Open access • Journal Article • DOI:10.1109/TASLP.2018.2811184

Analysis of Eigenvalue Decomposition-Based Late Reverberation Power Spectral Density Estimation — Source link

Ina Kodrasi, Simon Doclo

Institutions: University of Oldenburg

Published on: 01 Jun 2018 - IEEE Transactions on Audio, Speech, and Language Processing (Institute of Electrical and Electronics Engineers (IEEE))

Topics: Reverberation, Wiener filter and Estimator

Analysis of Eigenvalue Decomposition-Based Late Reverberation Power Spectral Density Estimation

Ina Kodrasi, *Member, IEEE,* Simon Doclo, *Senior Member, IEEE*

***Abstract*—Many speech dereverberation techniques require an estimate of the late reverberation power spectral density (PSD). State-of-the-art multi-channel methods for estimating the late reverberation PSD typically rely on 1) an estimate of the relative transfer functions (RTFs) of the target signal, 2) a model for the spatial coherence matrix of the late reverberation, and 3) an estimate of the reverberant speech or reverberant and noisy speech PSD matrix. The RTFs, the spatial coherence matrix, and the speech PSD matrix are all prone to modeling and estimation errors in practice, with the RTFs being particularly difficult to estimate accurately, especially in highly reverberant and noisy scenarios. Recently, we proposed an eigenvalue decomposition (EVD)-based late reverberation PSD estimator which does not require an estimate of the RTFs. In this paper, this EVD-based PSD estimator is further analyzed and its estimation accuracy and computational complexity is analytically compared to a state-of-the-art maximum likelihood (ML)-based PSD estimator. It is shown that for perfect knowledge of the RTFs, spatial coherence matrix, and reverberant speech PSD matrix, the ML-based and EVD-based PSD estimates are both equal to the true late reverberation PSD. In addition, it is shown that for erroneous RTFs but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix, the ML-based PSD estimate is larger than or equal to the true late reverberation PSD, whereas the EVD-based PSD estimate is obviously still equal to the true late reverberation PSD. Finally, it is shown that when modeling and estimation errors occur in all quantities, the ML-based PSD estimate is larger than or equal to the EVD-based PSD estimate. Simulation results for several realistic acoustic scenarios demonstrate the advantages of using the EVD-based PSD estimator in a multi-channel Wiener filter, yielding a significantly better performance than the ML-based PSD estimator.**

***Index Terms*—Dereverberation, PSD estimation, EVD, prewhitening, ML**

I. Introduction

In hands-free speech communication applications the recorded microphone signals are often corrupted by reverberation, which arises from the superposition of delayed and attenuated copies of the anechoic speech signal. While early reverberation may be desirable [1], late reverberation may degrade the perceived speech quality and intelligibility [2], [3] as well as the performance of automatic speech recognition systems [4], [5]. Hence, speech enhancement techniques which effectively suppress the late reverberation are required. In the last decades many single-channel and multi-channel dereverberation techniques have been proposed [6]–[8], with multi-channel techniques being generally preferred since they are able to exploit both the spectro-temporal and the spatial characteristics of the received microphone signals. Commonly used techniques for speech dereverberation are acoustic multi-channel equalization techniques [9]–[12], multi-channel linear prediction-based techniques [13]–[15], and the multi-channel Wiener filter (MWF) as well as various beamformer-postfilter structures [16]–[28]. The MWF is typically implemented as a minimum variance distortionless response (MVDR) beamformer followed by a single-channel Wiener postfilter [20]–[28]. Modeling the late reverberation as a spatially homogeneous sound field [20]–[28], the implementation of the MVDR beamformer and Wiener postfilter requires (among other quantities) an estimate of the spatial coherence matrix and of the power spectral density (PSD) of the late reverberation. While the spatial coherence matrix can be computed assuming a reasonable sound field model for the late reverberation (e.g., diffuse), estimating the late reverberation PSD is challenging.

To estimate the late reverberation PSD several single-channel estimators based on a temporal model of reverberation [29]–[31] and multi-channel estimators based on a model for the spatial coherence matrix of the late reverberation [22]–[28], [32] have been proposed. The multi-channel estimators can be classified as non-blocking-based estimators [23], [26], [28], where the target signal and late reverberation PSDs are jointly estimated, and blocking-based estimators [22], [24], [25], [27], [32], where the late reverberation PSD is estimated at the output of a blocking matrix aiming to block the target signal. For both classes of estimators, either maximum-likelihood (ML)-based estimators [23], [25]–[27] or estimators minimizing the Frobenius norm of an error PSD matrix [22], [24], [28] have been proposed. Whereas for noisy scenarios the ML-based estimators require an iterative optimization procedure, e.g. based on Newton's method [25], [26] or root finding [27], in noise-free scenarios a closed-form solution for the ML estimator can be derived [23], [27]. In [33] it has been analytically shown that the ML-based PSD estimator from [23] yields a higher PSD estimation accuracy than the PSD estimator based on the Frobenius norm in [22]. It should be realized that all multi-channel late reverberation PSD estimators in [22]–[28], [32] require an estimate of the relative transfer functions (RTFs) of the target signal from the reference microphone to all microphones. In addition, all estimators require a model for the spatial coherence matrix of the late reverberation and an estimate of the reverberant speech or reverberant and noisy speech PSD matrix. While

This work was supported in part by the Cluster of Excellence 1077 Hearing4All, funded by the German Research Foundation (DFG), and in part by the joint Lower Saxony-Israeli Project ATHENA, funded by the State of Lower Saxony.

I. Kodrasi was with the Department of Medical Physics and Acoustics, University of Oldenburg, 26129 Oldenburg, Germany. She is now with the Idiap Research Institute, 1920 Martigny, Switzerland (email: ina.kodrasi@idiap.ch).

S. Doclo is with the Department of Medical Physics and Acoustics, University of Oldenburg, 26129 Oldenburg, Germany (email: simon.doclo@uni-oldenburg.de).

the spatial coherence matrix can be computed assuming a diffuse sound field model [22]–[28] and the PSD matrix can be directly estimated from the received microphone signals, the RTFs may be more difficult to estimate accurately, particularly in highly reverberant and noisy scenarios. As experimentally validated in [34]–[36], erroneously estimated RTFs degrade the dereverberation performance of the speech enhancement system.

Recently, we proposed a multi-channel late reverberation PSD estimator which does not require an estimate of the RTFs [36], [37]. The late reverberation PSD is estimated using the eigenvalue decomposition (EVD) of the reverberant speech PSD matrix prewhitened with the spatial coherence matrix of the late reverberation. In this paper, we further analyze this EVD-based PSD estimator, providing novel insights in terms of 1) its estimation accuracy in comparison to the state-of-the-art ML-based PSD estimator from [23], 2) its computational complexity, and 3) its performance not only in reverberant scenarios as in [36], but also in the presence of additive noise. It is shown that when the true RTFs, spatial coherence matrix, and reverberant speech PSD matrix are known, the ML-based and EVD-based PSD estimators are equivalent and yield the true late reverberation PSD. Furthermore, it is shown that for erroneously estimated RTFs but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix, the ML-based PSD estimate is larger than or equal to the true late reverberation PSD, whereas the EVD-based PSD estimate is obviously still equal to the true late reverberation PSD. Finally, it is shown that when modeling and estimation errors occur in all quantities, the ML-based PSD estimate is larger than or equal to the EVD-based PSD estimate. On the one hand, when such errors result in an overestimation of the true late reverberation PSD for both estimators, the ML-based PSD estimation error is larger than or equal to the EVD-based PSD estimation error. On the other hand, when such errors result in an underestimation of the true late reverberation PSD for both estimators, the ML-based PSD estimation error is smaller than or equal to the EVD-based PSD estimation error. Simulation results for several realistic acoustic scenarios with different reverberation times and microphone configurations demonstrate the advantages of using the EVD-based PSD estimator in the MWF, yielding a significantly better performance than the ML-based PSD estimator.

The paper is organized as follows. In Section II the considered acoustic configuration and the used notation is introduced. In Section III the ML-based and EVD-based late reverberation PSD estimators are reviewed and analytical insights on the equivalence of both estimators are provided. In Section IV the impact of modeling and estimation errors in the RTFs, spatial coherence matrix, and reverberant speech PSD matrix on the estimation accuracy of the ML-based and EVD-based PSD estimators is theoretically analyzed. In addition, the computational complexity of the ML-based and EVD-based PSD estimators is compared. In Section V all analytical derivations are experimentally validated and the performance of the MWF using the ML-based and EVD-based PSD estimators in realistic acoustic scenarios is compared.

II. Configuration and Notation

Consider a reverberant and noisy acoustic system with a single speech source and $M \geq 2$ microphones, as depicted in Fig. 1. In the short-time Fourier transform (STFT) domain, the m-th microphone signal $Y_m(k,l)$ at frequency bin k and time frame index l is given by

$$Y_m(k,l) = \underbrace{X_{\mathrm{e},m}(k,l) + X_{\mathrm{r},m}(k,l)}_{X_m(k,l)} + V_m(k,l), \tag{1}$$

with $X_m(k,l)$ the reverberant speech component which consists of the direct and early reverberation component $X_{\mathrm{e},m}(k,l)$ and the late reverberation component $X_{\mathrm{r},m}(k,l)$, and $V_m(k,l)$ the noise component. In vector notation, the M-dimensional microphone signal vector $\mathbf{y}(k,l)$ can be written as

$$\mathbf{y}(k,l) = \underbrace{\mathbf{x}_{\mathrm{e}}(k,l) + \mathbf{x}_{\mathrm{r}}(k,l)}_{\mathbf{x}(k,l)} + \mathbf{v}(k,l), \tag{2}$$

with $\mathbf{y}(k,l) = [Y_1(k,l)\ Y_2(k,l)\ \ldots\ Y_M(k,l)]^T$ and $\mathbf{x}(k,l)$, $\mathbf{x}_{\mathrm{e}}(k,l)$, $\mathbf{x}_{\mathrm{r}}(k,l)$, and $\mathbf{v}(k,l)$ similarly defined. For a single source scenario, the direct and early reverberation component $\mathbf{x}_{\mathrm{e}}(k,l)$ can be expressed as

$$\mathbf{x}_{\mathrm{e}}(k,l) = S(k,l)\mathbf{d}(k,l), \tag{3}$$

where $S(k,l)$ denotes the target signal, i.e., direct and early reverberation component received at the reference microphone, and $\mathbf{d}(k,l) = [D_1(k,l)\ D_2(k,l)\ \ldots\ D_M(k,l)]^T$ denotes the M-dimensional vector of RTFs of the target signal from the reference microphone to all microphones. The target signal $S(k,l)$ is often defined as the direct component only, such that for calibrated microphones the RTF vector $\mathbf{d}(k,l)$ only depends on the direction of arrival (DOA) of the speech source and the microphone array geometry [22], [24]–[26], [28], [32].

Assuming that the components in (2) are mutually uncorrelated, the PSD matrix of the microphone signals $\mathbf{y}(k,l)$ is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,l) &= \mathcal{E}\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\} &(4)\\ &= \underbrace{\mathbf{\Phi}_{\mathbf{x}_{\mathrm{e}}}(k,l) + \mathbf{\Phi}_{\mathbf{x}_{\mathrm{r}}}(k,l)}_{\mathbf{\Phi}_{\mathbf{x}}(k,l)} + \mathbf{\Phi}_{\mathbf{v}}(k,l), &(5)\end{aligned}$$

where $\mathcal{E}$ denotes the expectation operator, $\mathbf{\Phi}_{\mathbf{x}}(k,l) = \mathcal{E}\{\mathbf{x}(k,l)\mathbf{x}^H(k,l)\}$ is the reverberant speech PSD matrix, $\mathbf{\Phi}_{\mathbf{x}_{\mathrm{e}}}(k,l) = \mathcal{E}\{\mathbf{x}_{\mathrm{e}}(k,l)\mathbf{x}_{\mathrm{e}}^H(k,l)\}$ is the direct and early reverberation PSD matrix, $\mathbf{\Phi}_{\mathbf{x}_{\mathrm{r}}}(k,l) = \mathcal{E}\{\mathbf{x}_{\mathrm{r}}(k,l)\mathbf{x}_{\mathrm{r}}^H(k,l)\}$

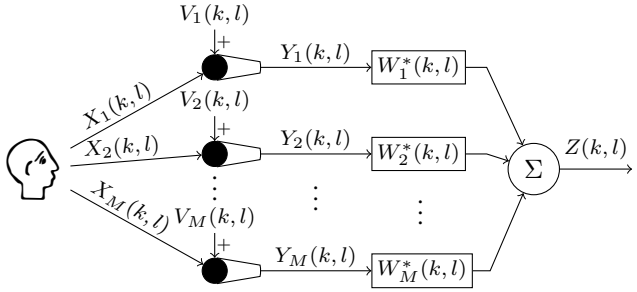

Fig. 1: Acoustic system configuration.

is the late reverberation PSD matrix, and $\mathbf{\Phi}_{\mathbf{v}}(k,l) = \mathcal{E}\{\mathbf{v}(k,l)\mathbf{v}^H(k,l)\}$ is the noise PSD matrix. The direct and early reverberation PSD matrix $\mathbf{\Phi}_{\mathbf{x}_\mathrm{e}}(k,l)$ is a rank-1 matrix given by (cf. (3))

$$\mathbf{\Phi}_{\mathbf{x}_\mathrm{e}}(k,l) = \Phi_\mathrm{s}(k,l)\mathbf{d}(k,l)\mathbf{d}^H(k,l), \tag{6}$$

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

$$\mathbf{\Phi}_{\mathbf{x}_\mathrm{r}}(k,l) = \Phi_\mathrm{r}(k,l)\mathbf{\Gamma}(k), \tag{7}$$

with $\Phi_\mathrm{r}(k,l)$ the time-varying PSD of the late reverberation and $\mathbf{\Gamma}(k)$ the spatial coherence matrix of the late reverberation which is assumed to be time-invariant. It is commonly assumed that the sound field modeling the late reverberation is diffuse [22]–[28], such that the spatial coherence matrix $\mathbf{\Gamma}(k)$ can be analytically computed based on the microphone array geometry [38]. Using (6) and (7), the reverberant speech PSD matrix $\mathbf{\Phi}_{\mathbf{x}}(k,l)$ can be expressed as

$$\mathbf{\Phi}_{\mathbf{x}}(k,l) = \Phi_\mathrm{s}(k,l)\mathbf{d}(k,l)\mathbf{d}^H(k,l) + \Phi_\mathrm{r}(k,l)\mathbf{\Gamma}(k). \tag{8}$$

Assuming that the reverberant speech PSD matrix is given by (8) is a commonly used assumption when deriving multi-channel late reverberation PSD estimators [21]–[28]. The analytical derivations provided in this paper are also based on this assumption. It should however be noted that (8) does not hold in practice, since 1) the late reverberation is typically not a perfect spatially homogeneous sound field and 2) the early and late reverberation components are not perfectly uncorrelated.

Given the M-dimensional filter vector $\mathbf{w}(k,l) = [W_1(k,l)\ W_2(k,l)\ \ldots\ W_M(k,l)]^T$, the output signal $Z(k,l)$ of the speech enhancement system in Fig. 1 is equal to the sum of the filtered microphone signals, i.e.,

$$Z(k,l) = \mathbf{w}^H(k,l)\mathbf{y}(k,l). \tag{9}$$

Dereverberation and noise reduction techniques aim at designing the filter $\mathbf{w}(k,l)$ such that the output signal $Z(k,l)$ is as close as possible to the target signal $S(k,l)$. Many such techniques require (among other quantities) an estimate of the late reverberation PSD $\Phi_\mathrm{r}(k,l)$, e.g., [19]–[21]. For conciseness, the frequency bin k and the frame index l are omitted in the remainder of this paper, unless explicitly required.

III. ML-based and EVD-based Late Reverberation Power Spectral Density Estimators

In this section, the ML-based late reverberation PSD estimator from [23] and the EVD-based late reverberation PSD estimator from [36] are reviewed and analytical insights on the equivalence of both estimators are provided. For simplicity, in the following we assume a noise-free scenario, i.e., $\mathbf{y} = \mathbf{x}$ and $\mathbf{\Phi}_{\mathbf{y}} = \mathbf{\Phi}_{\mathbf{x}}$. However, it should be noted that the considered late reverberation PSD estimators can also be used in a noisy scenario if an estimate of the reverberant speech PSD matrix $\mathbf{\Phi}_{\mathbf{x}}$ can be obtained (cf. Section V-E).

A. ML-based PSD estimator

In [23] an ML-based estimator for the late reverberation PSD has been derived, assuming the spectral coefficients of the components $\mathbf{x}_\mathrm{e}$ and $\mathbf{x}_\mathrm{r}$ to be circularly-symmetric complex Gaussian distributed. Maximizing the likelihood function computed using these distributions results in the late reverberation PSD estimate

$$\Phi_\mathrm{r}^{\mathrm{ml}} = \frac{1}{M-1}\mathrm{tr}\left\{\left(\mathbf{I} - \mathbf{d}\frac{\mathbf{d}^H\mathbf{\Gamma}^{-1}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}}\right)\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\right\}, \tag{10}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace operator and $\mathbf{I}$ denotes the $M \times M$-dimensional identity matrix. Estimating the late reverberation PSD using (10) requires knowledge of the RTF vector $\mathbf{d}$, the spatial coherence matrix $\mathbf{\Gamma}$, and the reverberant speech PSD matrix $\mathbf{\Phi}_{\mathbf{x}}$. While $\mathbf{\Gamma}$ can be computed assuming a diffuse sound field model and $\mathbf{\Phi}_{\mathbf{x}}$ can be estimated from the microphone signals, accurately estimating $\mathbf{d}$ may not be straightforward, particularly in highly reverberant and noisy environments. As will be analytically shown in Section IV-A, erroneous RTFs degrade the accuracy of the ML-based PSD estimate, hence resulting in a degradation of the dereverberation performance of the speech enhancement system (cf. simulation results in Section V-D).

B. EVD-based PSD estimator

Aiming to remove the dependency on the RTF vector $\mathbf{d}$, in [36] we proposed to estimate the late reverberation PSD using the eigenvalues of the reverberant speech PSD matrix prewhitened with the diffuse spatial coherence matrix. Using the Cholesky decomposition of the positive definite spatial coherence matrix $\mathbf{\Gamma}$, i.e.,

$$\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H, \tag{11}$$

with $\mathbf{L}$ an $M \times M$-dimensional lower triangular matrix, the prewhitened reverberant speech PSD matrix can be computed as

$$\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}} = \mathbf{L}^{-1}\mathbf{\Phi}_{\mathbf{x}}\mathbf{L}^{-H}. \tag{12}$$

Substituting (8) in (12), it can be observed that the matrix $\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}$ is equal to the sum of a rank-1 matrix and the identity matrix scaled by the late reverberation PSD, i.e.,

$$\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}} = \Phi_\mathrm{s}\underbrace{\mathbf{L}^{-1}\mathbf{d}}_{\mathbf{d}_\mathrm{w}}\underbrace{\mathbf{d}^H\mathbf{L}^{-H}}_{\mathbf{d}_\mathrm{w}^H} + \Phi_\mathrm{r}\mathbf{L}^{-1}\mathbf{\Gamma}\mathbf{L}^{-H} \tag{13}$$

$$= \Phi_\mathrm{s}\mathbf{d}_\mathrm{w}\mathbf{d}_\mathrm{w}^H + \Phi_\mathrm{r}\mathbf{I}, \tag{14}$$

with the vector $\mathbf{d}_\mathrm{w}$ introduced in order to simplify the notation. Due to the structure in (14), the eigenvalues of the matrix $\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}$ (arranged in descending order) are equal to

$$\lambda_1\{\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}\} = \sigma + \Phi_\mathrm{r}, \tag{15a}$$

$$\lambda_j\{\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}\} = \Phi_\mathrm{r}, \qquad j = 2,\ \ldots, M, \tag{15b}$$

with σ the only non-zero eigenvalue of the rank-1 matrix $\Phi_\mathrm{s}\mathbf{d}_\mathrm{w}\mathbf{d}_\mathrm{w}^H$. Based on (15), in [36] we proposed to estimate the late reverberation PSD using either any of the last $M-1$ eigenvalues of the matrix $\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}$, i.e.,

$$\Phi_{\mathrm{r},j}^{\mathrm{evd}} = \lambda_j\{\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}\}, \qquad j = 2,\ \ldots,\ M, \tag{16}$$

or the mean of the last $M-1$ eigenvalues of the matrix $\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}$, i.e.,

$$\Phi_{\mathrm{r},\mu}^{\mathrm{evd}} = \frac{1}{M-1}\left(\mathrm{tr}\left\{\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}\right\} - \lambda_1\left\{\mathbf{\Phi}_{\mathbf{x}}^{\mathrm{w}}\right\}\right), \tag{17}$$

with (17) derived using the fact that the trace of a matrix is equal to the sum of its eigenvalues. Obviously, when the true spatial coherence matrix $\mathbf{\Gamma}$ and the true reverberant speech PSD matrix $\mathbf{\Phi}_{\mathbf{x}}$ are known, the EVD-based PSD estimates in (16) and (17) are equal.

Estimating the late reverberation PSD using (16) or (17) only requires knowledge of the spatial coherence matrix $\mathbf{\Gamma}$ and the reverberant speech PSD matrix $\mathbf{\Phi}_{\mathbf{x}}$. Unlike the ML-based PSD estimator in (10), it is important to note that our EVD-based PSD estimator does not require an estimate of the RTF vector $\mathbf{d}$, which is advantageous in order to avoid propagation of RTF estimation errors into the late reverberation PSD estimate (cf. sensitivity analysis in Section IV-A and simulation results in Section V-D).

C. Equivalence of the ML-based and EVD-based PSD estimators

In the following, it is shown that when the true RTF vector $\mathbf{d}$, the true spatial coherence matrix $\mathbf{\Gamma}$, and the true reverberant speech PSD matrix $\mathbf{\Phi}_{\mathbf{x}}$ are known, the ML-based PSD estimate in (10) and the EVD-based PSD estimates in (16) and (17) are equivalent and equal to the true late reverberation PSD.

Since the trace is invariant under cyclic permutations, the ML-based PSD estimate in (10) can be written as

$$\Phi_{\mathrm{r}}^{\mathrm{ml}} = \frac{1}{M-1}\left(\mathrm{tr}\left\{\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\right\} - \frac{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}}\right). \tag{18}$$

Using $\mathbf{\Phi}_{\mathbf{x}}$ from (8), the terms in (18) can be simplified to

$$\mathrm{tr}\left\{\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\right\} = \Phi_{\mathrm{s}}\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d} + \Phi_{\mathrm{r}}M, \tag{19}$$

$$\frac{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}} = \Phi_{\mathrm{s}}\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d} + \Phi_{\mathrm{r}}. \tag{20}$$

Substituting (19) and (20) in (18), it can be observed that when the true $\mathbf{d}$, $\mathbf{\Gamma}$, and $\mathbf{\Phi}_{\mathbf{x}}$ are known, the ML-based PSD estimate is equal to the true late reverberation PSD, i.e.,

$$\Phi_{\mathrm{r}}^{\mathrm{ml}} = \frac{1}{M-1}\left(\Phi_{\mathrm{s}}\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d} + \Phi_{\mathrm{r}}M - \Phi_{\mathrm{s}}\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d} - \Phi_{\mathrm{r}}\right) \tag{21}$$

$$= \Phi_{\mathrm{r}}. \tag{22}$$

Clearly, when the true $\mathbf{\Gamma}$ and $\mathbf{\Phi}_{\mathbf{x}}$ are known, the EVD-based PSD estimates in (16) and (17) are also equal to the true late reverberation PSD (cf. (15)), i.e.,

$$\Phi_{\mathrm{r},j}^{\mathrm{evd}} = \Phi_{\mathrm{r},\mu}^{\mathrm{evd}} = \Phi_{\mathrm{r}}, \qquad j = 2,\ \ldots, M. \tag{23}$$

In summary, when the true RTF vector, spatial coherence matrix, and speech PSD matrix are known (which is rarely the case in practice, cf. Section IV), the ML-based and EVD-based estimators are equivalent and yield the true late reverberation PSD. It should be noted that this analytical result applies in practice only to scenarios where the late reverberation is a perfect spatially homogeneous sound field and the early and late reverberation components are perfectly uncorrelated.

IV. Impact of modeling and estimation errors on the ML-based and EVD-based PSD Estimators

The analysis in Section III-C is based on the assumption that the true RTF vector, spatial coherence matrix, and reverberant speech PSD matrix are known. In practice however, modeling and estimation errors typically occur in all quantities. First, the RTF vector may differ from the true RTF vector, e.g., due to DOA estimation errors in highly reverberant and noisy scenarios [39]–[42]. Second, since the spatial coherence matrix is typically computed assuming a perfectly diffuse sound field for the late reverberation whereas this is not the case in practice, it typically differs from the true spatial coherence matrix. Third, since the reverberant speech PSD matrix is typically estimated via recursive averaging of a single realization of the microphone signals or by subtracting the noise PSD matrix from the reverberant and noisy PSD matrix (cf. Section V), it will also typically differ from the true reverberant speech PSD matrix. In this section, we analyze the impact of modeling and estimation errors in the RTFs, spatial coherence matrix, and reverberant speech PSD matrix on the estimation accuracy of the ML-based and EVD-based PSD estimators. It should again be noted that the analytical results derived in this section apply in practice only to scenarios where the late reverberation is a perfect spatially homogeneous sound field and the early and late reverberation components are perfectly uncorrelated.

The estimated RTF vector, spatial coherence matrix, and reverberant speech PSD matrix are denoted by $\hat{\mathbf{d}}$, $\hat{\mathbf{\Gamma}}$, and $\hat{\mathbf{\Phi}}_{\mathbf{x}}$, respectively. Using the estimated quantities $\hat{\mathbf{d}}$, $\hat{\mathbf{\Gamma}}$, and $\hat{\mathbf{\Phi}}_{\mathbf{x}}$, the ML-based PSD estimate in (18) is given by

$$\hat{\Phi}_{\mathrm{r}}^{\mathrm{ml}} = \frac{1}{M-1}\left(\mathrm{tr}\left\{\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{\Gamma}}^{-1}\right\} - \frac{\hat{\mathbf{d}}^H\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}}{\hat{\mathbf{d}}^H\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}}\right). \tag{24}$$

Using $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ and the Cholesky decomposition of $\hat{\mathbf{\Gamma}}$, i.e.,

$$\hat{\mathbf{\Gamma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^H, \tag{25}$$

the estimated prewhitened reverberant speech PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}$ can be defined similarly to (12), i.e.,

$$\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}} = \hat{\mathbf{L}}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{L}}^{-H}, \tag{26}$$

such that the EVD-based PSD estimates in (16) and (17) are given by

$$\hat{\Phi}_{\mathrm{r},j}^{\mathrm{evd}} = \lambda_j\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\}, \qquad j = 2,\ \ldots,\ M, \tag{27}$$

$$\hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}} = \frac{1}{M-1}\left(\mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\} - \lambda_1\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\}\right). \tag{28}$$

In the presence of modeling and estimation errors in the spatial coherence matrix and reverberant speech PSD matrix, i.e., $\hat{\mathbf{\Gamma}} \neq \mathbf{\Gamma}$ and $\hat{\mathbf{\Phi}}_{\mathbf{x}} \neq \mathbf{\Phi}_{\mathbf{x}}$, the EVD-based PSD estimates in (27) and (28) are (typically) not equal, i.e., $\hat{\Phi}_{\mathrm{r},j}^{\mathrm{evd}} \neq \hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}$. The theoretical analysis in this section is conducted for the EVD-based PSD estimate $\hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}$ in (28). In the simulation results in Section V, the performance also when using the EVD-based PSD estimate $\hat{\Phi}_{\mathrm{r},2}^{\mathrm{evd}}$ ($j = 2$) in (27) will be investigated. It should be noted that $\hat{\Phi}_{\mathrm{r},2}^{\mathrm{evd}} \geq \hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}$, with equality holding when using $M = 2$ microphones.

In order to evaluate the estimation accuracy of the PSD estimators in (24) and (28), we define the ML-based and EVD-based PSD estimation errors $\xi_{\mathrm{r}}^{\mathrm{ml}}$ and $\xi_{\mathrm{r},\mu}^{\mathrm{evd}}$ as

$$\xi_{\mathrm{r}}^{\mathrm{ml}} = |\underbrace{\hat{\Phi}_{\mathrm{r}}^{\mathrm{ml}} - \Phi_{\mathrm{r}}}_{\delta_{\mathrm{r}}^{\mathrm{ml}}}|, \quad \xi_{\mathrm{r},\mu}^{\mathrm{evd}} = |\underbrace{\hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}} - \Phi_{\mathrm{r}}}_{\delta_{\mathrm{r},\mu}^{\mathrm{evd}}}|, \tag{29}$$

where $|\cdot|$ denotes the absolute value. If $\delta_{\mathrm{r}}^{\mathrm{ml}} > 0$ and $\delta_{\mathrm{r},\mu}^{\mathrm{evd}} > 0$, the estimators overestimate the true late reverberation PSD, whereas if $\delta_{\mathrm{r}}^{\mathrm{ml}} < 0$ and $\delta_{\mathrm{r},\mu}^{\mathrm{evd}} < 0$, the estimators underestimate the true late reverberation PSD.

A. Impact of erroneous RTFs

In the following, the estimation accuracy of the ML-based PSD estimator in (24) is analyzed for erroneous RTFs, i.e., $\hat{\mathbf{d}} \neq \mathbf{d}$, but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix, i.e., $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$ and $\hat{\mathbf{\Phi}}_{\mathbf{x}} = \mathbf{\Phi}_{\mathbf{x}}$. Since the EVD-based PSD estimator in (28) does not depend on the RTF vector, and hence, is not affected by RTF estimation errors, it always yields the true late reverberation PSD for $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$ and $\hat{\mathbf{\Phi}}_{\mathbf{x}} = \mathbf{\Phi}_{\mathbf{x}}$ (cf. (15)).

The ML-based PSD estimate in (24) using $\hat{\mathbf{d}} \neq \mathbf{d}$, $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$, and $\hat{\mathbf{\Phi}}_{\mathbf{x}} = \mathbf{\Phi}_{\mathbf{x}}$ is given by

$$\hat{\Phi}_{\mathrm{r}}^{\mathrm{ml}} = \frac{1}{M-1}\left(\mathrm{tr}\left\{\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\right\} - \frac{\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\hat{\mathbf{d}}}{\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{d}}}\right). \tag{30}$$

Substituting $\mathbf{\Phi}_{\mathbf{x}}$ from (8), the second term in (30) can be expressed as

$$\frac{\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\mathbf{\Phi}_{\mathbf{x}}\mathbf{\Gamma}^{-1}\hat{\mathbf{d}}}{\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{d}}} = \Phi_{\mathrm{s}}\frac{(\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\mathbf{d})^2}{\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{d}}} + \Phi_{\mathrm{r}}. \tag{31}$$

Using (19) and (31), the ML-based PSD estimate in (30) can be written as

$$\hat{\Phi}_{\mathrm{r}}^{\mathrm{ml}} = \underbrace{\frac{\Phi_{\mathrm{s}}}{M-1}\left(\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d} - \frac{(\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\mathbf{d})^2}{\hat{\mathbf{d}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{d}}}\right)}_{\delta_{\mathrm{r}}^{\mathrm{ml}}} + \Phi_{\mathrm{r}}, \tag{32}$$

with $\delta_{\mathrm{r}}^{\mathrm{ml}}$ the difference between the ML-based PSD estimate and the true PSD in the presence of RTF estimation errors. In the following, the Cauchy-Schwarz inequality is used to show that $\delta_{\mathrm{r}}^{\mathrm{ml}} \geq 0$ and $\xi_{\mathrm{r}}^{\mathrm{ml}} \geq 0$.

In order to simplify the notation, we use the vector $\mathbf{d}_{\mathrm{w}}$ in (14) and additionally define the vector $\hat{\mathbf{d}}_{\mathrm{w}}$ as

$$\hat{\mathbf{d}}_{\mathrm{w}} = \mathbf{L}^{-1}\hat{\mathbf{d}}, \tag{33}$$

such that the difference $\delta_{\mathrm{r}}^{\mathrm{ml}}$ in (32) can be expressed as

$$\delta_{\mathrm{r}}^{\mathrm{ml}} = \frac{\Phi_{\mathrm{s}}}{M-1}\frac{(\mathbf{d}_{\mathrm{w}}^H\mathbf{d}_{\mathrm{w}})(\hat{\mathbf{d}}_{\mathrm{w}}^H\hat{\mathbf{d}}_{\mathrm{w}}) - (\hat{\mathbf{d}}_{\mathrm{w}}^H\mathbf{d}_{\mathrm{w}})^2}{\hat{\mathbf{d}}_{\mathrm{w}}^H\hat{\mathbf{d}}_{\mathrm{w}}}. \tag{34}$$

Based on the Cauchy-Schwarz inequality, it can be shown that[1]

$$(\mathbf{d}_{\mathrm{w}}^H\mathbf{d}_{\mathrm{w}})(\hat{\mathbf{d}}_{\mathrm{w}}^H\hat{\mathbf{d}}_{\mathrm{w}}) - (\hat{\mathbf{d}}_{\mathrm{w}}^H\mathbf{d}_{\mathrm{w}})^2 > 0. \tag{35}$$

[1]It should be noted that since the RTF vectors $\mathbf{d}$ and $\hat{\mathbf{d}}$ are linearly independent, the vectors $\mathbf{d}_{\mathrm{w}}$ and $\hat{\mathbf{d}}_{\mathrm{w}}$ are also linearly independent, hence, the Cauchy-Schwarz inequality in (35) is sharp.

Given (35) and since $\Phi_{\mathrm{s}} \geq 0$, $M-1 > 0$, and $\hat{\mathbf{d}}_{\mathrm{w}}^H\hat{\mathbf{d}}_{\mathrm{w}} > 0$ (assuming that $\hat{\mathbf{d}}_{\mathrm{w}} \neq \mathbf{0}$, i.e., $\hat{\mathbf{d}} \neq \mathbf{0}$), we conclude that in the presence of RTF estimation errors

$$\delta_{\mathrm{r}}^{\mathrm{ml}} \geq 0 \;\text{ and }\; \xi_{\mathrm{r}}^{\mathrm{ml}} \geq 0, \tag{36}$$

with equality only holding when the target signal PSD is equal to zero, i.e., $\Phi_{\mathrm{s}} = 0$.

In summary, in the presence of RTF estimation errors but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix, the ML-based PSD estimate is larger than or equal to the true late reverberation PSD, whereas the EVD-based PSD estimate is obviously still equal to the true late reverberation PSD. Overestimation of the true late reverberation PSD will lead to undesired speech distortion when used in a speech enhancement algorithm, e.g., a postfilter. This derivation can hence be valuable in practice to decide against using the ML-based PSD estimate in applications where the RTF is difficult to estimate accurately and speech distortion is unacceptable.

B. Impact of modeling and estimation errors in all quantities

In the following, the estimation accuracy of the ML-based and EVD-based PSD estimators in (24) and (28) is analyzed in the presence of modeling and estimation errors in the RTFs, spatial coherence matrix, and reverberant speech PSD matrix, i.e., $\hat{\mathbf{d}} \neq \mathbf{d}$, $\hat{\mathbf{\Gamma}} \neq \mathbf{\Gamma}$, and $\hat{\mathbf{\Phi}}_{\mathbf{x}} \neq \mathbf{\Phi}_{\mathbf{x}}$. Note that the analytical results derived in this section also hold for scenarios when modeling and estimation errors occur only in one (or two) of the required quantities, with the remaining quantities (or quantity) perfectly estimated. In realistic acoustic scenarios however, modeling and estimation errors typically occur in all quantities.

Based on the Cholesky decomposition of $\hat{\mathbf{\Gamma}}$ in (25) and since the trace is invariant under cyclic permutations, the first term in the ML-based estimate in (24) can be written as

$$\mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{\Gamma}}^{-1}\} = \mathrm{tr}\{\hat{\mathbf{L}}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{L}}^{-H}\} = \mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\}. \tag{37}$$

In order to simplify the notation, we define the vector $\hat{\mathbf{u}}$ as

$$\hat{\mathbf{u}} = \hat{\mathbf{L}}^{-1}\hat{\mathbf{d}}, \tag{38}$$

and express the second term in (24) as

$$\frac{\hat{\mathbf{d}}^H\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}}{\hat{\mathbf{d}}^H\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}} = \frac{\hat{\mathbf{u}}^H\hat{\mathbf{L}}^{-1}\hat{\mathbf{\Phi}}_{\mathbf{x}}\hat{\mathbf{L}}^{-H}\hat{\mathbf{u}}}{\hat{\mathbf{u}}^H\hat{\mathbf{u}}} = \frac{\hat{\mathbf{u}}^H\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\hat{\mathbf{u}}}{\hat{\mathbf{u}}^H\hat{\mathbf{u}}}. \tag{39}$$

Substituting (37) and (39) in (24), the ML-based PSD estimate in the presence of modeling and estimation errors in all quantities can be written as

$$\hat{\Phi}_{\mathrm{r}}^{\mathrm{ml}} = \frac{1}{M-1}\left(\mathrm{tr}\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\} - \frac{\hat{\mathbf{u}}^H\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\hat{\mathbf{u}}}{\hat{\mathbf{u}}^H\hat{\mathbf{u}}}\right). \tag{40}$$

It is well known that the Rayleigh quotient of a matrix is bounded by its maximum eigenvalue, i.e.,

$$\frac{\hat{\mathbf{u}}^H\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\hat{\mathbf{u}}}{\hat{\mathbf{u}}^H\hat{\mathbf{u}}} \leq \lambda_1\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\}, \tag{41}$$

with equality only holding when the vector $\hat{\mathbf{u}}$ corresponds to the (scaled) eigenvector of $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}$ associated with its largest

eigenvalue $\lambda_1\{\hat{\mathbf{\Phi}}_{\mathbf{x}}^{\mathrm{w}}\}$. By comparing the PSD estimates in (28) and (40), it can now be observed that for erroneous RTFs, spatial coherence matrix, and reverberant speech PSD matrix, the ML-based PSD estimate is larger than or equal to the EVD-based PSD estimate, i.e.,

$$\hat{\Phi}_{\mathrm{r}}^{\mathrm{ml}} \geq \hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}. \tag{42}$$

In order to compare the ML-based and EVD-based PSD estimation errors $\xi_{\mathrm{r}}^{\mathrm{ml}}$ and $\xi_{\mathrm{r},\mu}^{\mathrm{evd}}$, we use (42) and distinguish between the following cases:

- If the true late reverberation PSD is overestimated by both estimators, i.e., if $\delta_{\mathrm{r}}^{\mathrm{ml}} > 0$ and $\delta_{\mathrm{r},\mu}^{\mathrm{evd}} > 0$, the ML-based PSD estimation error is larger than or equal to the EVD-based PSD estimation error, i.e., $\xi_{\mathrm{r}}^{\mathrm{ml}} \geq \xi_{\mathrm{r},\mu}^{\mathrm{evd}}$. As already mentioned, overestimation of the true late reverberation PSD is particularly detrimental to the speech quality, since it results in speech distortion.
- If the true late reverberation PSD is underestimated by both estimators, i.e., if $\delta_{\mathrm{r}}^{\mathrm{ml}} < 0$ and $\delta_{\mathrm{r},\mu}^{\mathrm{evd}} < 0$, the ML-based PSD estimation error is smaller than or equal to the EVD-based PSD estimation error, i.e., $\xi_{\mathrm{r}}^{\mathrm{ml}} \leq \xi_{\mathrm{r},\mu}^{\mathrm{evd}}$. Underestimation of the true late reverberation PSD results in an unnecessary amount of residual reverberation in the output signal of the speech enhancement algorithms, while preserving the speech quality.
- If the true late reverberation PSD is overestimated by the ML-based estimator but underestimated by the EVD-based estimator, i.e., if $\delta_{\mathrm{r}}^{\mathrm{ml}} > 0$ and $\delta_{\mathrm{r},\mu}^{\mathrm{evd}} < 0$, no conclusions can be drawn on the PSD estimation errors $\xi_{\mathrm{r}}^{\mathrm{ml}}$ and $\xi_{\mathrm{r},\mu}^{\mathrm{evd}}$.

These derivations can be valuable in practice to decide 1) to use the ML-based PSD estimate in applications where late reverberation suppression is more important than speech quality preservation, or 2) to use the EVD-based PSD estimate in applications where speech quality preservation is more important than late reverberation suppression.

C. Computational complexity

In this section, we provide some insights on the computational complexity of the ML-based and EVD-based PSD estimators. The computational complexity of the ML-based PSD estimator in (10) is dominated by matrix multiplication, for which the best known upper bound is $O(M^{2.373})$ [43]. In contrast, the dominating operation for the EVD-based PSD estimators in (16) and (17) is the computation of the eigenvalues using an EVD. Although many algorithms exist for computing the EVD, we consider the QR decomposition-based algorithm [44], which is one of the most widely used algorithms to compute eigenvalues. The complexity of the QR decomposition-based algorithm for Hermitian matrices is $O(M^3)$ [45], also when the matrix is first transformed into real tridiagonal form using Householder reflections [44]. However, it should be noted that the EVD-based PSD estimators in (16) or (17) require only a single eigenvalue, for which more efficient algorithms exist, e.g., based on subspace tracking [46].

V. Simulation Results

In this section, the impact of modeling and estimation errors in the RTFs, spatial coherence matrix, and reverberant speech PSD matrix on the ML-based and EVD-based PSD estimates is experimentally validated. In addition, the performance of the MWF using the considered PSD estimators is compared for several realistic acoustic scenarios with and without background noise. In Section V-A the considered acoustic systems, algorithmic settings, and instrumental performance measures are presented. In Section V-B the analytical results of Section III-C and Section IV are experimentally validated. For noise-free scenarios, Section V-C compares the dereverberation performance of the MWF using the ML-based and EVD-based PSD estimators when the true RTF vector is known, whereas Section V-D compares the dereverberation performance of the MWF using the ML-based and EVD-based PSD estimators in the presence of RTF estimation errors. For noisy scenarios, the dereverberation and noise reduction performance of the MWF using the ML-based and EVD-based PSD estimators is investigated in Section V-E. The computation of the required quantities (RTF vector, spatial coherence matrix, reverberant speech PSD matrix) as well as the MWF implementation is presented at the beginning of each section.

A. Acoustic systems, algorithmic settings, and instrumental performance measures

We have considered two acoustic systems with a single speech source and $M \in \{2, 4\}$ microphones. The first acoustic system AS_1 consists of a circular microphone array with a radius of 10 cm [47] and the second acoustic system AS_2 consists of a linear microphone array with an inter-microphone distance of 6 cm [48]. Table I presents the room reverberation time T_{60}, the DOA θ of the speech source, and the direct-to-reverberation ratio (DRR) for both considered acoustic systems. The sampling frequency is $f_s = 16$ kHz. The speech components are generated by convolving a 38 s long clean speech signal with the measured room impulse responses (RIRs). The noise components either consist of non-stationary diffuse babble noise, representing background noise typically encountered in large crowded rooms, or stationary uncorrelated noise, representing e.g. microphone self-noise. The speech-plus-noise signal is preceded by a 1 s long noise-only segment. The signals are processed in the STFT domain using a weighted overlap-add framework with a Hamming window, a frame size $N = 1024$ samples, and an overlap of 75%. 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 first microphone is arbitrarily selected as the reference microphone.

TABLE I: Characteristics of the considered acoustic systems.

Acoustic system	T_{60} [s]	θ	DRR [dB]
AS_1	0.73	$45°$	1.43
AS_2	1.25	$-15°$	-0.05

In order to evaluate the performance, we use the perceptual evaluation of speech quality (PESQ) measure [49], the frequency-weighted segmental signal-to-noise ratio (fSNR) [50], the cepstral distance (CD) [51], and the short-time objective intelligibility measure (STOI) [52]. These instrumental performance measures are intrusive measures generating a similarity score between a test signal and a reference signal. The reference signal used in this paper is the clean speech signal. The improvement in these instrumental measures, i.e., ΔPESQ, ΔfSNR, ΔCD, and ΔSTOI, is computed as the difference between the PESQ, fSNR, CD, and STOI values of the output signal and the reference microphone signal. Note that a positive ΔPESQ, ΔfSNR, and ΔSTOI and a negative ΔCD indicate a performance improvement.

B. Validation of analytical results

In this section, the analytical results of Sections III-C, IV-A, and IV-B are experimentally validated for the exemplary acoustic system AS_1 with $M = 4$ microphones. It is experimentally validated that when the true RTF vector, spatial coherence matrix, and reverberant speech PSD matrix are known, the ML-based and EVD-based PSD estimates are equal to the true late reverberation PSD. In addition, it is experimentally validated that for erroneous RTFs but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix, the ML-based PSD estimate is larger than or equal to the true late reverberation PSD, whereas the EVD-based PSD estimate is still equal to the true late reverberation PSD. Finally, it is experimentally validated that in the presence of modeling and estimation errors in all quantities, the ML-based PSD estimate is larger than or equal to the EVD-based PSD estimate.

In this section, the true quantities $\mathbf{d}$, $\mathbf{\Gamma}$, and $\mathbf{\Phi_x}$ are computed as follows. The true RTF vector $\mathbf{d}$ is computed using the true DOA $\theta = 45°$ of the speech source as

$$\mathbf{d} = [1 \; e^{-j2\pi f \tau_2(\theta)} \; \dots \; e^{-j2\pi f \tau_M(\theta)}]^T, \tag{43}$$

with f the frequency and $\tau_m(\theta)$ the relative time delay of arrival between the m-th microphone and the 1^{st} (reference) microphone. The true spatial coherence matrix $\mathbf{\Gamma}$ is computed from the late reverberation components as

$$\mathbf{\Gamma}_{p,q}(k) = \frac{\sum_{l=0}^{L-1} X_{\text{r},p}(k,l) X_{\text{r},q}^*(k,l)}{\sqrt{\left(\sum_{l=0}^{L-1} |X_{\text{r},p}(k,l)|^2\right)\left(\sum_{l=0}^{L-1} |X_{\text{r},q}(k,l)|^2\right)}}, \tag{44}$$

with $\mathbf{\Gamma}_{p,q}(k)$ the $\{p,q\}$-th element of $\mathbf{\Gamma}(k)$, L the total number of time frames, and the late reverberation components generated by convolving the clean speech signal with the late reverberant tail of the measured RIRs (and transforming the resulting signal to the STFT domain). Using $\mathbf{d}$ and $\mathbf{\Gamma}$, the true reverberant speech PSD matrix $\mathbf{\Phi_x}$ is computed as

$$\mathbf{\Phi_x} = \Phi_{\text{s}} \mathbf{d}\mathbf{d}^H + \Phi_{\text{r}} \mathbf{\Gamma}, \tag{45}$$

where the PSDs Φ_{s} and Φ_{r} are computed from the target signal $X_{\text{e},1}$ and the late reverberation component $X_{\text{r},1}$ using recursive averaging with a smoothing factor α as

$$\Phi_{\text{s}}(k,l) = \alpha |X_{\text{e},1}(k,l)|^2 + (1-\alpha)\Phi_{\text{s}}(k,l-1), \tag{46}$$
$$\Phi_{\text{r}}(k,l) = \alpha |X_{\text{r},1}(k,l)|^2 + (1-\alpha)\Phi_{\text{r}}(k,l-1). \tag{47}$$

The target signal $X_{\text{e},1}$ is generated by convolving the clean speech signal with the direct part of the RIR (and transforming the resulting signal to the STFT domain). The used smoothing factor is $\alpha = 0.67$, corresponding to a time constant of 40 ms. It should be noted that since the objective of this section is only to validate the analytical results, the true $\mathbf{\Phi_x}$ is computed as in (45) in order to ensure that (8) perfectly holds.

The estimated quantities $\hat{\mathbf{d}}$, $\hat{\mathbf{\Gamma}}$, and $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ are computed as follows. The estimated RTF vector $\hat{\mathbf{d}}$ is computed using an exemplary erroneous DOA $\hat{\theta} = 30°$. The estimated spatial coherence matrix $\hat{\mathbf{\Gamma}}$ is computed assuming a spherically diffuse sound field as [38]

$$\hat{\mathbf{\Gamma}}_{p,q}(k) = \text{sinc}\left(\frac{2\pi k f_s}{Nc} r_{pq}\right), \tag{48}$$

with $c = 340$ m/s the speed of sound and r_{pq} the distance between the p-th and q-th microphone. The reverberant speech PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ is estimated from the received microphone signals using recursive averaging as

$$\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,l) = \alpha \mathbf{X}(k,l)\mathbf{X}^H(k,l) + (1-\alpha)\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,l-1). \tag{49}$$

Equivalence of the ML-based and EVD-based PSD estimators ($\hat{\mathbf{d}} = \mathbf{d}$, $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$, $\hat{\mathbf{\Phi}}_{\mathbf{x}} = \mathbf{\Phi_x}$). Using the true RTF vector $\mathbf{d}$ in (43), the true spatial coherence matrix $\mathbf{\Gamma}$ in (44), and the true reverberant speech PSD matrix $\mathbf{\Phi_x}$ in (45), the late reverberation PSD is estimated using the ML-based estimator in (24) and the EVD-based PSD estimator in (28). Fig. 2 depicts the true late reverberation PSD Φ_{r}, the ML-based PSD estimate $\hat{\Phi}_{\text{r}}^{\text{ml}}$, and the EVD-based PSD estimate $\hat{\Phi}_{\text{r},\mu}^{\text{evd}}$ averaged over all time frames. Obviously, in this case both PSD estimates are equal to the true late reverberation PSD, confirming the derivations in Section III-C.

Impact of erroneous RTFs ($\hat{\mathbf{d}} \neq \mathbf{d}$, $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$, $\hat{\mathbf{\Phi}}_{\mathbf{x}} = \mathbf{\Phi_x}$). Using the estimated erroneous RTF vector $\hat{\mathbf{d}}$, the true spatial coherence matrix $\mathbf{\Gamma}$ in (44), and the true reverberant speech PSD matrix $\mathbf{\Phi_x}$ in (45), the late reverberation PSD is estimated using the ML-based estimator in (24) and the EVD-based PSD

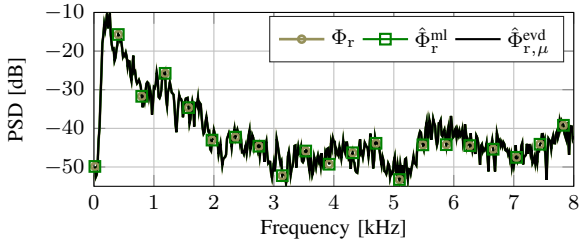

Fig. 2: True and estimated late reverberation PSDs averaged over all time frames when the true RTF vector, spatial coherence matrix, and reverberant speech PSD matrix are known (AS_1, $M = 4$).

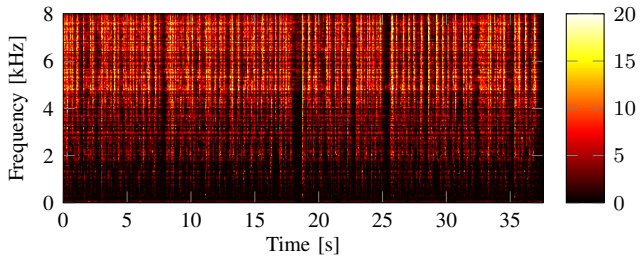

Fig. 3: Difference between the ML-based PSD estimate and the true late reverberation PSD in the presence of RTF estimation errors but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix (AS_1, $M = 4$).

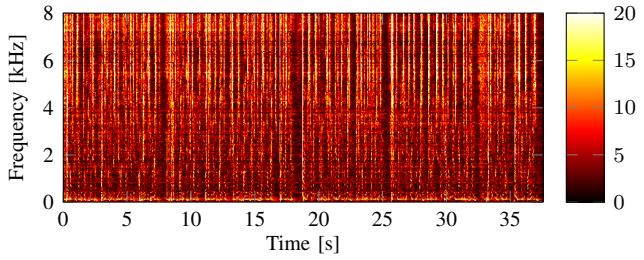

Fig. 4: Difference between the ML-based and EVD-based PSD estimates in the presence of modeling and estimation errors in all quantities (AS_1, $M = 4$).

estimator in (28). Clearly, the EVD-based estimator is not affected by RTF estimation errors and still yields the true late reverberation PSD Φ_{r} as in Fig. 2. Fig. 3 illustrates the difference between the ML-based PSD estimate and the true late reverberation PSD in dB, i.e., $10\log_{10}\hat{\Phi}_{\text{r}}^{\text{ml}} - 10\log_{10}\Phi_{\text{r}}$, for all time-frequency bins. For the sake of clarity, the maximum difference has been limited to 20 dB. It can be observed that in the presence of RTF estimation errors, the difference between the ML-based estimate and the true late reverberation PSD is always larger than or equal to 0, confirming the derivations in Section IV-A. For the considered scenario, it appears that the difference between the ML-based estimate and the true late reverberation PSD is larger at higher frequencies.

Impact of errors in all quantities ($\hat{\mathbf{d}} \neq \mathbf{d}$, $\hat{\mathbf{\Gamma}} \neq \mathbf{\Gamma}$, $\hat{\mathbf{\Phi}}_{\mathbf{x}} \neq \mathbf{\Phi}_{\mathbf{x}}$). Using the estimated erroneous RTF vector $\hat{\mathbf{d}}$, the estimated spatial coherence matrix $\hat{\mathbf{\Gamma}}$ in (48), and the estimated reverberant speech PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ in (49), the late reverberation PSD is estimated using the ML-based estimator in (24) and the EVD-based estimator in (28). Fig. 4 illustrates the difference between the ML-based and EVD-based PSD estimates in dB, i.e., $10\log_{10}\hat{\Phi}_{\text{r}}^{\text{ml}} - 10\log_{10}\hat{\Phi}_{\text{r},\mu}^{\text{evd}}$, for all time-frequency bins. For the sake of clarity, the maximum difference has been limited to 20 dB. It can be observed that in the presence of modeling and estimation errors in all quantities, the ML-based estimate is larger than or equal to the EVD-based PSD estimate, confirming the derivations in Section IV-B.

C. Dereverberation performance for perfectly estimated RTFs

In this section, the dereverberation performance of the MWF using different late reverberation PSD estimators is investigated for the noise-free case assuming that the RTF vector is perfectly estimated, i.e., assuming that the true DOA of the speech source is known. Both acoustic systems and configurations are considered.

The MWF is implemented as an MVDR beamformer $\mathbf{w}_{\text{MVDR}}$ followed by a single-channel Wiener postfilter G applied to the MVDR output, i.e.,

$$\mathbf{w}_{\text{MWF}} = \underbrace{\frac{\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}}{\hat{\mathbf{d}}^H\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}}}_{\mathbf{w}_{\text{MVDR}}} \underbrace{\frac{\hat{\Phi}_{\text{s}}}{\hat{\Phi}_{\text{s}} + \frac{\hat{\Phi}_{\text{r}}}{\hat{\mathbf{d}}^H\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{d}}}}}_{G}, \tag{50}$$

with $\hat{\mathbf{\Gamma}}$ the diffuse spatial coherence matrix computed as in (48), $\hat{\mathbf{d}} = \mathbf{d}$ the RTF vector computed using the true DOA of the speech source, $\hat{\Phi}_{\text{r}}$ the estimated late reverberation PSD, and $\hat{\Phi}_{\text{s}}$ the target signal PSD estimated using the decision directed approach [53]. Using $\hat{\mathbf{d}} = \mathbf{d}$, $\hat{\mathbf{\Gamma}}$, and $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ estimated using recursive averaging as in (49), three different estimates are considered for the late reverberation PSD $\hat{\Phi}_{\text{r}}$, i.e., the ML-based PSD estimate $\hat{\Phi}_{\text{r}}^{\text{ml}}$ in (24), the EVD-based PSD estimate $\hat{\Phi}_{\text{r},\mu}^{\text{evd}}$ in (28), and the EVD-based PSD estimate $\hat{\Phi}_{\text{r},2}^{\text{evd}}$ ($j = 2$) in (27). Note that $\hat{\Phi}_{\text{r},2}^{\text{evd}} \geq \hat{\Phi}_{\text{r},\mu}^{\text{evd}}$, with equality holding for $M = 2$ microphones.

Figs. 5 and 6 depict the ΔPESQ, ΔfSNR, ΔCD, and ΔSTOI values obtained using the MWF with different PSD estimators for both acoustic systems and configurations. For completeness, the performance of the MVDR beamformer implemented as in (50) is also depicted. As expected, the performance of the MVDR beamformer and the MWF improves with increasing number of microphones for both acoustic systems. In addition, for both acoustic systems it can be observed that the MWF using any of the considered late reverberation PSD estimates improves the performance in comparison to the MVDR beamformer. When comparing the performance of the MWF for the different late reverberation PSD estimates, it can be observed that the performance is in general rather similar independently of the used PSD estimate. In terms of ΔPESQ, the ML-based PSD estimate $\hat{\Phi}_{\text{r}}^{\text{ml}}$ yields a slightly better performance for acoustic system AS_1, whereas the EVD-based PSD estimate $\hat{\Phi}_{\text{r},2}^{\text{evd}}$ yields a slightly better performance for acoustic system AS_2. In terms of ΔfSNR, ΔCD, and ΔSTOI, the EVD-based PSD estimate $\hat{\Phi}_{\text{r},\mu}^{\text{evd}}$ yields a slightly better performance than $\hat{\Phi}_{\text{r}}^{\text{ml}}$ and $\hat{\Phi}_{\text{r},2}^{\text{evd}}$ for both acoustic systems.

In summary, these simulation results show the applicability of the EVD-based PSD estimator for dereverberation, yielding a similar or slightly better performance than the state-of-the-art ML-based PSD estimator when the true RTFs are known.

D. Dereverberation performance for erroneous RTFs

In this section, the dereverberation performance of the MWF using different late reverberation PSD estimators is investigated for the noise-free case assuming that the RTF vector is erroneously estimated, i.e., using an erroneous DOA

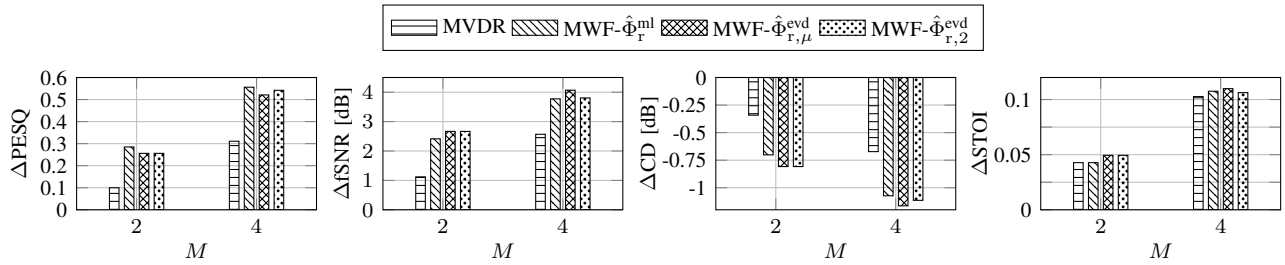

Fig. 5: Dereverberation performance of the MVDR beamformer and the MWF for acoustic system AS_1 using the true RTF vector: (a) ΔPESQ, (b) ΔfSNR, (c) ΔCD, and (d) ΔSTOI.

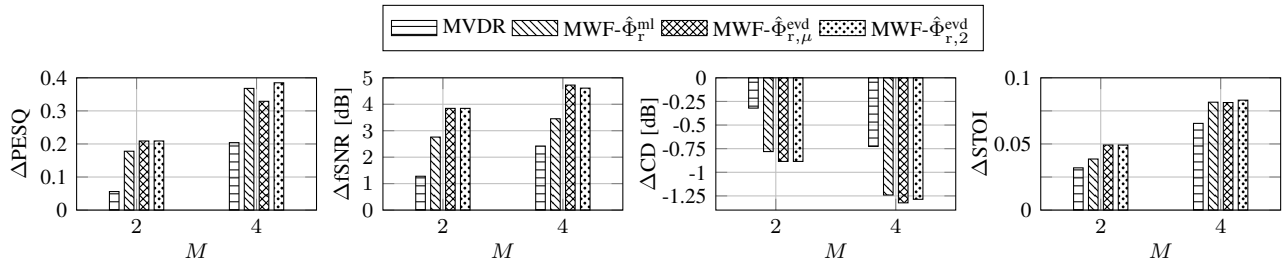

Fig. 6: Dereverberation performance of the MVDR beamformer and the MWF for acoustic system AS_2 using the true RTF vector: (a) ΔPESQ, (b) ΔfSNR, (c) ΔCD, and (d) ΔSTOI.

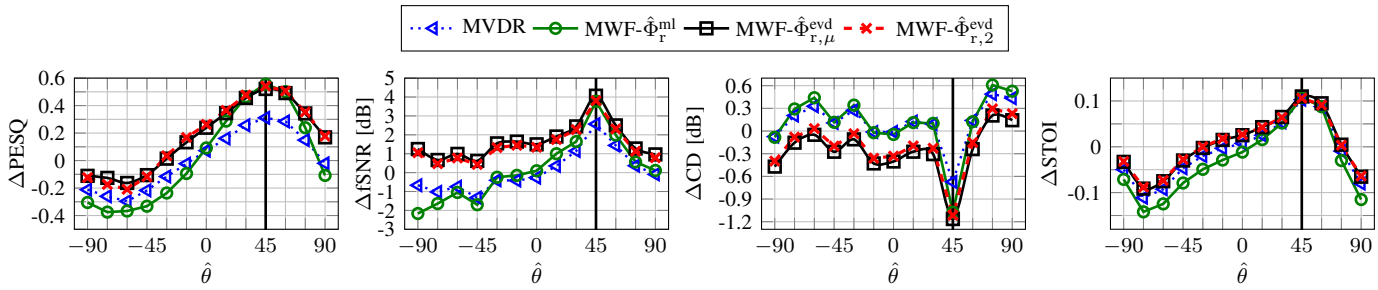

Fig. 7: Dereverberation performance of the MVDR beamformer and the MWF for acoustic system AS_1 with $M = 4$ microphones using erroneous RTF vectors: (a) ΔPESQ, (b) ΔfSNR, (c) ΔCD, and (d) ΔSTOI.

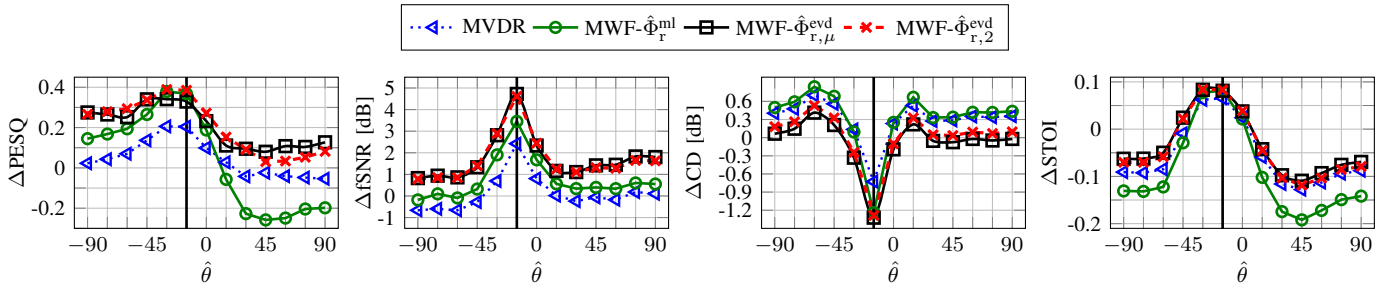

Fig. 8: Dereverberation performance of the MVDR beamformer and the MWF for acoustic system AS_2 with $M = 4$ microphones using erroneous RTF vectors: (a) ΔPESQ, (b) ΔfSNR, (c) ΔCD, and (d) ΔSTOI.

of the speech source. We consider both acoustic systems and $M = 4$ microphones. The MWF is implemented as in (50), with the estimated RTF vector $\hat{\mathbf{d}}$ computed based on several erroneous DOAs

$$\hat{\theta} \in \{-90^\circ,\ -75^\circ,\ \ldots,\ 90^\circ\}, \tag{51}$$

and the remaining quantities computed as in Section V-C. It should be noted that independently of the estimator used for the late reverberation PSD, the MWF implemented as in (50) is sensitive to RTF estimation errors due to the sensitivity of the MVDR beamformer to RTF estimation errors. However, as will be shown, a significantly higher sensitivity of the MWF is observed when the late reverberation PSD estimator is also affected by RTF estimation errors.

Figs. 7 and 8 depict the ΔPESQ, ΔfSNR, ΔCD, and ΔSTOI values obtained using the MWF with different late

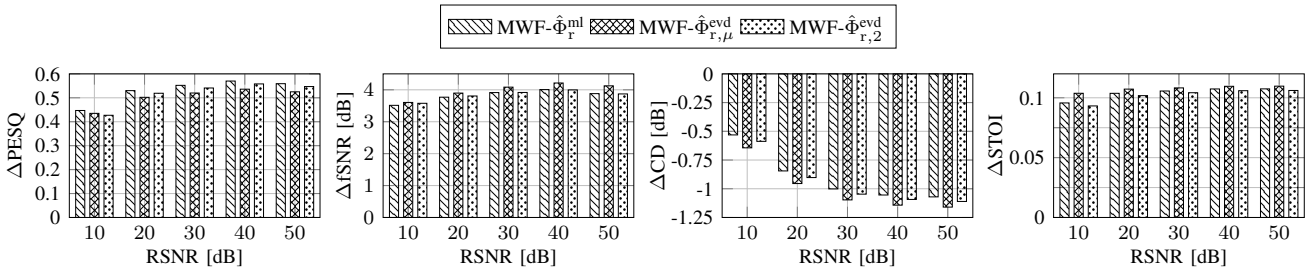

Fig. 9: Dereverberation and noise reduction performance of the MWF in the presence of non-stationary diffuse babble noise (AS_1, $M = 4$).

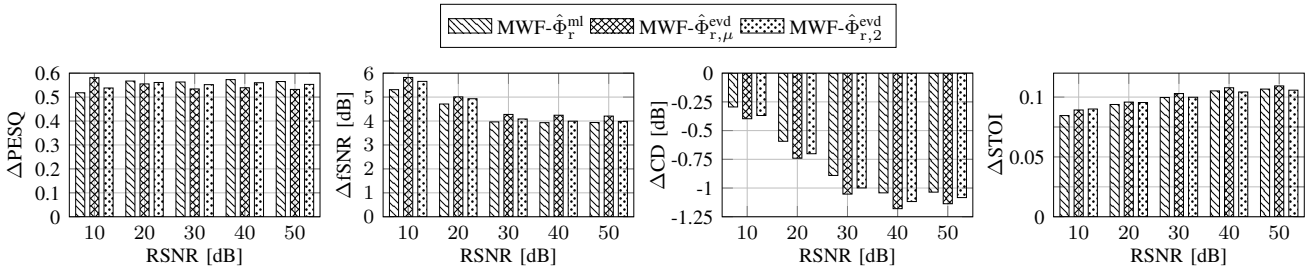

Fig. 10: Dereverberation and noise reduction performance of the MWF in the presence of stationary uncorrelated noise (AS_1, $M = 4$).

reverberation PSD estimators for both acoustic systems in the presence of RTF estimation errors. In addition, the performance obtained using the true RTF vector (i.e., $\hat{\theta} = 45°$ for acoustic system AS_1 and $\hat{\theta} = -15°$ for acoustic system AS_2) is depicted. For completeness, the performance of the MVDR beamformer implemented as in (50) is also presented. As expected, it can be observed that the performance of the MVDR beamformer deteriorates in the presence of RTF estimation errors in terms of all instrumental performance measures. Since the MWF is equivalent to an MVDR beamformer followed by a single-channel Wiener postfilter, cf. (50), it can be observed that RTF estimation errors yield a performance deterioration also for the MWF using any of the considered late reverberation PSD estimates. However, since the ML-based PSD estimate additionally relies on the RTF vector, Figs. 7 and 8 clearly show that the performance of the MWF using the ML-based PSD estimate is substantially lower than using any of the proposed EVD-based PSD estimates. For large DOA estimation errors, the performance of the MWF using the ML-based PSD estimate can even be worse than the performance of the MVDR beamformer, illustrating that PSD estimation errors arising due to RTF estimation errors have a large impact on the MWF performance. When comparing the performance of the MWF using the EVD-based PSD estimates $\hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}$ and $\hat{\Phi}_{\mathrm{r},2}^{\mathrm{evd}}$, Figs. 7 and 8 show that the performance of the MWF is rather similar independently of the EVD-based PSD estimate used. In terms of ΔPESQ, both PSD estimates achieve a very similar performance for acoustic system AS_1, while $\hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}$ appears to achieve a slightly better performance than $\hat{\Phi}_{\mathrm{r},2}^{\mathrm{evd}}$ for acoustic system AS_2 (particularly for $\hat{\theta} \in \{45°, \ldots, 90°\}$). In terms of ΔfSNR and ΔCD, $\hat{\Phi}_{\mathrm{r},\mu}^{\mathrm{evd}}$ appears to achieve a slightly better performance than $\hat{\Phi}_{\mathrm{r},2}^{\mathrm{evd}}$ for both acoustic systems. In terms of ΔSTOI, both PSD estimates achieve a very similar performance for both acoustic systems.

In summary, the presented simulation results show that compared to the ML-based PSD estimator, the EVD-based PSD estimator yields a similar dereverberation performance when the true RTF vector is known and a significantly better dereverberation performance in the presence of RTF estimation errors, making the EVD-based PSD estimator an advantageous PSD estimator to use in realistic reverberant scenarios.

E. Dereverberation and noise reduction performance

In this section, the dereverberation and noise reduction performance of the MWF using the ML-based and EVD-based PSD estimates is investigated for two different additive noise scenarios, i.e., non-stationary spherically diffuse babble noise simulated using [54] and stationary temporally and spatially uncorrelated noise. The broadband reverberant signal-to-noise ratio (RSNR) for both considered noisy scenarios is varied between 10 dB and 50 dB. We consider acoustic system AS_1 with $M = 4$ microphones and compute the RTF vector using the true DOA $\theta = 45°$ of the speech source as in (43).

For the diffuse noise scenario, the ML-based and EVD-based PSD estimators can be readily used to estimate the joint late reverberation and noise PSD by prewhitening the noisy PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}$, such that the MWF can be implemented as in the noise-free scenario in Section V-C. The PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}$ can be estimated from the received microphone signals using recursive averaging, similarly as in (49). The MWF can then be implemented as in (50) using the estimated joint late reverberation and noise PSD, the diffuse spatial coherence matrix $\hat{\mathbf{\Gamma}}$ in (48) (modeling both late reverberation and noise), and the RTF vector $\hat{\mathbf{d}}$ computed using the true DOA.

For the uncorrelated noise scenario, the MWF is implemented as

$$\mathbf{w}_{\text{MWF}} = \underbrace{\frac{\left(\hat{\Phi}_{\text{r}}\hat{\mathbf{\Gamma}} + \hat{\mathbf{\Phi}}_{\mathbf{v}}\right)^{-1}\hat{\mathbf{d}}}{\hat{\mathbf{d}}^H\left(\hat{\Phi}_{\text{r}}\hat{\mathbf{\Gamma}} + \hat{\mathbf{\Phi}}_{\mathbf{v}}\right)^{-1}\hat{\mathbf{d}}}}_{\mathbf{w}_{\text{MVDR}}} \underbrace{\frac{\hat{\Phi}_{\text{s}}}{\hat{\Phi}_{\text{s}} + \left(\hat{\mathbf{d}}^H\left(\hat{\Phi}_{\text{r}}\hat{\mathbf{\Gamma}} + \hat{\mathbf{\Phi}}_{\mathbf{v}}\right)^{-1}\hat{\mathbf{d}}\right)^{-1}}}_{G}, \tag{52}$$

with $\hat{\mathbf{\Phi}}_{\mathbf{v}}$ the estimated noise PSD matrix. Assuming stationary noise, $\hat{\mathbf{\Phi}}_{\mathbf{v}}$ is estimated from L_v noise-only frames (corresponding to 1 s) as

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(k) = \frac{1}{L_v}\sum_{l=0}^{L_v-1}\mathbf{V}(k,l)\mathbf{V}^H(k,l). \tag{53}$$

Furthermore, the PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ required for the late reverberation PSD estimators is computed as

$$\hat{\mathbf{\Phi}}_{\mathbf{x}} = \hat{\mathbf{\Phi}}_{\mathbf{y}} - \hat{\mathbf{\Phi}}_{\mathbf{v}}. \tag{54}$$

Since the reverberant speech PSD matrix in (54) may not be positive semi-definite, particularly at low input RSNRs, the matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ is forced to be positive semi-definite by setting its negative eigenvalues to 0. The MWF is then implemented as in (52) using the estimated late reverberation PSD, the diffuse spatial coherence matrix $\hat{\mathbf{\Gamma}}$ in (48), the noise PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{v}}$ in (53), and the RTF vector $\hat{\mathbf{d}}$ computed using the true DOA.

For different broadband RSNRs, Figs. 9 and 10 depict the ΔPESQ, ΔfSNR, ΔCD, and ΔSTOI values obtained using the MWF with the ML-based and EVD-based PSD estimates for the diffuse and uncorrelated noise scenarios, respectively. It can be observed that in terms of all instrumental performance measures, all PSD estimators yield a similarly large dereverberation and noise reduction performance. In terms of ΔPESQ, the ML-based PSD estimate generally yields a slightly better performance than the EVD-based PSD estimates. In terms of ΔfSNR, ΔCD, and ΔSTOI, the EVD-based PSD estimate $\hat{\Phi}_{\text{r},\mu}^{\text{evd}}$ consistently yields a slightly better performance than $\hat{\Phi}_{\text{r}}^{\text{ml}}$ and $\hat{\Phi}_{\text{r},2}^{\text{evd}}$.

In summary, these simulation results show the applicability of the EVD-based PSD estimator in realistic reverberant and noisy scenarios, yielding a similar or slightly better performance than the state-of-the-art ML-based PSD estimator when the true RTFs are known.

VI. Conclusion

In this paper, the recently proposed EVD-based late reverberation PSD estimator has been analyzed and its estimation accuracy has been analytically compared to a state-of-the-art ML-based PSD estimator. It has been shown that when the true RTFs, late reverberation spatial coherence matrix, and reverberant speech PSD matrix are known, the ML-based and EVD-based PSD estimators are equivalent and yield the true late reverberation PSD. Furthermore, it has been shown that in the presence of RTF estimation errors but perfect knowledge of the spatial coherence matrix and reverberant speech PSD matrix, the ML-based PSD estimate is larger than or equal to the true late reverberation PSD, whereas the EVD-based PSD estimate is still equal to the true late reverberation PSD. Finally, it has been shown that in the presence of modeling and estimation errors in all quantities (which is typically the case in practice), the ML-based PSD estimate is larger than or equal to the EVD-based PSD estimate. Simulation results for several realistic reverberant acoustic scenarios have demonstrated that compared to the ML-based PSD estimator, the EVD-based PSD estimator yields a similar dereverberation performance when the true RTF vector is known and a significantly better dereverberation performance in the presence of RTF estimation errors. In addition, it has been experimentally validated that the EVD-based PSD estimator can also be successfully used in reverberant and noisy scenarios, as long as an estimate of the reverberant speech PSD matrix can be obtained. Conveniently, if the noise can also be modeled as a diffuse sound field, an estimate of the reverberant speech PSD matrix is not required and the EVD-based estimator can be readily used to estimate the joint late reverberation and noise PSD.

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, Jun. 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, Jul. 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, Sep. 2014, pp. 333–337.

[4] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[5] F. Xiong, B. T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-end technologies for robust ASR in reverberant environments – spectral enhancement-based dereverberation and auditory modulation filterbank features," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, Aug. 2015.

[6] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, Eindhoven, Netherlands, Jun. 2007.

[7] P. A. Naylor and N. D. Gaubitch, Eds., *Speech dereverberation*. London, UK: Springer, 2010.

[8] I. Kodrasi, "Dereverberation and noise reduction techniques based on acoustic multi-channel equalization," Ph.D. dissertation, University of Oldenburg, Oldenburg, Germany, Dec. 2015.

[9] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[10] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, Jun. 2014.

[11] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, Apr. 2015.

[12] ——, "Sparsity-promoting acoustic multi-channel equalization techniques," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 7, pp. 1512–1525, Jul. 2017.

[13] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[14] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[15] A. Jukić, T. Van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[16] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[17] 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, Sep. 2001, pp. 31–34.

[18] 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.

[19] 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.

[20] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, Feb. 2015.

[21] A. Kuklasiński, "Multi-channel dereverberation for speech intelligibility improvement in hearing aid applications," Ph.D. dissertation, Aalborg University, Aalborg, Denmark, Sep. 2016.

[22] 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, Sep. 2013.

[23] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. European Signal Processing Conference*, Lisbon, Portugal, Sep. 2014, pp. 61–65.

[24] 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.

[25] 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.

[26] 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.

[27] 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, Sep. 2016.

[28] 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, Sep. 2016, pp. 1123–1127.

[29] 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.

[30] 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, Sep. 2009.

[31] 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, Sep. 2016.

[32] 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.

[33] A. Kuklasinski, S. Doclo, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberant and noisy conditions," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 599–603.

[34] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Multichannel Wiener filter for speech dereverberation in hearing aids - sensitivity to DoA errors," in *Proc. AES 60th Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.

[35] A. Kuklasiński and J. Jensen, "Multichannel Wiener filters in binaural and bilateral hearing aids – speech intelligibility improvement and robustness to DoA errors," *Journal of Audio Engineering Society*, vol. 65, no. 1/2, pp. 8–16, Jan./Feb. 2017.

[36] 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.

[37] ——, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, San Francisco, USA, Mar. 2017, pp. 116–120.

[38] 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.

[39] B. D. Rao and K. V. S. Hari, "Performance analysis of Root-Music," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1939–1949, Dec. 1989.

[40] B. Ottersten, M. Viberg, and T. Kailath, "Performance analysis of the total least squares ESPRIT algorithm," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1122–1135, May 1991.

[41] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Performance evaluation of sparse source separation and DOA estimation with observation vector clustering in reverberant environments," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Paris, France, Sep. 2006.

[42] J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen, "DOA estimation of audio sources in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 176–180.

[43] A. M. Davie and A. J. Stothers, "Improved bound for complexity of matrix multiplication," *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, vol. 143, no. 2, pp. 351–369, Apr. 2013.

[44] G. Golub and C. Van Loan, *Matrix Computations*. Baltimore, USA: The John Hopkins University Press, 1996.

[45] J. M. Ortega and H. F. Kaiser, "The LL^T and QR methods for symmetric diagonal matrices," *The Computer Journal*, vol. 6, no. 1, pp. 99–101, Jan. 1963.

[46] B. Yang, "Projection approximation subspace tracking," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 95–107, Jan. 1995.

[47] 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.

[48] 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.

[49] 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 P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[50] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

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

[52] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Texas, USA, Mar. 2010, pp. 4214–4217.

[53] 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.

[54] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

Ina Kodrasi (S'11-M'16) received the Master of Science degree in Communications, Systems and Electronics in 2010 from Jacobs University Bremen, Germany and the PhD degree in 2015 from the University of Oldenburg, Germany. From 2010 to 2015 she worked as an early stage researcher at the Signal Processing Group of the University of Oldenburg. From 2010 to 2011 she was also with the Fraunhofer Institute for Digital Media Technology (IDMT), Project group Hearing, Speech and Audio Technology in Oldenburg where she worked on microphone array beamforming. From 2015 to 2017 she was a postdoctoral researcher at the Signal Processing Group of the University of Oldenburg in the field of speech dereverberation and noise reduction. Since December 2017 she is a postdoctoral researcher at the Idiap Research Institute in Martigny, Switzerland working in the field of signal processing for clinical applications.

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, Belgium, in 1997 and 2003. From 2003 to 2007 he was a Postdoctoral Fellow with the Research Foundation Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Cognitive Systems Laboratory (McMaster University, Canada). From 2007 to 2009 he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group in Leuven, Belgium. Since 2009 he is a full professor at the University of Oldenburg, Germany, and scientific advisor for the project group Hearing, Speech and Audio Technology of the 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. Prof. Doclo received 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 Jingdong Chen, Jacob Benesty, Arden Huang). He is 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. Prof. Doclo served as guest editor for several special issues (IEEE Signal Processing Magazine, Elsevier Signal Processing) and is associate editor for IEEE/ACM Transactions on Audio, Speech and Language Processing and EURASIP Journal on Advances in Signal Processing.