Square Root-Based Multi-Source Early PSD Estimation and Recursive RETF Update in Reverberant Environments by Means of the Orthogonal Procrustes Problem

Thomas Dietzen, Simon Doclo, *Senior Member, IEEE*, Marc Moonen, *Fellow, IEEE*, and Toon van Waterschoot, *Member, IEEE*

***Abstract*—Multi-channel short-time Fourier transform (STFT) domain-based processing of reverberant microphone signals commonly relies on power-spectral-density (PSD) estimates of early source images, where early refers to reflections contained within the same STFT frame. State-of-the-art approaches to multi-source early PSD estimation, given an estimate of the associated relative early transfer functions (RETFs), conventionally minimize the approximation error defined with respect to the early correlation matrix, requiring non-negative inequality constraints on the PSDs. Instead, we here propose to factorize the early correlation matrix and minimize the approximation error defined with respect to the early-correlation-matrix square root. The proposed minimization problem—constituting a generalization of the so-called orthogonal Procrustes problem—seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, whereby non-negative inequality constraints become redundant. A solution is obtained iteratively, requiring one singular value decomposition (SVD) per iteration. The estimated unitary matrix and early PSD square roots further allow to recursively update the RETF estimate, which is not inherently possible in the conventional approach. An estimate of the said early-correlation-matrix square root itself is obtained by means of the generalized eigenvalue decomposition (GEVD), where we further propose to restore non-stationarities by desmoothing the generalized eigenvalues in order to compensate for inevitable recursive averaging. Simulation results indicate fast convergence of the proposed multi-source early PSD estimation approach in only one iteration if initialized appropriately, and better performance as compared to the conventional approach. A MATLAB implementation is available.**

***Index Terms*—Early PSD estimation, RETF estimation, orthogonal Procrustes problem, unitary constraint, singular value decomposition (SVD), generalized eigenvalue decomposition (GEVD).**

Manuscript received June 15, 2019; revised December 24, 2019; accepted January 7, 2020. Date of publication January 15, 2020; date of current version January 30, 2020. This work was supported in part by the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal fund C2-16-00449; VLAIO O&O Project no. HBC.2017.0358; EU FP7-PEOPLE Marie Curie Initial Training Network funded by the European Commission under Grant Agreement no. 316969; the European Union's Horizon 2020 research and innovation program/ERC Consolidator Grant no. 773268. This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Huseyin Hacihabiboglu. *(Corresponding author: Thomas Dietzen.)*

T. Dietzen and T. van Waterschoot are with the KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics and ETC Technology Cluster Electrical Engineering, 3001 Leuven, Belgium (e-mail: thomas.dietzen@esat.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be).

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

M. Moonen is with the KU Leuven, ESAT, STADIUS, 3001 Leuven, Belgium (e-mail: marc.moonen@esat.kuleuven.be).

Digital Object Identifier 10.1109/TASLP.2020.2966891

I. Introduction

In many multi-microphone signal processing applications, the recorded microphone signals constitute a mixture of several spatially diverse components, originating from different sources, bearing reverberation and noise. As far as speech is concerned, early reflections are perceived jointly with the direct component and are said to colorize and reinforce it, while late reverberant components deteriorate the perceived quality and intelligibility [1]. In order to process the mixture, e.g., for the purpose of speech enhancement [2]–[5], many techniques heavily rely on power spectral density (PSD) estimates [6]–[16] of the various mixture components or the direct-to-reverberant ratio (DRR) [17], [18], for instance in the spectral gain computation of the multi-channel Wiener filter (MWF) [4].

In recent years, a number of multi-microphone approaches to PSD estimation have been proposed, which rely on a spatial correlation matrix model in the short-time Fourier transform (STFT) domain [6]–[18]. In order to estimate the PSDs of the mixture components, i.e. the early speech PSDs, late reverberant PSDs, and/or noise PSDs, the associated spatial parameters of the correlation matrix model are assumed to be known or estimated beforehand, namely the direction(s) of arrival (DoA(s)) or alternatively the relative early transfer function(s)[1] (RETF(s)) associated to the source(s) [6]–[14], [18], the spatial coherence matrix of the noise and/or the late reverberant component, where in particular the latter is commonly modeled as a spatially diffuse sound field [7]–[9], [11]–[19]. It should be noted that the majority of these approaches consider a single source [6], [9], [11]–[15], [18], while only some consider multiple sources [7], [8], [10], [16], which is the focus of this paper.

[1] The RETF vector can be thought of as a generalization of the DoA steering vector in reverberant environments, which models level and phase differences across microphones due to both the direct component and early reflections.

Given such spatial knowledge, different estimation approaches may be taken. In [9], [11], the early speech and late reverberant PSD estimates are obtained by maximum-likelihood estimation, where in [9], both are estimated jointly, and in [11], the late reverberant PSD estimation relies on blocking the early speech component. Other estimators seeking the PSD of a particular mixture component rely on Frobenius-norm minimization of the approximation error defined with respect to the associated correlation matrix component [6]–[8], [10], [12], [13], [16], [18]. Specifically, in [6], the speech PSD is estimated by minimizing the approximation error to the speech-only correlation matrix component (while reverberation is not considered). In a similar manner, in [7], considering multiple sources, the early PSDs are estimated from the early correlation matrix component. In [8], the late reverberant PSD is estimated from a blocking-based correlation matrix, generated by blocking the direct components, while the multiple early PSDs are estimated according to [7]. Likewise, one may also jointly estimate the PSDs of multiple mixture components, i.e. one may jointly estimate early speech PSD(s), the late reverberant PSD, and/or the noise PSD(s) [10], [12], [13], [18]. In [16], joint estimation of all mixture component PSDs and the RETFs is proposed by jointly minimizing a number of approximation errors defined over several frames, during which the RETFs are assumed to be stationary. Note that the PSD estimates based on these optimization problems are not inherently guaranteed to be non-negative, requiring either non-negative thresholding, or, alternatively, non-negative inequality constraints. In [15] instead, the estimation of the late reverberant PSD is based on a subspace decomposition, outperforming the late reverberant PSD estimators in [8], [9], [11].

In this contribution, we are mainly concerned with early PSD estimation and recursive RETF updates for multiple sources in reverberant environments, given initial estimates of the associated RETFs. Instead of minimizing the approximation error defined with respect to the early correlation matrix as in the manner of [7], [10], [12], [13], however, we propose to factorize the early correlation matrix and minimize the approximation error defined with respect to the early-correlation-matrix square root. Instead of directly estimating the early PSDs, the proposed minimization problem seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, making non-negative thresholding or non-negative inequality constraints redundant. The proposed minimization problem constitutes a generalization [23] of the so-called orthogonal Procrustes problem [24], [25] and may be solved iteratively, requiring one singular value decomposition (SVD) per iteration. The estimated unitary matrix and early PSD square roots further allow us to recursively update the RETF estimate, which is not inherently possible in the conventional approach. An estimate of the said early-correlation-matrix square root itself is obtained from an estimate of the microphone signal correlation matrix and the diffuse coherence matrix by means of the generalized eigenvalue decomposition (GEVD). Hereat, in order to compensate for the inevitable recursive averaging in the microphone-signal-correlation-matrix estimation, we further propose to restore non-stationarities by desmoothing the generalized eigenvalues. Simulation results indicate fast convergence of the proposed multi-source early PSD estimation approach in only one iteration if initialized appropriately, and better performance as compared to the conventional approach in terms of the relative squared PSD estimation error and the signal-to-interference ratio [26] measuring the source-component separation. A MATLAB implementation is available at [27]. An application of the proposed algorithm in a Kalman filter-based speech enhancement approach can be found in [28].

The remainder of this paper is organized as follows. In Section II, we introduce the signal model. Given an estimate of the early correlation matrix component, some state-of-the-art approaches to early PSD estimation are reviewed in Section III, while the proposed approach, given an estimate of the early-correlation-matrix square root, is presented in Section IV. In Section V, we discuss the estimation of the required early correlation matrix component and its factorization. The proposed approach is evaluated in Section VI, followed by a conclusion in Section VII.

II. Signal Model

Throughout the paper, we use the following notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}$ and $\mathbf{0}$ denote identity and zero matrices, $\mathbf{i}$ and $\mathbf{1}$ denote the first column of $\mathbf{I}$ and a vector of ones, respectively, $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathrm{E}[\mathbf{A}]$, and $\hat{\mathbf{A}}$ denote the transpose, the complex conjugate transpose or Hermitian, the expected value and an estimate of a matrix $\mathbf{A}$. The operation $\mathrm{diag}[\mathbf{A}]$ creates a column vector from the diagonal elements of a square matrix $\mathbf{A}$, $\mathrm{Diag}[\mathbf{a}]$ and $\mathrm{Diag}[\mathbf{a}^T]$ create a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal, $\mathrm{Diagg}[\mathbf{A}] = \mathrm{Diag}\big[\mathrm{diag}[\mathbf{A}]\big]$ zeros the off-diagonal elements of $\mathbf{A}$, and $\mathrm{tr}[\mathbf{A}]$ denotes the trace of $\mathbf{A}$. For non-negative $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{a}^{1/2} \in \mathbb{C}^N$ and $\mathbf{a}^{H/2} = (\mathbf{a}^{1/2})^H$ denote a complex vector with arbitrary complex argument that satisfy $\mathrm{Diag}[\mathbf{a}^{H/2}]\mathbf{a}^{1/2} = \mathbf{a}$, and hence $\big|\mathbf{a}^{1/2}\big| = \sqrt{\mathbf{a}}$, with absolute value and non-negative square-root applied element-wise. The operation $\max[\mathbf{a}_1, \mathbf{a}_2]$ returns a vector of the element-wise maxima of $\mathbf{a}_1$ and $\mathbf{a}_2$. $\|\mathbf{A}\|_F$ denotes the Frobenius norm of $\mathbf{A}$, whereas $\|\mathbf{a}\|_2$ denotes the Euclidian norm of $\mathbf{a}$. Row i and column j of $\mathbf{A}$ are denoted as $[\mathbf{A}]_{i,:}$ and $[\mathbf{A}]_{:,j}$, respectively, the element at their intersection as $[\mathbf{A}]_{i,j}$, and submatrices spanning rows i_1 to i_2 or columns j_1 to j_2 as $[\mathbf{A}]_{i_1:i_2,:}$ and $[\mathbf{A}]_{:,j_1:j_2}$, respectively. $\Re[a]$ and $\Im[a]$ denote the real and imaginary part of $a \in \mathbb{C}$.

In the STFT domain, with l and k indexing the frame and the frequency bin, respectively, let $x_m(l, k)$ with $m = 1, \ldots, M$ denote the mth microphone signal, with M the number of microphones. In the following, we treat all frequency bins independently and hence omit the frequency index. We define the stacked microphone signal vector $\mathbf{x}(l) \in \mathbb{C}^M$,

$$\mathbf{x}(l) = \begin{pmatrix} x_1(l) & \cdots & x_M(l) \end{pmatrix}^T \tag{1}$$

composed of the reverberant signal components $\mathbf{x}_n(l)$ with $n = 1, \ldots, N$ originating from N point sources, defined equivalently to (1), i.e.

$$\mathbf{x}(l) = \sum_{n=1}^{N} \mathbf{x}_n(l). \tag{2}$$

Each reverberant signal component $\mathbf{x}_n(l)$ may be decomposed into the early component $\mathbf{x}_{n|e}(l)$ containing the direct component and early reflections and the late reverberant component $\mathbf{x}_{n|\ell}(l)$ containing late reflections, i.e.

$$\mathbf{x}_n(l) = \mathbf{x}_{n|e}(l) + \mathbf{x}_{n|\ell}(l), \tag{3}$$

which are assumed to have distinct spatial properties as outlined below. Early reflections are assumed to arrive within the same frame, with the early components in $\mathbf{x}_{n|e}(l)$ related by the RETF in $\mathbf{h}_n(l) \in \mathbb{C}^M$, i.e.

$$\mathbf{x}_{n|e}(l) = \mathbf{h}_n(l) s_n(l). \tag{4}$$

Here, without loss of generality, the RETF $\mathbf{h}_n(l)$ is assumed to be relative to the first microphone, i.e. $\mathbf{i}^T\mathbf{h}_n(l) = [\mathbf{h}_n(l)]_1 = 1$, and $s_n(l) = [\mathbf{x}_{n|e}(l)]_1$ denotes the early component in the first microphone originating from the nth source, in the following referred to as early source image. We define the stacked RETF matrix $\mathbf{H}(l) \in \mathbb{C}^{M\times N}$, yielding

$$\mathbf{H}(l) = \begin{pmatrix}\mathbf{h}_1(l) & \cdots & \mathbf{h}_N(l)\end{pmatrix}, \tag{5}$$

$$\mathbf{i}^T\mathbf{H}(l) = [\mathbf{H}(l)]_{1,:} = \mathbf{1}^T. \tag{6}$$

Similarly, we stack $s_n(l)$ into $\mathbf{s}(l) \in \mathbb{C}^N$, i.e.

$$\mathbf{s}(l) = \begin{pmatrix}s_1(l) & \cdots & s_N(l)\end{pmatrix}^T, \tag{7}$$

such that the sum of the early components $\mathbf{x}_{n|e}(l)$ is expressed more compactly as

$$\sum_{n=1}^{N} \mathbf{x}_{n|e}(l) = \mathbf{H}(l)\mathbf{s}(l). \tag{8}$$

Further, we assume that $\mathbf{x}_{n|e}(l)$ and $\mathbf{x}_{n|\ell}(l)$ are mutually uncorrelated within frame l. Let $\mathbf{\Psi}_x(l) = \mathrm{E}[\mathbf{x}(l)\mathbf{x}^H(l)] \in \mathbb{C}^{M\times M}$ denote the microphone signal correlation matrix, and let the early and late reverberant correlation matrix $\mathbf{\Psi}_{x_e}(l)$ and $\mathbf{\Psi}_{x_\ell}(l)$ be similarly defined. With (3)–(8), we then find

$$\mathbf{\Psi}_x(l) = \mathbf{\Psi}_{x_e}(l) + \mathbf{\Psi}_{x_\ell}(l), \tag{9}$$

wherein $\mathbf{\Psi}_{x_e}(l)$ generally has rank N and is expressed by

$$\boxed{\mathbf{\Psi}_{x_e}(l) = \mathbf{H}(l)\mathbf{\Phi}_s(l)\mathbf{H}^H(l),} \tag{10}$$

$$\mathbf{\Phi}_s(l) = \mathrm{Diag}[\boldsymbol{\varphi}_s(l)], \tag{11}$$

$$\boldsymbol{\varphi}_s(l) = (\varphi_{s_1}(l) \cdots \varphi_{s_N}(l))^T, \tag{12}$$

with $\varphi_{s_n}(l)$ denoting the PSD of the early source image $s_n(l)$. Note that applying (6) to (10)–(11) while using $\mathbf{1}^T\mathbf{\Phi}_s(l)\mathbf{1} = \mathbf{1}^T\boldsymbol{\varphi}_s(l)$, we find that

$$\boxed{\mathbf{i}^T\mathbf{\Psi}_{x_e}(l)\mathbf{i} = [\mathbf{\Psi}_{x_e}(l)]_{1,1} = \mathbf{1}^T\boldsymbol{\varphi}_s(l),} \tag{13}$$

i.e. the sum of the early PSDs $\varphi_{s_n}(l)$ equals $[\mathbf{\Psi}_{x_e}(l)]_{1,1}$. Assuming that $\mathbf{x}_{n|\ell}(l)$ may be modeled as diffuse [7]–[9], [11]–[16], [19] with coherence matrix $\mathbf{\Gamma} \in \mathbb{C}^{M\times M}$, which can be computed from the microphone array geometry [19] and is therefore considered to be known in the remainder, we write $\mathbf{\Psi}_{x_\ell}(l)$ as

$$\mathbf{\Psi}_{x_\ell}(l) = \varphi_{x_\ell}(l)\mathbf{\Gamma}, \tag{14}$$

$$\text{with} \quad \varphi_{x_\ell}(l) = \sum_{n=1}^{N} \varphi_{x_{n|\ell}}(l), \tag{15}$$

and $\varphi_{x_{n|\ell}}(l)$ denoting the PSD of the late reverberant component $\mathbf{x}_{n|\ell}(l)$. The PSDs $\boldsymbol{\varphi}_s(l)$ and $\varphi_{x_\ell}(l)$ may be highly non-stationary, especially if the point sources are speech sources, while the associated coherence matrices $\mathbf{h}_n(l)\mathbf{h}_n^H(l)$ and $\mathbf{\Gamma}$ are commonly assumed to be comparably slowly time-varying or even time-invariant.

Note that with (14)–(15), one could easily include further diffuse components, e.g., diffuse babble noise, without formally changing the signal model. However, since in this paper, we are mainly concerned with the estimation of the early PSDs $\boldsymbol{\varphi}_s(l)$ and the recursive updating of the estimate of the RETFs $\mathbf{H}(l)$, we restrict the discussion and simulations, cf. Section VI, to the example of late reverberation for the sake of conciseness.

Further, note that while the above signal model is commonly and effectively used [7]–[9], [11]–[16] due to its simplicity, it may be said to be deficient in a number of aspects. The assumption that $\mathbf{x}_{n|e}(l)$ and $\mathbf{x}_{n|\ell}(l)$ are mutually uncorrelated within frame l may be violated due to overlapping windows in the STFT-processing or source signals remaining correlated over several frames. The assumption that $\mathbf{\Psi}_{x_e}(l)$ in (10) has rank N implicitly relies on the assumption that the frequency bins can be treated independently, ignoring cross-bin dependencies [29]. Finally, related to that, there may be components that can be modeled neither by the rank-N component $\mathbf{\Psi}_{x_e}(l)$ in (10) nor by the diffuse component $\mathbf{\Psi}_{x_\ell}(l)$ in (14), depending on the geometry and physical properties of the acoustic environment.

In the remainder, as we mostly consider the single frame l only, we also drop the frame index for conciseness and refer back to it only where necessary, namely when we differentiate the frames l and $l-1$ in recursive equations.

III. Early PSD Estimation Based on the Early Correlation Matrix

In this section, we discuss some state-of-the-art approaches [7], [10], [12], [13], [16] to the estimation of the early PSDs $\boldsymbol{\varphi}_s$ based on the signal model in (10)–(13). In the following, we refer to (10)–(13) as the conventional signal model. We develop our discussion from the premise that estimates $\hat{\mathbf{\Psi}}_{x_e}$ and $\hat{\mathbf{H}}$ of the early correlation matrix $\mathbf{\Psi}_{x_e}$ and the RETFs $\mathbf{H}$ in (10) are readily available. Throughout the paper, despite being irrelevant to the approaches discussed in this section, we consider $\hat{\mathbf{\Psi}}_{x_e}$ to generally have rank N, similar to $\mathbf{\Psi}_{x_e}$. A rank-N estimator of $\mathbf{\Psi}_{x_e}$ is described in Section V. Further, we assume that $\hat{\mathbf{H}}$ satisfies $\mathbf{i}^T\hat{\mathbf{H}} = \mathbf{1}^T$, cf. $\mathbf{H}$ in (6).

Given the estimates of an early correlation matrix $\hat{\mathbf{\Psi}}_{x_e}$ and the therein superimposed coherence matrices $\hat{\mathbf{h}}_n\hat{\mathbf{h}}_n^H$, one can estimate the associated PSDs φ_{s_n}, cf. (5), (10)–(11), as described in [7], [10], [12], [13].[2] Adopting this approach, we define the

[2] In [7] as in our case, point-source coherence matrices of rank one are considered, while in [10], [12], [13], without rendering a difference in the principle approach, general coherence matrices are considered.

approximation error as a function of $\boldsymbol{\varphi}_s$ as

$$\mathbf{E}_c(\boldsymbol{\varphi}_s) = \hat{\boldsymbol{\Psi}}_{x_e} - \hat{\mathbf{H}}\,\mathrm{Diag}[\boldsymbol{\varphi}_s]\hat{\mathbf{H}}^H, \tag{16}$$

where the subscript c stands for conventional. The early PSDs $\boldsymbol{\varphi}_s$ can then be estimated by Frobenius-norm minimization of the approximation error followed by non-negative thresholding, i.e.

$$\hat{\boldsymbol{\varphi}}'_s = \arg\min_{\boldsymbol{\varphi}_s} \left\|\mathbf{E}_c(\boldsymbol{\varphi}_s)\right\|_F^2, \tag{17}$$

$$\hat{\boldsymbol{\varphi}}_s = \max[\hat{\boldsymbol{\varphi}}'_s, \, \mathbf{0}]. \tag{18}$$

The non-negative thresholding in (18) is necessary as the elements of $\hat{\boldsymbol{\varphi}}'_s$ in (17) may in fact be negative, conflicting with the notion of $\boldsymbol{\varphi}_s$ being a vector of PSDs. If $\hat{\mathbf{H}}^H\hat{\mathbf{H}}$ has full rank, which (without sufficiency) requires $N \leq M$, the problem in (17) has a unique solution given by

$$\hat{\boldsymbol{\varphi}}'_s = \mathbf{A}_{c_0}^{-1}\mathbf{b}_{c_0}, \tag{19}$$

where $\mathbf{A}_{c_0} \in \mathbb{R}^{N\times N}$ and $\mathbf{b}_{c_0} \in \mathbb{R}^N$ are defined by

$$[\mathbf{A}_{c_0}]_{n,n'} = |\hat{\mathbf{h}}_n^H\hat{\mathbf{h}}_{n'}|^2, \tag{20}$$

$$\mathbf{b}_{c_0} = \mathrm{diag}[\hat{\mathbf{H}}^H\hat{\boldsymbol{\Psi}}_{x_e}\hat{\mathbf{H}}]. \tag{21}$$

Alternatively, instead of simple thresholding after solving (17), one can solve the minimization problem subject to the non-negative inequality constraint $\boldsymbol{\varphi}_s \geq \mathbf{0}$, as proposed in [16]. In addition to this, one can further impose a soft constraint on $\mathbf{1}^T\boldsymbol{\varphi}_s$ corresponding to (13), i.e. one can define the soft-constraint error as a function of $\boldsymbol{\varphi}_s$,

$$\begin{aligned} e_c(\boldsymbol{\varphi}_s) &= [\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1} - \mathbf{1}^T\boldsymbol{\varphi}_s \\ &= [\mathbf{E}_c(\boldsymbol{\varphi}_s)]_{1,1}. \end{aligned} \tag{22}$$

Note that $e_c(\boldsymbol{\varphi}_s)$ in (22) is independent of $\hat{\mathbf{H}}$, and so its penalization may be useful in case of RETF estimation errors. The resulting minimization problem can then be written as

$$\boxed{\begin{aligned} \hat{\boldsymbol{\varphi}}_s = \arg\min_{\boldsymbol{\varphi}_s} &\left\|\mathbf{E}_c(\boldsymbol{\varphi}_s)\right\|_F^2 + \alpha\left|e_c(\boldsymbol{\varphi}_s)\right|^2 \\ \text{s.t.} \quad &\boldsymbol{\varphi}_s \geq \mathbf{0}, \end{aligned}} \tag{23}$$

where α is the penalty factor. For $\alpha \to \infty$, a hard constraint $\mathbf{1}^T\boldsymbol{\varphi}_s = [\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}$ is introduced, which is however not desirable due to potential estimation errors in $[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}$. Note that in [16], instead of a soft constraint, a box constraint on $\mathbf{1}^T\boldsymbol{\varphi}_s$ has been used. For the sake of comparison to the algorithm proposed in Section IV, however, we restrict our discussion to the soft constraint. The problem in (23) is convex, but does not have a closed-form solution due to the inequality constraint $\boldsymbol{\varphi}_s \geq \mathbf{0}$. A well-suited but computationally simple solver for problems of this kind is the proximal gradient method [30], [31], which obtains a solution by iterating the below set of equations until convergence is reached,

$$\hat{\boldsymbol{\varphi}}_s'^{(i)} = \hat{\boldsymbol{\varphi}}_s^{(i-1)} + \mu\big(\mathbf{b}_c - \mathbf{A}_c\hat{\boldsymbol{\varphi}}_s^{(i-1)}\big), \tag{24}$$

$$\hat{\boldsymbol{\varphi}}_s^{(i)} = \max[\hat{\boldsymbol{\varphi}}_s'^{(i)}, \mathbf{0}], \tag{25}$$

where i is the iteration index, μ the step-size, and $\mathbf{b}_c - \mathbf{A}_c\hat{\boldsymbol{\varphi}}_s^{(i-1)}$ the gradient with $\mathbf{A}_c \in \mathbb{R}^{N\times N}$ and $\mathbf{b}_c \in \mathbb{R}^N$ defined by

$$\mathbf{A}_c = \mathbf{A}_{c_0} + \alpha\mathbf{1}\mathbf{1}^T, \tag{26}$$

$$\mathbf{b}_c = \mathbf{b}_{c_0} + \alpha[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}\mathbf{1}, \tag{27}$$

and $\mathbf{A}_{c_0}$ and $\mathbf{b}_{c_0}$ defined in (20)–(21). As initial value, it is straight-forward to choose $\hat{\boldsymbol{\varphi}}_s^{(0)} = \mathbf{A}_c^{-1}\mathbf{b}_c$, which yields the global minimum if $\hat{\boldsymbol{\varphi}}_s^{(0)} \geq \mathbf{0}$. In this case therefore, convergence is reached after one iteration of (24)–(25). In any case, for $\hat{\boldsymbol{\varphi}}_s^{(0)} = \mathbf{A}_c^{-1}\mathbf{b}_c$ and $\alpha = 0$, the estimate obtained after one iteration of (24)–(25) corresponds to the estimate defined by (17)–(19). We therefore use (23) as a reference for comparison in the remainder. In the following, we refer to (23) as the conventional minimization problem (conventional MP).

IV. Early PSD Estimation and Recursive RETF Update Based on the Early-Correlation- Matrix Square Root

In this section, in order to estimate the early PSDs $\boldsymbol{\varphi}_s$, instead of defining the approximation error to be minimized with respect to $\hat{\boldsymbol{\Psi}}_{x_e}$ as in (16), we propose to define the approximation error with respect to the square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} \in \mathbb{C}^{M\times N}$ of $\hat{\boldsymbol{\Psi}}_{x_e}$, satisfying $\hat{\boldsymbol{\Psi}}_{x_e} = \hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\hat{\boldsymbol{\Psi}}_{x_e}^{H/2}$. As to be shown, instead of directly estimating the diagonal of $\boldsymbol{\Phi}_s = \mathrm{Diag}[\boldsymbol{\varphi}_s]$, the resulting minimization problem now consists in estimating a unitary matrix $\boldsymbol{\Omega} \in \mathbb{C}^{N\times N}$ and the diagonal of $\boldsymbol{\Phi}_s^{1/2} = \mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$, which constitutes a generalization [23] of the so-called orthogonal Procrustes problem [24], [25]. Since the early PSDs herein are represented by $\boldsymbol{\varphi}_s = \mathrm{Diag}[\boldsymbol{\varphi}_s^{H/2}]\boldsymbol{\varphi}_s^{1/2}$, the corresponding estimate $\hat{\boldsymbol{\varphi}}_s$ is guaranteed to be non-negative, such that a non-negative inequality constraint as in (23) is not required. Further, we show that the obtained estimates $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ can be used to recursively update the RETF estimate $\hat{\mathbf{H}}$, which is not inherently possible from the estimate $\hat{\boldsymbol{\varphi}}_s$ given by (23).

In Section IV-A, as a pre-requisite to our derivation, we discuss the factorization of the conventional signal model in (10)–(13), yielding the square-root signal model. In Section IV-B, based upon the square-root signal model and given the estimates $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$, we then define and solve the square-root minimization problem (square-root MP). In Section IV-C, we discuss the recursive updating of the RETF estimate $\hat{\mathbf{H}}$.

A. Early-Correlation-Matrix Factorization

We consider the factorization of the rank-N matrices on both sides of (10). On the left-hand side of (10), we define the square root $\boldsymbol{\Psi}_{x_e}^{1/2} \in \mathbb{C}^{M\times N}$ such that $\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Psi}_{x_e}^{H/2} = \boldsymbol{\Psi}_{x_e}$. Note that the product is invariant to right-multiplication of a particular square root with any unitary matrix, and so $\boldsymbol{\Psi}_{x_e}^{1/2}$ is not unique. On the right-hand side of (10), with $\boldsymbol{\varphi}_s^{1/2} \in \mathbb{C}^N$ and $\mathrm{Diag}[\boldsymbol{\varphi}_s^{H/2}]\boldsymbol{\varphi}_s^{1/2} = \boldsymbol{\varphi}_s$, we define the square roots $\boldsymbol{\Phi}_s^{1/2} = \mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$ and $\mathbf{H}\boldsymbol{\Phi}_s^{1/2}$ such that $\mathbf{H}\boldsymbol{\Phi}_s^{1/2}\boldsymbol{\Phi}_s^{H/2}\mathbf{H}^H = \mathbf{H}\boldsymbol{\Phi}_s\mathbf{H}^H$. Note that while the magnitude of the elements in $\boldsymbol{\varphi}_s^{1/2} \in \mathbb{C}^N$

is well-defined, namely by $\left|\boldsymbol{\varphi}_s^{1/2}\right| = \sqrt{\boldsymbol{\varphi}_s}$, their complex argument can be chosen arbitrarily, and so $\boldsymbol{\Phi}_s^{1/2}$ is not unique. The non-uniqueness of both square roots implies that while their respective products on both sides of (10) coincide, the said square roots themselves generally do not, i.e. we have $\boldsymbol{\Psi}_{x_e}^{1/2} \neq \mathbf{H}\boldsymbol{\Phi}_s^{1/2}$. Hence, for a particular $\boldsymbol{\Psi}_{x_e}^{1/2}$ and $\boldsymbol{\Phi}_s^{H/2}$, we introduce the unitary matrix $\boldsymbol{\Omega} \in \mathbb{C}^{N\times N}$, which is such that $\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega}$ and $\mathbf{H}\boldsymbol{\Phi}_s^{1/2}$ do coincide, i.e. we summarize

$$\boxed{\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega} = \mathbf{H}\boldsymbol{\Phi}_s^{1/2},} \tag{28}$$

$$\boldsymbol{\Phi}_s^{1/2} = \mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}], \tag{29}$$

$$\boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I}, \tag{30}$$

where right-multiplying each side of (28) with its Hermitian yields (10). At this point, in order to stress the meaning of (28)–(30), we add that the column vectors $[\boldsymbol{\Psi}_{x_e}^{1/2}]_{:,n}$ and $[\mathbf{H}\boldsymbol{\Phi}_s^{1/2}]_{:,n} = \mathbf{h}_n\varphi_s^{1/2}$ form generally different bases[3] of the same vector space, and hence $\boldsymbol{\Omega}$ implements a change of basis.

Applying (6) to (28)–(29) and noting that $\mathbf{1}^T\,\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}] = \boldsymbol{\varphi}_s^{T/2}$, we find that $\boldsymbol{\varphi}_s^{1/2}$ and $\boldsymbol{\Omega}$ satisfy

$$\boxed{\mathbf{i}^T\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega} = [\boldsymbol{\Psi}_{x_e}^{1/2}\boldsymbol{\Omega}]_{1,:} = \boldsymbol{\varphi}_s^{T/2},} \tag{31}$$

where right-multiplying each side of (31) with its Hermitian yields (13). We further note that if $\boldsymbol{\Omega}$ was known for a given square root $\boldsymbol{\Psi}_{x_e}^{1/2}$, then $\boldsymbol{\varphi}_s^{1/2}$ could be obtained from (31) immediately. In the following, we refer to (28)–(31) as the square-root signal model.

B. *Orthogonal Procrustes-Based Early PSD Estimate*

In this section, based on the square-root signal model in (28)–(31), we seek unitary and diagonal estimates $\hat{\boldsymbol{\Omega}}$ and $\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{1/2}]$ of $\boldsymbol{\Omega}$ and $\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$. Similarly to Section III, we develop our discussion from the premise that estimates $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$ of the early-correlation-matrix square root $\boldsymbol{\Psi}_{x_e}^{1/2}$ and the RETF $\mathbf{H}$ in (28) are readily available, with $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ generally of rank N and $\mathbf{i}^T\hat{\mathbf{H}} = \mathbf{1}^T$. An estimator of $\boldsymbol{\Psi}_{x_e}^{1/2}$ is described in Section V-B, while Section IV-C describes a recursive update scheme for $\hat{\mathbf{H}}$.

Similarly to Section III, now based on the square-root signal model in (28)–(30) instead of the conventional signal model in (10), we define the approximation error as a function of $\boldsymbol{\Omega}$ and $\boldsymbol{\varphi}_s^{1/2}$, i.e.

$$\mathbf{E}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2}) = \hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega} - \hat{\mathbf{H}}\,\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}], \tag{32}$$

which is akin to $\mathbf{E}_c(\boldsymbol{\varphi}_s)$ in (16), and where the subscript sq stands for square root. Further, now based on the square-root signal model in (31) instead of the conventional signal model in (13), we define a soft-constraint error as a function of $\boldsymbol{\Omega}$ and $\boldsymbol{\varphi}_s^{1/2}$,

$$\begin{aligned}\mathbf{e}_{sq}^T(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2}) &= [\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}\boldsymbol{\Omega}]_{1,:} - \boldsymbol{\varphi}_s^{T/2} \\ &= [\mathbf{E}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})]_{1,:},\end{aligned} \tag{33}$$

which is akin to $e_c(\boldsymbol{\varphi}_s)$ in (22). Similarly to $e_c(\boldsymbol{\varphi}_s)$, also $\mathbf{e}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})$ in (33) is independent of $\hat{\mathbf{H}}$, and so its penalization may be useful in case of RETF estimation errors. While $e_c(\boldsymbol{\varphi}_s)$ defines an error on the sum of the early PSDs, however, $\mathbf{e}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})$ instead defines an error on each of the early PSD square roots and is therefore more informative. Based on (32), (33), and the unitary constraint in (30), we define the minimization problem,

$$\boxed{\begin{aligned}&\{\hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\varphi}}_s^{1/2}\} = \\ &\qquad \underset{\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2}}{\arg\min} \left\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})\right\|_F^2 + \alpha\left\|\mathbf{e}_{sq}(\boldsymbol{\Omega}, \boldsymbol{\varphi}_s^{1/2})\right\|_2^2 \\ &\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I},\end{aligned}} \tag{34}$$

which is akin to the conventional MP in (23) and referred to as the square-root minimization problem (square-root MP) in the following. While the unitary constraint in (34) does not have an equivalent in (23), the inequality constraint $\boldsymbol{\varphi}_s \geq \mathbf{0}$ used in (23) is not required in (34), as in the square-root signal model, we find that $\boldsymbol{\varphi}_s = \mathrm{Diag}[\boldsymbol{\varphi}_s^{H/2}]\boldsymbol{\varphi}_s^{1/2}$, and therefore the corresponding estimate $\hat{\boldsymbol{\varphi}}_s$ is guaranteed to be non-negative. Problems of the kind as in (34), i.e. Frobenius-norm minimization problems seeking a unitary and a diagonal matrix, here $\boldsymbol{\Omega}$ and $\mathrm{Diag}[\boldsymbol{\varphi}_s^{1/2}]$, constitute a generalization [23] of the so-called orthogonal Procrustes problem [23]–[25], which seeks a unitary matrix only. As outlined in the following, under a specific rank condition, the orthogonal Procrustes problem has a unique closed-form solution, which is found by means of the SVD [24], [25]. The generalized orthogonal Procrustes problem, on the contrary, does not have a unique closed-form solution, but can be solved iteratively [23]. In particular, along the lines of [23], we propose to solve (34) by alternatingly (re-)estimating $\boldsymbol{\Omega}$ and $\boldsymbol{\varphi}_s^{1/2}$ until convergence is reached, namely by solving the orthogonal Procrustes problem and the soft-constrained convex problem, respectively,

$$\begin{aligned}\hat{\boldsymbol{\Omega}}^{(i)} &= \underset{\boldsymbol{\Omega}}{\arg\min}\left\|\mathbf{E}_{sq}(\boldsymbol{\Omega}, \hat{\boldsymbol{\varphi}}_s^{1/2|(i-1)})\right\|_F^2 \\ &\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I},\end{aligned} \tag{35}$$

$$\begin{aligned}\hat{\boldsymbol{\varphi}}_s^{1/2|(i)} &= \underset{\boldsymbol{\varphi}_s^{1/2}}{\arg\min}\left\|\mathbf{E}_{sq}(\hat{\boldsymbol{\Omega}}^{(i)}, \boldsymbol{\varphi}_s^{1/2})\right\|_F^2 \\ &\qquad + \alpha\left\|\mathbf{e}_{sq}(\hat{\boldsymbol{\Omega}}^{(i)}, \boldsymbol{\varphi}_s^{1/2})\right\|_2^2,\end{aligned} \tag{36}$$

where the soft constraint is applied in (36) only, i.e. once per iteration. Using (32), by expansion of the Frobenius norm in (35), it is easily shown [24], [25] that (35) is equivalent to

$$\begin{aligned}\hat{\boldsymbol{\Omega}}^{(i)} &= \underset{\boldsymbol{\Omega}}{\arg\max}\,\Re\left[\mathrm{tr}[\boldsymbol{\Omega}\mathbf{C}_{\mathrm{sq}}^{(\mathrm{i-1})}]\right] \\ &\qquad \text{s.t.} \quad \boldsymbol{\Omega}\boldsymbol{\Omega}^H = \mathbf{I},\end{aligned} \tag{37}$$

[3] A particular case is obtained for $N = 1$, where $\boldsymbol{\Omega}$ and $\boldsymbol{\Phi}_s^{1/2}$ are scalar, while $\mathbf{H} = \mathbf{h}$ and $\boldsymbol{\Psi}_{x_e}^{1/2} = \boldsymbol{\psi}_{x_e}^{1/2}$ are proportional column vectors. In this case, given an estimate $\hat{\boldsymbol{\psi}}_{x_e}^{1/2}$, we may even estimate $\mathbf{h}$ by $\hat{\mathbf{h}} = \hat{\boldsymbol{\psi}}_{x_e}^{1/2}/[\hat{\boldsymbol{\psi}}_{x_e}^{1/2}]_1$, satisfying $[\hat{\mathbf{h}}]_1 = 1$, cf. (6). In essence, despite somewhat different derivation and terminology, this is equivalent to the approach taken in subspace-based single-source RETF estimation [20].

$$\text{with} \quad \mathbf{C}^{(i-1)} = \mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2|(i-1)}]\hat{\mathbf{H}}^H \hat{\mathbf{\Psi}}_{x_e}^{1/2}. \tag{38}$$

If $\mathbf{C}^{(i-1)}$ has full rank, which (without sufficiency) requires $N \leq M$, the problem in (35) has a unique closed-form solution, which is based on the SVD of $\mathbf{C}^{H|(i-1)}$. Precisely, if we decompose $\mathbf{C}^{H|(i-1)}$ as

$$\mathbf{C}^{H|(i-1)} = \mathbf{U}_L \mathbf{\Sigma} \mathbf{U}_R^H, \tag{39}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ is a diagonal matrix of singular values and both $\mathbf{U}_L \in \mathbb{C}^{N \times N}$ and $\mathbf{U}_R \in \mathbb{C}^{N \times N}$ are unitary, then $\hat{\mathbf{\Omega}}^{(i)}$ is given by

$$\hat{\mathbf{\Omega}}^{(i)} = \mathbf{U}_L \mathbf{U}_R^H. \tag{40}$$

With (32) and (33), the solution to (36) is easily found as

$$\hat{\boldsymbol{\varphi}}_s^{1/2|(i)} = \mathbf{A}_{sq}^{-1} \mathbf{b}_{sq}^{(i)}, \tag{41}$$

with $\mathbf{A}_{sq} \in \mathbb{R}^{N \times N}$ and $\mathbf{b}_{sq}^{(i)} \in \mathbb{C}^N$ defined by

$$\mathbf{A}_{sq} = \mathrm{Diagg}[\hat{\mathbf{H}}^H \hat{\mathbf{H}}] + \alpha \mathbf{I}, \tag{42}$$

$$\begin{aligned} \mathbf{b}_{sq}^{(i)} &= \mathrm{diag}[\hat{\mathbf{H}}^H (\mathbf{I} + \alpha \mathbf{i}\mathbf{i}^T) \hat{\mathbf{\Psi}}_{x_e}^{1/2} \hat{\mathbf{\Omega}}^{(i)}] \\ &= \mathrm{diag}[\hat{\mathbf{H}}^H \hat{\mathbf{\Psi}}_{x_e}^{1/2} \hat{\mathbf{\Omega}}^{(i)}] + \alpha [\hat{\mathbf{\Psi}}_{x_e}^{1/2} \hat{\mathbf{\Omega}}^{(i)}]_{1,:}^T. \end{aligned} \tag{43}$$

The set of equations (42)–(43) is akin to (26)–(27) for the conventional MP. Note that for $\alpha \to \infty$, the soft constraint in the square-root MP in (36) becomes a hard constraint and, moreover, solely determines $\hat{\boldsymbol{\varphi}}_s^{1/2|(i)}$, namely as $\hat{\boldsymbol{\varphi}}_s^{1/2|(i)} = [\hat{\mathbf{\Psi}}_{x_e}^{1/2} \hat{\mathbf{\Omega}}^{(i)}]_{1,:}^T$ according to (41)–(43). This is not the case for the soft constraint in the conventional MP in (23).

Note that since the problem in (34) is non-convex, the iteration in (35)–(36) is not guaranteed to converge to a global minimum [23]. The initial value $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$ of the iteration can, e.g., be chosen based on the sum constraint in (13) as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{[\hat{\mathbf{\Psi}}_{x_e}]_{1,1}/N}\,\mathbf{1}$, or based on the comparably lowly complex estimator in (17)–(18), here denoted by $\hat{\boldsymbol{\varphi}}_{s|c_0}$, as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{\hat{\boldsymbol{\varphi}}_{s|c_0}}$. Here, the latter provides faster convergence, cf. Section VI-A4.

C. Recursive RETF Update

Based upon the square-root model in (28), the estimates $\hat{\mathbf{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ obtained as discussed in Section IV-B can also be used to recursively update the RETF estimate $\hat{\mathbf{H}}$. Note that in doing so, the quality of the estimates $\hat{\mathbf{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ on the one hand and $\hat{\mathbf{H}}$ on the other hand depend upon each other, such that particular means (as outlined below) have to be taken in order to limit error amplification from frame to frame and thereby avoid divergence.

In the following, we differentiate the prior and posterior estimates $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}^+$, and propose to simply propagate the posterior in the previous frame to the prior in the current frame, i.e.

$$\hat{\mathbf{H}}(l) = \hat{\mathbf{H}}^+(l{-}1). \tag{44}$$

In each frame, we use $\hat{\mathbf{H}}$ to obtain $\hat{\mathbf{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ with (35)–(36), and then use $\hat{\mathbf{\Omega}}$ and $\hat{\boldsymbol{\varphi}}_s^{1/2}$ to obtain $\hat{\mathbf{H}}^+$, where we again resort to the square-root signal model in (28). To this end, we define the approximation error as a function of $\mathbf{H}$,

$$\mathbf{E}_{sq}(\mathbf{H}) = \hat{\mathbf{\Psi}}_{x_e}^{1/2} \hat{\mathbf{\Omega}} - \mathbf{H}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{1/2}] \tag{45}$$

which is similar to (32). Based upon (45) and the constraint in (6), we define the minimization problem,

$$\boxed{\begin{aligned} \hat{\mathbf{H}}^+ = \arg\min_{\mathbf{H}} & \left\|\mathbf{E}_{sq}(\mathbf{H})\right\|_F^2 + \left\|(\hat{\mathbf{H}} - \mathbf{H})\,\mathrm{Diag}\left[\sqrt{\boldsymbol{\beta}}\right]\right\|_F^2 \\ \text{s.t.} \quad & \mathbf{i}^T \mathbf{H} = \mathbf{1}^T, \end{aligned}} \tag{46}$$

where the penalty term $\left\|(\hat{\mathbf{H}} - \mathbf{H})\,\mathrm{Diag}\left[\sqrt{\boldsymbol{\beta}}\right]\right\|_F^2$ relates to Levenberg-Marquardt regularization [32], [33] in that it penalizes deviation from the previous (i.e., the prior) estimate $\hat{\mathbf{H}}$ and thus limits error amplification from frame to frame. Here, we leave $\boldsymbol{\beta}$ subject to tuning as described in the following. In this respect, recall that according to (28), both $\mathbf{\Psi}_{x_e}^{1/2}$ and $\mathbf{H}$ span the same column space. However, due to modeling and estimation errors, this is not necessarily true for the corresponding estimates $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$. In particular, if the nth source image has a comparably low early PSD φ_{s_n} or is inactive, then the associated subspace dimension will not be well or not at all be represented in $\hat{\mathbf{\Psi}}_{x_e}^{1/2}$, and both $[\hat{\mathbf{\Omega}}_s]_{:,n}$ and $\hat{\varphi}_{s_n}^{1/2}$ may exhibit comparably large estimation errors. Further, the estimate $\hat{\varphi}_{s_n}^{1/2}$ may contain residual late reverberation due to erroneous separation of $\hat{\mathbf{\Psi}}_x$ into $\hat{\mathbf{\Psi}}_{x_e}$ and $\mathbf{\Psi}_{x_\ell}$, cf. (9), Section V and Section VI. In such a case, one would preferably rely on the prior estimate $\hat{\mathbf{h}}_n$ instead of updating based on $[\hat{\mathbf{\Omega}}_s]_{:,n}$ and $\hat{\varphi}_{s_n}^{1/2}$. Considering the solution to (46), which is given by

$$[\hat{\mathbf{H}}^+]_{1,:} = \mathbf{1}^T, \tag{47}$$

$$\begin{aligned} [\hat{\mathbf{H}}^+]_{2:M,:} = l\big[\big(&\hat{\mathbf{\Psi}}_{x_e}^{1/2} \hat{\mathbf{\Omega}}\,\mathrm{Diag}[\hat{\boldsymbol{\varphi}}_s^{H/2}] + \hat{\mathbf{H}}\,\mathrm{Diag}[\boldsymbol{\beta}]\big) \\ &\cdot \mathrm{Diag}^{-1}[\hat{\boldsymbol{\varphi}}_s + \boldsymbol{\beta}]\big]_{2:M,:}, \end{aligned} \tag{48}$$

we indeed find that the smaller $\hat{\varphi}_{s_n}$ as compared to $\beta_n = [\boldsymbol{\beta}]_n$, the more $\hat{\mathbf{h}}_n^+$ relies on $\hat{\mathbf{h}}_n$, as desired. In order to further increase robustness against modeling and estimation errors, source inactivity and residual late reverberation in $\hat{\varphi}_{s_n}$, we propose to make β_n time-varying with binary values. More precisely, we base β_n on the power ratio

$$\boldsymbol{\xi} = \hat{\boldsymbol{\varphi}}_s / (\mathbf{1}^T \hat{\boldsymbol{\varphi}}_s + \varphi_{reg}), \tag{49}$$

where $\xi_n = [\boldsymbol{\xi}]_n \in [0,\, 1]$. Here, φ_{reg} can be used for regularization, e.g., we may choose $\varphi_{reg} = \varphi_{x_\ell}$ in order to limit ξ_n in frames where pre-dominantly late reverberation is estimated. Given ξ_n, we set β_n as

$$\beta_n \begin{cases} = \beta & \text{if} \quad \xi_n \geq \xi_{th}, \\ \to \infty & \text{else}, \end{cases} \tag{50}$$

and thereby resort to $\hat{\mathbf{h}}_n^+ = \hat{\mathbf{h}}_n$ if ξ_n is smaller than the predefined threshold ξ_{th}. The value β, used if $\xi_n \geq \xi_{th}$, should scale in relation to the dynamic range of φ_{s_n} and can be chosen depending on the (estimated) probability density function of the complex STFT coefficients s_n, cf. Section VI.

Note that in order to start the recursion defined by (44), (35)–(36), and (46), an initial estimate $\hat{\mathbf{H}}(0)$ is required, which may be based on, e.g., initial single-source RETF estimates acquired from segments with mutual-exclusively active sources [20], or some initial knowledge or estimates of the associated DoAs [7], [21], [22].

V. Subspace-Based Early Correlation Matrix Estimation

In Section III and Section IV, we respectively assumed that the early-correlation-matrix estimate $\hat{\boldsymbol{\Psi}}_{x_e}$ and its square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ of rank N are available. In this section, we discuss how to obtain these estimates from the microphone signals $\mathbf{x}$. We estimate $\boldsymbol{\Psi}_x = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$ by recursively averaging $\mathbf{x}\mathbf{x}^H$, yielding the *smooth* estimate $\hat{\boldsymbol{\Psi}}_{x|sm}$ and its equally *smooth* subspace representation based on the GEVD. From the latter, we first define a *desmoothed* estimate $\hat{\boldsymbol{\Psi}}_x$ based on desmoothed generalized eigenvalues, and second extract the early component $\hat{\boldsymbol{\Psi}}_{x_e}$ and its square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$.

In Section V-A, we introduce the subspace model of $\boldsymbol{\Psi}_x$. In Section V-B, we obtain the smooth and desmoothed estimates $\hat{\boldsymbol{\Psi}}_{x|sm}$ and $\hat{\boldsymbol{\Psi}}_x$, respectively. In Section V-C, given $\hat{\boldsymbol{\Psi}}_x$, we then retrieve subspace-based rank-N estimates $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$.

A. Correlation Matrix Subspace Decomposition

In each frame l, we define the GEVD [34] of $\boldsymbol{\Psi}_x$ and the diffuse coherence matrix $\boldsymbol{\Gamma}$, cf. (14), i.e.

$$\boldsymbol{\Psi}_x \mathbf{P} = \boldsymbol{\Gamma}\mathbf{P}\boldsymbol{\Lambda}_x, \tag{51}$$

$$\text{with} \quad \boldsymbol{\Lambda}_x = \mathrm{Diag}[\boldsymbol{\lambda}_x], \tag{52}$$

where $\boldsymbol{\lambda}_x \in \mathbb{R}^M$ comprises the generalized eigenvalues, and the columns of $\mathbf{P} \in \mathbb{C}^{M\times M}$ comprise the associated generalized eigenvectors. In the GEVD, the generalized eigenvectors in $\mathbf{P}$ are uniquely defined up to a scaling factor, and for any factorization $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Gamma}^{H/2}$, we find that $\boldsymbol{\Gamma}^{H/2}\mathbf{P}$ is column-wise orthogonal due to $\boldsymbol{\Psi}_x = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$ being Hermitian. In the following, without loss of generality, we assume the eigenvectors to be scaled such that $\boldsymbol{\Gamma}^{H/2}\mathbf{P}$ is unitary, i.e.

$$\mathbf{P}^H\boldsymbol{\Gamma}\mathbf{P} = \mathbf{I}, \tag{53}$$

and therefore, combining (51) and (53),

$$\mathbf{P}^H\boldsymbol{\Psi}_x\mathbf{P} = \boldsymbol{\Lambda}_x. \tag{54}$$

An alternative, but mathematically equivalent formulation to the GEVD in (51) is given by the EVD of the pre-whitened matrix $\boldsymbol{\Psi}'_x = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Psi}_x\boldsymbol{\Gamma}^{-H/2}$ [15], [20], [35], which is defined by $\boldsymbol{\Psi}'_x\mathbf{P}' = \mathbf{P}'\boldsymbol{\Lambda}'_x$. By comparison with (51), we find $\boldsymbol{\Lambda}'_x = \boldsymbol{\Lambda}_x$ and $\mathbf{P}' = \boldsymbol{\Gamma}^{H/2}\mathbf{P}$, provided that the respective (generalized) eigenvalues are sorted in the same order, and the (generalized) eigenvectors are scaled accordingly.

For convenience of presentation, assume that the generalized eigenvalues in $\boldsymbol{\lambda}_x$ are sorted in a descending order, and the generalized eigenvectors in $\mathbf{P}$ are sorted accordingly. Then, inserting $\boldsymbol{\Psi}_x = \boldsymbol{\Psi}_{x_e} + \boldsymbol{\Psi}_{x_\ell}$ with $\boldsymbol{\Psi}_{x_\ell} = \varphi_{x_\ell}\boldsymbol{\Gamma}$, cf. (9) and (14), into (54) while making use of (53) yields

$$\boldsymbol{\Lambda}_x = \mathbf{P}^H\boldsymbol{\Psi}_{x_e}\mathbf{P} + \varphi_{x_\ell}\mathbf{I}, \tag{55}$$

wherein $\boldsymbol{\Psi}_{x_e}$ and in consequence $\mathbf{P}^H\boldsymbol{\Psi}_{x_e}\mathbf{P}$ generally have rank N, and the latter in addition is diagonal, i.e. if $N < M$ we have

$$\mathbf{P}^H\boldsymbol{\Psi}_{x_e}\mathbf{P} = \begin{pmatrix} \boldsymbol{\Lambda}_{x_e} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{56}$$

$$\text{with} \quad \boldsymbol{\Lambda}_{x_e} = \mathrm{Diag}[\boldsymbol{\lambda}_{x_e}], \tag{57}$$

and $\boldsymbol{\lambda}_{x_e} \in \mathbb{R}^N$.

B. Recursive Correlation Matrix Estimation and Desmoothing

We compute a smooth estimate $\hat{\boldsymbol{\Psi}}_{x|sm}$ of $\boldsymbol{\Psi}_x$ by recursively averaging $\mathbf{x}\mathbf{x}^H$ using some pre-defined forgetting factor $\zeta \in (0,\,1)$, i.e.

$$\hat{\boldsymbol{\Psi}}_{x|sm}(l) = \zeta\hat{\boldsymbol{\Psi}}_{x|sm}(l{-}1) + (1{-}\zeta)\mathbf{x}(l)\mathbf{x}^H(l), \tag{58}$$

and perform the GEVD $\hat{\boldsymbol{\Psi}}_{x|sm}\hat{\mathbf{P}} = \boldsymbol{\Gamma}\hat{\mathbf{P}}\hat{\boldsymbol{\Lambda}}_{x|sm}$ similar to (51)–(54), with $\hat{\mathbf{P}}$ an estimate of $\mathbf{P}$ and $\hat{\boldsymbol{\Lambda}}_{x|sm} = \mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{x|sm}]$ a smooth estimate of $\boldsymbol{\Lambda}_x$. Note that in order to excite all subspace dimensions and the associated generalized eigenvalues and hence to achieve a meaningful decomposition, $\hat{\boldsymbol{\Psi}}_{x|sm}$ needs to be well-conditioned, and so ζ must be sufficiently close to one. As discussed in Section II, the PSDs φ_{s_n} and φ_{x_ℓ} may be highly non-stationary, while the associated coherence matrices $\mathbf{h}_n\mathbf{h}_n^H$ and $\boldsymbol{\Gamma}$ are commonly assumed to be comparably slowly time-varying or even time-invariant. In theory, a linear combination of the PSDs φ_{s_n} and φ_{x_ℓ} is rendered by the *unknown* generalized eigenvalues $\boldsymbol{\lambda}_x$ of $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Gamma}$, i.e. also $\boldsymbol{\lambda}_x$ may be highly non-stationary. In contrast, due to the (inevitable) recursive averaging in (58), the *computed* generalized eigenvalues $\hat{\boldsymbol{\lambda}}_{x|sm}$ of $\hat{\boldsymbol{\Psi}}_{x|sm}$ and $\boldsymbol{\Gamma}$ are slowly time-varying if ζ is sufficiently large, i.e. non-stationarities are to some extent smoothed, and so would be PSD estimates based on $\hat{\boldsymbol{\lambda}}_{x|sm}$ or $\hat{\boldsymbol{\Psi}}_{x|sm}$. While smooth PSD estimates are commonly used in some applications (e.g., in the computation of spectral gains in speech enhancement [2]), others exploit non-stationarities (such as, e.g., the Kalman filter [36], where PSD estimates of the observation noise act as a regularization term in the recursive update of the state estimate [28]). Depending on the application, we therefore propose to restore non-stationarities by desmoothing $\hat{\boldsymbol{\lambda}}_{x|sm}$, yielding an estimate $\hat{\boldsymbol{\lambda}}_x$ of $\boldsymbol{\lambda}_x$.

To this end, we note that the recursive averaging in (58) can be considered an element-wise filtering operation with $\mathbf{x}(l)\mathbf{x}^H(l)$ as the input, $\hat{\boldsymbol{\Psi}}_{x|sm}(l)$ as the output, and the (all-pole) z-domain transfer function given by $(1-\zeta)/(1-\zeta z^{-1})$. Therefore, in order to desmooth $\hat{\boldsymbol{\lambda}}_{x|sm}(l)$, we propose to apply the corresponding (all-zero) inverse transfer function given by $(1-\zeta z^{-1})/(1-\zeta)$ followed by non-negative thresholding, i.e.

$$\hat{\boldsymbol{\lambda}}'_x(l) = \frac{\hat{\boldsymbol{\lambda}}_{x|sm}(l) - \zeta\hat{\boldsymbol{\lambda}}_{x|sm}(l{-}1)}{1-\zeta}, \tag{59}$$

$$\hat{\boldsymbol{\lambda}}_x(l) = \max[\hat{\boldsymbol{\lambda}}'_x(l),\,\mathbf{0}], \tag{60}$$

where the thresholding in (60) avoids negative eigenvalue estimates, which otherwise may appear in a limited number of frames due to modeling and estimation errors. Note that the desmoothing operation requires the associated generalized eigenvalues in $\hat{\boldsymbol{\lambda}}_{x|sm}(l)$ and $\hat{\boldsymbol{\lambda}}_{x|sm}(l{-}1)$ to be sorted correspondingly. This can be ensured by sorting $\hat{\mathbf{P}}(l)$ such that $\hat{\mathbf{P}}^{H}(l{-}1)\boldsymbol{\Gamma}\hat{\mathbf{P}}(l) \approx \mathbf{I}$, cf. (53), and $\hat{\boldsymbol{\lambda}}_{x|sm}(l)$ accordingly, which can be done easily for large ζ and the therewith slowly time-varying GEVD [27]. Alternatively, recursive sorting can be avoided if the GEVD is estimated recursively, e.g., by means of the power method [37], [38]. One may then define the corresponding desmoothed estimate $\hat{\boldsymbol{\Psi}}_x$ via its decomposition

$$\hat{\boldsymbol{\Psi}}_x\hat{\mathbf{P}} = \boldsymbol{\Gamma}\hat{\mathbf{P}}\hat{\boldsymbol{\Lambda}}_x, \tag{61}$$

$$\text{with} \quad \hat{\boldsymbol{\Lambda}}_x = \mathrm{Diag}[\hat{\boldsymbol{\lambda}}_x], \tag{62}$$

where $\hat{\mathbf{P}}$ remains unchanged.

C. Early Correlation Matrix Estimation and Factorization

Given $\hat{\mathbf{P}}$ and $\hat{\boldsymbol{\Lambda}}_x$ in (61)–(62), we now retrieve the subspace-based rank-N estimates $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$. To this end, based on (55)–(57), we note that $\boldsymbol{\lambda}_{x_e}$ can be estimated as

$$\hat{\boldsymbol{\lambda}}_{x_e} = [\hat{\boldsymbol{\lambda}}_x]_{1:N} - \hat{\varphi}_{x_\ell}\mathbf{1}, \tag{63}$$

where $\hat{\varphi}_{x_\ell}$ in turn is obtained by averaging the last $M-N$ generalized eigenvalues in $[\hat{\boldsymbol{\lambda}}_x]_{N+1:M}$ [15]. Considering (56)–(57), given $\hat{\boldsymbol{\Lambda}}_{x_e} = \mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{x_e}]$ from (63) and $\hat{\mathbf{P}}^{-1} = \hat{\mathbf{P}}^{H}\boldsymbol{\Gamma}$ from (53), we can define a rank-N estimate of $\boldsymbol{\Psi}_{x_e}$ as

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_{x_e} &= \boldsymbol{\Gamma}\hat{\mathbf{P}}\begin{pmatrix}\hat{\boldsymbol{\Lambda}}_{x_e} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\hat{\mathbf{P}}^{H}\boldsymbol{\Gamma}\\ &= \boldsymbol{\Gamma}[\hat{\mathbf{P}}]_{:,1:N}\hat{\boldsymbol{\Lambda}}_{x_e}[\hat{\mathbf{P}}]_{:,1:N}^{H}\boldsymbol{\Gamma}.\end{aligned} \tag{64}$$

From (64), we can further easily derive a square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ as

$$\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} = \boldsymbol{\Gamma}[\hat{\mathbf{P}}]_{:,1:N}\hat{\boldsymbol{\Lambda}}_{x_e}^{1/2} \tag{65}$$

$$\text{with} \quad \hat{\boldsymbol{\Lambda}}_{x_e}^{1/2} = \mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{x_e}^{1/2}], \tag{66}$$

with arbitrary complex arguments of the elements in $\hat{\boldsymbol{\lambda}}_{x_e}^{1/2}$.

Note that as opposed to the order presented in Section V-B and this section, we may also apply desmoothing only after obtaining a smooth estimate of the early correlation matrix and its square root, which showed to yield comparable results in our simulations.

VI. Simulations

In this section, we compare the algorithms based on the conventional and the square-root MP as presented in Section III and Section IV, respectively. We assume that an (initial) RETF estimate $\hat{\mathbf{H}}$ is available, and that $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ are obtained as described in Section V.

Apart from estimation errors in $\hat{\boldsymbol{\Psi}}_{x_e}$, $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$, and $\hat{\mathbf{H}}$, the performance of both algorithms is subject to modeling errors, cf. Section II. Unfortunately, due to the model deficiencies in (9)–(15), exact and observable ground truth early PSDs $\boldsymbol{\varphi}_s$ and ground truth RETFs $\mathbf{H}$ do not exist in a practical setup based on realistic acoustic data. Therefore, in order to yield a broader understanding of the algorithms' behavior, we perform two kinds of simulations. In the first kind, instead of generating time-domain data and estimating $\boldsymbol{\Psi}_x$ in the STFT domain, we generate $\hat{\boldsymbol{\Psi}}_x = \boldsymbol{\Psi}_x$ directly based on (9)–(14) and assumed geometric and physical properties, i.e. $\hat{\boldsymbol{\Psi}}_x$ is free of modeling and estimation errors. This way, we are able to define exact ground truth early PSDs $\boldsymbol{\varphi}_s$ and ground truth RETFs $\mathbf{H}$ that can be used to define exact performance measures. Further, the estimates $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ obtained as described in Section V will be free of estimation errors, such that the performance of both algorithms depends on the RETF estimation error in $\hat{\mathbf{H}}$ and the algorithmic settings in Section III and Section IV only. We refer to these simulations as the *model-based-data* case. In the second kind of simulations, we generate acoustic data in the time domain from recorded speech signals and measured room impulse responses (RIRs), and estimate $\boldsymbol{\Psi}_x$ in the STFT domain. This way, the setup becomes more practical, however, evaluation becomes less trivial in terms of the definition of performance measures, such that we need to define and rely on an approximate ground truth early PSD $\tilde{\boldsymbol{\varphi}}_s$ as a reference. We refer to these simulations as the *acoustic-data* case. The model-based-data case and the acoustic-data case are discussed in Section VI-A and Section VI-B, respectively.

A. Model-Based Data

We define our performance measures in Section VI-A1, discuss the data-generation in Section VI-A2, the algorithmic settings in Section VI-A3, and the evaluation results in Section VI-A4.

1) Performance Measures: We define the RETF estimation error,

$$\mathbf{E}_H = \hat{\mathbf{H}} - \mathbf{H}, \tag{67}$$

where $\mathbf{i}^T\mathbf{E}_H = [\mathbf{E}_H]_{1,:} = \mathbf{0}^T$ since both $\mathbf{H}$ and $\hat{\mathbf{H}}$ satisfy (6), and based on that the relative squared RETF estimation error,

$$\varepsilon_H = 10\log_{10}\frac{\mathrm{tr}[\mathbf{E}_{\mathrm{H}}^{\mathrm{H}}\mathbf{E}_{\mathrm{H}}]}{\mathrm{tr}[\mathbf{H}^{\mathrm{H}}\mathbf{H}] - \mathrm{N}}\mathrm{dB}, \tag{68}$$

where we subtract N in the denominator in order to compensate for the fact that the first row of $\mathbf{H}$ is known. Since the early PSDs $\boldsymbol{\varphi}_s$ are already a second-order property of the underlying signal $\mathbf{s}$, we define the PSD estimation error with respect to the non-negative square root of $\hat{\boldsymbol{\varphi}}_s$ and $\boldsymbol{\varphi}_s$, i.e.

$$\mathbf{e}_{\varphi_s} = \sqrt{\hat{\boldsymbol{\varphi}}_s} - \sqrt{\boldsymbol{\varphi}_s}, \tag{69}$$

and based on that the relative squared PSD estimation error,

$$\varepsilon_{\varphi_s} = 10\log_{10}\frac{\mathbf{e}_{\varphi_s}^T\mathbf{e}_{\varphi_s}}{\mathbf{1}^T\boldsymbol{\varphi}_s}\,\mathrm{dB}. \tag{70}$$

2) Data Generation: Let $\hat{\boldsymbol{\Psi}}_x$ be available and free of modeling and estimation errors, i.e. we have $\hat{\boldsymbol{\Psi}}_x = \boldsymbol{\Psi}_x$ with $\boldsymbol{\Psi}_x$ adhering to (9)–(14). We generate $\boldsymbol{\Psi}_x$ based on assumed geometric and physical properties. We assume a linear microphone array of $M = 5$ microphones with inter-microphone distance of 8 cm and the speed of sound to be 340 m/s. Further, we assume $N = 3$ sources, positioned at $(-30, 0, 60)^\circ$ relative to the broadside direction of the microphone array. The RETFs $\mathbf{H}$ are generated assuming omnidirectional microphones of equal gain as well as free- and far-field propagation for the early components, i.e. $\mathbf{H}$ depends on the DoAs only and is fully defined by the corresponding phase shifts between microphones. The estimate $\hat{\mathbf{H}}$ is generated by adding an error component $\mathbf{E}_H$ according to (67), where the elements $[\mathbf{E}_H]_{:,2:M}$ are drawn from independent complex Gaussian distributions, yielding a particular ε_H according to (68). The diffuse coherence matrix $\boldsymbol{\Gamma}$ is computed assuming a spherical-isotropic sound field. The early PSDs $\boldsymbol{\varphi}_s$ are generated in the following manner. We draw the real and imaginary parts of the elements of $\mathbf{s}$ from independent Laplace distributions, which is a commonly assumed distribution for STFT coefficients of speech [39], [40], i.e. we have $\Re[s_n] \sim (1/b)e^{-2|\Re[s_n]|/b}$ and $\Im[s_n] \sim (1/b)e^{-2|\Im[s_n]|/b}$, where the scaling parameter b is referred to as diversity. Then, we define $\boldsymbol{\varphi}_s = \mathrm{Diag}[\mathbf{s}^H]\mathbf{s}$, i.e. $\boldsymbol{\varphi}_s$ is the squared magnitude of $\mathbf{s}$. Given the above, we set $\boldsymbol{\Psi}_{x_e} = \mathbf{H}\boldsymbol{\varphi}_s\mathbf{H}^H$ according to (10). Note that since $\hat{\boldsymbol{\Psi}}_x = \boldsymbol{\Psi}_x$ is free of modeling errors, where $\boldsymbol{\Psi}_x = \boldsymbol{\Psi}_{x_e} + \boldsymbol{\Psi}_{x_\ell}$ with $\boldsymbol{\Psi}_{x_\ell} = \varphi_{x_\ell}\boldsymbol{\Gamma}$, cf. (9) and (14), the component $\boldsymbol{\Psi}_{x_e}$ can be perfectly estimated from $\hat{\boldsymbol{\Psi}}_x$ by means of the GEVD as described in Section V-C, yielding $\hat{\boldsymbol{\Psi}}_{x_e} = \boldsymbol{\Psi}_{x_e}$ independently of φ_{x_ℓ}. Further, note that next to $\mathbf{H}$ and $\boldsymbol{\varphi}_s$, via the GEVD, also $\boldsymbol{\Gamma}$ influences the shape of the square root $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} = \boldsymbol{\Psi}_{x_e}^{1/2}$ in the sense of defining the basis for a given vector space, cf. Section V-C. For each data-point in the evaluation, cf. Section VI-A4, we simulate 2^{14} realizations of $\hat{\boldsymbol{\Psi}}_{x_e}$, $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ and $\hat{\mathbf{H}}$.

3) Algorithmic Settings: In the model-based-data case, as opposed to the acoustic-data case, cf. Section VI-B3, the sampling frequency and STFT-processing parameters are irrelevant since we generate $\hat{\boldsymbol{\Psi}}_x$ directly in the STFT-domain, cf. Section VI-A2. Regardless, we simulate frequencies up to $f = 8$ kHz, corresponding to a virtual sampling frequency of $f_s = 16$ kHz. The soft-constraint penalty factor α in the conventional MP in (23) and the square-root MP in (34) is simulated in the range $\alpha \in [10^{-3}, 10^{5}]$. We perform at most $i_{\max} = 20$ iterations of the associated iterative algorithms in (24)–(25) and (35)–(36). All but one of our simulations consider a single frame l only. In the one simulation considering recursive behavior, we do not update $\hat{\mathbf{H}}$ for the conventional MP, but we do update $\hat{\mathbf{H}}$ recursively for the square-root MP as described in Section IV-C. In the latter case, in (49), since $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2} = \boldsymbol{\Psi}_{x_e}^{1/2}$ is free of modeling and estimation errors and therefore free of residual late reverberation, cf. Section VI-A2, we set $\varphi_{reg} = 0$. In (50), the threshold ξ_{th} is set as $10 \log_{10} \xi_{th} = -2$ dB and β is set as $\beta = 20b^2$, with b the diversity of the Laplace distributions used in the generation of $\boldsymbol{\varphi}_s$, cf. Section VI-A2.

4) Results: Fig. 1 shows the PSD estimation performance in terms of the relative squared PSD estimation error ε_{φ_s} for different values of the relative squared RETF estimation error ε_H for the algorithms based on the conventional MP [—●—] and the square-root MP [——] with $\alpha = 10^3$ at $f = 2$ kHz within a single frame l. In this figure and similar ones in the following, the graphs denote medians over all 2^{14} realizations, cf. Section VI-A2, and the shaded areas denote the range from the first to the third quartile. As can be seen, for both the conventional MP and the square-root MP, ε_{φ_s} increases at a rate of about 10 dB per 10 dB increase in ε_H until roughly $\varepsilon_H = 0$ dB and $\varepsilon_H = 5$ dB is reached, respectively, after which ε_{φ_s} begins to saturate. This saturation is due to the fact that both algorithms yield non-negative estimates $\hat{\boldsymbol{\varphi}}_s \geq \mathbf{0}$, which limits the estimation error at high values of ε_H. The square-root MP outperforms the conventional MP by at least 5.7 dB for $\varepsilon_H \leq 0$ dB, and by somewhat less for $\varepsilon_H \geq 5$ dB.

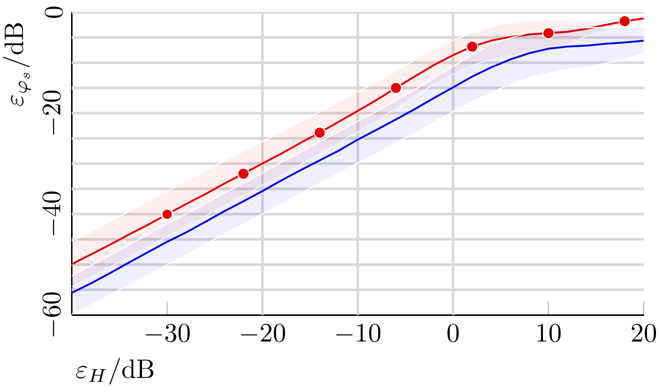

Fig. 1. ε_{φ_s} versus ε_H for conventional MP [—●—] and square-root MP [——] with $\alpha = 10^3$ at $f = 2$ kHz.

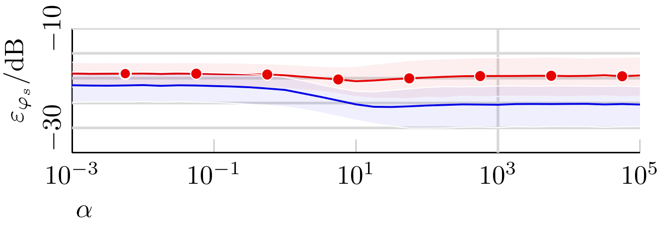

Fig. 2. ε_{φ_s} versus α for conventional MP [—●—] and square-root MP [——] at $\varepsilon_H = -10$ dB and $f = 2$ kHz.

Fig. 2 illustrates ε_{φ_s} for different values of the soft constraint penalty factor α for the conventional MP [—●—] and the square-root MP [——] at $\varepsilon_H = -10$ dB and $f = 2$ kHz within a single frame l. We note that while α hardly impacts the performance of the conventional MP, we generally reach larger improvements for higher values of α in the square-root MP. Recall that the soft constraint in the conventional MP is scalar-based, cf. (22), while the soft constraint in the square-root MP is vector-based, cf. (33), and is therefore more informative. The square-root MP outperforms the conventional MP by 2.5 dB at low values of α, and by 5.7 dB at high values of α. Interestingly, for both algorithms, despite $\hat{\boldsymbol{\Psi}}_{x_e}$ and $\hat{\boldsymbol{\Psi}}_{x_e}^{1/2}$ being free of estimation errors, the minimum of ε_{φ_s} does not occur at the highest values of α, but at around $\alpha = 10^1$. As compared to higher values, the improvement is however mild.

Fig. 3 illustrates ε_{φ_s} for different frequencies f for the conventional MP [—●—] and the square-root MP [——] with $\alpha = 10^3$ at (a) $\varepsilon_H = 0$ dB, (b) $\varepsilon_H = -10$ dB, and (c) $\varepsilon_H = -20$ dB within a single frame l. Note that at some frequencies, due to spatial

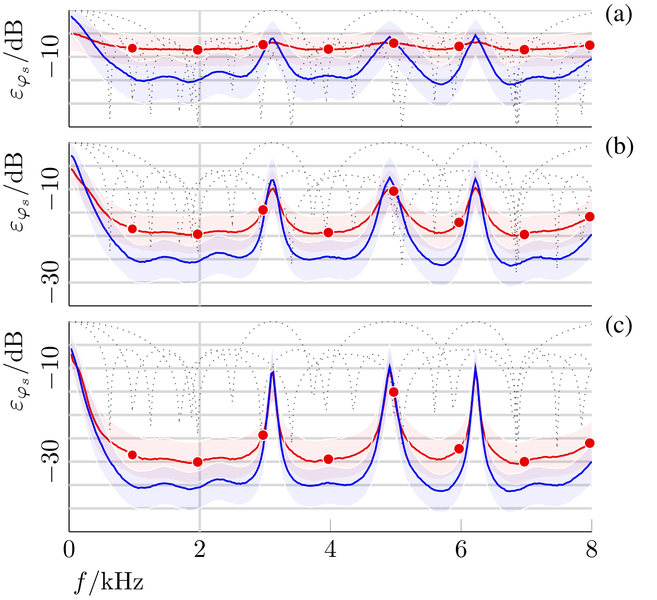

Fig. 3. ε_{φ_s} versus f for conventional MP [—●—] and square-root MP [——] with $\alpha = 10^3$ at (a) $\varepsilon_H = 0$ dB, (b) $\varepsilon_H = -10$ dB, and (c) $\varepsilon_H = -20$ dB. The graphs denoted by [⋯⋯] correspond to $10\log_{10}|\mathbf{h}_n^H\mathbf{h}_{n'}|/M$ dB for $n' \neq n$.

aliasing, which occurs for two different DoAs if their phase difference in each microphone is a multiple of 2π, the two corresponding DoA-based RETFs in $\mathbf{H}$, cf. Section VI-A2, will be identical, and therefore $\mathbf{H}$ itself and consequently also $\boldsymbol{\Psi}_{x_e}$ and $\boldsymbol{\Psi}_{x_e}^{1/2}$ will be rank-deficient. In our setup, this situation occurs for $f \in \{3.11, 4.91, 6.22\}$ kHz, cf. also the dotted lines [⋯⋯] corresponding to $10\log_{10}|\mathbf{h}_n^H\mathbf{h}_{n'}|/M$ dB for $n' \neq n$, which reach 0 dB if $\mathbf{h}_{n'} = \mathbf{h}_n$. As expected, by comparing Fig. 3(a) to Fig. 3(c), neither of the two algorithms performs well in the proximity of these frequencies, independent of ε_H. At other frequencies, however, the square-root MP outperforms the conventional MP by roughly 5 to 7 dB.

Fig. 4 demonstrates the effect in the median of the initial estimate $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$ on the convergence behavior in terms of the relative squared PSD estimation error $\varepsilon_{\varphi_s}^{(i)}$ at iteration i for different values of ε_H of the iterative algorithm in (35)–(36) solving the square-root MP with $\alpha = 10^3$ at $f = 2$ kHz. The initial value is based on (a) the sum constraint in (13) as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{[\hat{\boldsymbol{\Psi}}_{x_e}]_{1,1}/N}\,\mathbf{1}$, and (b) the estimator in (17)–(18), here denoted by $\hat{\boldsymbol{\varphi}}_{s|c_0}$, as $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)} = \sqrt{\hat{\boldsymbol{\varphi}}_{s|c_0}}$. In both cases, the algorithm converges to almost the same final value of ε_{φ_s}. However, we find that in (a), convergence is reached at around $i = 3$ to $i = 4$, while in (b), due to the improved initial estimate, convergence is reached at $i = 1$ already. Hence, while the computation of the initial estimate in (b) is somewhat more expensive, we save 2 to 3 iterations as compared to (a).

Fig. 5 demonstrates the recursive behavior in terms of (a) $\varepsilon_H(l+r)$ and (b) $\varepsilon_{\varphi_s}(l+r)$ with r the recursion index for the conventional MP [—●—] and the square-root MP [——] with $\alpha = 10^3$ at $f = 2$ kHz and $\varepsilon_H(l) = 0$ dB. Here, the source positioned at $-30°$ transitions to $-40°$ at $r = 32$, resulting in a transient change in the otherwise constant RETF $\mathbf{H}$. While no update of the estimate $\hat{\mathbf{H}}$ is performed for the conventional MP, we do update $\hat{\mathbf{H}}$ recursively for square-root MP as described in Section IV-C. For the conventional MP, we expectably find that $\varepsilon_H(l+r)$ and $\varepsilon_{\varphi_s}(l+r)$ remain constant except for a transient increase of 6.8 dB and 3.2 dB at $r = 33$, respectively. For the square-root MP, due to the recursive update of $\hat{\mathbf{H}}$, we find that $\varepsilon_H(l+r)$ and $\varepsilon_{\varphi_s}(l+r)$ decrease by 5.2 dB and 4.7 dB over the course of the first 32 recursions, followed by an increase of 11.2 dB and 6.1 dB at $r = 33$, respectively, and a subsequent decrease at roughly the same rate.

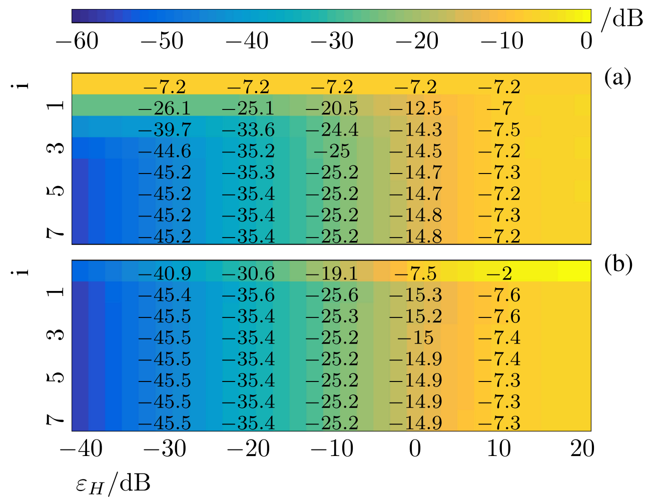

Fig. 4. $\varepsilon_{\varphi_s}^{(i)}$ versus ε_H and i for square-root MP with $\alpha = 10^3$ and $\hat{\boldsymbol{\varphi}}_s^{1/2|(0)}$ based upon (a) the sum constraint in (13) and (b) the estimator in (17)–(18) at $f = 2$ kHz.

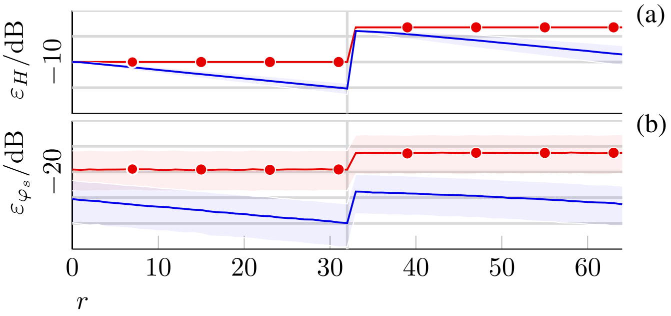

Fig. 5. (a) $\varepsilon_H(l+r)$ and (b) $\varepsilon_{\varphi_s}(l+r)$ versus r for conventional MP [—●—] and square-root MP [——] with $\alpha = 10^3$ at $f = 2$ kHz and $\varepsilon_H(l) = 0$ dB if $\mathbf{H}$ changes at $r = 32$ and remains constant otherwise.

B. Acoustic Data

We define the performance measures in Section VI-B1, discuss the acoustic scenario in Section VI-B2, the algorithmic settings in Section VI-B3, and the evaluation results in Section VI-B4.

1) Performance Measures: In the acoustic-data case, due to the model deficiencies in (9)–(15), cf. Section II, exact and observable ground truth early PSDs $\boldsymbol{\varphi}_s$ and ground truth RETFs $\mathbf{H}$ do unfortunately not exist, and so the performance measures in (67)–(70) cannot be used. However, one may define approximate

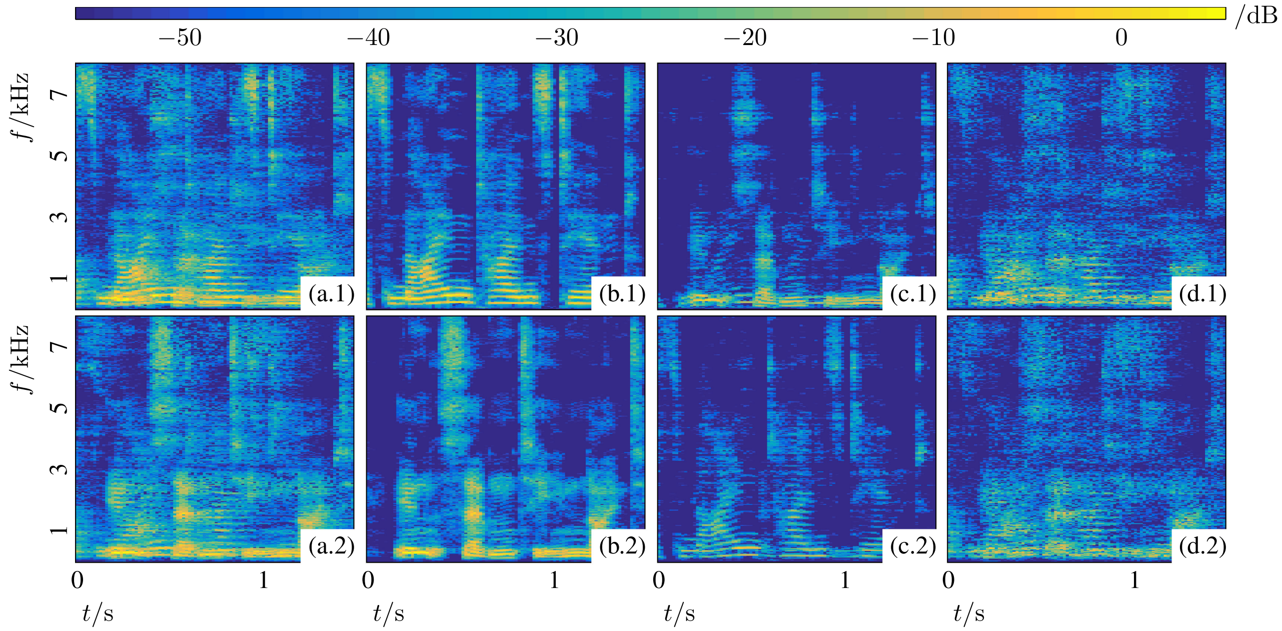

Fig. 6. Exemplary spectrograms depicting $\hat{\varphi}_{s_n}$ in (a.n), $\bar{\varphi}_{s_n}$ in (b.n), $e^{2|int}_{\varphi_{s_n}}$ in (c.n), and $e^{2|art}_{\varphi_{s_n}}$ in (d.n), with $\bar{\varphi}_{s_n}$, $e^{2|int}_{\varphi_{s_n}}$, and $e^{2|art}_{\varphi_{s_n}}$ obtained from the decomposition of $\sqrt{\hat{\varphi}_{s_n}}$, cf. Section VI-B1. The reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$ originate from a female and a male speaker at −30° and 60°, respectively, and the estimate $\hat{\varphi}_{s_n}$ is obtained by means of the square-root MP.

ground truth early PSDs $\tilde{\boldsymbol{\varphi}}_s$ as a reference for evaluation. To this end, given the source signals and RIRs of a particular acoustic scenario, cf. Section VI-B2, we convolve the source signals with only the early part of the RIR to the first microphone and transform to the STFT-domain, yielding $\tilde{\mathbf{s}}$, and set $\tilde{\boldsymbol{\varphi}}_s = \text{Diag}[\tilde{\mathbf{s}}^H]\tilde{\mathbf{s}}$, i.e. $\tilde{\boldsymbol{\varphi}}_s$ is the squared magnitude,[4] of $\tilde{\mathbf{s}}$. Note that the definition of the early part of the RIR is somewhat arbitrary due to the weighted and overlapping windows in the STFT-processing. For STFT windows of N_{STFT} samples with 50% overlap, one may, e.g., choose the first N_{STFT} or the first $N_{STFT}/2$ taps of the RIR. Here, we have chosen the first N_{STFT} samples corresponding to 32 ms, cf. Section VI-B3. In our setup, we have found that different choices result in quantitatively different performance, but not qualitatively different conclusions.

Given a segment of L frames of $\tilde{\boldsymbol{\varphi}}_s$ and $\hat{\boldsymbol{\varphi}}_s$, we decompose $\sqrt{\hat{\boldsymbol{\varphi}}_s}$ according to [26] as

$$\sqrt{\hat{\boldsymbol{\varphi}}_s} = \sqrt{\bar{\boldsymbol{\varphi}}_s} + \mathbf{e}^{int}_{\varphi_s} + \mathbf{e}^{art}_{\varphi_s}, \qquad (71)$$

where $\sqrt{\bar{\varphi}_{s_n}}$ is the component of $\sqrt{\hat{\varphi}_{s_n}}$ associated to $\sqrt{\tilde{\varphi}_{s_n}}$, i.e. the correctly estimated component, $e^{int}_{\varphi_{s_n}} = [\mathbf{e}^{int}_{\varphi_s}]_n$ contains components associated to $\sqrt{\tilde{\varphi}_{s_{n'}}}$ with $n' \neq n$, i.e. erroneously estimated leakage or interference components across sources, and $e^{art}_{\varphi_{s_n}} = [\mathbf{e}^{art}_{\varphi_s}]_n$ contains components not associated to any $\sqrt{\tilde{\varphi}_{s_n}}$, i.e. erroneously estimated artifact components. Exemplary spectrograms illustrating the decomposition in (71) are shown in Fig. 6, cf. also the discussion in Section VI-B4.

Given L frames of $\sqrt{\bar{\boldsymbol{\varphi}}_s}$, $\mathbf{e}^{int}_{\varphi_s}$ and $\mathbf{e}^{art}_{\varphi_s}$, we define the signal-to-interference ratio $SIR(\kappa)$, the signal-to-artifacts ratio $SAR(\kappa)$, and the signal-to-distortion ratio $SDR(\kappa)$ per third-octave band κ along the lines of [26] as

$$SIR(\kappa) = 10\log_{10}\frac{\sum_{k,l}\left\|\sqrt{\bar{\boldsymbol{\varphi}}_s}(k,l)\right\|_2^2}{\sum_{k,l}\left\|\mathbf{e}^{int}_{\varphi_s}(k,l)\right\|_2^2}\,\text{dB}, \qquad (72)$$

$$SAR(\kappa) = 10\log_{10}\frac{\sum_{k,l}\left\|\sqrt{\bar{\boldsymbol{\varphi}}_s}(k,l) + \mathbf{e}^{int}_{\varphi_s}(k,l)\right\|_2^2}{\sum_{k,l}\left\|\mathbf{e}^{art}_{\varphi_s}(k,l)\right\|_2^2}\,\text{dB}, \qquad (73)$$

$$SDR(\kappa) = 10\log_{10}\frac{\sum_{k,l}\left\|\sqrt{\bar{\boldsymbol{\varphi}}_s}(k,l)\right\|_2^2}{\sum_{k,l}\left\|\mathbf{e}^{int}_{\varphi_s}(k,l) + \mathbf{e}^{art}_{\varphi_s}(k,l)\right\|_2^2}\,\text{dB}, \qquad (74)$$

with $k = k_\kappa^-, \ldots, k_\kappa^+$ and k_κ^- and k_κ^+ the frequency-bin indices of the lower and upper band limits of third-octave-band κ, and $l = 0, \ldots, L-1$.

The decomposition in (71) relies on a segment of L frames of $\tilde{\boldsymbol{\varphi}}_s$ and $\hat{\boldsymbol{\varphi}}_s$ and is done in the following manner. Let $\underline{\hat{\boldsymbol{\varphi}}}_{s_n}$ be a vector stacking the early PSD estimates φ_{s_n} of source n over L observed frames, i.e. $\underline{\hat{\boldsymbol{\varphi}}}_{s_n} = (\hat{\varphi}_{s_n}(0) \quad \cdots \quad \hat{\varphi}_{s_n}(L{-}1))^T$, and let $\underline{\tilde{\boldsymbol{\varphi}}}_{s_n}$, $\underline{\bar{\boldsymbol{\varphi}}}_{s_n}$, $\underline{\mathbf{e}}^{int}_{\varphi_{s_n}}$ and $\underline{\mathbf{e}}^{art}_{\varphi_{s_n}}$ be defined equivalently, such that $\sqrt{\underline{\hat{\boldsymbol{\varphi}}}_{s_n}} = \sqrt{\underline{\bar{\boldsymbol{\varphi}}}_{s_n}} + \underline{\mathbf{e}}^{int}_{\varphi_{s_n}} + \underline{\mathbf{e}}^{art}_{\varphi_{s_n}}$, similarly to (71). Then, we perform the orthonormal projection of each individual vector $\sqrt{\underline{\hat{\boldsymbol{\varphi}}}_{s_n}}$ onto the one-dimensional subspace spanned by the corresponding vector $\sqrt{\underline{\tilde{\boldsymbol{\varphi}}}_{s_n}}$, yielding $\sqrt{\underline{\bar{\boldsymbol{\varphi}}}_{s_n}}$ with $\sqrt{\underline{\bar{\boldsymbol{\varphi}}}_{s_n}} \propto \sqrt{\underline{\tilde{\boldsymbol{\varphi}}}_{s_n}}$, as well as onto the N-dimensional subspace spanned by all N vectors $\sqrt{\underline{\tilde{\boldsymbol{\varphi}}}_{s_n}}$, yielding $\sqrt{\underline{\bar{\boldsymbol{\varphi}}}_{s_n}} + \underline{\mathbf{e}}^{int}_{\varphi_{s_n}}$, which then allows us to

[4] If subspace-based desmoothing, cf. Section V-B is not applied in the computation of $\hat{\boldsymbol{\varphi}}_s$, one can instead choose a recursively averaged version of the squared magnitude as a reference.

explicitly compute $\underline{\mathbf{e}}^{int}_{\varphi_{s_n}}$ and $\underline{\mathbf{e}}^{art}_{\varphi_{s_n}}$. For further details, we refer the interested reader to [26].

2) Acoustic Scenario: We use RIRs of 0.61 s reverberation time to a physical linear microphone array of $M = 5$ microphones with an inter-microphone distance of 8 cm [41], similar to the assumed microphone array in Section VI-A2. We simulate $N = 2$ sources, using female and male speech [42] as source signals. The sources are assigned to two out of three possible source positions in 2 m distance of the microphone array at $\{-30, 0, 60\}^\circ$ relative to the broad-side direction, yielding six different speaker-source-position combinations. From the two source signal files, we randomly select 32 segments of 5 s each. Per segment-pair, we generate microphone signals for each speaker-source-position combination.

3) Algorithmic Settings: In the acoustic-data case, the sampling frequency is $f_s = 16$ kHz, and the STFT-analysis and synthesis is based on square-root Hann windows of $N_{STFT} = 512$ samples (corresponding to 32 ms) with 50% overlap, resulting in $L = 312$ frames per segment. The desmoothed correlation matrix estimate $\hat{\mathbf{\Psi}}_x$ (cf. Section V-A and Section V-B) is computed using $\zeta = e^{-N_{STFT}/2f_s\tau}$ with $\tau = 160$ ms. As in Section VI-A2, $\mathbf{\Gamma}$ is computed assuming a spherical-isotropic sound field. Given $\hat{\mathbf{\Psi}}_x$ and $\mathbf{\Gamma}$, we compute the estimates $\hat{\varphi}_{x_\ell}$, $\hat{\mathbf{\Psi}}_{x_e}$ and $\hat{\mathbf{\Psi}}^{1/2}_{x_e}$ as described in Section V-C. We assume that the DoAs are known [7], [21], [22], and compute the (initial) estimate $\hat{\mathbf{H}}$ based on that. Note that in a reverberant environment, where the free-field assumption does not hold, the RETFs are generally not only defined by the DoA, but also by early reflections, and therefore we generally have $\hat{\mathbf{H}} \neq \mathbf{H}$ in our setup. Similarly to the model-based data case, cf. Section VI-A3, the penalty factor α in the conventional MP in (23) and the square-root MP in (34) is simulated in the range $\alpha \in [10^{-3}, 10^5]$. We perform at most $i_{\max} = 20$ iterations of the associated iterative algorithms in (24)–(25) and (35)–(36). While we do not update $\hat{\mathbf{H}}$ for the conventional MP in Section III, we consider two cases for the square-root MP in Section IV, namely first where we do not update $\hat{\mathbf{H}}$, and second where we update $\hat{\mathbf{H}}$ recursively as described in Section IV-C. In the latter case, in (49), since $\hat{\mathbf{\Psi}}^{1/2}_{x_e}$ is subject to modeling and estimation errors and contains residual late reverberation, we set $\varphi_{reg} = \hat{\varphi}_{x_\ell}$. In (50), the threshold ξ_{th} is again set as $10\log_{10}\xi_{th} = -2$ dB and β is set per third-octave band κ as $\beta(\kappa) = 20\hat{b}^2(\kappa)$, with $\hat{b}(\kappa)$ pre-defined as the diversity of the Laplace distributions fitted to the real and imaginary parts of the STFT coefficients of a training signal within third-octave band κ. Here, the training signal is generated from the entire female and male speech source signals, cf. Section VI-B2, by convolving the early part of the RIR of the first microphone corresponding to a source at 2 m distance at 0° relative to the broadside direction, cf. also the similar segment-wise definition of the reference signal $\tilde{s}_n$ in Section VI-B1. Note that while $\hat{b}(\kappa)$ is pre-computed using all STFT coefficients of both male and female speech within third-octave band κ, the actual distributions vary across speakers, across source positions, across individual frequency bins, and across individual segments, cf. also Section VI-B2.

4) Results: Before discussing the performance of the conventional MP and the square-root MP in terms of the measures *SIR*, *SAR*, and *SDR*, we first consider the examplary spectrograms in Fig. 6 visualizing the decomposition of $\sqrt{\hat{\boldsymbol{\varphi}}_s}$ upon which these measures are based. In this example, the microphone signals $\mathbf{x}$ and the reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$ originate from a female and a male speaker at -30° and 60°, respectively, and the estimates $\hat{\varphi}_{s_1}$ and $\hat{\varphi}_{s_2}$ in Fig. 6(a.1) and Fig. 6(a.2) are obtained by means of the square-root MP. The correctly estimated components $\bar{\varphi}_{s_1}$ and $\bar{\varphi}_{s_2}$ in Fig. 6(b.1) and Fig. 6(b.2) are frequency-bin-wise scaled versions of the reference PSDs $\tilde{\varphi}_{s_1}$ and $\tilde{\varphi}_{s_2}$, respectively, cf. Section VI-B1. As can be seen, the leakage or interference components in $e^{2|int}_{\varphi_{s_1}}$ and $e^{2|int}_{\varphi_{s_2}}$ in Fig. 6(c.1) and Fig. 6(c.2) relate to the opposing reference PSDs, cf. Fig. 6(b.2) and Fig. 6(b.1), respectively. Finally, the artifact components $e^{2|art}_{\varphi_{s_1}}$ and $e^{2|art}_{\varphi_{s_2}}$ in Fig. 6(d.1) and Fig. 6(d.2) do not relate to any of the reference PSDs, but rather to residual late reverberation in the estimate $\hat{\mathbf{\Psi}}^{1/2}_{x_e}$, cf. also Section V-C, which is due to modeling errors in (9)–(14) and a potential deviation of the late reverberant sound field from the spatial coherence matrix $\mathbf{\Gamma}$. Note that in $e^{2|art}_{\varphi_{s_1}}$ and $e^{2|art}_{\varphi_{s_2}}$, the energy is concentrated in the same spectro-temporal regions, indicating a similar spatial sound field of these components.

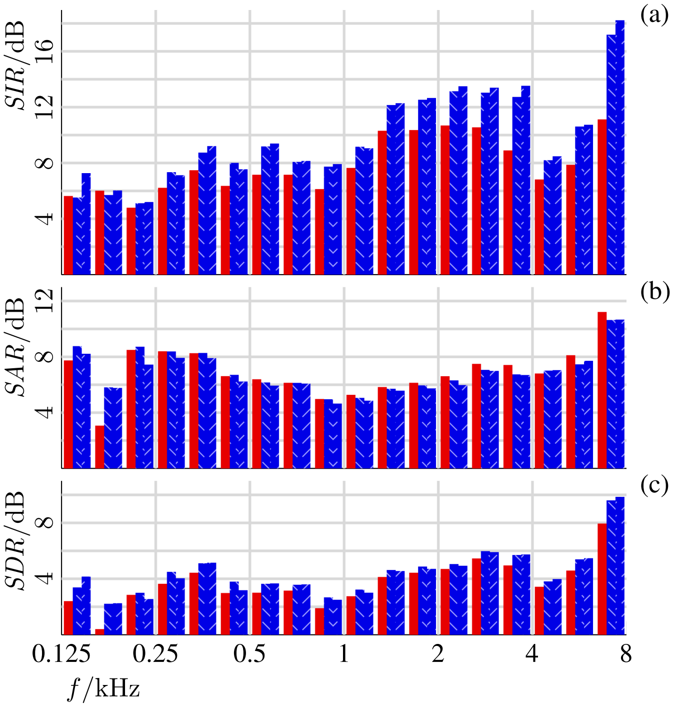

Fig. 7. (a) *SIR*, (b) *SAR*, and (c) *SDR* in third-octave bands for conventional MP [🟥], square-root MP without recursive RETF update [▧], and square-root MP with recursive RETF update [▨].

Fig. 7 shows the median over all segments and speaker-source-combinations, cf. Section VI-B2, of (a) *SIR*, (b) *SAR*, and (c) *SDR* in third-octave bands for the conventional MP [🟥], the square-root MP without recursive RETF update [▧], and the square-root MP with recursive RETF update [▨]. Here, in each third-octave band κ, we have selected $\alpha(\kappa)$ such that $SIR(\kappa)$ is maximized for each algorithm, i.e. the figure indicates their upper performance limit in terms of $SIR(\kappa)$ with respect to the tuning of $\alpha(\kappa)$. Note that in our setup, selecting $\alpha(\kappa)$ to maximize $SAR(\kappa)$ or $SDR(\kappa)$ does not lead to qualitatively substantial differences. For the conventional MP, we have found values of $\alpha(\kappa) \ll 1$ to be preferable in all third-octave bands κ, indicating that the soft-constraint penalty in (23) is not very useful in practice. For the square-root MP, with and without recursive RETF update, we have found $\alpha(\kappa) \gg 1$ to be preferable in third-octave bands below 0.5 kHz, and $\alpha(\kappa) \leq 1$ to be preferable above 0.5 kHz. From Fig. 7(a), we find that the square-root MP clearly outperforms the conventional MP in terms of SIR in third-octave bands above 0.25 kHz, with improvements of 1 dB to 6 dB, indicating better source-component separation performance. Further, for the square-root MP, we find that the recursive RETF update mildly improves the performance by up to 1 dB. Recall that the initial RETF estimate $\hat{\mathbf{H}}$ is based on the correct DoAs, but does not consider early reflections, cf. Section VI-B3. From Fig. 7(b), we note that for all algorithms, we have $SAR(\kappa) < SIR(\kappa)$ in third-octave bands above 0.5 kHz, indicating comparably strong residual late reverberation. The square-root MP performs slightly worse than the conventional MP in terms of SAR in third-octave bands above 0.25 kHz, with degradations of less than 1 dB. In the square-root MP, recursive RETF updating results in minor differences only. As can be seen from Fig. 7(c), we find that the square-root MP outperforms the conventional MP in terms of SDR, however, due to the comparably strong residual late reverberation, by much less than in terms of SIR. Again, in the square-root MP, recursive RETF updating results in minor differences only.

VII. Conclusion

We have discussed early PSD estimation and recursive RETF updates in the STFT domain for multiple sources in reverberant environments, based on a commonly used multi-microphone correlation matrix model, given (initial) RETF estimates. State-of-the-art approaches to early PSD estimation minimize the approximation error with respect to an estimate of the early correlation matrix, referred to as conventional MP. Instead, we here have factorized the early correlation matrix model and minimized the approximation error with respect to an estimate of the early-correlation-matrix square root, which we referred to as the square-root MP. The square-root MP seeks a unitary matrix and the square roots of the early PSDs up to an arbitrary complex argument, and therewith constitutes a generalization of the orthogonal Procrustes problem. As opposed to the conventional MP, non-negative inequality constraints are not required in the square-root MP. The square-root MP may be solved iteratively, requiring one SVD per iteration. Based on the estimated unitary matrix and early PSD square roots, we are further able to recursively update the RETF estimate, which is not inherently possible in the conventional approach. The respectively required estimates of the early correlation matrix and the early-correlation-matrix square root may be obtained from an estimate of the microphone signal correlation matrix and the diffuse coherence matrix by means of the GEVD. Hereat, in order to compensate for inevitable recursive averaging, we have restored non-stationarities by desmoothing the generalized eigenvlaues.

In order to evaluate the proposed approach, we have performed two kinds of simulations. In the first kind, the data is generated based on the microphone signal correlation matrix model and assumed geometric and physical properties, excluding modeling errors from the evaluation. This is referred to as model-based-data case. In the second kind, the data is generated from recorded speech and measured RIRs, creating a more practical setup. This is referred to as acoustic-data case. In the model-based-data case, the simulation results indicate better performance of the square-root MP as compared to the conventional MP in terms of the relative squared PSD estimation error. If initialized accordingly, the square-root MP can be solved in only one iteration. In the acoustic-data case, the simulation results indicate better performance of the square-root MP as compared to the conventional MP in terms of the source-component separation measured by the signal-to-interference ratio. Both the square-root MP and the conventional MP suffer somewhat from residual late reverberation in the early-correlation-matrix estimate.

References

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

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

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

[4] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[6] H. Q. Dam, S. Y. Low, H. H. Dam, and S. Nordholm, "Space constrained beamforming with source PSD updates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* Montreal, QC, Canada, May 2004, pp. 93–96.

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

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

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

[10] 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 Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 380–384.

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

[12] 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. 24th Eur. Signal Process. Conf.*, Budapest, Hungary, Aug. 2016, pp. 1123–1127.

[13] I. Kodrasi and S. Doclo, "Joint late reverberation and noise power spectral density estimation in a spatially homogeneous noise field," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 441–445.

[14] S. Braun *et al.*, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[15] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1102–1114, Jun. 2018.

[16] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multi-microphone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, Jul. 2019.

[17] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[18] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2374–2384, Nov. 2011.

[19] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 204–210, Jul. 2000.

[20] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[21] J. Scheuing and B. Yang, "Correlation-based TDOA-estimation for multiple sources in reverberant environments," in *Speech and Audio Processing in Adverse Environments*. Berlin, Germany: Springer, 2008, pp. 381–416.

[22] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. Norwood, MA, USA: Artech House, 2010.

[23] R. Everson, "Orthogonal, but not orthonormal, Procrustes problems," Tech. Rep., Laboratory for Applied Mathematics, City Univ. New York, New York, NY, USA, and Mount Sinai Medical School, New York, NY, USA, 1998.

[24] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966.

[25] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 635–650, Mar. 2002.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] T. Dietzen, "GitHub repository: Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," Jul. 2019, [Online]. Available: https://github.com/tdietzen/SQRT-PSD-RETF

[28] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, to be published, doi: 10.1109/TASLP.2020.2966869.

[29] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, Apr. 2007.

[30] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[31] N. Antonello, L. Stella, P. Patrinos, and T. van Waterschoot, "Proximal gradient algorithms: Applications in signal processing," ESAT-STADIUS Tech. Rep. TR 17-112, KU Leuven, Leuven, Belgium, Jan. 2018.

[32] T. van Waterschoot, G. Rombouts, and M. Moonen, "Optimally regularized adaptive filtering algorithms for room acoustic signal enhancement," *Signal Process.*, vol. 88, no. 3, pp. 594–611, Mar. 2008.

[33] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA, USA: MIT Press, 1986.

[34] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[35] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 544–548.

[36] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NJ, USA: Prentice-Hall., 2002.

[37] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins Univ. Press: Baltimore, MD, USA, 2012, vol. 3.

[38] M. Tammen, I. Kodrasi, and S. Doclo, "Complexity reduction of eigenvalue decomposition-based diffuse power spectral density estimators using the power method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 451–455.

[39] M. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5-2, pp. 845–856, Sep. 2005.

[40] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, May 2005.

[41] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC 2014)*, Antibes – Juan les Pins, France, Sep. 2014, pp. 313–317.

[42] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.

Thomas Dietzen received the Dipl.Ing. degree from Kaiserslautern University, Kaiserslautern, Germany, in 2011, and the Ph.D. degree from KU Leuven, Leuven, Belgium, in 2019. He is currently a Postdoctoral Researcher with KU Leuven. Between 2012 and 2014, he was a Research Assistant with the University of Heidelberg, Heidelberg, Germany, and with the Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany. From 2014 to 2017, he was a Doctoral Researcher with NXP Semiconductors Belgium NV, Leuven, Belgium. His research interests include room acoustic modeling and signal enhancement in adverse acoustic conditions, specifically on spatio-temporal adaptive filtering and power-spectral-density estimation.

Mr. Dietzen has served as a Reviewer for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, the IEEE SIGNAL PROCESSING LETTERS, and the *EURASIP Journal on Audio, Speech, and Music Processing*.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from Katholieke Universiteit Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders, Electrical Engineering Department, Katholieke Universiteit Leuven and the Cognitive Systems Laboratory, McMaster University, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors, Leuven, Belgium. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and Scientific Advisor for the Division Hearing, Speech and Audio Technology with the Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks, and hearing aid processing.

Prof. Doclo was the recipient of the several best paper awards (International Workshop on Acoustic Echo and Noise Control in 2001, EURASIP Signal Processing in 2003, IEEE Signal Processing Society in 2008, and VDE Information Technology Society in 2019). He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing, and the EAA Technical Committee on Audio Signal Processing. He was and is involved in several large-scale national and European research projects (ITN DREAMS, Cluster of Excellence Hearing4all, and CRC Hearing Acoustics). He was a Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in 2013 and a Chair of the ITG Conference on Speech Communication in 2018. In addition, he served as a Guest Editor for several special issues, such as IEEE SIGNAL PROCESSING MAGAZINE, *Elsevier Signal Processing*, and was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*.

Marc Moonen (Fellow, IEEE) is currently a Full Professor with the Electrical Engineering Department, KU Leuven, Leuven, Belgium, where he is also heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL, and audio signal processing.

Prof. Moonen is a Fellow of the European Association for Signal Processing (EURASIP). He was the recipient of the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with P. Vandaele), the 2004 Alcatel Bell (Belgium) Award (with R. Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He was also the recipient of Journal best paper awards from the IEEE TRANSACTIONS ON SIGNAL PROCESSING (with G. Leus and D. Giacobello) and *Elsevier Signal Processing* (with S. Doclo). He was a Chairman of the IEEE Benelux Signal Processing Chapter, from 1998 to 2002, a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and the President of EURASIP, from 2007 to 2008 and from 2011 to 2012. He has served as an Editor-in-Chief for the *EURASIP Journal on Applied Signal Processing*, from 2003 to 2005, an Area Editor for Feature Articles in the IEEE SIGNAL PROCESSING MAGAZINE, from 2012 to 2014, and has been a member of the Editorial Board for the *Signal Processing*, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II, the IEEE SIGNAL PROCESSING MAGAZINE, *Integration the VLSI Journal*, *EURASIP Journal on Wireless Communications and Networking*, and *EURASIP Journal on Advances in Signal Processing*.

Toon van Waterschoot (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from KU Leuven, Leuven, Belgium, in 2001 and 2009, respectively. He is currently with KU Leuven as an Associate Professor and Consolidator Grantee of the European Research Council. He has also held teaching and research positions with the Delft University of Technology, Delft, The Netherlands and the University of Lugano, Lugano, Switzerland. His research interests include signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction.

Mr. van Waterschoot has been serving as an Associate Editor for the *Journal of the Audio Engineering Society* and for the *EURASIP Journal on Audio, Music, and Speech Processing*, and as a Guest Editor for the *Elsevier Signal Processing*. He is a Director of the European Association for Signal Processing (EURASIP), a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, a member of the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair of the 60th AES International Conference in Leuven, Belgium, in 2016, and has been serving on the Organizing Committee of the European Conference on Computational Optimization since 2016, the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics since 2017, and the 28th European Signal Processing Conference since 2020. He is a member of EURASIP, ASA, and AES.