberation. However, the properties and the performance of this large variety of PSD estimators are unclear and have never been compared in a unified framework. In this paper, we provide an overview and comparison of the current state-of-the-art reverberation PSD estimators. The obtained results provide a guideline for choosing an estimator for a specific use-case, and reveal strengths and weaknesses of existing estimators that can drive further research and developments.

In Section II, we present the signal model assuming a single source per time-frequency bin and derive the MWF to estimate the desired signal, requiring an estimate of the reverberation PSD. Section III reviews coherence-based direct estimators with and without blocking, Section IV the coherence-based indirect PSD estimators, and Section V reviews temporal model-based PSD estimators. The relations and differences between the estimators are discussed in Section VI. A common weakness of the spatial coherence-based estimators is a systematic bias at high direct-to-diffuse ratios (DDRs). Therefore, we propose a bias compensation method depending on the DDR in Section VII. A comprehensive experimental evaluation using controlled and realistic simulations is presented in Section VIII, where we analyze the error of the estimated PSDs as well as the resulting performance of the spatial filter using these estimates. Finally, the paper is concluded in Section IX.

II. Problem Formulation

A. Signal Model

We assume that the sound field is captured by an array of M omni-directional microphones with an arbitrary geometry. The microphone signals given in the STFT domain $Y_m(k,n)$, $m \in \{1,\dots,M\}$ are stacked into the vector $\mathbf{y}(k,n) = [Y_1(k,n),\dots,Y_M(k,n)]^{\mathrm{T}}$, where k and n denote the frequency and time frame indices. We describe the sound field using a parametric signal model, where the microphone signal vector is given by

$$\mathbf{y}(k,n) = \mathbf{a}(k)X(k,n) + \mathbf{d}(k,n) + \mathbf{v}(k,n), \tag{1}$$

where $X(k,n)$ denotes the desired signal component as received by a reference microphone, $\mathbf{a}(k) = [A_1(k),\dots,A_M(k)]^{\mathrm{T}}$ is a vector containing the acoustic relative transfer functions (RTFs) $A_m(k)$ of the desired signal from the reference microphone to all M microphones, $\mathbf{d}(k,n)$ is the reverberation, and $\mathbf{v}(k,n)$ is the additive noise. Throughout this paper, we assume that the RTFs $\mathbf{a}(k)$ are time-invariant, but in general they can also be time-varying. Note that the desired signal component $X(k,n)$ is often modeled only as the direct sound, ignoring the early reflections arriving within the same STFT frame as the direct sound. The component $\mathbf{d}(k,n)$ models the late reverberation, which is assumed to be uncorrelated with the desired speech component $X(k,n)$. The component $\mathbf{v}(k,n)$ models stationary or slowly time-varying additive noise components such as sensor noise and ambient noise.

For typical STFT window lengths of 20 to 30 ms, the three additive components in (1) can be assumed to be mutually uncorrelated, and the PSD matrix of the microphone signals $\mathbf{y}(k,n)$ is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) &= E\left\{\mathbf{y}(k,n)\,\mathbf{y}^{\mathrm{H}}(k,n)\right\} \\ &= \phi_x(k,n)\mathbf{a}(k)\mathbf{a}^{\mathrm{H}}(k) + \mathbf{\Phi}_{\mathbf{d}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n),\end{aligned} \tag{2}$$

where $E\{\cdot\}$ is the expectation operator, $\phi_x(k,n) = E\{|X(k,n)|^2\}$ is the PSD of the desired signal at the reference microphone, $\mathbf{\Phi}_{\mathbf{d}}(k,n) = E\{\mathbf{d}(k,n)\,\mathbf{d}^{\mathrm{H}}(k,n)\}$ denotes the late reverberation PSD matrix, and $\mathbf{\Phi}_{\mathbf{v}}(k,n) = E\{\mathbf{v}(k,n)\,\mathbf{v}^{\mathrm{H}}(k,n)\}$ denotes the noise PSD matrix.

We assume that the late reverberation PSD matrix can be modeled as a spatially homogenous and isotropic sound field with a time-varying power. Therefore, the late reverberation PSD matrix $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ can be described by a time-invariant coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$, which is scaled by the time-varying late reverberation PSD $\phi_d(k,n)$ [36], i.e.,

$$\mathbf{\Phi}_{\mathbf{d}}(k,n) = \phi_d(k,n)\,\mathbf{\Gamma}_{\mathrm{d}}(k). \tag{3}$$

The time-invariant coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$ can be determined in advance from the microphone array configuration. In the case of a free-field microphone array in a spherical or cylindrical diffuse field, there exist analytic expressions for the spatial coherence, i.e. the *sinc* or *Bessel* functions, respectively [44], whereas e.g., for directional microphones [45] or in a hearing aid setup [46], the spatial coherence function is more complex to describe. A widely used model for the spatial coherence of late reverberation uses the spherical diffuse field assumption, where the $\{i,j\}$-th element of the spatial coherence matrix for omnidirectional microphones is given by [44]

$$\mathbf{\Gamma}_{\mathrm{d}}^{(i,j)}(k) = \mathrm{sinc}\left(2\pi \frac{k\,f_{\mathrm{s}}}{N_{\mathrm{FFT}}\,c}\|\mathbf{r}_i - \mathbf{r}_j\|_2\right), \tag{4}$$

where $\mathrm{sinc}(\cdot) = \frac{\sin(\cdot)}{(\cdot)}$, the vector $\mathbf{r}_m$ denotes the position of the m-th microphone, f_{s} denotes the sampling frequency, N_{FFT} is the FFT length and c is the speed of sound.

Although in most considered late reverberation PSD estimation methods, the coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$ is assumed to be given by (4), the method in [40] also allows to estimate this matrix from the observed signals, which could be advantageous when the reverberant sound field differs from a theoretical diffuse field, e.g., in rooms with strongly non-homogenous or partially non-reflecting boundaries.

B. Desired Signal Estimation

To estimate the desired signal $X(k,n)$, we apply a complex valued spatial filter $\mathbf{w}(k,n)$ to the microphone signals, i.e.,

$$\widehat{X}(k,n) = \mathbf{w}^{\mathrm{H}}(k,n)\,\mathbf{y}(k,n). \tag{5}$$

By minimizing the mean-squared error (MSE) cost-function

$$J_{\mathrm{MWF}}(\mathbf{w}) = \mathrm{E}\left\{|\mathbf{w}^{\mathrm{H}}(k,n)\mathbf{y}(k,n) - X(k,n)|^2\right\} \tag{6}$$

we obtain the well-known MWF, which is given by

$$\mathbf{w}_{\mathrm{MWF}} = \big[\phi_x\mathbf{a}\mathbf{a}^{\mathrm{H}} + \underbrace{\phi_d\mathbf{\Gamma}_{\mathrm{d}} + \mathbf{\Phi}_{\mathbf{v}}}_{\mathbf{\Phi}_{\mathbf{d}+\mathbf{v}}}\big]^{-1}\mathbf{a}\phi_x, \tag{7}$$

where $\mathbf{\Phi}_{\mathbf{d}+\mathbf{v}}(k,n)$ denotes the interference PSD matrix. The frequency and time frame indices k and n are omitted here and in the following equations for brevity, wherever possible.

The MWF can be split into a minimum variance distortionless response (MVDR) beamformer, denoted by $\mathbf{w}_{\mathrm{MVDR}}(k,n)$, and a single-channel Wiener post-filter, denoted by $W_{\mathrm{WF}}(k,n)$, i.e. [47]

$$\mathbf{w}_{\mathrm{MWF}} = \underbrace{\frac{\mathbf{\Phi}_{\mathbf{d}+\mathbf{v}}^{-1}\mathbf{a}}{\mathbf{a}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{d}+\mathbf{v}}^{-1}\mathbf{a}}}_{\mathbf{w}_{\mathrm{MVDR}}}\;\underbrace{\frac{\xi}{\xi+1}}_{W_{\mathrm{WF}}}, \tag{8}$$

where $\xi = \frac{\phi_x}{[\mathbf{a}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{d}+\mathbf{v}}^{-1}\mathbf{a}]^{-1}}$ is the a priori signal-to-interference ratio of the MVDR output signal, which can be estimated using the decision-directed approach [48].

The aim in this paper is to investigate different estimation methods for the PSD $\phi_d(k,n)$, which determines the late reverberation PSD matrix using the model (3) together with $\mathbf{\Gamma}_{\mathrm{d}}(k)$. We assume the RTF vector $\mathbf{a}(k)$ and the noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ to be known. In practice, both have to be estimated as well, which is beyond the scope of this paper. Popular noise PSD estimation methods are, for example, [49]–[52], and for DOA estimation the reader is referred to [53], [54].

III. Coherence-Based Direct PSD Estimators

The *coherence-based direct reverberation PSD estimators* comprise blocking based methods (Section III-A) and non-blocking based methods (Section III-B). These estimators have in common that they are exclusively based on the spatial coherence model (3) with the signal model (1). Therefore, these estimators depend on the diffuse coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$ and on the RTF vector $\mathbf{a}(k)$.

A. Blocking-Based Methods

The blocking-based methods use a set of $J = M-1$ signals, which are generated by canceling the desired sound from the microphone signals using a blocking matrix. The J-dimensional signal vector $\mathbf{u}$ is obtained as

$$\mathbf{u} = \mathbf{B}^{\mathrm{H}}\,\mathbf{y}, \tag{9}$$

where the blocking matrix $\mathbf{B}$ of dimension $M \times J$ has to fulfill the constraint

$$\mathbf{B}^{\mathrm{H}}\,\mathbf{a} = \mathbf{0}_{J\times 1}. \tag{10}$$

Possible choices for the blocking matrix are discussed in [21], [34], [55]. In this work, we use the *eigenspace blocking matrix* given by [55]

$$\mathbf{B} = \left[\mathbf{I}_{M\times M} - \mathbf{a}(\mathbf{a}^{\mathrm{H}}\mathbf{a})^{-1}\mathbf{a}^{\mathrm{H}}\right]\mathbf{I}_{M\times J}, \tag{11}$$

where $\mathbf{I}_{M\times J}$ is a truncated identity matrix. As a consequence of using (1) and (3) with (9) and (10), it follows that the PSD matrix of the blocking output signals $\mathbf{u}(k,n)$ is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{u}} &= \mathbf{B}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{y}}\mathbf{B}\\ &= \underbrace{\phi_x\mathbf{B}^{\mathrm{H}}\mathbf{a}\mathbf{a}^{\mathrm{H}}\mathbf{B}}_{\mathbf{0}_{J\times J}} + \phi_d\,\underbrace{\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_{\mathrm{d}}\,\mathbf{B}}_{\widetilde{\mathbf{\Gamma}}_{\mathrm{d}}} + \underbrace{\mathbf{B}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{B}}_{\widetilde{\mathbf{\Phi}}_{\mathbf{v}}}.\end{aligned} \tag{12}$$

Note that $\widetilde{\mathbf{\Gamma}}_{\mathrm{d}}(k)$ and $\widetilde{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$ denote the corresponding second-order statistics after applying the blocking matrix.

In Sections III-A1 and III-A2 ML methods are used to estimate the late reverberation PSD, where in Section III-A1 the

elements of an error matrix are assumed as random variables, and in Section III-A2, the elements of the vector $\mathbf{u}(k,n)$ are assumed to be random variables.

1) PSD Matrix-Based Least-Squares Method With Blocking: In [21], the error matrix between the estimated PSD matrix $\widehat{\mathbf{\Phi}}_{\mathbf{u}}(k,n)$ and its model is defined as

$$\mathbf{\Phi}_{\mathbf{e}} = \widehat{\mathbf{\Phi}}_{\mathbf{u}} - \left[\widetilde{\mathbf{\Phi}}_{\mathbf{v}} + \phi_d\, \widetilde{\mathbf{\Gamma}}_{\mathrm{d}}\right], \tag{13}$$

The off-diagonal elements of the error matrix $\mathbf{\Phi}_{\mathbf{e}}(k,n)$ are assumed to be drawn from independent zero-mean complex Gaussian distributions with equal variance [21]. The solution to this ML problem is in both cases obtained by solving the least-squares problem of minimizing the squared Frobenius norm of the error matrix

$$\widehat{\phi}_d = \underset{\phi_d}{\operatorname{argmin}} ||\mathbf{\Phi}_{\mathbf{e}}||_{\mathrm{F}}^2, \tag{14}$$

where $||\cdot||_{\mathrm{F}}^2$ denotes the Frobenius norm, and is given by

$$\widehat{\phi}_d = \frac{\operatorname{tr}\left\{\widetilde{\mathbf{\Gamma}}_{\mathrm{d}}^{\mathrm{H}}\left(\widehat{\mathbf{\Phi}}_{\mathbf{u}} - \widetilde{\mathbf{\Phi}}_{\mathbf{v}}\right)\right\}}{\operatorname{tr}\left\{\widetilde{\mathbf{\Gamma}}_{\mathrm{d}}^{\mathrm{H}}\widetilde{\mathbf{\Gamma}}_{\mathrm{d}}\right\}}, \tag{15}$$

where $\operatorname{tr}\{\cdot\}$ denotes the trace operator. An estimate of $\mathbf{\Phi}_{\mathbf{u}}(k,n)$ can be obtained by using recursive averaging, i.e., $\widehat{\mathbf{\Phi}}_{\mathbf{u}}(k,n) = \beta\widehat{\mathbf{\Phi}}_{\mathbf{u}}(k,n-1) + (1-\beta)\mathbf{u}(k,n)\mathbf{u}^{\mathrm{H}}(k,n)$, where β denotes the forgetting factor.

2) ML Using Blocking Output Signals: The methods presented in [35] and [24] both start from the assumption that the elements of the microphone signal vector $\mathbf{y}(k,n)$ are zero-mean complex Gaussian random variables

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Phi}_{\mathbf{y}}\right). \tag{16}$$

In [35], the ML problem is solved by an iterative Newton method, whereas in [24] a filtered version of the signals $\mathbf{u}(k,n)$ are used and the ML problem is solved by a root-finding method.

a) Solution using root-finding method [24]: In the ML estimator using a root finding method, the blocking output signals $\mathbf{u}(k,n)$ are filtered to diagonalize $\widetilde{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$ in (12). Specifically, a whitening matrix $\mathbf{D}(k,n)$ of dimension $J \times J$ defined as the Cholesky factor of the inverse of $\widetilde{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$, i, e. $\widetilde{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(k,n) = \mathbf{D}(k,n)\mathbf{D}^{\mathrm{H}}(k,n)$, yielding

$$\mathbf{z} = \mathbf{D}^{\mathrm{H}}\mathbf{B}^{\mathrm{H}}\mathbf{y} \tag{17}$$

and its PSD matrix is given by

$$\mathbf{\Phi}_{\mathbf{z}} = \mathbf{D}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{u}}\mathbf{D} = \phi_d\widetilde{\widetilde{\mathbf{\Gamma}}}_{\mathrm{d}} + \mathbf{I}, \tag{18}$$

with $\widetilde{\widetilde{\mathbf{\Gamma}}}_{\mathrm{d}} = \mathbf{D}^{\mathrm{H}}\widetilde{\mathbf{\Gamma}}_{\mathrm{d}}\mathbf{D} = \mathbf{D}^{\mathrm{H}}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_{\mathrm{d}}\mathbf{B}\mathbf{D}$.

As a result of the described whitening of $\widetilde{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$, the matrices $\mathbf{\Phi}_{\mathbf{z}}(k,n)$ and $\widetilde{\widetilde{\mathbf{\Gamma}}}_{\mathrm{d}}(k,n)$ can be diagonalized using the same unitary matrix $\mathbf{C}(k,n)$, i.e.,

$$\mathbf{\Phi}_{\mathbf{z}} = \mathbf{C}\mathbf{\Lambda}_{\mathbf{z}}\mathbf{C}^{\mathrm{H}},\ \widetilde{\widetilde{\mathbf{\Gamma}}}_{\mathrm{d}} = \mathbf{C}\mathbf{\Lambda}_{\mathbf{\Gamma}}\mathbf{C}^{\mathrm{H}}, \tag{19}$$

where the orthonormal columns of $\mathbf{C}(k,n)$ are the eigenvectors, and where $\mathbf{\Lambda}_{\mathbf{z}}(k,n)$ and $\mathbf{\Lambda}_{\mathbf{\Gamma}}(k,n)$ are diagonal matrices containing the eigenvalues of $\mathbf{\Phi}_{\mathbf{z}}(k,n)$ and $\widetilde{\widetilde{\mathbf{\Gamma}}}_{\mathrm{d}}(k,n)$, respectively. Due to (18), these eigenvalues are related as $\lambda_{\mathbf{z},j} = \phi_d\lambda_{\mathbf{\Gamma},j} + 1$, where $\lambda_{\mathbf{z},j}(k,n)$ and $\lambda_{\mathbf{\Gamma},j}(k,n)$ denote the j-th eigenvalue of $\mathbf{\Phi}_{\mathbf{z}}(k,n)$ and $\widetilde{\widetilde{\mathbf{\Gamma}}}_{\mathrm{d}}(k,n)$, respectively.

Given the the filtered blocking output signals $\mathbf{z}(k,n)$ in (18), with $\mathbf{z}(k,n) \sim f_{\mathbf{z}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{z}})$, the ML estimate of ϕ_d is given by

$$\widehat{\phi}_d = \underset{\phi_d}{\arg\max}\ \log f_{\mathbf{z}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{z}}), \tag{20}$$

where $f(\boldsymbol{\mu}, \mathbf{\Phi})$ denotes the complex Gaussian likelihood function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Phi}$. By setting the derivative of the log-likelihood function to zero and exploiting the diagonal structure of the involved matrices (for more details, see [24]), we obtain the polynomial

$$p(\phi_d) = \sum_{j=1}^{J} p_j(\phi_d), \quad \text{where}$$

$$p_j(\phi_d) = \left(\phi_d - \frac{g_j - 1}{\lambda_{\mathbf{\Gamma},j}}\right)\prod_{\ell=1}^{J,\,\ell\neq j}\left(\phi_d + \frac{1}{\lambda_{\mathbf{\Gamma},\ell}}\right)^2, \tag{21}$$

where $g_j(k,n)$ denotes the j-th diagonal element of $\mathbf{C}^{\mathrm{H}}\mathbf{D}^{\mathrm{H}}\widehat{\mathbf{\Phi}}_{\mathbf{u}}\mathbf{D}\mathbf{U}$. It has been shown in [24] that the root of the polynomial $p(\phi_d)$ yielding the highest value of the likelihood (20) is the ML estimate $\widehat{\phi}_d(k,n)$.

b) Solution using Newton's method: To solve the ML estimation problem [35]

$$\widehat{\phi}_d = \underset{\phi_d}{\arg\max}\ \log f_{\mathbf{u}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{u}}), \tag{22}$$

Newton's method is used to derive an iterative search [56]

$$\phi_d^{(\ell+1)} = \phi_d^{(\ell)} - \frac{\mathcal{D}\left(\phi_d^{(\ell)}\right)}{\mathcal{H}\left(\phi_d^{(\ell)}\right)}, \tag{23}$$

where ℓ denotes the iteration index, and $\mathcal{D}(\phi_d)$ and $\mathcal{H}(\phi_d)$ are the gradient and the Hessian of the log-likelihood

$$\mathcal{D}(\phi_d) \equiv \frac{\partial \log f_{\mathbf{u}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{u}})}{\partial \phi_d}, \tag{24}$$

$$\mathcal{H}(\phi_d) \equiv \frac{\partial^2 \log f_{\mathbf{u}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{u}})}{\partial \phi_d^2}. \tag{25}$$

As shown in [35], the gradient is equal to

$$\mathcal{D}(\phi_d) = J\ \operatorname{tr}\left\{\left(\mathbf{\Phi}_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^{\mathrm{H}} - \mathbf{I}\right)\mathbf{\Phi}_{\mathbf{u}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{u}}}{\partial \phi_d}\right\}, \tag{26}$$

with $\frac{\partial \mathbf{\Phi}_{\mathbf{u}}}{\partial \phi_d} = \mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_{\mathrm{d}}\mathbf{B}$, whereas the Hessian matrix is equal to

$$\mathcal{H}(\phi_d) = -J\ \operatorname{tr}\left\{\mathbf{\Phi}_{\mathbf{u}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{u}}}{\partial \phi_d}\mathbf{\Phi}_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{u}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{u}}}{\partial \phi_d}\right.$$
$$\left. + \left(\mathbf{\Phi}_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^{\mathrm{H}} - \mathbf{I}\right)\mathbf{\Phi}_{\mathbf{u}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{u}}}{\partial \phi_d}\mathbf{\Phi}_{\mathbf{u}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{u}}}{\partial \phi_d}\right\}. \tag{27}$$

As shown in [35], the Newton update (23) can be computed efficiently by re-arranging (26) and (27), using an eigenvalue

decomposition of $\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_{\mathrm{d}}\mathbf{B}$ and exploiting the resulting diagonal matrices. In practice, the matrix $\mathbf{u}(k,n)\mathbf{u}^{\mathrm{H}}(k,n)$ is substituted by the smoothed version $\widehat{\mathbf{\Phi}}_{\mathbf{u}}(k,n)$. The Newton iterations are initialized with $\phi_d^{(0)}(k,n) = \epsilon\frac{1}{J}\,\mathrm{tr}\{\widehat{\mathbf{\Phi}}_{\mathbf{u}}(k,n)\}$, where ϵ is a small positive value. The Newton algorithm is stopped if the estimate at iteration $\ell = \ell_{\text{stop}}$ reaches a predefined lower or upper bound, or if a convergence threshold is reached, and the estimate is obtained by $\widehat{\phi}_d(k,n) = \phi_d^{(\ell_{\text{stop}})}(k,n)$.

3) Diffuse Beamformers [36], [37]: Thiergart *et al.* developed several beamformers that aim at extracting the late reverberation, modeled as diffuse sound, while blocking the desired sound. As our preliminary experiments unveiled almost identical performance across those beamformers in terms of late reverberation PSD estimation, we present the most elegant here: The beamformer proposed in [37] minimizes the noise under the linear constraints of blocking the desired sound and not distorting the average transfer function of the diffuse sound, i.e.,

$$\mathbf{w}_{\mathrm{d}} = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{w} \tag{28a}$$

subject to

$$\mathbf{w}^{\mathrm{H}}\mathbf{a} = 0 \tag{28b}$$

$$\mathbf{w}^{\mathrm{H}}\boldsymbol{\gamma}_1 = 1, \tag{28c}$$

where $\boldsymbol{\gamma}_1$ is the first column of $\mathbf{\Gamma}_{\mathrm{d}}$. The analytic solution to (28) is equal to an LCMV filter [37].

The late reverberation PSD can then be estimated by subtracting the PSD of the filtered noise components from the PSD of the filter input signals normalized by the filtered diffuse coherence [36], i.e.

$$\widehat{\phi}_d = \max\left\{\frac{\mathbf{w}_{\mathrm{d}}^{\mathrm{H}}\widehat{\mathbf{\Phi}}_{\mathbf{y}}\mathbf{w}_{\mathrm{d}} - \mathbf{w}_{\mathrm{d}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{w}_{\mathrm{d}}}{\mathbf{w}_{\mathrm{d}}^{\mathrm{H}}\mathbf{\Gamma}_{\mathrm{d}}\mathbf{w}_{\mathrm{d}}},\ 0\right\}, \tag{29}$$

where the $\max\{\cdot\}$ operation is introduced to avoid negative PSD estimates. The input PSD matrix is recursively estimated using $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) = \beta\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n-1) + (1-\beta)\mathbf{y}(k,n)\mathbf{y}^{\mathrm{H}}(k,n)$.

B. Non-Blocking Based Methods

In contrast to the blocking based methods from Section III-A, methods within the class discussed in this section do not rely on blocking the desired sound component. Instead, they jointly estimate the desired and late reverberation PSDs. Although the method presented in [38] also falls into this category, it is excluded here, as it does not consider additive noise. Nevertheless, it is worthwhile to note that if the noise component $\mathbf{v}(k,n)$ is zero, the solution from [38] provides a closed-form solution to the problem of jointly estimating $\phi_x(k,n)$ and $\phi_d(k,n)$ in the ML sense. In Section III-B1, the Newton method is used to obtain the ML estimates of desired and late reverberation PSDs by assuming the diffuse coherence $\mathbf{\Gamma}_{\mathrm{d}}(k)$ to be known, whereas the method reviewed in Section III-B2, can also estimate $\mathbf{\Gamma}_{\mathrm{d}}(k)$ from the data using an EM algorithm. The original EM method is described in Section III-B2a, whereas in Section III-B2b we assume that $\mathbf{\Gamma}_{\mathrm{d}}(k)$ is known. In Section III-B3, the desired and diffuse PSDs are estimated jointly in the least-squares sense.

1) ML Using Newton's Method: In [39], a ML method to jointly estimate $\phi_x(k,n)$ and $\phi_d(k,n)$ is proposed under the assumption that the diffuse coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$ is known.

By defining $\mathbf{p}(k,n) = [\phi_x(k,n)\,,\ \phi_d(k,n)]^{\mathrm{T}}$ as the unknown parameter set and assuming $\mathbf{y}(k,n) \sim f_{\mathbf{y}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{y}})$, the ML estimate of $\mathbf{p}(k,n)$ given $\mathbf{y}(k,n)$ can be found by the Newton method [56] using

$$\mathbf{p}^{(\ell+1)} = \mathbf{p}^{(\ell)} - \boldsymbol{\mathcal{H}}^{-1}\left(\mathbf{p}^{(\ell)}\right)\boldsymbol{\delta}\left(\mathbf{p}^{(\ell)}\right), \tag{30}$$

where $\boldsymbol{\delta}(\mathbf{p})$ is the gradient of the log-likelihood, and $\boldsymbol{\mathcal{H}}(\mathbf{p})$ is the corresponding Hessian matrix, i.e.,

$$\boldsymbol{\delta}(\mathbf{p}) \equiv \frac{\partial \log f_{\mathbf{y}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{y}})}{\partial \mathbf{p}} \tag{31}$$

$$\boldsymbol{\mathcal{H}}(\mathbf{p}) \equiv \frac{\partial^2 \log f_{\mathbf{y}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{y}})}{\partial \mathbf{p}\partial \mathbf{p}^{\mathrm{T}}}, \tag{32}$$

where $f(\mathbf{y};\,\mathbf{p})$ is the p.d.f. of the microphone signal vector. The gradient $\boldsymbol{\delta}(\mathbf{p}) \equiv [\delta_x(\mathbf{p}),\ \delta_d(\mathbf{p})]^{\mathrm{T}}$ is a 2-dimensional vector with elements

$$\delta_i(\mathbf{p}) = M\ \mathrm{tr}\left\{\left(\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{y}\mathbf{y}^{\mathrm{H}} - \mathbf{I}\right)\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_i}\right\}, \tag{33}$$

where $i \in \{x, d\}$, $\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_x} = \mathbf{a}\mathbf{a}^{\mathrm{H}}$, and $\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_d} = \mathbf{\Gamma}_{\mathrm{d}}$.

The Hessian is a symmetric 2×2 matrix:

$$\boldsymbol{\mathcal{H}}(\mathbf{p}) \equiv \begin{bmatrix} \mathcal{H}_{xx}(\mathbf{p}) & \mathcal{H}_{dx}(\mathbf{p}) \\ \mathcal{H}_{xd}(\mathbf{p}) & \mathcal{H}_{dd}(\mathbf{p}) \end{bmatrix}. \tag{34}$$

with the elements

$$\begin{aligned}\mathcal{H}_{ij}(\mathbf{p}) = -M\ \mathrm{tr}\Bigg\{&\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_j}\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{y}\mathbf{y}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_i} \\ &+ \left(\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{y}\mathbf{y}^{\mathrm{H}} - \mathbf{I}\right)\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_j}\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial \mathbf{\Phi}_{\mathbf{y}}}{\partial \phi_i}\Bigg\},\end{aligned} \tag{35}$$

where $i, j \in \{x, d\}$.

In practice, the matrix $\mathbf{y}\mathbf{y}^{\mathrm{H}}$ in (33) and (35) is replaced by the smoothed version $\widehat{\mathbf{\Phi}}_{\mathbf{y}}$. The algorithm is initialized with $\phi_x^{(0)} = \phi_d^{(0)} = \epsilon\frac{1}{M}\,\mathrm{tr}\{\widehat{\mathbf{\Phi}}_{\mathbf{y}}\}$. The Newton algorithm is stopped if the estimates at iteration ℓ reach a predefined lower or upper bound, or if a convergence threshold is reached.

2) ML Using the EM Method: The EM algorithm proposed in [40] is a batch algorithm that provides estimates of the desired sound PSD $\phi_x(k,n)$, the RTF vector $\mathbf{a}(k)$, the late reverberation PSD $\phi_d(k,n)$ and the late reverberation coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$. For consistency with the other methods, we assume that $\mathbf{a}(k)$ is known and therefore is not estimated within the EM. In Section III-B2a, we describe the method proposed in [40]. In Section III-B2b, the method is modified by assuming prior knowledge of the coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$ to investigate the effect of estimating $\mathbf{\Gamma}_{\mathrm{d}}(k)$.

a) ML-EM with unknown reverberation coherence matrix: The desired and diffuse sound components are concatenated in the hidden data vector

$$\mathbf{q}(k,n) \triangleq \begin{bmatrix} X(k,n) & \mathbf{d}^{\mathrm{T}}(k,n) \end{bmatrix}^{\mathrm{T}}. \tag{36}$$

Using this definition, equation (1) can be rewritten as

$$\mathbf{y}(k,n) = \mathbf{H}(k,n)\,\mathbf{q}(k,n) + \mathbf{v}(k,n), \tag{37}$$

where the matrix $\mathbf{H}(k,n) \triangleq [\mathbf{a}(k,n),\, \mathbf{I}_{M\times M}]$. The desired parameter set is

$$\boldsymbol{\theta}(k) = \left\{\bar{\boldsymbol{\phi}}_x(k),\ \bar{\boldsymbol{\phi}}_d(k),\ \boldsymbol{\Gamma}_{\mathrm{d}}(k)\right\}, \tag{38}$$

where $\bar{\boldsymbol{\phi}}_x(k) = [\phi_x(k,1),\ldots,\phi_x(k,N)]^{\mathrm{T}}$ and $\bar{\boldsymbol{\phi}}_d(k) = [\phi_d(k,1),\ldots,\phi_d(k,N)]^{\mathrm{T}}$, with N being the number of frames. By concatenating the hidden data vectors of all time frames $1,\ldots,N$ to $\bar{\mathbf{q}}(k) = [\mathbf{q}^{\mathrm{T}}(k,1),\ldots,\mathbf{q}^{\mathrm{T}}(k,N)]^{\mathrm{T}}$, and defining $\bar{\mathbf{y}}(k)$ similarly, the conditional expectation of the log-likelihood function can be deduced as

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(\ell)}\right) = E\left\{\log f_{\bar{\mathbf{y}}}(\mathbf{0},\boldsymbol{\Phi}_{\bar{\mathbf{y}}}) \;\middle|\; \bar{\mathbf{y}}(k);\,\boldsymbol{\theta}^{(\ell)}\right\}, \tag{39}$$

where $\boldsymbol{\theta}^{(\ell)}$ is the parameter-set estimate at iteration ℓ.

For implementing the E-step, it is sufficient to estimate $\widehat{\mathbf{q}}(k,n) \triangleq E\{\mathbf{q}(k,n)|\mathbf{y}(k,n);\boldsymbol{\theta}^{(\ell)}\}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbf{q}}(k,n) \triangleq E\{\mathbf{q}(k,n)\mathbf{q}^{\mathrm{H}}(k,n)|\mathbf{y}(k,n);\boldsymbol{\theta}^{(\ell)}\}$ being the first- and second-order statistics of the hidden-data given the measurements, respectively. Assuming that $\mathbf{y}(k,n)$ and $\mathbf{q}(k,n)$ in (37) are Gaussian random vectors, $\mathbf{q}(k,n)$ can be estimated by the optimal linear estimator [40]

$$\begin{aligned}\widehat{\mathbf{q}} &= E\left\{\mathbf{q}\mathbf{y}^{\mathrm{H}}\right\} \times \left(E\left\{\mathbf{y}\mathbf{y}^{\mathrm{H}}\right\}\right)^{-1}\mathbf{y}\\ &= \boldsymbol{\Phi}_{\mathbf{q}}^{(\ell)}\mathbf{H}^{\mathrm{H}}\left(\boldsymbol{\Phi}_{\mathbf{y}}^{(\ell)}\right)^{-1}\mathbf{y}\end{aligned} \tag{40}$$

with

$$\boldsymbol{\Phi}_{\mathbf{q}}^{(\ell)}(k,n) = \begin{bmatrix}\phi_x^{(\ell)}(k,n) & \mathbf{0}_{1\times M}\\ \mathbf{0}_{M\times 1} & \phi_d^{(\ell)}(k,n)\,\boldsymbol{\Gamma}_{\mathrm{d}}^{(\ell)}(k)\end{bmatrix}, \tag{41}$$

and $\boldsymbol{\Phi}_{\mathbf{y}}^{(\ell)} = \mathbf{H}\boldsymbol{\Phi}_{\mathbf{q}}^{(\ell)}\mathbf{H}^{\mathrm{H}} + \boldsymbol{\Phi}_{\mathbf{v}}$. The matrix

$$\widehat{\boldsymbol{\Psi}}_{\mathbf{q}}(k,n) \triangleq \begin{bmatrix}\widehat{|X(k,n)|^2} & \widehat{X(k,n)\mathbf{d}^{\mathrm{H}}(k,n)}\\ \widehat{X^*(k,n)\mathbf{d}(k,n)} & \widehat{\mathbf{d}(k,n)\mathbf{d}^{\mathrm{H}}(k,n)}\end{bmatrix}, \tag{42}$$

can be obtained by [40]

$$\widehat{\boldsymbol{\Psi}}_{\mathbf{q}} = \widehat{\mathbf{q}}\,\widehat{\mathbf{q}}^{\mathrm{H}} + \boldsymbol{\Phi}_{\mathbf{q}}^{(\ell)} - \boldsymbol{\Phi}_{\mathbf{q}}^{(\ell)}\mathbf{H}^{\mathrm{H}}\left(\boldsymbol{\Phi}_{\mathbf{y}}^{(\ell)}\right)^{-1}\mathbf{H}\boldsymbol{\Phi}_{\mathbf{q}}^{(\ell)}. \tag{43}$$

Maximizing $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(\ell)})$ with relation to the problem parameters constitutes the M-step, i.e. [40]

1. $$\phi_x^{(\ell+1)}(k,n) = \widehat{|X(k,n)|^2} \tag{44}$$

2. $$\boldsymbol{\Gamma}_{\mathrm{d}}^{(\ell+1)}(k) = \frac{1}{N}\sum_{n=1}^{N}\frac{\widehat{\mathbf{d}(k,n)\mathbf{d}^{\mathrm{H}}(k,n)}}{\phi_d^{(\ell)}(k,n)} \tag{45}$$

3. $$\phi_d^{(\ell+1)}(k,n) = \frac{1}{M}\,\mathrm{tr}\left\{\widehat{\mathbf{d}(k,n)\mathbf{d}^{\mathrm{H}}(k,n)}\left(\boldsymbol{\Gamma}_{\mathrm{d}}^{(\ell+1)}(k)\right)^{-1}\right\}. \tag{46}$$

The EM iterations are initialized with $\phi_x^{(0)}(k,n) = \epsilon_x\,\frac{1}{M}\,\mathrm{tr}\{\widehat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,n)\}$ and $\phi_d^{(0)}(k,n) = \epsilon_d\,\frac{1}{M}\,\mathrm{tr}\{\widehat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,n)\}$, where $\epsilon_x > \epsilon_d$, and $\boldsymbol{\Gamma}_{\mathrm{d}}^{(0)}(k)$ is initialized with (4).

b) ML-EM with known reverberation coherence matrix: By assuming that the model for $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$ given by (4) holds, the method described in Section III-B2a needs to be modified only by omitting (45) in the M-step and using the a priori known spatial coherence matrix instead.

3) PSD Matrix-Based Least-Squares Method: By matching $\widehat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,n)$, which can be estimated from the microphone signals, and its model given in (2), the problem at hand can be formulated as a system of M^2 equations in two unknown variables [41]. Since there are more equations than variables, the vector $\mathbf{p}(k,n) = [\phi_x(k,n)\,,\,\phi_d(k,n)]^{\mathrm{T}}$ that minimizes the total squared error can be found by minimizing the squared Frobenius norm as

$$\widehat{\mathbf{p}} = \underset{\mathbf{p}}{\mathrm{argmin}}\left\|\underbrace{\widehat{\boldsymbol{\Phi}}_{\mathbf{y}} - \left(\phi_x\mathbf{a}\mathbf{a}^{\mathrm{H}} + \phi_d\boldsymbol{\Gamma}_{\mathrm{d}} + \boldsymbol{\Phi}_{\mathbf{v}}\right)}_{\boldsymbol{\Phi}_{\mathrm{LS}}}\right\|_{\mathrm{F}}^2, \tag{47}$$

where $\boldsymbol{\Phi}_{\mathrm{LS}}(k,n)$ is the error matrix. Following some algebraic steps, the cost function in (47) can be written as

$$||\boldsymbol{\Phi}_{\mathrm{LS}}||_{\mathrm{F}}^2 = \mathbf{p}^{\mathrm{T}}\mathbf{A}\mathbf{p} - 2\mathbf{b}^{\mathrm{T}}\mathbf{p} + C, \tag{48}$$

where $C(k,n)$ is independent of $\mathbf{p}(k,n)$, and $\mathbf{A}(k,n)$ and $\mathbf{b}(k,n)$ are defined as

$$\mathbf{A} \equiv \begin{bmatrix}\left(\mathbf{a}^{\mathrm{H}}\mathbf{a}\right)^2 & \mathbf{a}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{a}\\ \mathbf{a}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{a} & \mathrm{tr}\left\{\boldsymbol{\Gamma}_{\mathrm{d}}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{d}}\right\}\end{bmatrix} \tag{49}$$

$$\mathbf{b} \equiv \begin{bmatrix}\Re\left\{\mathbf{a}^{\mathrm{H}}\left(\widehat{\boldsymbol{\Phi}}_{\mathbf{y}} - \boldsymbol{\Phi}_{\mathbf{v}}\right)\mathbf{a}\right\}\\ \Re\left\{\mathrm{tr}\left\{\left(\widehat{\boldsymbol{\Phi}}_{\mathbf{y}} - \boldsymbol{\Phi}_{\mathbf{v}}\right)\boldsymbol{\Gamma}_{\mathrm{d}}^{\mathrm{H}}\right\}\right\}\end{bmatrix}. \tag{50}$$

Since the cost function $||\boldsymbol{\Phi}_{\mathrm{LS}}(k,n)||_{\mathrm{F}}^2$ in (48) has a quadratic form, setting its gradient w.r.t. $\mathbf{p}(k,n)$ to zero yields

$$\widehat{\mathbf{p}}(k,n) = \mathbf{A}^{-1}(k,n)\mathbf{b}(k,n). \tag{51}$$

Note that this method is related to the method presented in Section III-A1. Both methods minimize the Frobenius norm of an error matrix, where the desired sound is blocked in the first method, whereas the late reverberation and desired sound PSDs are estimated jointly in the second method.

IV. Coherence-Based Indirect PSD Estimators

While all methods in Section III directly estimate the late reverberation PSD, we consider indirect estimators in this section. Within this class, we focus on methods using an estimate of the CDR to estimate the PSD of the diffuse sound, i.e. late reverberation. These estimators rely on the fact that the desired signal $X(k,n)$ is coherent across all microphones.

The CDR as defined in [27], [42] for the microphone pair $i,j \in \{1,\ldots,M\}$ is given by

$$\mathrm{CDR}_{i,j}(k,n) = \frac{\phi_x(k,n)\,|A_i(k)|\,|A_j(k)|}{\phi_d(k,n)}. \tag{52}$$

The CDR can be estimated using various methods, e.g., [42], [57]. To limit the number of estimators under test, we restrict ourselves to the "proposed 2" CDR estimator described in [42], which was reported to perform best across the considered CDR estimators. The CDR estimator requires knowledge of the RTFs

$\mathbf{a}(k)$ and the diffuse coherence matrix $\mathbf{\Gamma}_{\mathrm{d}}(k)$. Furthermore, we compensate for the additive noise as proposed in [45].

To take all microphones into account, we average the CDR estimate for each microphone pair [30], [31]

$$\widehat{\mathrm{CDR}}(k,n) = \frac{1}{\mathcal{M}} \sum_{i,j \in \mathcal{M}} \frac{\widehat{\mathrm{CDR}}_{i,j}(k,n)}{|A_i(k)|\,|A_j(k)|}, \tag{53}$$

where the set $\mathcal{M}$ contains all microphone pair combinations. Given an estimate of the CDR, and exploiting the diffuse homogeneity, the late reverberation PSD is obtained by [21]

$$\widehat{\phi}_d(k,n) = \frac{\frac{1}{M}\,\mathrm{tr}\left\{\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) - \mathbf{\Phi}_{\mathbf{v}}(k,n)\right\}}{\frac{1}{M}\mathbf{a}^{\mathrm{H}}(k)\mathbf{a}(k)\,\widehat{\mathrm{CDR}}(k,n) + 1}. \tag{54}$$

If the desired sound $X(k,n)$ is modeled as a plane wave such that $|A_m(k)| = 1, \ \forall m$, then (52)-(54) can be simplified.

V. Temporal Model-Based PSD Estimators

Instead of modeling the late reverberation as a diffuse sound field, estimators within the fourth class exploit the temporal structure of reverberation, such that they can be applied to each individual microphone signal.

A. Statistical Temporal Model

In [15], [16] it was proposed to model the impulse response by an exponentially decaying random process per frequency band. Using this model, the late reverberation PSD of the m-th microphone $\phi_d^{(m)}$ can be estimated depending on two a priori required parameters, namely the frequency-dependent reverberation time $T_{60}(k)$ and the inverse direct-to-reverberation ratio (DRR) $\kappa(k)$, by [16]

$$\begin{aligned}\widehat{\phi}_d^{(m)}(k,n) &= [1-\kappa(k)]e^{-2\alpha(k)RN_{\mathrm{D}}}\widehat{\phi}_d^{(m)}(k,n-N_{\mathrm{D}}) \\ &+ \kappa(k)e^{-2\alpha(k)RN_{\mathrm{D}}}\left[\widehat{\phi}_y^{(m)}(k,n-N_{\mathrm{D}}) - \phi_v^{(m)}(k,n-N_{\mathrm{D}})\right],\end{aligned} \tag{55}$$

where N_{D} corresponds to the number of frames between the direct sound and the start time of the late reverberation, $\alpha(k) = 3\ln(10)/(T_{60}(k)f_{\mathrm{s}})$ is the reverberation decay constant, R is the hop-size, and $\widehat{\phi}_y^{(m)}(k,n)$ and $\phi_v^{(m)}(k,n)$ are the diagonal entries of $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$, respectively. Following the spatial homogeneity assumption of the late reverberation, we can spatially average the PSD estimates across all microphones as

$$\widehat{\phi}_d(k,n) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\phi}_d^{(m)}(k,n). \tag{56}$$

B. Convolutive Transfer Function Based Methods

Using the convolutive transfer function (CTF) approximation [58] per frequency band, the m-th microphone signal can be described by

$$Y_m(k,n) = \sum_{\ell=0}^{L} H_m(k,\ell)X_m(k,n-\ell) + V_m(k,n), \tag{57}$$

where $X_m(k,n)$ is the direct speech signal at the m-th microphone, $H_m(k,\ell)$ for $\ell \in \{0,\ldots,L\}$ are the CTFs and L is the required number of frames to model the reverberation. By using a relative CTF formulation [43], we can re-interpret $X_m(k,n)$ as the speech component in the m-th microphone containing some early reflections and $H_m(k,\ell)$ as the relative CTFs such that $H_m(k,0) = 1$. By modeling the coefficients $H_m(k,\ell)$ by a first-order Markov random variable, $H_m(k,\ell)$ can be estimated using a Kalman filter for $\ell \in \{1,\ldots,L\}$, and past frames of $X_m(k,n)$ can be estimated using an auxiliary Wiener filter. Using the estimates $\widehat{H}_m(k,\ell)$ and $\widehat{X}_m(k,n)$, the late reverberation PSD in the m-th microphone can be estimated by

$$\widehat{\phi}_d^{(m)}(k,n) = E\left\{\left|\sum_{\ell=N_{\mathrm{D}}}^{L}\widehat{H}_m(k,\ell)\widehat{X}_m(k,n-\ell)\right|^2\right\}, \tag{58}$$

where N_{D} again denotes the start time frame of the late reverberation. The expectation in (58) can be approximated by a recursive average. As with the previous single-channel estimator in Section V-A, the microphone-specific PSDs $\widehat{\phi}_d^{(m)}(k,n)$ can be spatially averaged using (56).

VI. Discussion

An overview of the different classes of estimators along with important properties are shown in Table I. The first two columns indicate the section numbers and short acronyms of the methods. Discriminative properties of the methods are whether they

- exploit a spatial coherence model,
- require prior knowledge of the spatial coherence of the late reverberation $\mathbf{\Gamma}_{\mathrm{d}}(k)$,
- exploit a temporal structure model,
- additionally/inherently deliver an estimate of the desired sound PSD $\phi_x(k,n)$,
- are online or batch processing methods, and the type of solution (closed-form, iterative, recursive, etc.),
- yield a high or low computational complexity in terms of the real-time factor.

The real-time factor, i.e. the processing time per time frame, was measured running MATLAB R2016b on a 3.1 GHz Intel Core i5 processor. Although the implementations were not optimized for runtime, the real-time factors give a good indication of the computational complexity. Algorithms with a real-time factor <1 are not complex and therefore easy to implement also on less powerful devices, whereas algorithms with a real-time factor >1 require more powerful processors and strong optimization to be able to run in real-time. It can be observed that mainly the methods that do not have a closed-form solution, are rather complex. For the *CTF* method, the complexity depends on the filter length L. The parameter settings for each algorithm are described in Section VIII-A.

TABLE I
CLASSIFICATION AND PROPERTIES OF LATE REVERBERATION PSD ESTIMATORS

Section	Method	exploits spatial coherence model	requires knowledge of spatial coherence	exploits temporal structure	estimates ϕ_x	processing / solution	real-time factor
III-A1	Blocking PSD LS [21]	✓	✓	✗	✗	online / closed-form	0.8
III-A2a	Blocking ML root [24]	✓	✓	✗	✓	online / polyn. rooting	8.8
III-A2b	Blocking ML Newton [35]	✓	✓	✗	✗	online / iterative	3.7
III-A3	BF LCMV [37]	✓	✓	✗	✗	online / closed-form	0.5
III-B1	ML Newton [39]	✓	✓	✗	✓	online / iterative	6.5
III-B2a	ML-EM est. coh. [40]	✓	✗	✗	✓	batch / iterative	15.1
III-B2b	ML-EM diff. coh. [40]	✓	✓	✗	✓	batch / iterative	14.9
III-B3	PSD LS [41]	✓	✓	✗	✓	online / closed-form	0.6
IV	CDR [42]	✓	✓	✗	✗	online / closed-form	0.7
V-A	LRSV [16]	✗	✗	✓	✗	online / closed-form	0.4
V-B	CTF [43]	✗	✗	✓	✗	online / recursive	8.5

While the coherence-based methods and the *LRSV* method practically instantaneously deliver useful PSD estimates without delay, the *CTF* method requires a short initial convergence phase of 1-2 s before providing accurate estimates as shown in [43]. An exception is the ML-EM that is a batch method, which requires a larger amount of data (in the range of several seconds) before providing a result, and is therefore not useful for online processing.

It is interesting that the estimator pairs *blocking PSD LS* (Section III-A1) and *PSD LS* (Section III-B3), and *blocking ML Newton* (Section III-A2b) and *ML Newton* (Section III-B1), use the same mathematical solution methods, while the first ones use a blocking of the desired sound whereas the latter ones jointly estimate late reverberation and desired sound PSDs.

Note that all spatial coherence-based methods (Sections III and IV) can be used also to estimate the PSD of non-reverberation related sound fields, such as possibly non-stationary diffuse noise or ambient sounds like babble noise. However, spatial coherence-based estimators cannot discriminate between reverberation originating from a speech signal and other diffuse sounds, if the reverberation and the other diffuse components have the same spatial coherence. Methods exploiting temporal structures of reverberation such as the ones imposing a model on the reverberant tail (Section V-A) or exploiting the CTF model (Section V-B) can discriminate between reverberation and other diffuse sound fields. Consequently, these methods are not suitable to estimate the PSD of general diffuse sound fields.

Furthermore, all spatial coherence-based methods require prior knowledge or estimates of the RTFs of the desired sound $\mathbf{a}(k)$, whereas the temporal model-based reverberation PSD estimators in Section V work independently per channel and require some temporal information like the T_{60} or the length of the relative CTF L.

VII. BIAS COMPENSATION

As will be shown in Section VIII, the spatial coherence-based estimators under test are severely biased in high DDR conditions. Since overestimation of the late reverberation PSD is especially harmful to the audio quality as it causes speech distortion when used for dereverberation, we propose a simple compensation method using the correction factor $c_d(k,n) = f_c(\widehat{\phi}_{\mathrm{d}}, \widehat{\mathrm{DDR}})$ as a function of the estimated DDR. A similar compensation method was proposed in [59] in the context of noise reduction.

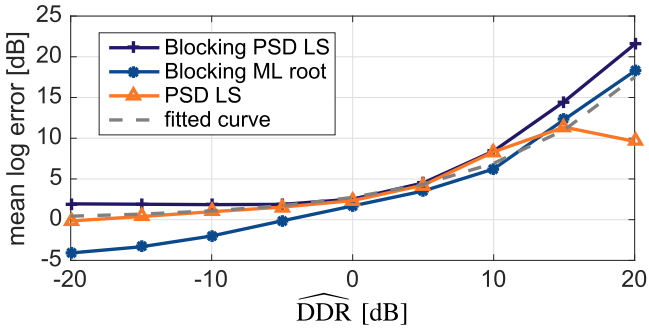

Fig. 1. Mean log PSD error and fitted exponential compensation function depending on the estimated DDR obtained using simulated RIRs.

As a proof of concept and without claiming optimality, we fit an exponential function to the mean logarithmic PSD estimation error of the three coherence-based estimators depending on the estimated DDR as shown in Fig. 1, where the logarithmic PSD estimation error will be defined in (65). We approximate the error using the function

$$c_{\mathrm{d}}(\widehat{\mathrm{DDR}}) = a \cdot e^{b 10 \log_{10} \widehat{\mathrm{DDR}}}, \tag{59}$$

where the bias function c_{d} is obtained in dB and the DDR is estimated by $\widehat{\mathrm{DDR}} = \frac{\widehat{\phi}_y - \widehat{\phi}_{\mathrm{d}} - \widehat{\phi}_v}{\widehat{\phi}_{\mathrm{d}}}$. By using MATLAB's `fit()` function we fit the exponential function (59) to the average error of the three coherence-based estimators within the range $\widehat{\mathrm{DDR}} = [-20, 20]$ dB as shown in Fig. 1 as an example. The such obtained values are $a = 2.735$ and $b = 0.0928$. Fig. 1 shows the used error data and the fitted curve $c_{\mathrm{d}}(\widehat{\mathrm{DDR}})$.

The compensated PSD $\widehat{\phi}_d^{\mathrm{comp}}$ is then obtained by multiplying the estimated PSD $\widehat{\phi}_d$ with the inverse linearized bias function

$$\widehat{\phi}_d^{\mathrm{comp}} = 10^{-c_{\mathrm{d}}(\widehat{\mathrm{DDR}})/10}\, \widehat{\phi}_d. \tag{60}$$

VIII. Evaluation

In this section, we evaluate the performance of the estimators reviewed in Sections III and V for different acoustic setups. Sections VIII-A and VIII-C discuss the used simulation parameters, signal generation and performance measures. In Section VIII-D we first use a controlled setup for only spatial coherence-based estimators using artificial white noise signals in a stationary diffuse noise field. The reverberant PSD estimators discussed in Section V are excluded from this first evaluation since they are not suitable for this scenario. The ML-EM with unknown coherence (Section III-B2a) is also omitted from this first evaluation, as in this case the assumed coherence model perfectly fits the data. Second, an evaluation using speech and measured room impulse responses (RIRs) and recorded noise is conducted in Section VIII-E. Finally in Sections VIII-F and VIII-G, measured RIRs are used to confirm the results in realistic environments.

A. Acoustic Setup and Simulation Parameters

In all simulations and measurements, we used a uniform circular array with a radius of 10 cm and $M = 8$ omnidirectional microphones. In Section VIII-D, the desired sound component $X(k,n)$ was generated as a plane wave, while the late reverberation component $\mathbf{d}(k,n)$ was generated as a stationary diffuse field. In Section VIII-E, realistic signals were generated using measured RIRs and recorded noise from the REVERB challenge database [60] and speech data from [61]. The REVERB database provides in total 12 acoustic conditions: three different rooms with $T_{60} \approx \{0.3, 0.6, 0.7\}$ s, where in each room two different source angles at two distances (0.5 m and 2 m) were measured. The speech data featured 3 female and 3 male speakers with a total length of over 2 minutes.

The signals were sampled with a sampling frequency of $f_{\mathrm{s}} = 16$ kHz, and analyzed using an STFT with 50% overlapping square-root Hann windows of length 32 ms, and $N_{\mathrm{FFT}} = 1024$. The stationary noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ and the RTF vector $\mathbf{a}(k)$ of the desired sound were assumed to be known in advance. The noise PSD matrix was computed during speech absence and the RTF vector was obtained from the direct sound peak of the RIRs. Therefore, by assuming the source in the far-field, the RTF vector $\mathbf{a}(k,n)$ corresponds to simple delays and is referred to as *steering vector* in the following. The recursive smoothing parameter for estimating the PSD matrices was set to $\beta = 0.73$, which corresponds to an exponential smoothing with a time constant of 50 ms. Algorithm specific parameters were chosen as follows: $N_{\mathrm{D}} = 1$ frames, $T_{60}(k)$ was set according to the fullband reverberation time for each room, the CTF length L was chosen according to the fullband T_{60} in each room, $\epsilon = 0.01$, $\epsilon_x = 0.5$, and $\epsilon_d = 0.1$. The iterative Newton and EM algorithms were halted after a maximum of 10 iterations, even if the convergence threshold of $|\phi_d^{(\ell)} - \phi_d^{(\ell-1)}| < 10^{-12}$ was not reached.

B. Signal Generation

The definition of the ground truth, i.e. the oracle PSD $\phi_d(k,n)$, is very important as it significantly influences the results. In Section VIII-D, we utilize a highly controlled test scenario. All signal components were generated using stationary white noise in the time domain: the direct sound component at the reference microphone $X(k,n)$ with PSD $\phi_x(k,n)$, the diffuse sound component $\mathbf{d}(k,n)$ with PSD $\phi_d^{(m)}(k,n) = \phi_d(k,n)$ by imposing the diffuse long-term coherence given by (4) on white noise signals using the method proposed in [62], and the additive noise component $\mathbf{v}(k,n)$ with PSD $\phi_v^{(m)}(k,n) = \phi_v(k,n)$. To obtain the test signals, the different stationary signal components of 10 s length were summed up depending on the reverberant signal-to-noise ratio (RSNR)

$$\mathrm{RSNR} = \frac{\sum_{k,n} \left(\phi_x(k,n) + \phi_d(k,n)\right)}{\sum_{k,n} \phi_v(k,n)}, \tag{61}$$

and the DDR

$$\mathrm{DDR} = \frac{\sum_{k,n} \phi_x(k,n)}{\sum_{k,n} \phi_d(k,n)}. \tag{62}$$

Although it is often assumed that the transition between early reflections and late reverberation starts around 50 ms after the direct sound, we chose this transition smaller to find a fair common ground truth for the coherence-based and temporal models: Unlike the temporal model based methods, the coherence-based model does not have a control parameter to define the start time of the late reverberation. Therefore, the only reasonable option is to define $N_{\mathrm{D}} = 1$, i.e. that the late reverberation starts one frame shift (in our case 16 ms) after the direct sound. In Sections VIII-E and VIII-G, the reverberant signals $\mathbf{a}(k)X(k,n) + \mathbf{d}(k,n)$ were generated by convolving non-reverberant speech signals with measured RIRs from the REVERB database. The time-domain representation of the oracle reverberation component for evaluation purposes $\mathbf{d}(k,n)$ was then obtained by convolution of the non-reverberant test speech signal with windowed RIRs containing only the late part of the reverberation, starting 16 ms after the direct sound. The time-domain representation of the additive noise $\mathbf{v}(k,n)$ was pink noise, in order to maintain an approximately constant RSNR per frequency band to the speech.

The oracle reverberation PSD used as a target for evaluation is the spatially averaged instantaneous late reverberation power, i.e.,

$$\overline{\phi}_d(k,n) = \frac{\mathbf{d}^{\mathrm{H}}(k,n)\mathbf{d}(k,n)}{M}. \tag{63}$$

For the speech enhancement evaluation in Sections VIII-F and VIII-G, in addition to using the theoretical diffuse coherence given by (4), we also used the oracle coherence matrix of the late reverberation, where the (i,j)-th element was computed by

$$\overline{\mathbf{\Gamma}}_{\mathrm{d}}^{(i,j)}(k) = \frac{\sum_{n=1}^{N} D_i(k,n) D_j^*(k,n)}{\sqrt{\left(\sum_{n=1}^{N} |D_i(k,n)|^2\right)\left(\sum_{n=1}^{N} |D_j(k,n)|^2\right)}}, \tag{64}$$

where $D_m(k,n)$ is the m-th element of $\mathbf{d}(k,n)$.

The oracle desired signal for evaluation purposes required in Sections VIII-F and VIII-G is defined as the direct sound at the reference microphone, and was obtained by convolving only the

windowed direct path of the reference microphone RIR with the anechoic speech signal.

C. Performance Measures

1) Logarithmic PSD Estimation Error: To evaluate the estimation accuracy of the various PSD estimators, we employ the bin-wise logarithmic error

$$e(k,n) = 10\log_{10}\frac{\widehat{\phi}_d(k,n)}{\overline{\phi}_d(k,n)}, \tag{65}$$

which directly reflects over- and underestimation as positive and negative values in dB, respectively. The log error is analyzed statistically in terms of its mean μ_e and the lower and upper semi-variance [63]

$$\sigma^2_{e,\mathrm{l}} = \frac{1}{|\mathcal{T}_\mathrm{l}|}\sum_{\{k,n\}\in\mathcal{T}_\mathrm{l}}\left(e(k,n)-\mu_e\right)^2,\ \mathcal{T}_\mathrm{l}:\ e(k,n)\le\mu_e \tag{66a}$$

$$\sigma^2_{e,\mathrm{u}} = \frac{1}{|\mathcal{T}_\mathrm{u}|}\sum_{\{k,n\}\in\mathcal{T}_\mathrm{u}}\left(e(k,n)-\mu_e\right)^2,\ \mathcal{T}_\mathrm{u}:\ e(k,n)>\mu_e, \tag{66b}$$

where the sets of time-frequency bins $\mathcal{T}_\mathrm{l}$ and $\mathcal{T}_\mathrm{u}$ contain all bins below or above the mean, respectively. Therefore, a log error with zero mean and small semi-standard deviations $\sigma_{e,\mathrm{l}}$ and $\sigma_{e,\mathrm{u}}$ is desired. In the following figures, the mean is represented by symbols (circle, square, etc.) and the semi-standard deviations are indicated by whisker bars.

2) Speech Enhancement Measures: We also assess the influence of the various PSD estimates on the dereverberation performance of the MWF using (5), (8) and (3) by employing several speech enhancement measures.

To assess the perceptual similarity between the MWF output signal $\widehat{X}(k,n)$ and the oracle desired signal $X(k,n)$, we employ the *Cepstral Distance* (CD) [64] and the *Perceptual Evaluation of Speech Quality* (PESQ) measure [65]. The amount of perceived reverberation is quantified by the *normalized Speech-to-Reverberation-Modulation Energy Ratio* (SRMR) [66]. Furthermore, we compute the segmental *interference reduction* (IR) and *speech distortion index* (SDI) [67] as

$$\mathrm{IR} = \frac{1}{\mathcal{T}}\sum_{n\in\mathcal{T}} 10\log_{10}\frac{\sum_{t=(n-1)T}^{nT} s^2_{dv,\mathrm{in}}(t)}{\sum_{t=(n-1)T}^{nT} s^2_{dv,\mathrm{MWF}}(t)} \tag{67}$$

$$\mathrm{SDI} = \frac{1}{\mathcal{T}}\sum_{n\in\mathcal{T}} \frac{\sum_{t=(n-1)T}^{nT}\left(s_{x,\mathrm{MWF}}(t)-s_{x,\mathrm{in}}(t)\right)^2}{\sum_{t=(n-1)T}^{nT} s^2_{x,\mathrm{in}}(t)}, \tag{68}$$

where $s_{dv,\mathrm{in}}(t)$ and $s_{dv,\mathrm{MWF}}(t)$ are the time-domain representations of late reverberation plus noise at the reference microphone and at the MWF output, respectively, $s_{x,\mathrm{in}}(t)$ and $s_{x,\mathrm{MWF}}(t)$ are the time-domain representations of direct sound at the reference microphone and the MWF output, t is the sample index, T is the number of samples corresponding to a segment of 20 ms and the set $\mathcal{T}$ contains only time segments where speech is active, determined by an ideal voice activity detector. For the perceptually motivated measures, we compute the improvement with respect to the unprocessed reference microphone signal, indicated by ΔCD, ΔPESQ and ΔSRMR.

D. Evaluation of Spatial Coherence-Based PSD Estimators for Stationary Diffuse Noise

For a stationary diffuse field the results of the logarithmic PSD estimation error as described in Section VIII-C is shown in Fig. 2 over varying RSNRs with a fixed DDR at 10 dB. It can be observed that at higher RSNRs, all estimators have a small variance, and their error means are close to 0 dB, which means a small estimation error. *Blocking ML Newton* is performing slightly worse in medium RSNRs compared to the other estimators due to a small increase in its variance. At low RSNRs the estimators *Blocking ML root*, *Blocking ML Newton*, and *ML Newton* are the most robust as mainly their variance increases, but the mean stays close to 0 dB. *ML-EM* and *CDR* are very robust in terms of their variance in low RSNRs, but their mean increases, so they become biased to positive values.

Fig. 3 shows the dependency of the log error on the DDR while keeping the RSNR fixed at 15 dB. At first it may seem surprising that all coherence-based estimators show a large error at high positive DDRs. However, this can be explained by the fact that if the direct sound component dominates in the observed signal, it becomes difficult to accurately estimate the comparatively weak diffuse sound component.

Figs. 4 and 5 show the influence of DOA estimation errors for two DDRs (25 dB and 0 dB, respectively) at a RSNR of 15 dB. In this simulation, the steering vector $\mathbf{a}(k)$ used to compute the MWF (8) was computed with an angular offset, while the actual source position was kept constant. From these results it can be observed that an offset in the steered DOA increases the variance and the mean shifts to positive values (overestimation). At high DDR, the impact of steering errors is very prominent (see Fig. 4), whereas at DDR $= 0$ dB and lower, the DOA error influence is minor (Fig. 5). However, fortunately, at high DRRs the underlying assumptions of the typical signal models used for steering vector estimation are matched more accurately. This can mitigate the influence of steering vector estimation errors at high DRR in practice.

In this section we showed that the coherence-based PSD estimators work well at high RSNR, low DDR and low steering errors, but they exhibit weaknesses in low RSNR, high DDR and in the presence of steering errors. Note that a high DRR occurs when the T_{60} is small or when the source-array distance is small.

E. Evaluation of PSD Estimators for Late Reverberation

Fig. 6 shows the log error obtained using measured RIRs in the room with $T_{60} = 0.7$ s for varying RSNR. In contrast to Section VIII-D, this experiment includes all PSD estimators. The trends and the relative behavior between the estimators confirms the results from the controlled stationary experiment in Fig. 2, but the variances are much larger due to model mismatches. However, the coherence-based methods show a positive bias of their mean. The temporal model-based estimators *LRSV* and *CTF* show a different behavior and

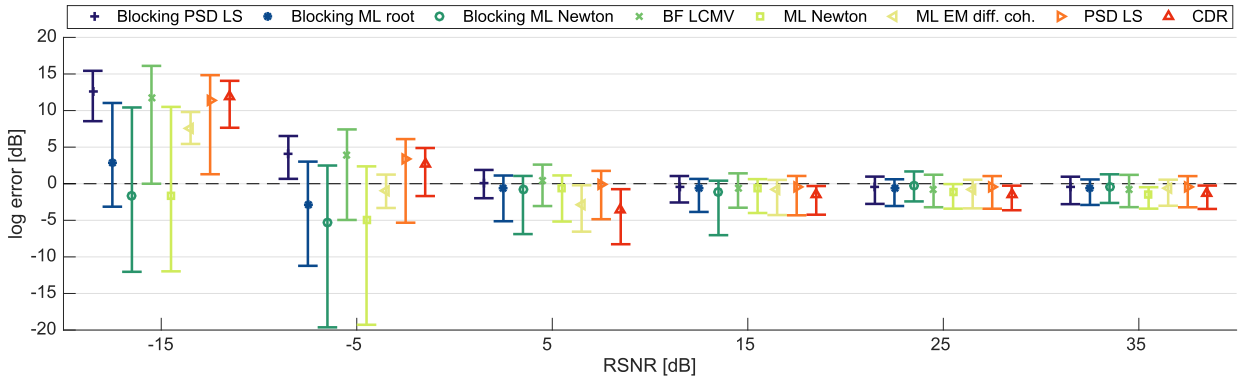

Fig. 2. Mean and standard deviation of log error for artificial stationary sound field with DDR = 10 dB.

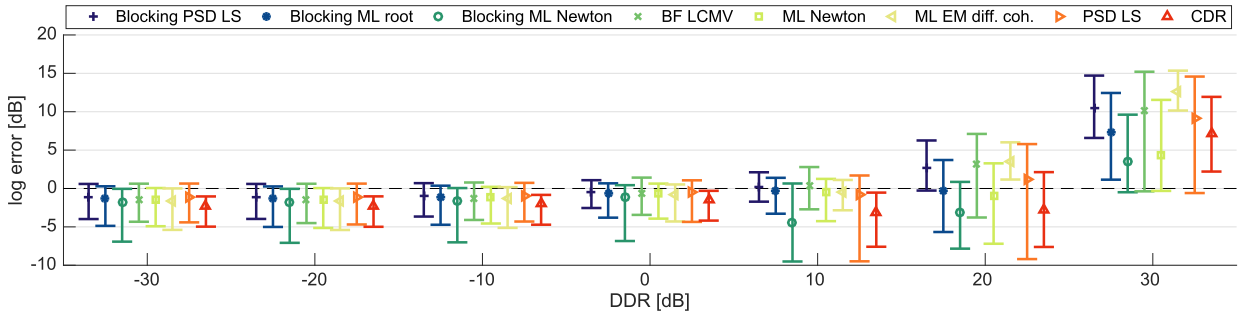

Fig. 3. Mean and standard deviation of log error for artificial stationary sound field with RSNR = 15 dB.

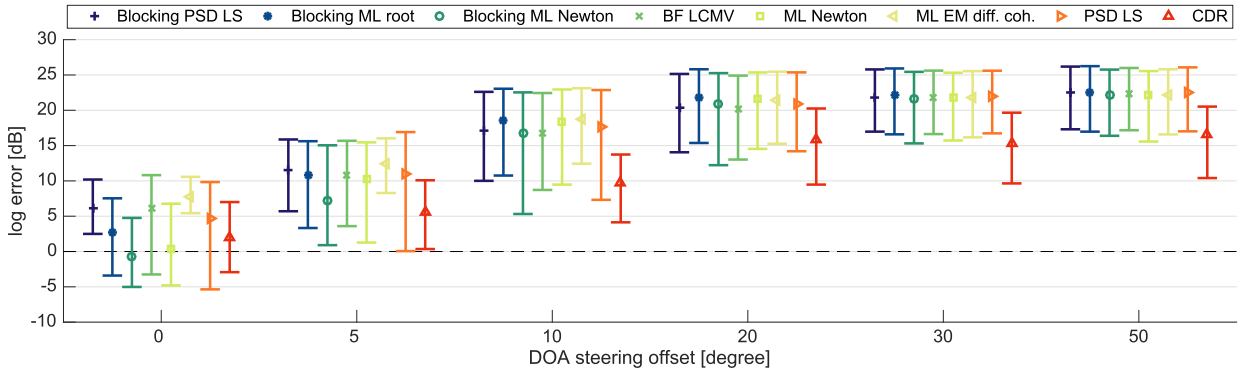

Fig. 4. Mean and standard deviation of log error for artificial stationary diffuse field for DOA offset with RSNR = 15 dB and DDR = 25 dB.

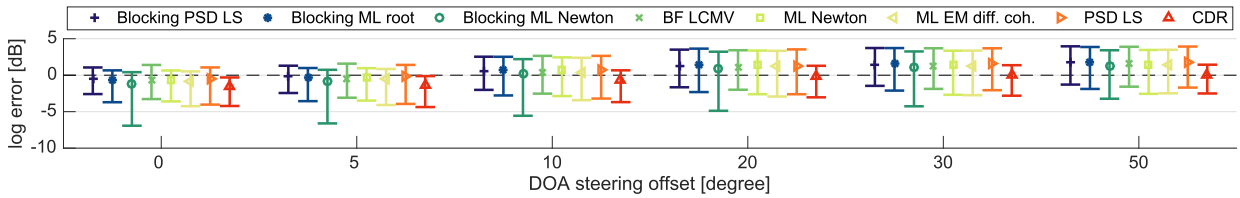

Fig. 5. Mean and standard deviation of log error for artificial stationary diffuse field for DOA offset with RSNR = 15 dB and DDR = 0 dB.

generally show a lower mean log error compared to the spatial coherence-based estimators. Therefore, the temporal model-based estimators yield less overestimation, which is beneficial in terms of speech distortion. The CTF-based estimator is among the most robust estimators at low RSNRs. The error of the *ML-EM with diffuse coherence* and the *ML-EM with estimated coherence* yield the same variance, but surprisingly, the mean of *ML-EM with diffuse coherence* is slightly closer to zero dB.

Fig. 7 shows the log error for varying T_{60} and source distances at RSNR of 15 dB. We can observe that the log error decreases towards higher reverberation times and larger source distances, i.e. for decreasing DRR. This confirms the trends from Fig. 3.

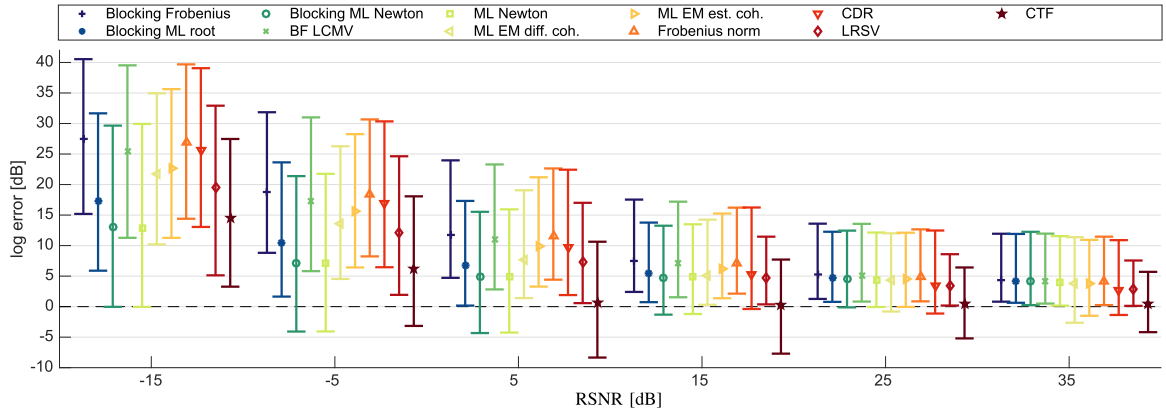

Fig. 6. Mean and standard deviation of log error for $T_{60} = 0.7$ s.

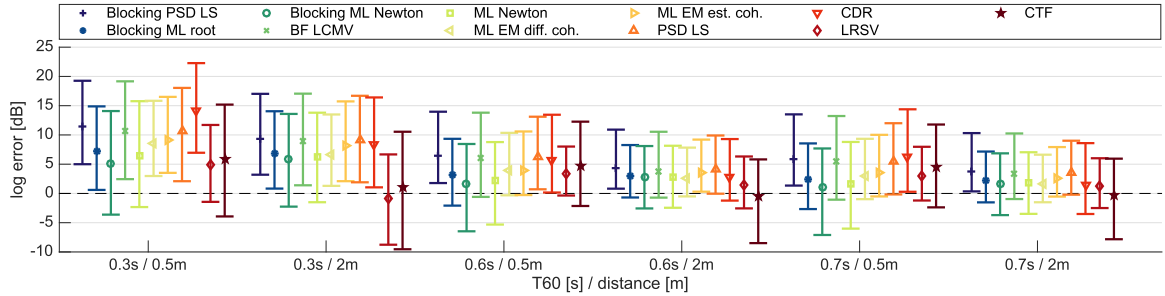

Fig. 7. Mean and standard deviation of log error in rooms with varying T_{60} and source distances and RSNR $= 15$ dB.

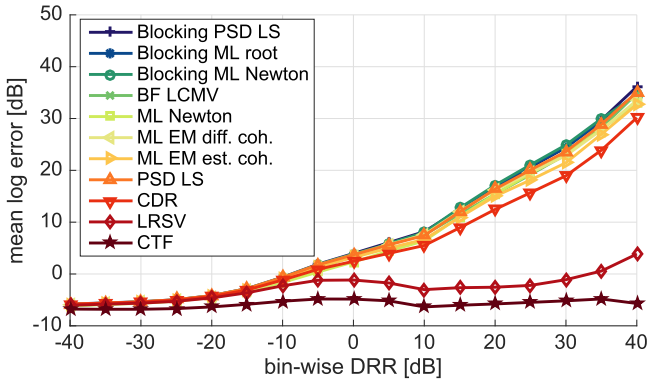

Fig. 8. Mean log error depending on the bin-wise DRR using measured RIRs.

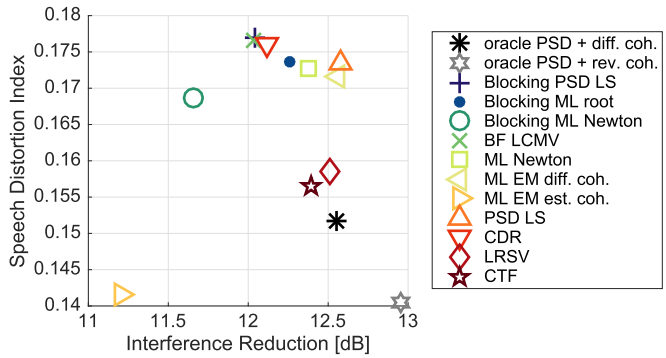

Fig. 9. Speech distortion vs. interference reduction for RSNR $= 15$ dB.

In Fig. 8, the mean log error is grouped depending on the true bin-wise (local) DRR in steps of 5 dB using the data from all acoustic conditions shown in Fig. 7. It is interesting to note that all coherence-based PSD estimators show a similar behavior in contrast to the temporal model-based PSD estimators LRSV and CTF. The latter two are more robust against overestimation at high DRRs, which we expect to result in less speech distortion.

The trends shown in the previous section for the coherence-based estimators can be confirmed when used to estimate the late reverberation PSD. The temporal model-based estimators yield more underestimation and less overestimation than the spatial coherence-based estimators. For the tested rooms, the coherence-based methods show a positive bias of the mean error compared to the temporal model based methods.

F. Performance of the Spatial Filter Using the Late Reverberation PSD Estimates

In this subsection we investigate the performance of the MWF using the various PSD estimates. As there are no significant differences between the spatial coherence-based estimators observable at higher RSNRs (c.f. Fig. 6), we present results here only for RSNR $= 15$ dB. In this experiment, the oracle late reverberation PSD is used either with the diffuse coherence (4) or with the oracle late reverberation coherence (64) to investigate the mismatch effect of the diffuse field model.

Fig. 9 shows the interference reduction (IR) vs. the speech distortion index (SDI) as computed by (67) and (68). The optimal point lies in the lower right corner. The best performance is obtained by using the *oracle PSD with oracle reverberation coherence*, which has a clear advantage over the *oracle PSD*

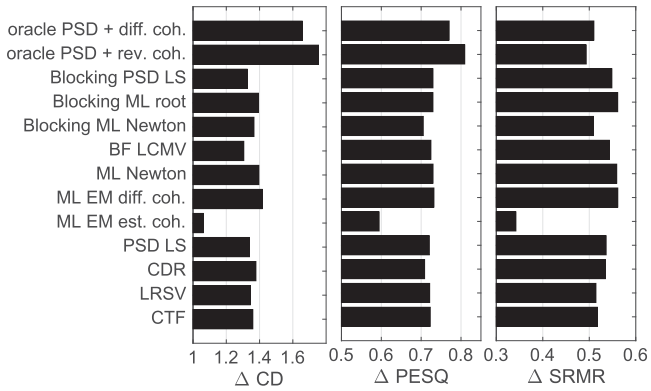

Fig. 10. Improvement of perceptual measures for RSNR = 15 dB.

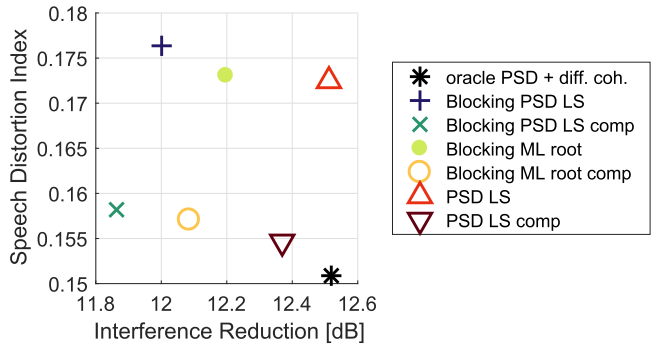

Fig. 11. Speech distortion vs. interference reduction without and with bias compensation, RSNR = 15 dB.

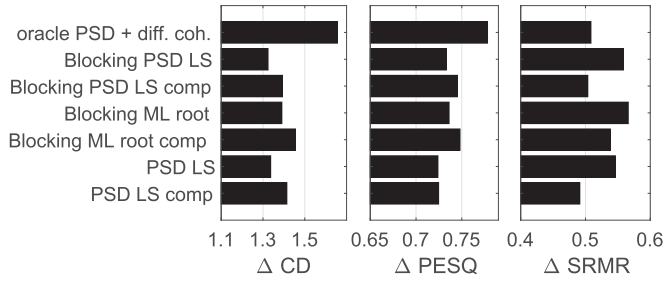

Fig. 12. Improvement of perceptual measures without and with bias compensation, RSNR = 15 dB.

with theoretical diffuse coherence matrix. The closest performing estimators to the oracle PSD are the temporal model-based methods *LRSV* and *CTF*. Among the coherence-based methods, the *PSD LS*, *ML-EM diff. coh.*, *ML Newton* and *Blocking ML root* perform slightly better than *Blocking PSD LS*, *BF LCMV* and *CDR*. The *Blocking ML Newton* has a lower SDI at the expense of less IR, while the *ML-EM est. coh.* surprisingly performs worse with a low IR.

The improvement of CD, PESQ and SRMR compared to the unprocessed reference microphone is shown in Fig. 10 (higher values are better). While the *oracle PSD with reverberation coherence* clearly achieves the best performance, the results for most estimators are rather close, except the *ML-EM with estimated coherence*, which clearly performs worse. This could be explained by the fact that the estimated coherence is not accurate enough, since the ML-EM using the theoretical coherence yields the best performance of all estimators in terms of CD, PESQ and SRMR.

Note that the estimators with the best values in Fig. 10 are not necessarily the best sounding ones as this judgement is highly subjective, and also the speech distortion shown in Fig. 9 plays a large role. Subjective listening to the processed signals confirms that some estimators produce very similar results, while others can sound very different[1] as represented in Fig. 10. Perceptual differences between the estimators are more prominent at lower RSNRs, while at RSNRs above 25 dB, perceptual differences between the coherence-based estimators become almost indistinguishable (see Fig. 6). The tradeoff between speech distortion and interference reduction (see Fig. 9) is clearly audible, which can be a guide on which estimator to choose depending on subjective preference and application. While Fig. 9 suggests that the temporal model estimators are superior, it has to be kept in mind that these rely on information about the reverberation time which is challenging to estimate in practice [68], while the coherence-based estimators rely on information about the DOA, which is commonly easier to estimate in practice. The best performing coherence-based estimator with low complexity is the *PSD LS*.

[1]Sound examples can be found online at www.audiolabs-erlangen.de/resources/2017-COMPARE-PSD-ESTIMATORS

G. Evaluation of Bias Compensation for Coherence-Based Reverberation PSD Estimators

In this section, we evaluate the bias compensation method for selected coherence-based estimators. The compensation function shown in Fig. 1 was trained using RIRs simulated by the image method [69], while for the following evaluation, the measured RIRs from the REVERB dataset and different speech and noise data were used.

The results for some selected coherence-based estimators without and with the proposed bias compensation function are shown in Figs. 11 and 12. We can see in Fig. 11 that for those estimators the bias compensation method proposed in Section VII significantly reduces the speech distortion, while sacrificing only a small amount of interference reduction. The perceptual measures in Fig. 12 show an improvement of CD and PESQ by using bias compensation, while the SRMR slightly suffers from bias compensation. Informal subjective listening confirms that the speech distortion can be reduced by the proposed bias compensation method.

IX. Conclusion

We reviewed and classified a variety of late reverberation PSD estimators that can be used for dereverberation. The majority of estimators is based on a spatial coherence model, but also estimators exploiting temporal models have been investigated. It was shown in extensive controlled and realistic experiments that differences between the spatial coherence-based estimators are rather small, where only a few estimators have limitations and achieve results below average. We showed that all spa-

tial coherence-based estimators under test suffer from the same issues, i.e., overestimation in high DDR and low RSNR conditions. Temporal model based estimators are less biased in high DRR, and mostly yield less speech distortion, but also less interference reduction. Furthermore, we proposed a method to compensate the systematic overestimation of the spatial coherence-based estimators in high DDR conditions, which greatly reduced the speech distortion. Using this bias compensation, similar results to the temporal model based estimators can be achieved using spatial coherence-based estimators with low complexity and without information about the room acoustics. Future work could be to develop a PSD estimator that combines spatial and temporal models.

REFERENCES

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 331–342, 2006.

[2] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[4] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951–1968, Oct. 1998.

[5] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[7] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1379–1390, Sep. 2014.

[8] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 680–693, Apr. 2016.

[9] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[10] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[11] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1369–1380, Jul. 2013.

[12] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[13] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[14] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Proc. IEEE 15th Int. Conf. Digit. Signal Process.*, Cardiff, U.K., Jul. 2007, pp. 607–610.

[15] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoust.*, vol. 87, pp. 359–366, 2001.

[16] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sep. 2009.

[17] X. Bao and J. Zhu, "An improved method for late-reverberant suppression based on statistical models," *Speech Commun.*, vol. 55, no. 9, pp. 932–940, Oct. 2013.

[18] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[19] C. Marro, Y. Mahieux, and K. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.

[20] E. A. P. Habets, "Multi-microphone spectral enhancement," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. New York, NY, USA: Springer, 2010.

[21] S. Braun and E. A. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 34, pp. 1–14, Dec. 2015.

[22] B. Cauchi *et al.*, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 61, p. 1–12, 2015.

[23] O. Schwartz, S. Gannot, and E. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 240–251, Jan. 2015.

[24] A. Kuklasinski, S. Doclo, S. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[25] E. A. P. Habets and S. Gannot, "Dual-microphone speech dereverberation using a reference signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, USA, Apr. 2007, vol. IV, pp. 901–904.

[26] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 669–673.

[27] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc. Eur. Signal Process. Conf.*, Barcelona, Spain, 2011, pp. 1347–1351.

[28] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "Diffuseness estimation with high temporal resolution via spatial coherence between virtual first-order microphones," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2011, pp. 217–220.

[29] A. Schwarz, K. Reindl, and W. Kellermann, "A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dec. 2012, pp. 113–116.

[30] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[31] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Commun.*, vol. 49, no. 7–8, pp. 657–666, Jul. 2007.

[32] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 1–5.

[33] K. Reindl *et al.*, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 726–745, 2013.

[34] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using maxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[35] O. Schwartz, S. Braun, S. Gannot, and E. Habets, "Maximum likelihood estimation of the late reverberant power spectral density in noisy environments," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New York, NY, USA, Oct. 2015, pp. 1–5.

[36] O. Thiergart, M. Taseska, and E. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[37] O. Thiergart and E. Habets, "Extracting reverberant sound using a linearly constrained minimum variance spatial filter," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 630–634, Mar. 2014.

[38] A. Kuklasinski, S. Doclo, S. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. Eur. Signal Process. Conf.*, Lisbon, Portugal, Sep. 2014, pp. 61–65.

[39] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 151–155.

[40] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multi-microphone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, Sep. 2016.
[41] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm," in *Proc. Eur. Signal Process. Conf.*, Aug. 2016, pp. 1123–1127.
[42] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.
[43] S. Braun, B. Schwartz, S. Gannot, and E. A. P. Habets, "Late reverberation PSD estimation for single-channel dereverberation using relative convolutive transfer functions," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Xi'an, China, Sep. 2016, pp. 1–5.
[44] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, Nov. 1962.
[45] 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. Amer.*, vol. 132, no. 4, pp. 2337–2346, 2012.
[46] M. Jeub, M. Dorbecker, and P. Vary, "A semi-analytical model for the binaural coherence of noise fields," *IEEE Signal Process. Lett.*, vol. 18, no. 3, pp. 197–200, Mar. 2011.
[47] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.
[48] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[49] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[50] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.
[51] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[52] F. Heese and P. Vary, "Noise PSD estimation by logarithmic baseline tracing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 4405–4409.
[53] Z. Chen, G. K. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. London, U.K.: Artech House, 2010.
[54] T. E. Tuncer and B. Friedlander, Eds.,*Classical and Modern Direction-of-Arrival Estimation*. Burlington, VT, USA: Academic, 2009.
[55] S. Markovich-Golan, S. Gannot, and I. Cohen, "A sparse blocking matrix for multiple constraints GSC beamformer," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 197–200.
[56] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[57] 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. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 309–312.
[58] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.
[59] J. Eaton, M. Brookes, and P. A. Naylor, "A comparison of non-intrusive SNR estimation algorithms and the use of mapping functions," in *Proc. Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.
[60] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 7, pp. 1–19, Jan. 2016.
[61] E. B. Union, Sound quality assessment material recordings for subjective tests, 1988. [Online]. Available: http://tech.ebu.ch/publications/sqamcd
[62] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.
[63] A. Hald, *Statistical Theory With Engineering Applications*, 1st ed., Hoboken, NJ, USA: Wiley, 1952.
[64] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 242–248, Feb. 1988.
[65] 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 Recommendation P.862, Feb. 2001.
[66] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Antibes, France, Sep. 2014, pp. 55–59.
[67] M. S. Brandstein and D. B. Ward, Eds.,*Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.
[68] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.
[69] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

Sebastian Braun received the M.Sc. degree in electrical engineering and sound engineering from the University of Music and Dramatic Arts Graz and the Technical University Graz, Graz, Austria, in 2012. He then joined the International Audio Laboratories Erlangen (a joint institution of Friedrich-Alexander Universität Erlangen-Nürnberg and Fraunhofer IIS) as a Ph.D. candidate in the field of acoustic signal processing. His research interests include spatial audio processing, spatial filtering, speech enhancement (dereverberation, noise reduction, echo cancelation, feedback cancelation, and automatic gain control), adaptive filtering, and binaural processing techniques.

Adam Kuklasiński received the M.Sc. degree in acoustics from Adam Mickiewicz University, Poznań, Poland, and the Ph.D. degree in digital signal processing from Aalborg University, Aalborg, Denmark, in 2012 and 2016, respectively. During his Ph.D study, he was a Marie Skłodowska-Curie Fellow in the ITN-DREAMS project. He is currently an Audio Research Engineer with Zylia, Poznań, Poland. His scientific interests include statistical signal processing, speech dereverberation, and binaural cue preservation in hearing aids.

Ofer Schwartz received the B.Sc. (*cum laude*) and M.Sc. degrees in electrical engineering from Bar-Ilan University, Ramat Gan, Israel, in 2010 and 2013, respectively. He is currently working toward the Ph.D. degree in electrical engineering at the Speech and Signal Processing Laboratory, Faculty of Engineering, Bar-Ilan University. In 2017, he joined the Audio Department, CEVA-DSP, Herzelia, Israel, as a Senior Algorithm Researcher and Developer. His research interests include statistical signal processing and in particular noise reduction and dereverberation using microphone arrays and speaker localization and tracking.

Oliver Thiergart received the Dipl.-Ing. (M.Sc.) degree in mediatechnology from Ilmenau University of Technology, Ilmenau, Germany, in 2008, and the Ph.D. degree in parametric spatial sound processing from the International Audio Laboratories Erlangen, Erlangen, Germany, in 2015. He then joined the Audio Department, Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany. In 2011, he became a Member of the International Audio Laboratories Erlangen. He is currently a Researcher with the Fraunhofer Institute for Integrated Circuits. His research interests include parametric spatial sound processing, microphone arrays, spatial filtering, and parameter estimation.

Emanuël A. P. Habets (S'02–M'07–SM'11) received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, University of Technology, Eindhoven, The Netherlands, in 2002 and 2007, respectively. He is an Associate Professor with the International Audio Laboratories Erlangen (a joint institution of Friedrich-Alexander Universität Erlangen-Nürnberg and Fraunhofer Institute for Integrated Circuits (IIS)), and the Head of the Spatial Audio Research Group, Fraunhofer IIS, Erlangen, Germany.

From 2007 to 2009, he was a Postdoctoral Fellow with the Technion—Israel Institute of Technology and with the Bar-Ilan University, Ramat Gan, Israel. From 2009 to 2010, he was a Research Fellow with the Communication and Signal Processing Group, Imperial College London, U.K. His research activities center around audio and acoustic signal processing, and include spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, and echo reduction), and sound localization and tracking.

Dr. Habets is currently a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the Vice-Chair of the EURASIP Special Area Team on Acoustic, Sound and Music Signal Processing, and the Editor-in-Chief of the *EURASIP Journal on Audio, Speech, and Music Processing.* He was a Member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control, Eindhoven, The Netherlands, the General Co-Chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, and the General Co-Chair of the 2014 International Conference on Spatial Audio, Erlangen, Germany. He was a Member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013–2015), a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and the EURASIP JOURNAL ON ADVANCES IN SIGNAL PROCESSING, and an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (2013–2017). He was the recipient, with S. Gannot and I. Cohen, of the 2014 IEEE Signal Processing Letters Best Paper Award.

Simon Doclo (S'95–M'03–SM'13) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders, Electrical Engineering Department, Katholieke Universiteit Leuven, and the Cognitive Systems Laboratory, McMaster University, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group, Leuven, Belgium. Since 2009, he is has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and a Scientific Advisor for the project group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, active noise control, acoustic sensor networks and hearing aid processing. He is a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and the EAA Technical Committee on Audio Signal Processing. He was a Guest Editor for several special issues (the IEEE SIGNAL PROCESSING MAGAZINE, *Elsevier Signal Processing*) and currently is an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*. He was the recipient of the Master Thesis Award of the Royal Flemish Society of Engineers in 1997 (with Erik De Clippel), the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with Marc Moonen), and the IEEE Signal Processing Society 2008 Best Paper Award (with J. Chen, J. Benesty, and A. Huang).

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (*summa cum laude*) from the Technion—Israel Institute of Technology, Haifa, Israel, in 1986, and the M.Sc. (*cum laude*) and Ph.D. degrees from Tel-Aviv University, Tel Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he held a Postdoctoral position with the Department of Electrical Engineering, KU Leuven, Leuven, Belgium. From 2002 to 2003, he held a Research and Teaching position with the Faculty of Electrical Engineering, Technion–Israel Institute of Technology. He is currently, a Full Professor with the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, where he is heading the Speech and Signal Processing Laboratory and the Signal Processing Track. His research interests include multimicrophone speech processing and specifically distributed algorithms for *ad hoc* microphone arrays for noise reduction and speaker separation, dereverberation, single microphone speech enhancement, and speaker localization and tracking. He was an Associate Editor for the *EURASIP Journal of Advances in Signal Processing* during 2003–2012, and an Editor for several special issues on multimicrophone speech processing of the same journal. He was a Guest Editor for the Elsevier *Speech Communication* and *Signal Processing* journals. He was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2009–2013, and the Area Chair for the same journal during 2013–2017. He is currently a Moderator for arXiv in the field of audio and speech processing. He is also a Reviewer for many IEEE journals and conferences. Since January 2010, he has been a Member of the Audio and Acoustic Signal Processing technical committee of the IEEE. Since January 2017, he has been the Committee Chair. Since 2005, he has been a Member of the technical and steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) and was the General Co-Chair of the IWAENC held in Tel-Aviv, Israel in August 2010. He was the General Co-Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in October 2013. He was selected (with colleagues) to present tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013, and EUSIPCO 2013 and was a keynote speaker for IWAENC 2012 and LVA/ICA 2017. He was the recipient of the Bar-Ilan University outstanding lecturer award in 2010 and 2014. He is also a corecipient of ten best paper awards.

Jesper Jensen is a Senior Scientist with Oticon A/S, Smørum, Denmark, where he is responsible for scouting and development of signal processing concepts for hearing instruments. He is also a Professor with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark. He is also a cohead of the Centre for Acoustic Signal Processing Research, Aalborg University. His research interests include acoustic signal processing, including signal retrieval from noisy observations, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing. His work on speech intelligibility prediction received the 2017 IEEE Signal Processing Society's best paper award.