JOINT LATE REVERBERATION AND NOISE POWER SPECTRAL DENSITY ESTIMATION IN A SPATIALLY HOMOGENEOUS NOISE FIELD

Ina Kodrasi[⋆†]*, Simon Doclo*[⋆]

[⋆]University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All, Oldenburg, Germany
[†]Idiap Research Institute, Speech and Audio Processing Group, Martigny, Switzerland
{ina.kodrasi,simon.doclo}@uni-oldenburg.de

ABSTRACT

Many multi-channel dereverberation and noise reduction techniques such as the multi-channel Wiener filter (MWF) require an estimate of the late reverberation and noise power spectral densities (PSDs). State-of-the-art multi-channel methods for estimating the late reverberation PSD typically assume that the noise PSD matrix is known. Instead of assuming that the noise PSD matrix is known, in this paper we model the noise as a spatially homogeneous sound field with an unknown time-varying PSD and a known time-invariant spatial coherence matrix. Based on this model, two joint estimators of the late reverberation and noise PSDs are proposed, i.e., a non-blocking-based estimator which simultaneously estimates the target signal, late reverberation, and noise PSDs, and a blocking-based estimator which first estimates the late reverberation and noise PSDs at the output of a blocking matrix aiming to block the target signal. Experimental results show that the proposed blocking-based estimator yields the best performance when used in an MWF, even resulting in a similar or better performance than a state-of-the-art blocking-based estimator of the late reverberation PSD which assumes that the noise PSD matrix is known.

Index Terms— PSD estimation, late reverberation, noise, MWF, least-squares

1. INTRODUCTION

In many hands-free speech communication applications, the recorded microphone signals do not only contain the desired speech signal, but also attenuated and delayed copies of the desired speech signal due to reverberation, as well as additive noise. While early reverberation may be desirable [1], late reverberation and noise may degrade the perceived quality and hinder the intelligibility of speech [2, 3]. Hence, effective dereverberation and noise reduction techniques are required.

A commonly used dereverberation and noise reduction technique is the multi-channel Wiener filter (MWF), which aims at minimizing the mean-square error between the output signal and the target signal [4–6]. The implementation of the MWF requires (among other parameters) an estimate of the late reverberation and noise power spectral densities (PSDs). To estimate the late reverberation PSD, several single-channel estimators based on a temporal model of reverberation [7–9] as well as multi-channel estimators based on a diffuse sound field model for the late reverberation [10–18] have been proposed. To the best of our knowledge, state-of-the-art multi-channel late reverberation PSD estimators estimate the late reverberation PSD assuming that an estimate of the noise PSD matrix is available. The noise PSD matrix is typically estimated from the microphone signals during speech pauses detected by means of a voice activity detector (VAD) [19, 20], generally requiring the noise PSD to be rather time-invariant. However, in many acoustic scenarios, e.g., in highly reverberant environments, speech pauses may rarely occur, making the estimation of the noise PSD matrix challenging. In addition, in many acoustic scenarios the noise PSD can be time-varying, e.g., when the noise consists of microphone self-noise in a system with the input gain automatically adjusted during operation using an automatic gain control.

Instead of assuming that an estimate of the noise PSD matrix is available, in this paper we model the noise as a spatially homogeneous sound field with a time-varying PSD and assume that only knowledge of the time-invariant spatial coherence matrix is available. Two alternative joint estimators of the late reverberation and noise PSDs are proposed, i.e., a non-blocking-based estimator which simultaneously estimates the target signal, late reverberation, and noise PSDs, and a blocking-based estimator which first estimates only the late reverberation and noise PSDs at the output of a blocking matrix aiming to block the target signal. The proposed PSD estimators can be viewed as extensions of the PSD estimators in [10] and [16], where only the target signal and late reverberation PSDs are estimated assuming that an estimate of the noise PSD matrix is available. Simulation results for several realistic acoustic scenarios show that the proposed blocking-based PSD estimator yields the best performance when used in an MWF, also yielding a similar or better performance than the PSD estimator in [10] which assumes that the noise PSD matrix is known.

2. SIGNAL MODEL AND ASSUMPTIONS

Consider a reverberant and noisy multi-channel acoustic system with a single speech source and M microphones. In the short-time Fourier transform (STFT) domain, the M-dimensional vector of the received microphone signals $\mathbf{y}(k,l) = [Y_1(k,l)\ \dots\ Y_M(k,l)]^T$ at frequency bin k and frame index l is given by

$$\mathbf{y}(k,l) = \underbrace{\mathbf{x}_{\mathrm{e}}(k,l) + \mathbf{x}_{\mathrm{r}}(k,l)}_{\mathbf{x}(k,l)} + \mathbf{v}(k,l), \tag{1}$$

with $\mathbf{x}_{\mathrm{e}}(k,l)$ the direct and early reverberation component, $\mathbf{x}_{\mathrm{r}}(k,l)$ the late reverberation component, $\mathbf{x}(k,l)$ the reverberant component, and $\mathbf{v}(k,l)$ the noise component. The vectors $\mathbf{x}_{\mathrm{e}}(k,l)$, $\mathbf{x}_{\mathrm{r}}(k,l)$, $\mathbf{x}(k,l)$, and $\mathbf{v}(k,l)$ are defined similarly as $\mathbf{y}(k,l)$. The direct and early reverberation component $\mathbf{x}_{\mathrm{e}}(k,l)$ can be expressed as

$$\mathbf{x}_{\mathrm{e}}(k,l) = S(k,l)\mathbf{d}(k), \tag{2}$$

with $S(k,l)$ the target signal, i.e., the direct and early reverberation component received at the reference microphone, and $\mathbf{d}(k) =$

This work was supported by the Cluster of Excellence Hearing4All, funded by the German Research Foundation (DFG), and the joint Lower Saxony-Israeli Project ATHENA, funded by the State of Lower Saxony.

$[D_1(k)\ \ldots\ D_M(k)]^T$ the M-dimensional vector of relative transfer functions (RTFs) of the target signal between the reference microphone and all microphones. The target signal $S(k,l)$ is often defined as the direct component only, such that the RTF vector $\mathbf{d}(k)$ only depends on the direction of arrival (DOA) of the speech source and the microphone array geometry [10–14, 16]. For conciseness, the frequency index k is omitted in the remainder of this paper.

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

$$\mathbf{\Phi}_{\mathbf{y}}(l) = \mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} = \underbrace{\mathbf{\Phi}_{\mathbf{x}_\mathrm{e}}(l) + \mathbf{\Phi}_{\mathbf{x}_\mathrm{r}}(l)}_{\mathbf{\Phi}_{\mathbf{x}}(l)} + \mathbf{\Phi}_{\mathbf{v}}(l), \quad (3)$$

where $\mathcal{E}$ denotes the expectation operator, $\mathbf{\Phi}_{\mathbf{x}_\mathrm{e}}(l)$ is the direct and early reverberation PSD matrix, $\mathbf{\Phi}_{\mathbf{x}_\mathrm{r}}(l)$ is the late reverberation PSD matrix, $\mathbf{\Phi}_{\mathbf{x}}(l)$ is the reverberant PSD matrix, and $\mathbf{\Phi}_{\mathbf{v}}(l)$ is the noise PSD matrix. The PSD matrix $\mathbf{\Phi}_{\mathbf{x}_\mathrm{e}}(l)$ can be expressed as (cf. (2))

$$\mathbf{\Phi}_{\mathbf{x}_\mathrm{e}}(l) = \Phi_\mathrm{s}(l)\mathbf{d}\mathbf{d}^H, \quad (4)$$

with $\Phi_\mathrm{s}(l)$ the time-varying PSD of the target signal, i.e., $\Phi_\mathrm{s}(l) = \mathcal{E}\{|S(l)|^2\}$. Modeling the late reverberation as a diffuse sound field [10–18], the PSD matrix $\mathbf{\Phi}_{\mathbf{x}_\mathrm{r}}(l)$ can be expressed as

$$\mathbf{\Phi}_{\mathbf{x}_\mathrm{r}}(l) = \Phi_\mathrm{r}(l)\mathbf{\Gamma}, \quad (5)$$

with $\Phi_\mathrm{r}(l)$ the time-varying PSD of the late reverberation and $\mathbf{\Gamma}$ the spatial coherence matrix of a diffuse sound field, which can be analytically computed based on the microphone array geometry [21]. Modeling the additive noise as a spatially homogeneous sound field, the noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(l)$ can be expressed as

$$\mathbf{\Phi}_{\mathbf{v}}(l) = \Phi_\mathrm{v}(l)\mathbf{\Psi}, \quad (6)$$

with $\Phi_\mathrm{v}(l)$ the time-varying noise PSD and $\mathbf{\Psi}$ the spatial coherence matrix of the noise, which is assumed to be time-invariant. In the presence of spatially uncorrelated noise (e.g., microphone self-noise), $\mathbf{\Psi} = \mathbf{I}$, with $\mathbf{I}$ the $M \times M$-dimensional identity matrix. Using (4), (5), and (6), the PSD matrix $\mathbf{\Phi}_{\mathbf{y}}(l)$ is equal to

$$\mathbf{\Phi}_{\mathbf{y}}(l) = \Phi_\mathrm{s}(l)\mathbf{d}\mathbf{d}^H + \Phi_\mathrm{r}(l)\mathbf{\Gamma} + \Phi_\mathrm{v}(l)\mathbf{\Psi}. \quad (7)$$

Given the filter vector $\mathbf{w}(l) = [W_1(l)\ \ldots\ W_M(l)]^T$, the output signal $Z(l)$ of the speech enhancement system is equal to the sum of the filtered microphone signals, i.e., $Z(l) = \mathbf{w}^H(l)\mathbf{y}(l)$. Dereverberation and noise reduction techniques aim at designing the filter $\mathbf{w}(l)$ such that the output signal $Z(l)$ is as close as possible to the target signal $S(l)$. A widely used dereverberation and noise reduction technique is the MWF, which aims at minimizing the mean-square error between $Z(l)$ and $S(l)$ [4–6]. The MWF is typically implemented as a minimum variance distortionless response (MVDR) beamformer $\mathbf{w}_{\text{MVDR}}(l)$ followed by a single-channel Wiener postfilter $G(l)$ [10, 12–18], i.e.,

$$\mathbf{w}_{\text{MWF}}(l) = \underbrace{\frac{[\hat{\Phi}_\mathrm{r}(l)\mathbf{\Gamma} + \hat{\Phi}_\mathrm{v}(l)\mathbf{\Psi}]^{-1}\mathbf{d}}{\mathbf{d}^H[\hat{\Phi}_\mathrm{r}(l)\mathbf{\Gamma} + \hat{\Phi}_\mathrm{v}(l)\mathbf{\Psi}]^{-1}\mathbf{d}}}_{\mathbf{w}_{\text{MVDR}}(l)} \underbrace{\frac{\hat{\rho}(l)}{1+\hat{\rho}(l)}}_{G(l)}, \quad (8)$$

with $\hat{\Phi}_\mathrm{r}(l)$ and $\hat{\Phi}_\mathrm{v}(l)$ denoting the estimated late reverberation and noise PSDs respectively and $\hat{\rho}(l)$ denoting the estimated target-to-late reverberation and noise ratio (TRNR) at the output of the MVDR beamformer. The TRNR can be estimated as

$$\hat{\rho}(l) = \frac{\hat{\Phi}_\mathrm{s}(l)}{\hat{\Phi}_{\mathrm{rn}}(l)}, \quad (9)$$

with $\hat{\Phi}_\mathrm{s}(l)$ denoting the estimated target signal PSD and $\hat{\Phi}_{\mathrm{rn}}(l) = \{\mathbf{d}^H[\hat{\Phi}_\mathrm{r}(l)\mathbf{\Gamma} + \hat{\Phi}_\mathrm{v}(l)\mathbf{\Psi}]^{-1}\mathbf{d}\}^{-1}$ the estimated residual late reverberation and noise PSD at the output of the MVDR beamformer. Alternatively, $\hat{\rho}(l)$ can be estimated using the decision directed approach as [16, 22]

$$\hat{\rho}_{\text{DD}}(l) = \beta\frac{|Z(l-1)|^2}{\hat{\Phi}_{\mathrm{rn}}(l-1)} + (1-\beta)\frac{\hat{\Phi}_\mathrm{s}(l)}{\hat{\Phi}_{\mathrm{rn}}(l)}, \quad (10)$$

with β a smoothing parameter. As can be observed in (8), (9), and (10), the implementation of the MWF requires estimates of the time-varying target signal, late reverberation, and noise PSDs. The objective of this paper is to derive estimates $\hat{\Phi}_\mathrm{s}(l)$, $\hat{\Phi}_\mathrm{r}(l)$, and $\hat{\Phi}_\mathrm{v}(l)$, assuming that the RTF vector $\mathbf{d}$, the diffuse spatial coherence matrix $\mathbf{\Gamma}$, and the noise spatial coherence matrix $\mathbf{\Psi}$ are known. The RTF vector can be constructed based on a DOA estimate, the diffuse spatial coherence matrix can be constructed based on the microphone array geometry, and the noise spatial coherence matrix can be constructed assuming a reasonable sound field model for the noise.

3. JOINT TARGET SIGNAL, LATE REVERBERATION, AND NOISE PSD ESTIMATORS

To the best of our knowledge, state-of-the-art multi-channel PSD estimators do not explicitly model the noise as a spatially homogeneous sound field and only derive target signal and late reverberation PSDs estimates $\hat{\Phi}_\mathrm{s}(l)$ and $\hat{\Phi}_\mathrm{r}(l)$ assuming that an estimate of the noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(l)$ is available [10–18]. The noise PSD matrix is typically estimated from the microphone signals during speech pauses detected by means of a VAD [19, 20], generally requiring the noise PSD $\Phi_\mathrm{v}(l)$ to be time-invariant. Instead of assuming that an estimate of the noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(l)$ is available, in this paper we assume that only knowledge of the noise spatial coherence matrix $\mathbf{\Psi}$ is available and propose a non-blocking-based and a blocking-based estimator of the target signal PSD $\Phi_\mathrm{s}(l)$, the late reverberation PSD $\Phi_\mathrm{v}(l)$, and the noise PSD $\Phi_\mathrm{v}(l)$. The proposed PSD estimators can be viewed as extensions of the PSD estimators in [10] and [16], where only estimates of $\Phi_\mathrm{s}(l)$ and $\Phi_\mathrm{r}(l)$ are derived assuming that the noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(l)$ is known.

3.1. Non-blocking-based PSD estimator

In the following we propose to simultaneously estimate the target signal, late reverberation, and noise PSDs using the signal model in (7) and an estimate of the PSD matrix $\mathbf{\Phi}_{\mathbf{y}}(l)$. An estimate of $\mathbf{\Phi}_{\mathbf{y}}(l)$ can be directly obtained from the microphone signals using recursive averaging as

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(l) = \alpha\mathbf{y}(l)\mathbf{y}^H(l) + (1-\alpha)\hat{\mathbf{\Phi}}_{\mathbf{y}}(l-1), \quad (11)$$

with α a smoothing factor. Matching (11) to (7) and since the matrices $\mathbf{d}\mathbf{d}^H$, $\mathbf{\Gamma}$, and $\mathbf{\Psi}$ are known, a system of $M(M+1)/2$ equations with three unknowns $\Phi_\mathrm{s}(l)$, $\Phi_\mathrm{r}(l)$, and $\Phi_\mathrm{v}(l)$ arises[1]. For $M \geq 3$, the system of equations is overdetermined and an estimate of the unknown PSDs $\Phi_\mathrm{s}(l)$, $\Phi_\mathrm{r}(l)$, and $\Phi_\mathrm{v}(l)$ can be obtained by minimizing the least-squares cost function[2]

$$J_\mathrm{n}(l) = \|\hat{\mathbf{\Phi}}_{\mathbf{y}}(l) - \Phi_\mathrm{s}(l)\mathbf{d}\mathbf{d}^H - \Phi_\mathrm{r}(l)\mathbf{\Gamma} - \Phi_\mathrm{v}(l)\mathbf{\Psi}\|_F^2, \quad (12)$$

[1] Note that since the matrices $\hat{\mathbf{\Phi}}_{\mathbf{y}}(l)$, $\mathbf{d}\mathbf{d}^H$, $\mathbf{\Gamma}$, and $\mathbf{\Psi}$ are symmetric, matching (11) to (7) yields $M(M+1)/2$ equations instead of M^2 equations.

[2] Note that this non-blocking-based least-squares cost function has already been used in [23] in the context of noise reduction only, in order to estimate the PSDs of different spatially homogeneous noise fields.

with $\|\cdot\|_F$ the matrix Frobenius norm. Setting the derivative of (12) with respect to $\Phi_{\mathrm{s}}(l)$, $\Phi_{\mathrm{r}}(l)$, and $\Phi_{\mathrm{v}}(l)$ to 0 results in a system of equations which can be written as

$$\underbrace{\begin{bmatrix} (\mathbf{d}^H\mathbf{d})^2 & \mathbf{d}^H\mathbf{\Gamma}\mathbf{d} & \mathbf{d}^H\mathbf{\Psi}\mathbf{d} \\ \mathbf{d}^H\mathbf{\Gamma}\mathbf{d} & \mathrm{tr}\{\mathbf{\Gamma}^H\mathbf{\Gamma}\} & \mathrm{tr}\{\mathbf{\Gamma}^H\mathbf{\Psi}\} \\ \mathbf{d}^H\mathbf{\Psi}\mathbf{d} & \mathrm{tr}\{\mathbf{\Gamma}^H\mathbf{\Psi}\} & \mathrm{tr}\{\mathbf{\Psi}^H\mathbf{\Psi}\} \end{bmatrix}}_{\mathbf{A}_{\mathrm{n}}} \underbrace{\begin{bmatrix} \hat{\Phi}_{\mathrm{s,n}}(l) \\ \hat{\Phi}_{\mathrm{r,n}}(l) \\ \hat{\Phi}_{\mathrm{v,n}}(l) \end{bmatrix}}_{\hat{\boldsymbol{\phi}}_{\mathrm{n}}(l)} = \underbrace{\begin{bmatrix} \mathbf{d}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(l)\mathbf{d} \\ \mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{y}}^H(l)\mathbf{\Gamma}\} \\ \mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{y}}^H(l)\mathbf{\Psi}\} \end{bmatrix}}_{\mathbf{p}_{\mathrm{n}}(l)}, \tag{13}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace operator and the quantities $\mathbf{A}_{\mathrm{n}}$, $\hat{\boldsymbol{\phi}}_{\mathrm{n}}(l)$, and $\mathbf{p}_{\mathrm{n}}(l)$ have been introduced in order to simplify the notation. The solution to (13) is given by

$$\hat{\boldsymbol{\phi}}_{\mathrm{n}}(l) = \mathbf{A}_{\mathrm{n}}^{-1}\mathbf{p}_{\mathrm{n}}(l), \tag{14}$$

with the proposed target signal PSD estimate $\hat{\Phi}_{\mathrm{s,n}}(l)$ being the first element of $\hat{\boldsymbol{\phi}}_{\mathrm{n}}(l)$, late reverberation PSD estimate $\hat{\Phi}_{\mathrm{r,n}}(l)$ being the second element of $\hat{\boldsymbol{\phi}}_{\mathrm{n}}(l)$, and noise PSD estimate $\hat{\Phi}_{\mathrm{v,n}}(l)$ being the third element of $\hat{\boldsymbol{\phi}}_{\mathrm{n}}(l)$.

3.2. Blocking-based PSD estimator

In the following we propose an alternative PSD estimator which first estimates the late reverberation and noise PSDs using reference signals at the output of a blocking matrix aiming to block the target signal. Based on the estimated late reverberation and noise PSDs, the target signal PSD is then estimated in a second step.

In order to block the target signal, an $M \times (M-1)$-dimensional blocking matrix $\mathbf{B}$ is constructed such that

$$\mathbf{B}^H\mathbf{d} = \mathbf{0}, \tag{15}$$

and a set of $M-1$ reference signals $\tilde{\mathbf{u}}(l)$ containing only late reverberation and noise is generated as $\tilde{\mathbf{u}}(l) = \mathbf{B}^H\mathbf{y}(l)$. There exist many blocking matrices which satisfy (15). In this paper, the blocking matrix is computed from the first $M-1$ columns of the matrix $\mathbf{T}$ defined as

$$\mathbf{T} = \mathbf{I} - \frac{\mathbf{d}\mathbf{d}^H}{\|\mathbf{d}\|_2^2}. \tag{16}$$

Based on (7) and (15), the PSD matrix of the reference signals at the blocking matrix output can be expressed as

$$\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l) = \mathcal{E}\{\tilde{\mathbf{u}}(l)\tilde{\mathbf{u}}^H(l)\} = \Phi_{\mathrm{r}}(l)\underbrace{\mathbf{B}^H\mathbf{\Gamma}\mathbf{B}}_{\tilde{\mathbf{\Gamma}}} + \Phi_{\mathrm{v}}(l)\underbrace{\mathbf{B}^H\mathbf{\Psi}\mathbf{B}}_{\tilde{\mathbf{\Psi}}}. \tag{17}$$

The matrices $\tilde{\mathbf{\Gamma}}$ and $\tilde{\mathbf{\Psi}}$ can be computed using the known spatial coherence matrices $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ and an estimate $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}(l)$ of the PSD matrix $\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l)$ can be directly obtained from the reference signals similarly to (11). Matching the estimated PSD matrix $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}(l)$ to (17) gives rise to a system of $M(M-1)/2$ equations with two unknowns $\Phi_{\mathrm{r}}(l)$ and $\Phi_{\mathrm{v}}(l)$[3]. For $M \geq 3$, the system of equations is overdetermined and an estimate of $\Phi_{\mathrm{r}}(l)$ and $\Phi_{\mathrm{v}}(l)$ can be obtained by minimizing the least-squares cost function

$$J_{\mathrm{b}}(l) = \|\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}(l) - \Phi_{\mathrm{r}}(l)\tilde{\mathbf{\Gamma}} - \Phi_{\mathrm{v}}(l)\tilde{\mathbf{\Psi}}\|_F^2. \tag{18}$$

Setting the derivative of (18) with respect to $\Phi_{\mathrm{r}}(l)$ and $\Phi_{\mathrm{v}}(l)$ to 0 yields a system of equations which can be written as

$$\underbrace{\begin{bmatrix} \mathrm{tr}\{\tilde{\mathbf{\Gamma}}^H\tilde{\mathbf{\Gamma}}\} & \mathrm{tr}\{\tilde{\mathbf{\Gamma}}^H\tilde{\mathbf{\Psi}}\} \\ \mathrm{tr}\{\tilde{\mathbf{\Gamma}}^H\tilde{\mathbf{\Psi}}\} & \mathrm{tr}\{\tilde{\mathbf{\Psi}}^H\tilde{\mathbf{\Psi}}\} \end{bmatrix}}_{\mathbf{A}_{\mathrm{b}}} \underbrace{\begin{bmatrix} \hat{\Phi}_{\mathrm{r,b}}(l) \\ \hat{\Phi}_{\mathrm{v,b}}(l) \end{bmatrix}}_{\hat{\boldsymbol{\phi}}_{\mathrm{b}}(l)} = \underbrace{\begin{bmatrix} \mathrm{tr}\{\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}^H(l)\tilde{\mathbf{\Gamma}}\} \\ \mathrm{tr}\{\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}^H(l)\tilde{\mathbf{\Psi}}\} \end{bmatrix}}_{\mathbf{p}_{\mathrm{b}}(l)}, \tag{19}$$

[3]Note that since the matrices $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}(l)$, $\tilde{\mathbf{\Gamma}}$, and $\tilde{\mathbf{\Psi}}$ are symmetric, matching $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}(l)$ to (7) yields $M(M-1)/2$ equations instead of $(M-1)^2$ equations.

where the quantities $\mathbf{A}_{\mathrm{b}}$, $\hat{\boldsymbol{\phi}}_{\mathrm{b}}(l)$, and $\mathbf{p}_{\mathrm{b}}(l)$ have been introduced in order to simplify the notation. The solution to (19) is given by

$$\hat{\boldsymbol{\phi}}_{\mathrm{b}}(l) = \mathbf{A}_{\mathrm{b}}^{-1}\mathbf{p}_{\mathrm{b}}(l), \tag{20}$$

with the proposed blocking-based late reverberation PSD estimate $\hat{\Phi}_{\mathrm{r,b}}(l)$ being the first element of $\hat{\boldsymbol{\phi}}_{\mathrm{b}}(l)$ and the noise PSD estimate $\hat{\Phi}_{\mathrm{v,b}}(l)$ being the second element of $\hat{\boldsymbol{\phi}}_{\mathrm{b}}(l)$. Using the late reverberation and noise PSD estimates $\hat{\Phi}_{\mathrm{r,b}}(l)$ and $\hat{\Phi}_{\mathrm{v,b}}(l)$, the blocking-based target signal PSD can be estimated as

$$\hat{\Phi}_{\mathrm{s,b}}(l) = \frac{1}{\mathbf{d}^H\mathbf{d}}\mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{y}}(l) - \hat{\Phi}_{\mathrm{r,b}}(l)\mathbf{\Gamma} - \hat{\Phi}_{\mathrm{v,b}}(l)\mathbf{\Psi}\}. \tag{21}$$

It should be noted that if the signal model in (7) perfectly holds, the non-blocking-based estimator proposed in Section 3.1 and the blocking-based estimator proposed in this section would result in the same PSD estimates. In practice however, the signal model in (7) does not perfectly hold since the early and late reverberation components are not perfectly uncorrelated, the late reverberation is not perfectly diffuse, and the noise cannot be typically perfectly modeled by a spatially homogeneous sound field. Furthermore, estimating the matrices $\mathbf{\Phi}_{\mathbf{y}}(l)$ and $\mathbf{\Phi}_{\tilde{\mathbf{u}}}(l)$ by recursive averaging of a single realization of the signals will not yield the expected value operator. As a result, the proposed PSD estimators yield different PSD estimates in practice. As will be shown in Section 4, using the blocking-based PSD estimates in an MWF yields a better performance than using the non-blocking-based PSD estimates.

4. EXPERIMENTAL RESULTS

In this section, we investigate the dereverberation and noise reduction performance of the MWF using the proposed PSD estimators and two alternative versions to compute the TRNR. More precisely, we investigate the performance of the MWF implemented using

- the proposed non-blocking-based estimator with the TRNR estimated as in (9), which will be referred to as NBB,
- the proposed non-blocking-based estimator with the TRNR estimated as in (10), which will be referred to as NBB-DD,
- the proposed blocking-based estimator with the TRNR estimated as in (9), which will be referred to as BB, and
- the proposed blocking-based estimator with the TRNR estimated as in (10), which will be referred to as BB-DD.

In addition, the performance of the BB and BB-DD methods will be compared to the performance of the MWF implemented using the target signal and late reverberation PSD estimates from [10], where it is assumed that an estimate of the noise PSD matrix is available.

4.1. Setup and instrumental measures

We consider three multi-channel acoustic systems with a single speech source and $M = 4$ microphones. The first acoustic system consists of a linear microphone array with an inter-sensor distance of 3 cm [24], the second acoustic system consists of a circular microphone array with a radius of 10 cm [25], and the third acoustic system consists of a linear microphone array with an inter-sensor distance of 6 cm [26]. Table 1 presents the reverberation time T_{60}, the DOA θ of the speech source, and the direct-to-reverberation ratio (DRR) for each acoustic system. The speech components are generated by convolving a 38 s long clean speech signal with measured room impulse responses at a sampling frequency $f_s = 16$ kHz. The noise components consist of stationary uncorrelated noise with a broadband reverberant signal-to-noise ratio (RSNR) between 10 dB and 40 dB. The reverberant speech-plus-noise signal is preceded

Table 1: Characteristics of the considered acoustic systems.

Acoustic system	T_{60} [s]	θ	DRR [dB]
1	0.61	90°	-0.76
2	0.73	45°	1.43
3	1.25	-15°	-0.04

by a 1 s long noise-only segment such that when using the PSD estimator from [10], the noise PSD matrix can be estimated from the noise-only segment. The signals are processed using a weighted overlap-add STFT framework with a frame size of 1024 samples and an overlap of 75%. The first microphone is arbitrarily selected as the reference microphone. The target signal is defined as the direct component only, such that the RTF vector can be computed based on the DOA of the speech source.

The PSD matrices $\hat{\mathbf{\Phi}}_{\mathbf{y}}(l)$ and $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{u}}}(l)$ are estimated as in (11) with a smoothing factor α corresponding to a time constant of 40 ms. The diffuse spatial coherence matrix $\mathbf{\Gamma}$ is computed based on the microphone array geometry and the noise spatial coherence matrix is set to $\mathbf{\Psi} = \mathbf{I}$. The smoothing parameter in (10) is set to $\beta = 0.98$ and the minimum gain of the single-channel Wiener postfilter is set to -17 dB. For the estimator from [10], the noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}$ is estimated as

$$\hat{\mathbf{\Phi}}_{\mathbf{v}} = \frac{1}{L_v} \sum_{l=1}^{L_v} \mathbf{v}(l)\mathbf{v}^H(l), \tag{22}$$

with L_v being the total number of noise-only segments.

The performance is evaluated in terms of the improvement in frequency-weighted segmental SNR (ΔfwSSNR) [27] and log-likelihood ratio (ΔLLR) [27] between the output signal and the reference microphone signal. The fwSSNR and LLR measures are intrusive measures comparing the signal being evaluated to a reference signal. The reference signal used in this paper is the anechoic speech signal. It should be noted that a positive ΔfwSSNR and a negative ΔLLR indicate a performance improvement.

4.2. Performance of the proposed estimators

In this section the performance of NBB, NBB-DD, BB, and BB-DD is investigated for all considered RSNRs and acoustic systems. The presented performance measures are averaged over all considered acoustic systems.

Fig. 1 depicts the performance of all considered techniques in terms of ΔfwSSNR and ΔLLR. It can be observed that, as expected, for all considered techniques the performance improvement decreases as the RSNR increases. Furthermore, it can be observed that in terms of both performance measures and for all considered

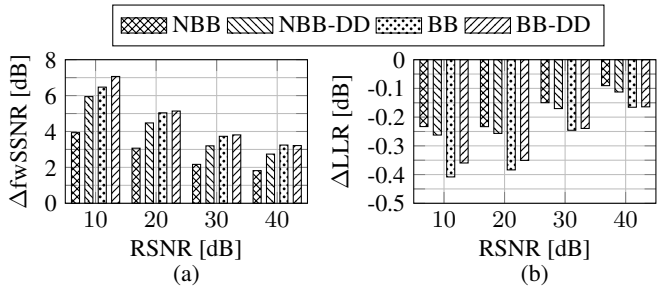

Fig. 1: MWF performance using the proposed PSD estimators.

Table 2: Average performance of the MWF using the proposed blocking-based estimator and the estimator from [10] which assumes that the noise PSD matrix is known (RSNR = 10 dB).

	BB	BR	BB-DD	BR-DD
ΔfwSSNR [dB]	**6.47**	5.31	**7.07**	6.53
ΔLLR [dB]	**-0.41**	-0.33	**-0.36**	-0.31

RSNRs, a larger performance improvement is obtained when using NBB-DD instead of NBB, suggesting that smoothing the TRNR estimate using the decision directed approach is particularly important when using non-blocking-based PSD estimates. In addition, it can be observed that BB and BB-DD outperform NBB and NBB-DD for all considered RSNRs. While BB-DD yields the highest ΔfwSSNR, BB results in the highest ΔLLR. Informal listening tests suggest that BB-DD yields a better perceptual quality than BB, with BB introducing more musical noise and signal artifacts than BB-DD.

In summary, the presented results show that for the considered acoustic scenarios, the proposed blocking-based PSD estimates yield a better performance than the non-blocking-based PSD estimates.

4.3. Performance of the proposed blocking-based estimator and the state-of-the-art estimator from [10]

In this section, the performance of BB and BB-DD is compared to the performance of the estimator from [10], which uses a blocking matrix and only estimates the target signal and late reverberation PSDs, assuming that an estimate of the noise PSD matrix is available. The noise PSD matrix is estimated as in (22) and the MWF is implemented using $\hat{\mathbf{\Phi}}_{\mathbf{v}}$ (instead of $\hat{\Phi}_{\mathrm{v}}(l)\mathbf{\Psi}$ in (8)) with the TRNR estimated as in (9) or (10). Using [10] with the TRNR estimated as in (9) will be referred to as BR, whereas using [10] with the TRNR estimated as in (10) will be referred to as BR-DD. Due to space constraints, only the performance for RSNR = 10 dB is presented and similarly as before, the performance is averaged over all considered acoustic systems.

Table 2 depicts the performance of the considered techniques in terms of ΔfwSSNR and ΔLLR. It can be observed that BB and BB-DD result in a similar or better performance than BR and BR-DD, respectively. It should be noted that the noise PSD matrix estimate used for BR and BR-DD is rather accurate, since the noise is stationary and all noise-only segments are used to compute the PSD matrix. The presented results show that the proposed blocking-based estimator manages to remove the assumption that the noise PSD matrix is known and additionally estimates the noise PSD without hindering the dereverberation and noise reduction performance.

5. CONCLUSION

In this paper joint estimators for the late reverberation and noise PSDs have been derived, removing the assumption made by state-of-the-art late reverberation PSD estimators that the noise PSD matrix is known. Modeling the noise as a spatially homogeneous sound field with an unknown time-varying PSD and a known time-invariant spatial coherence matrix, we have derived a non-blocking-based and a blocking-based joint estimator of the late reverberation and noise PSDs. Simulation results show that the proposed blocking-based PSD estimator yields the best performance when used in an MWF, also yielding a similar or better performance than a state-of-the-art blocking-based late reverberation PSD estimator which assumes that the noise PSD matrix is known.

6. REFERENCES

[1] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, June 2003.

[2] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, July 2006.

[3] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sept. 2014, pp. 333–337.

[4] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, Sept. 2001, pp. 31–34.

[5] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 945–958, May 2013.

[6] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[7] K. Lebart and J. M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, May-Jun. 2001.

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

[9] 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. International Workshop on Acoustic Echo and Noise Control*, Xi'an, China, Sept. 2016.

[10] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sept. 2013.

[11] O. Thiergart and E. A. P. Habets, "Extracting reverberant sound using a linearly constrained minimum variance spatial filter," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 630–634, May 2014.

[12] S. Braun and E. A. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Applied Signal Processing*, vol. 2015, no. 1, Dec. 2015.

[13] O. Schwartz, S. Braun, S. Gannot, and E. A. P. Habets, "Maximum likelihood estimation of the late reverberant power spectral density in noisy environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2015.

[14] 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 International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 151–155.

[15] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1595–1608, Sept. 2016.

[16] 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. European Signal Processing Conference*, Budapest, Hungary, Sept. 2016, pp. 1123–1127.

[17] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, Mar. 2017, pp. 611–615.

[18] I. Kodrasi and S. Doclo, "Multi-channel late reverberation power spectral density estimation based on nuclear norm minimization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2017, pp. 101–105.

[19] J. Ramiŕez, J. C. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, Apr. 2004.

[20] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, Jan. 2010.

[21] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1732–1736, Nov. 1962.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[23] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 380–384.

[24] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sept. 2014, pp. 313–317.

[25] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, Jan. 2016.

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

[27] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*, Prentice-Hall, New Jersey, USA, 1988.