Robust Joint Estimation of Multimicrophone Signal Model Parameters

Andreas I. Koutrouvelis, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen

***Abstract*—One of the biggest challenges in multimicrophone applications is the estimation of the parameters of the signal model, such as the power spectral densities (PSDs) of the sources, the early (relative) acoustic transfer functions of the sources with respect to the microphones, the PSD of late reverberation, and the PSDs of microphone-self noise. Typically, existing methods estimate subsets of the aforementioned parameters and assume some of the other parameters to be known *a priori*. This may result in inconsistencies and inaccurately estimated parameters and potential performance degradation in the applications using these estimated parameters. So far, there is no method to jointly estimate all the aforementioned parameters. In this paper, we propose a robust method for jointly estimating all the aforementioned parameters using confirmatory factor analysis. The estimation accuracy of the signal-model parameters thus obtained outperforms existing methods in most cases. We experimentally show significant performance gains in several multimicrophone applications over state-of-the-art methods.**

***Index Terms*—Confirmatory factor analysis, dereverberation, joint diagonalization, multimicrophone, source separation, speech enhancement.**

I. Introduction

MICROPHONE arrays (see e.g., [1] for an overview) are used extensively in many applications, such as source separation [2]–[6], multi-microphone noise reduction [1], [7]–[13], dereverberation [14]–[19], sound source localization [20]–[23], and room geometry estimation [24], [25]. All the aforementioned applications are based on a similar multi-microphone signal model, typically depending on the following parameters: i) the early relative acoustic transfer functions (RATFs) of the sources with respect to the microphones, ii) the power spectral densities (PSDs) of the early components of the sources, iii) the PSD of the late reverberation, and, iv) the PSDs of the microphone-self noise. Other parameters, like the target cross power spectral density matrix (CPSDM), the noise CPSDM, source locations and room geometry information, can be inferred from (combinations of) the above mentioned parameters. Often, none of these parameters are known *a priori*, while estimation is challenging. Often, only a subset of the parameters is estimated, see e.g., [14]–[17], [19], [26]–[30], typically requiring rather strict assumptions with respect to stationarity and/or knowledge of the remaining parameters.

In [15], [17] the target source PSD and the late reverberation PSD are jointly estimated assuming that the early RATFs of the target with respect to all microphones are known and all the remaining noise components (e.g., interferers) are stationary in time intervals typically much longer than a time-frame. In [19], [26], [31], it was shown that the method in [15], [17] may lead to inaccurate estimates of the late reverberation PSD, when the early RATFs of the target include estimation errors. In [19], [26], a more accurate estimator for only the late reverberation PSD was proposed, independent of early RATF estimation errors.

The methods proposed in [27], [28] do not assume that some noise components are stationary like in [17], but assume that the total noise CPSDM has a constant [27] or slow-varying [28] structure over time (i.e., it can be written as an unknown scaling parameter multiplied with a constant spatial structure matrix). This may not be realistic in practical acoustical scenarios, where different interfering sources change their power and location across time more rapidly and with different patterns. Moreover, these methods do not separate the late reverberation from the other noise components and only differentiate between the target source PSD and the overall noise PSD. As in [17], these methods assume that the early RATFs of the target are known. In [28], the structure of the noise CPSDM is estimated only in target-absent time-frequency tiles using a voice activity detector (VAD), which may lead to erroneous estimates if the spatial structure matrix of the noise changes during target-presence.

In [30], the early RATFs and the PSDs of all sources are estimated using the expectation maximization (EM) method [32]. This method assumes that only one source is active per time-frequency tile and the noise CPSDM (excluding the contributions of the interfering point sources) is estimated assuming it is time-invariant. Due to the time-varying nature of the late reverberation (included in the noise CPSDM), this assumption is often violated. This method does not estimate the time-varying PSD of the late reverberation, neither the PSDs of the microphone-self noise.

While the aforementioned methods focus on estimation of just one or several of the required model parameters, the method

Manuscript received October 12, 2018; revised March 14, 2019; accepted April 9, 2019. Date of publication April 15, 2019; date of current version May 7, 2019. This work was supported by the Oticon Foundation and NWO, the Dutch Organisation for Scientific Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W. H. Khong. (*Corresponding author: Andreas I. Koutrouvelis.*)

A. I. Koutrouvelis, R. C. Hendriks, and R. Heusdens are with the Department of Microelectronics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft 2628, CD, The Netherlands (e-mail: a.koutrouvelis@tudelft.nl; r.c.hendriks@tudelft.nl; r.heusdens@tudelft.nl).

J. Jensen is with the Oticon A/S, Smørum 2765, Denmark, and also with the Electronic Systems Department, Aalborg University, Aalborg 9100, Denmark (e-mail: jesj@oticon.com).

Digital Object Identifier 10.1109/TASLP.2019.2911167

presented in [4] jointly estimates the early RATFs of the sources, the PSDs of the sources and the PSDs of the microphone-self noise. Unlike [30], the method in [4] does not assume single source activity per time-frequency tile and, thus, it is applicable to more general acoustic scenarios. The method in [4] is based on the non-orthogonal joint-diagonalization of the noisy CPSDMs. This method unfortunately does not guarantee non-negative estimated PSDs and, thus, the obtained target CPSDM may not be positive semidefinite resulting in performance degradation. Moreover, this approach does not estimate the PSD of the late reverberation. In conclusion, most methods only focus on the estimation of a subset of the required model parameters and/or rely on assumptions which may be invalid and/or impractical.

In this paper, we propose a method which jointly estimates all the aforementioned parameters of the multi-microphone signal model. The proposed method is based on confirmatory factor analysis (CFA) [33]–[36]—a statistical theory more known and successfully applied in the field of psychology [36]—and on the non-orthogonal joint-diagonalization principle introduced in [4]. The combination of these two theories and the adjustment to the multi-microphone case gives us a robust method, which is applicable for temporally and spatially non-stationary sources. Unlike the methods in [15], [17], [19], [26]–[28], the proposed method uses linear constraints to reduce the feasibility set of the parameter space and thus increase robustness. Moreover, the proposed method guarantees positive estimated PSDs and, thus, positive semidefinite target and noise CPSDMs thanks to the CFA framework. Although generally applicable, in this manuscript, we will compare the performance of the proposed method with state-of-the-art approaches in the context of source separation and dereverberation.

The large number of parameters that are jointly estimated using the proposed method comes with its challenges. In this paper, we provide several identifiability conditions which should be satisfied in order to obtain reliable estimates of the parameters. For instance, we need to guarantee that the system of equations is sufficiently over-determined (i.e., more equations than parameters) in order to fit accurately the signal model to the noisy estimated CPSDM. The over-determination is achieved either by increasing the number of equations or by omitting the estimation of some parameters which in some acoustic scenarios do not play a significant role. For instance, if the late reverberation is low and the number of equations is small, we may skip estimating the late reverberation parameter and the estimation accuracy of the remaining parameters might be improved due to the increased over-determination. In this paper, we examine scenarios with different levels of reverberation and we experimentally show the trade-off between over-determination and estimation accuracy.

The remaining part of this paper is organized as follows. In Sec. II, the signal model, notation and used assumptions are introduced. In Sec. III, we review the CFA theory and its relation to the non-orthogonal joint diagonalization principle. In Sec. IV, the proposed method is introduced. In Sec. V, we introduce several constraints to increase the robustness of the proposed method. In Sec. VI, we discuss the implementation and practicality of the proposed method. In Sec. VII, we conduct experiments in several multi-microphone applications using the proposed method and existing state-of-the-art approaches. In Sec. VIII, we draw conclusions.

II. Preliminaries

A. Notation

We use lower-case letters for scalars, bold-face lower-case letters for vectors, and bold-face upper-case letters for matrices. A matrix $\mathbf{A}$ can be expressed as $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$, where $\mathbf{a}_i$ is its i-th column. The elements of a matrix $\mathbf{A}$ are denoted as a_{ij}. We use the operand $\mathrm{tr}(\cdot)$ to denote the trace of a matrix, $\mathrm{E}[\cdot]$ to denote the expected value of a random variable, $\mathrm{diag}(\mathbf{A}) = [a_{11}, \cdots, a_{mm}]^T$ to denote the vector formed from the diagonal of a matrix $\mathbf{A} \in \mathbb{C}^{m \times m}$, and $||\cdot||_F^2$ to denote the Frobenius norm of a matrix. We use $\mathrm{Diag}(\mathbf{v})$ to form a square diagonal matrix with diagonal $\mathbf{v}$. A hermitian positive semi-definite matrix is expressed as $\mathbf{A} \succeq 0$, where $\mathbf{A} = \mathbf{A}^H$ and its eigenvalues are real and non-negative. The cardinality of a set is denoted as $|\cdot|$. The minimum element of a vector $\mathbf{v}$ is obtained via the operation $\min(\mathbf{v})$.

B. Signal Model

Consider an M-element microphone array of arbitrary structure within a possibly reverberant enclosure, in which there are r acoustic point sources (target and interfering sources). The i-th microphone signal (in the short-time Fourier transform (STFT) domain) is modeled as

$$y_i(t,k) = \sum_{j=1}^{r} e_{ij}(t,k) + \sum_{j=1}^{r} l_{ij}(t,k) + v_i(t,k), \qquad (1)$$

where k is the frequency-bin index; t the time-frame index; e_{ij} and l_{ij} the early and late components of the j-th point source, respectively; and v_i denotes the microphone self-noise. The early components include the line of sight and a few initial strong reflections. The late components describe the effect of the remaining reflections and are usually referred to as late reverberation. The j-th early component is given by

$$e_{ij}(t,k) = a_{ij}(\beta,k) s_j(t,k), \qquad (2)$$

where $a_{ij}(\beta,k)$ is the corresponding RATF with respect to the i-th microphone, $s_j(t,k)$ the j-th point-source at the reference microphone, β is the index of a *time-segment*, which is a collection of *time-frames*. That is, we assume that the source signal can vary faster (from time-frame to time-frame) than the early RATFs, which are assumed to be constant over multiple time-frames (which we call a time-segment). By stacking all microphone recordings into vectors, the multi-microphone signal model is given by

$$\mathbf{y}(t,k) = \sum_{j=1}^{r} \underbrace{\mathbf{a}_j(\beta,k) s_j(t,k)}_{\mathbf{e}_j(t,k)} + \underbrace{\sum_{j=1}^{r} \mathbf{l}_j(t,k)}_{\mathbf{l}(t,k)} + \mathbf{v}(t,k) \in \mathbb{C}^{M \times 1}, \qquad (3)$$

where $\mathbf{y}(t,k) = [y_1(t,k), \cdots, y_M(t,k)]^T$ and all the other vectors can be similarly represented. If we assume that all sources

in (3) are mutually uncorrelated and stationary within a time-frame, the signal model of the CPSDM of the noisy recordings is given by

$$\mathbf{P}_{\mathbf{y}}(t,k) = \sum_{j=1}^{r} \mathbf{P}_{\mathbf{e}_j}(t,k) + \mathbf{P}_{\mathbf{l}}(t,k) + \mathbf{P}_{\mathbf{v}}(k) \in \mathbb{C}^{M\times M}, \tag{4}$$

where $\mathbf{P}_{\mathbf{e}_j}(t,k) = p_j(t,k)\mathbf{a}_j(\beta,k)\mathbf{a}_j^H(\beta,k), p_j = E[|s_j(t,k)|^2]$ is the PSD of the j-th source at the reference microphone, $\mathbf{P}_{\mathbf{l}}(t,k)$ the CPSDM of the late reverberation and $\mathbf{P}_{\mathbf{v}}(k)$ is a diagonal matrix, which has as its diagonal elements the PSDs of the microphone-self noise. Note that $p_j(t,k)$ and $\mathbf{P}_{\mathbf{l}}(t,k)$ are time-frame varying, while the microphone-self noise PSDs are typically time-invariant. The CPSDM model in (4) can be re-written as

$$\mathbf{P}_{\mathbf{y}}(t,k) = \mathbf{P}_{\mathbf{e}}(t,k) + \mathbf{P}_{\mathbf{l}}(t,k) + \mathbf{P}_{\mathbf{v}}(k), \tag{5}$$

where $\mathbf{P}_{\mathbf{e}}(t,k) = \mathbf{A}(\beta,k)\mathbf{P}(t,k)\mathbf{A}^H(\beta,k)$ and $\mathbf{A}(\beta,k) \in \mathbb{C}^{M\times r}$ is commonly referred to as mixing matrix and has as its columns the early RATFs of the sources. As we work with RATFs, the row of $\mathbf{A}(\beta,k)$ corresponding to the reference microphone is equal to a vector with only ones. Moreover, $\mathbf{P}(t,k)$ is a diagonal matrix, where $\text{diag}\,(\mathbf{P}(t,k)) = [p_1(t,k),\cdots,p_r(t,k)]^T$.

C. Late Reverberation Model

A commonly used assumption (adopted in this paper) is that the late reverberation CPSDM has a known spatial structure, $\mathbf{\Phi}(k)$, which is time-invariant but varying over frequency [14], [17]. Under this constant spatial-structure assumption, $\mathbf{P}_{\mathbf{l}}(t,k)$ is modeled as [14], [17]

$$\mathbf{P}_{\mathbf{l}}(t,k) = \gamma(t,k)\mathbf{\Phi}(k), \tag{6}$$

with $\gamma(t,k)$ the PSD of the late reverberation which is unknown and needs to be estimated. By combining (5), and (6), we obtain the final CPSDM model given by

$$\mathbf{P}_{\mathbf{y}}(t,k) = \mathbf{P}_{\mathbf{e}}(t,k) + \gamma(t,k)\mathbf{\Phi}(k) + \mathbf{P}_{\mathbf{v}}(k). \tag{7}$$

There are several existing methods [15], [17]–[19], [26] to estimate $\gamma(t,k)$ under the assumption that $\mathbf{\Phi}(k)$ is known. There are mainly two methodologies of obtaining $\mathbf{\Phi}(k)$. The first is to use many pre-calculated impulse responses measured around the array as in [7]. The second is to use a model which is based on the fact that the off-diagonal elements of $\mathbf{\Phi}(k)$ depend on the distance between every microphone pair. The distances between any two microphone pairs is described by the symmetric microphone-distance matrix $\mathbf{D}$ with elements d_{ij} which is the distance between microphones i and j. Two commonly used models for the spatial structure are the cylindrical and spherical isotropic noise fields [10], [37]. The cylindrical isotropic noise field is accurate for rooms where the ceiling and the floor are more absorbing than the walls. These models are accurate for sufficiently large rooms [10].

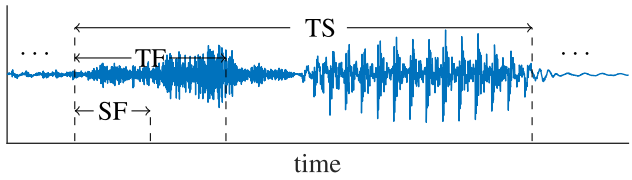


Fig. 1. Splitting time into time-segments (TS), time-frames (TF), and sub-frames (SF).

D. Estimation of CPSDMs Using Subframes

The estimation of $\mathbf{P}_{\mathbf{y}}(t,k)$ for the t-th time-frame, is achieved using multiple overlapping *sub-frames* (each of length N) within the t-th time-frame. The length of a time-frame is $\mathcal{T} \gg N$. Let $\mathbf{y}_\theta(t,k)$ the noisy DFT coefficients at the θ-th sub-frame and k-th frequency-bin of the t-th time-frame. The FFT length, K, is selected as the next power of two that is larger than N. The set of all used sub-frames within the t-th time-frame is denoted by Θ_t, and the number of used sub-frames is $|\Theta_t|$. We assume that the noisy microphone signals within a time-frame are stationary and, thus, we can estimate the noisy CPSDM using the sample CPSDM, i.e.,

$$\hat{\mathbf{P}}_{\mathbf{y}}(t,k) = \frac{1}{|\Theta_t|}\sum_{\theta\in\Theta_t} \mathbf{y}_\theta(t,k)\mathbf{y}_\theta^H(t,k). \tag{8}$$

To summarize, we have introduced sub-frames, time-frames and time-segments, which split time at different levels of hierarchy. This is visualized in Fig. 1.

E. Problem Formulation

The goal of this paper is to jointly estimate the parameters $\mathbf{A}(\beta,k)$, $\mathbf{P}(t,k)$, $\gamma(t,k)$, and $\mathbf{P}_{\mathbf{v}}(k)$ for the β-th time-segment of the signal model in (7) by only having estimates of the noisy CPSDM matrices $\hat{\mathbf{P}}_{\mathbf{y}}(t,k)$ for all time-frames belonging to the β-th time-segment and depending on the exact method that we discuss, an estimate of $\mathbf{\Phi}(k)$ and/or $\mathbf{D}$. From now on, we will neglect time-frequency indices to simplify notation and only use time-frequency indices wherever needed to avoid ambiguity.

III. Confirmatory Factor Analysis

In this section we review the confirmatory factor analysis (CFA) method [33], [34], [36], the simultaneous CFA (SCFA) method [35] and their relationship to the non-orthogonal joint diagonalization method proposed in [4]. This review is beneficial for understanding the proposed method in Sec. IV. CFA aims at estimating the parameters of the following CPSDM model:

$$\mathbf{P}_{\mathbf{y}} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \mathbf{P}_{\mathbf{v}} \in \mathbb{C}^{M\times M}, \tag{9}$$

where $\mathbf{P}_{\mathbf{v}} = \text{Diag}([q_1,\cdots,q_M]^T)$ and $\mathbf{P} \succeq 0$. The signal model in (9) is different from our signal model that we presented in (5) and (7). At first, unlike (7), $\mathbf{P}$ is not necessarily a diagonal matrix in (9). Secondly, unlike (7), the model in (9) does not take into account the late reverberation component.

Since we have complex values in (9), there will be a real and an imaginary part and, thus, the number of equations and unknowns will be doubled compared to the case where (9) consists of

purely real values. Specifically, since $\mathbf{P_y} \succeq 0$, there are $M(M+1)/2$ complex-valued equations, which means $M(M+1)$ real-valued equations in total. Similarly, $\mathbf{P}$ has $r(r+1)$ unknowns, because $\mathbf{P} \succeq 0$, while the matrix $\mathbf{A}$ has $2Mr$ unknowns. Finally, the matrix $\mathbf{P_v}$ has M unknowns because it is a real diagonal matrix.

In CFA, some of the real and imaginary parts of the elements in $\mathbf{A}$ and $\mathbf{P}$ are fixed such that the remaining variables are uniquely identifiable (see below). Specifically, let Υ_R and $\mathcal{K}_R$ denote the sets of the selected indices of the matrices $\mathbf{A}$ and $\mathbf{P}$, respectively, where the real part of their elements are fixed to a known constant $\tilde{a}_{ij}^R$, and $\tilde{p}_{k\nu}^R$. Similarly, let Υ_I and $\mathcal{K}_I$ denote the sets of the selected indices of the matrices $\mathbf{A}$ and $\mathbf{P}$, respectively, where the imaginary part of their elements are fixed to a known constant $\tilde{a}_{\epsilon\zeta}^I$, and $\tilde{p}_{\eta\iota}^I$. Furthermore, let $\Upsilon = \Upsilon_R \cup \Upsilon_I$ and $\mathcal{K} = \mathcal{K}_R \cup \mathcal{K}_I$.

There are several existing CFA methods (see e.g., [36], for an overview). Most of these are special cases of the following general CFA problem

$$\begin{aligned}
\hat{\mathbf{A}}, \hat{\mathbf{P}}, \hat{\mathbf{P}}_{\mathbf{v}} &= \underset{\mathbf{A},\mathbf{P},\mathbf{P_v}}{\arg\min}\, F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P_y}) \\
\text{s.t.} \quad & \mathbf{P_y} = \mathbf{APA}^H + \mathbf{P_v}, \\
& \mathbf{P_v} = \text{Diag}([q_1, \cdots, q_M]^T) \in \mathbb{R}^{M\times M}, \\
& q_i \geq 0, i = 1, \cdots, M, \\
& \mathbf{P} \succeq 0, \\
& \Re(a_{ij}) = \tilde{a}_{ij}^R, \forall (i,j) \in \Upsilon^R, \\
& \Im(a_{\epsilon\zeta}) = \tilde{a}_{\epsilon\zeta}^I, \forall (\epsilon,\zeta) \in \Upsilon^I, \\
& \Re(p_{k\nu}) = \tilde{p}_{k\nu}^R, \forall (k,\nu) \in \mathcal{K}^R, \\
& \Im(p_{\eta\iota}) = \tilde{p}_{\eta\iota}^I, \forall (\eta,\iota) \in \mathcal{K}^I
\end{aligned} \tag{10}$$

with $F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P_y})$ a cost function, which is typically one of the following cost functions: maximum likelihood (ML), least squares (LS), or generalized least squares (GLS). That is,

$$F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P_y}) = \begin{cases} \text{(ML): } \log|\mathbf{P_y}| + \text{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right), & [34], \\ \text{(LS): } \frac{1}{2}||\mathbf{P_y} - \hat{\mathbf{P}}_{\mathbf{y}}||_F^2, & [36], [38], \\ \text{(GLS): } \frac{1}{2}||\hat{\mathbf{P}}_{\mathbf{y}}^{-\frac{1}{2}}(\mathbf{P_y} - \hat{\mathbf{P}}_{\mathbf{y}})\hat{\mathbf{P}}_{\mathbf{y}}^{-\frac{1}{2}}||_F^2, & [39], \end{cases} \tag{11}$$

where $\mathbf{P_y}$ is given in (9). Notice, that the problem in (10) is not convex (due to the non-convex term $\mathbf{APA}^H$) and may have multiple local minima.

In (10), for notational convenience, the cost function $F(\cdot)$ is written as a function of $\mathbf{P_y}$ which is a combination of the parameters that we want to estimate. In addition to $F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P_y})$ we also use the notation $F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{A}, \mathbf{P}, \mathbf{P_v})$ to explicitly express $F(\cdot)$ in terms of the desired model parameters.

There are two necessary conditions for the parameters of the CPSDM model in (9) to be uniquely identifiable.[1] The *first identifiability condition* states that the number of equations should be larger than the number of unknowns [36], [40]. There are $2Mr - |\Upsilon|$ unknowns due to $\mathbf{A}$, $r(r+1) - |\mathcal{K}|$ unknowns due to $\mathbf{P}$, and M unknowns due to $\mathbf{P_v}$. Thus, the first identifiability condition is given by

$$M(M+1) \geq 2Mr + r(r+1) - |\Upsilon| - |\mathcal{K}| + M. \tag{12}$$

The identifiability condition in (12) is not sufficient for guaranting unique identifiability [36]. Specifically, for any arbitary non-singular matrix $\mathbf{T} \in \mathbb{C}^{r\times r}$, we have $\mathbf{APA}^H = \tilde{\mathbf{A}}\tilde{\mathbf{P}}\tilde{\mathbf{A}}^H$, where $\tilde{\mathbf{A}} = \mathbf{AT}^{-1}$ and $\tilde{\mathbf{P}} = \mathbf{TPT}^H$, and, therefore [34]

$$F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{A}, \mathbf{P}, \mathbf{P_v}) = F(\hat{\mathbf{P}}_{\mathbf{y}}, \tilde{\mathbf{A}}, \tilde{\mathbf{P}}, \mathbf{P_v}). \tag{13}$$

This means that there are infinitly many optimal solutions ($\tilde{\mathbf{A}}, \tilde{\mathbf{P}} \succeq 0$) of the problem in (10). Since there are r^2 complex variables (i.e., $2r^2$ unknown imaginary and real parts) in $\mathbf{T}$, the *second identifiability condition* of the CPSDM model in (9) states that we need to fix at least $2r^2$ of the imaginary and real parts of the parameters in $\mathbf{A}$ and $\mathbf{P}$ [34], [40], i.e.,

$$|\Upsilon| + |\mathcal{K}| \geq 2r^2. \tag{14}$$

This second condition is necessary but not sufficient, since we need to fix the proper parameters and not just any $2r^2$ parameters [34], [40] such that $\mathbf{T} = \mathbf{I}$. For a general full-element $\mathbf{P}$, a recipe on which parameters to fix, in order to achieve unique identifiability, is provided in [34].

A. Simultaneous CFA (SCFA) in Multiple Time-Frames

The β-th time-segment consists of the following $|\mathcal{B}_\beta|$ time-frames: $t = \beta|\mathcal{B}_\beta| + 1, \cdots, (\beta+1)|\mathcal{B}_\beta|$, where $\mathcal{B}_\beta$ is the set of the time-frames in the β-th time-segment. For ease of notation, we can alternatively re-write this as $\forall t \in \mathcal{B}_\beta$. The problem formulation in (10) considered that the β-th time-segment consists of $|\mathcal{B}_\beta| = 1$ time-frame. Now we assume that we estimate $\hat{\mathbf{P}}_{\mathbf{y}}(t)$ for $|\mathcal{B}_\beta| \geq 1$ time-frames in the β-th time-segment. We also assume that $\forall (t_i, t_j) \in \mathcal{B}_\beta, \hat{\mathbf{P}}_{\mathbf{y}}(t_i) \neq \hat{\mathbf{P}}_{\mathbf{y}}(t_j)$, if $i \neq j$. Recall that the mixing matrix $\mathbf{A}$ is assumed to be static within a time-segment. Moreover, $\mathbf{P_v}$ is time-invariant and, thus, shared among different time-frames within the same time-segment. One can exploit these two facts in order to increase the ratio between the number of equations and the number of unknown parameters [33], [35] and thus satisfy the first and second identifiability conditions with less microphones compared to the CFA problem in (10). This can be done by solving the following general simultaneous CFA (SCFA) problem [35]

$$\begin{aligned}
\hat{\mathbf{A}}, \{\hat{\mathbf{P}}(t) : \forall t \in \mathcal{B}_\beta\}, \hat{\mathbf{P}}_{\mathbf{v}} &= \underset{\substack{\mathbf{A},\mathbf{P_v} \\ \{\hat{\mathbf{P}}(t):\forall t\in\mathcal{B}_\beta\}}}{\arg\min} \sum_{\forall \tau \in \mathcal{B}_\beta} F(\hat{\mathbf{P}}_{\mathbf{y}}(\tau), \mathbf{P_y}(\tau)) \\
\text{s.t.} \quad & \mathbf{P_y}(t) = \mathbf{AP}(t)\mathbf{A}^H + \mathbf{P_v}, \forall t \in \mathcal{B}_\beta, \\
& \mathbf{P_v} = \text{Diag}([q_1, \cdots, q_M]^T) \in \mathbb{R}^{M\times M},
\end{aligned}$$

[1] We say that the parameters of a function are uniquely identifiable if there is one-to-one relationship between the parameters and the function value.

$$
\begin{aligned}
&q_i \geq 0, i = 1, \cdots, M,\\
&\mathbf{P}(t) \succeq 0, \forall t \in \mathcal{B}_\beta,\\
&\Re(a_{ij}) = \tilde{a}_{ij}^R, \forall (i,j) \in \Upsilon^R,\\
&\Im(a_{\epsilon\zeta}) = \tilde{a}_{\epsilon\zeta}^I, \forall (\epsilon,\zeta) \in \Upsilon^I,\\
&\Re(p_{k\nu})(t) = \tilde{p}_{k\nu}^R(t), \forall (k,\nu) \in \mathcal{K}_t^R, \forall t \in \mathcal{B}_\beta,\\
&\Im(p_{\eta\iota})(t) = \tilde{p}_{\eta\iota}^I(t), \forall (\eta,\iota) \in \mathcal{K}_t^I, \forall t \in \mathcal{B}_\beta.
\end{aligned} \tag{15}
$$

The CFA problem in (10) is a special case of SCFA, when we select $|\mathcal{B}_\beta| = 1$. The first identifiability condition for the SCFA problem becomes

$$
|\mathcal{B}_\beta| M(M+1) \geq 2Mr + |\mathcal{B}_\beta| r(r+1) - |\Upsilon| - \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| + M. \tag{16}
$$

We conclude from (12) and (16) that the SCFA problem (for $|\mathcal{B}_\beta| > 1$) needs less microphones compared to the problem in (10) to satisfy the first identifiability condition, assuming both problems have the same number of sources. Moreover, the second identifiability condtion in the SCFA problem becomes

$$
|\Upsilon| + \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| \geq 2r^2. \tag{17}
$$

From (14) and (17), we conclude that the SCFA problem (for $|\mathcal{B}_\beta| > 1$) satisfies easier the second identifiability condition compared to the problem in (10), if both problems have the same number of sources and microphones.

B. Special Case (S)CFA: $\mathbf{P}(t)$ is Diagonal

A special case of (S)CFA, which is more suitable for the application at hand, is when $\mathbf{P}(t), \forall t \in \mathcal{B}_\beta$ are constrained to be diagonal. This is similar to our assumed signal model as expressed by (5) and (7). We refer to this special case as the diagonal (S)CFA problem. By constraining $\mathbf{P}(t)$ to be diagonal, the total number of fixed parameters in $\mathbf{A}, \mathbf{P}(t), \forall t \in \mathcal{B}_\beta$ is

$$
|\Upsilon| + \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| = |\Upsilon| + \underbrace{|\mathcal{B}_\beta|(r^2 - r) + |\mathcal{B}_\beta| r}_{|\mathcal{B}_\beta| r^2}. \tag{18}
$$

The term $|\mathcal{B}_\beta|(r^2 - r)$ is due to the constraints that set to zero the upper-diagonal off-diagonal complex elements. Since $\mathbf{P}(t) \succeq 0$, it is implied that the lower-diagonal off-diagonal complex elements are also set to zero. The term $|\mathcal{B}_\beta| r$ in (18) is due to the constraint $\mathbf{P}(t) \succeq 0$ which means that the r unknown diagonal elements in $\mathbf{P}(t)$ should be real and, thus, their imaginary part can be set to zero. It has been shown in [41], [42] that in this case, and for $r > 1$, the class of the only possible $\mathbf{T}$ is $\mathbf{T} = \mathbf{\Pi S}$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{S}$ is a scaling matrix, if the following condition is satisfied

$$
2\kappa_{\mathbf{A}} + \kappa_{\mathbf{Z}} \geq 2(r+1), \tag{19}
$$

where

$$
\mathbf{Z} = \left[\mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_{|\mathcal{B}_\beta|} \right], \quad \mathbf{z}_t = \operatorname{diag}\left(\mathbf{P}(t)\right), t \in \mathcal{B}_\beta, \tag{20}
$$

TABLE I
MAXIMUM r AS A FUNCTION OF VARYING M, $|\mathcal{B}_\beta|$ SUCH THAT (23) AND (22) ARE SATISFIED

M	$\lvert\mathcal{B}_\beta\rvert = 1$	$\lvert\mathcal{B}_\beta\rvert = 2$	$\lvert\mathcal{B}_\beta\rvert = 4$	$\lvert\mathcal{B}_\beta\rvert = 8$	$\lvert\mathcal{B}_\beta\rvert = 16$
2	1	2	3	4	5
4	2	4	7	11	14
8	2	8	15	25	38

and $\kappa_{\mathbf{A}}, \kappa_{\mathbf{Z}}$ are the Kruskal-ranks [41] of the matrices $\mathbf{A}$ and $\mathbf{Z}$, respectively. We conclude, that if (16) is satisfied, and there are at least $2r^2$ fixed real and imaginary parts of the variables in $\mathbf{A}$ and $\mathbf{P}(t), \forall t \in \mathcal{B}_\beta$, and the condition in (19) is satisfied, then the parameters of (9) (for $\mathbf{P}(t)$ diagonal) will be uniquely identifiable up to a possible scaling and/or permutation.

C. Diagonal SCFA versus Nonorthogonal Joint Diagonalization

The diagonal SCFA problem in Sec. III-B is very similar to the joint diagonalization method in [4], apart from the two positive semidefinite constraints that avoid improper solutions, and which are lacking in [4]. Finally, it is worth mentioning that the method proposed in [4] solves the scaling ambiguity by setting $\Re(a_{ii}) = 1, \Im(a_{ii}) = 0$ (corresponding to a varying reference microphone per-source), which means $2r$ fixed parameters in $\mathbf{A}$, i.e., $|\Upsilon| = 2r$. Thus, in [4], the total number of fixed parameters in $\mathbf{A}, \mathbf{P}(t), \forall t \in \mathcal{B}_\beta$ is given by

$$
|\Upsilon| + \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| = 2r + |\mathcal{B}_\beta| r^2. \tag{21}
$$

By combining (21) and (17), the second identifiability condition becomes

$$
2r + |\mathcal{B}_\beta| r^2 \geq 2r^2. \tag{22}
$$

Note that for $r \geq 1$, if $|\mathcal{B}_\beta| \geq 2$, the second identifiability condition is always satisfied, but the permutation ambiguity still exists and needs extra steps to be resolved [4]. However, for $r = 1$, the second identifiability condition is satisfied for $|\mathcal{B}_\beta| \geq 1$ and there are no permutation ambiguities. By combining (21), and (16), the first identifiability condition for the diagonal SCFA with $|\Upsilon| = 2r$ becomes

$$
|\mathcal{B}_\beta| M(M+1) \geq 2Mr + |\mathcal{B}_\beta| r - 2r + M. \tag{23}
$$

Table I shows what is the maximum r for a varying M and $|\mathcal{B}_\beta|$ such that both (23) and (22) hold.

IV. PROPOSED DIAGONAL SCFA PROBLEMS

In this section, we will propose two methods based on the diagonal SCFA problem from Sec. III-B to estimate the different signal model parameters in (7). Unlike the diagonal SCFA problem and the non-orthogonal joint diagonalization method in [4], the first proposed method also estimates the late reverberation PSD. The second proposed method skips the estimation of the late reverberation PSD and thus is more similar to the diagonal SCFA problem and the non-orthogonal joint diagonalization

method in [4]. Since we are using the early RATFs as columns of $\mathbf{A}$, we fix all the elements of the ρ-th row of $\mathbf{A}$ to be equal to 1, where ρ is the reference microphone index. Thus, unlike the method proposed in [4], which uses a varying reference microphone (i.e., $\Re(a_{ii}) = 1, \Im(a_{ii}) = 0$), we use a single reference microphone (i.e., $\Re(a_{\rho j}) = 1, \Im(a_{\rho j}) = 0$).

Although our proposed constraints $\Re(a_{\rho j}) = 1, \Im(a_{\rho j}) = 0$ will resolve the scaling ambiguity (described in Sec III-B), the permutation ambiguity (described in Sec III-B) still exists and needs extra steps to be resolved. In this paper, we do not focus on this problem and we assume that we know the perfect permutation matrix per time-frequency tile. The interested reader can find more information on how to solve permutation ambiguities in [4]–[6]. An exception occurs in the context of dereverberation where, typically, a single point source (i.e., $r = 1$) exists and, therefore, a single fixed complex-valued parameter in $\mathbf{A}$ is sufficient to solve both the permutation and scaling ambiguities.

A. Proposed Diagonal SCFA Problem

The proposed basic diagonal SCFA problem is based on the signal model in (7), which takes into account the late reverberation. Here we assume that we have computed a priori $\hat{\mathbf{\Phi}}$. The proposed diagonal SCFA problem is given by

$$\begin{aligned}
&\underset{\substack{\hat{\mathbf{P}}_{\mathbf{v}},\{\hat{\gamma}(t):\forall t\in\mathcal{B}_\beta\}}}{\hat{\mathbf{A}},\{\hat{\mathbf{P}}(t):\forall t\in\mathcal{B}_\beta\},} = \underset{\substack{\mathbf{A},\{\mathbf{P}(t):\forall t\in\mathcal{B}_\beta\},\\ \mathbf{P}_{\mathbf{v}},\{\gamma(t):\forall t\in\mathcal{B}_\beta\}}}{\arg\min} \sum_{\forall\tau\in\mathcal{B}_\beta} F(\hat{\mathbf{P}}_{\mathbf{y}}(\tau),\mathbf{P}_{\mathbf{y}}(\tau))\\
&\text{s.t.}\quad \mathbf{P}_{\mathbf{y}}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H + \gamma(t)\hat{\mathbf{\Phi}} + \mathbf{P}_{\mathbf{v}}, \forall t\in\mathcal{B}_\beta\\
&\quad \mathbf{P}_{\mathbf{v}} = \text{Diag}([q_1,\cdots,q_M]^T) \in \mathbb{R}^{M\times M},\\
&\quad q_i \geq 0, i = 1,\cdots,M,\\
&\quad \mathbf{P}(t) = \text{Diag}([p_1(t),\cdots,p_r(t)]^T) \in \mathbb{R}^{M\times M}, \forall t\in\mathcal{B}_\beta,\\
&\quad p_j(t) \geq 0, \forall t\in\mathcal{B}_\beta, j = 1,\cdots,r,\\
&\quad \gamma(t) \geq 0, \forall t\in\mathcal{B}_\beta,\\
&\quad \Re(a_{\rho j}) = 1, \text{ for } j = 1,\cdots,r,\\
&\quad \Im(a_{\rho j}) = 0, \text{ for } j = 1,\cdots,r. \qquad (24)
\end{aligned}$$

We will refer to the problem in (24) as the SCFA$_{\text{rev}}$ problem. The cost function of the SCFA$_{\text{rev}}$ problem depends on $\gamma(t)$. This means that we have $|\mathcal{B}_\beta|$ additional real-valued unknowns in (23). The first identifiability condition therefore becomes

$$|\mathcal{B}_\beta|M(M+1) \geq 2Mr + |\mathcal{B}_\beta|r - 2r + |\mathcal{B}_\beta| + M. \qquad (25)$$

A simplified version of the SCFA$_{\text{rev}}$ problem is obtained when the reverberation parameter γ is omitted. This problem therefore uses the signal model of (9) instead of (7). We will refer to this simplified problem as the SCFA$_{\text{no-rev}}$ problem. The only differences between the SCFA$_{\text{no-rev}}$ and the method proposed [4], is that in the SCFA$_{\text{no-rev}}$ we use a fixed reference microphone and positivity constraints for the PSDs.

Since, we have $2r$ fixed parameters in $\mathbf{A}$ corresponding to the reference microphone, in both proposed methods, the total number of fixed parameters in $\mathbf{A}$ and $\mathbf{P}(t), \forall t\in\mathcal{B}_\beta$ is the same as in (21). The second identifiability condition of all proposed methods is therefore the same as in (22).

B. SCFA$_{rev}$ versus SCFA$_{no\text{-}rev}$

Although the SCFA$_{\text{rev}}$ method typically fits a more accurate signal model to the noisy measurements compared to the SCFA$_{\text{no-rev}}$ method, it does not necessarily guarantee a better performance over the SCFA$_{\text{no-rev}}$ method. In other words, the *model-mismatch* error is not the only critical factor in achieving good performance. Another important factor is how *over-determined* is the system of equations to be solved is, i.e., what is the ratio of number of equations and number of unknowns. With respect to the over-determination factor, the SCFA$_{\text{no-rev}}$ method is more efficient, since it has less parameters to estimate, if $\mathcal{B}_\beta$ is the same in both methods. Consequently, the problem boils down to how much is the model-mismatch error and the over-determination. Thus, it is natural to expect that for not highly reverberant environments, the SCFA$_{\text{no-rev}}$ method may perform better than the SCFA$_{\text{rev}}$ method, while for highly reverberant environments the inverse may hold.

V. Robust Estimation of Parameters

In Secs. V-A–V-E, we propose additional constraints in order to increase the robustness of the initial versions of the two diagonal SCFA problems proposed in Sec. IV. The robustness is needed in order to overcome CPSDM estimation errors and model-mismatch errors. We use linear inequality constraints (mainly simple box constraints) on the parameters to be estimated. These constraints limit the feasibility set of the parameters to be estimated and avoid unreasonable values.

A less efficient alternative procedure to increase robustness would be to solve the proposed problems with a multi-start optimization technique such that a good local optimum will be obtained. Note that this procedure is more computational demanding and also (without the box constraints) does not guarantee estimated parameters that belong in a meaningful region of values.

A. Constraining the Summation of PSDs

If the model in (7) perfectly describes the acoustic scene, the sum of the PSDs of the point sources, late reverberation, and microphone self-noise at the reference microphone equals $p^{\mathbf{y}}_{\rho\rho}$ (where ρ is the reference microphone index and $p^{\mathbf{y}}_{\rho\rho}$ is the (ρ,ρ) element of $\mathbf{P}_{\mathbf{y}}$). That is,

$$||\text{diag}\,(\mathbf{P})\,||_1 + \gamma\phi_{\rho\rho} + q_\rho = p^{\mathbf{y}}_{\rho\rho}, \qquad (26)$$

where $\phi_{\rho\rho}$ is the ρ-th diagonal element of $\mathbf{\Phi}$. In practice, the model is not perfect and have an estimate $\hat{p}^{\mathbf{y}}_{\rho\rho}$. Therefore, the following box constraint is introduced, that is,

$$0 \leq ||\text{diag}\,(\mathbf{P})\,||_1 + \gamma\hat{\phi}_{\rho\rho} + q_\rho \leq \delta_1\hat{p}^{\mathbf{y}}_{\rho\rho}, \qquad (27)$$

where δ_1 is a constant which implicitly depends on the variance of the estimator of $p^{\mathbf{y}}_{\rho\rho}$ and which effectively controls the underestimation or overestimation of the PSDs. This box constraint can be used to improve the robustness of the SCFA$_{\text{rev}}$ problem,

but cannot be used by the $\text{SCFA}_{\text{no-rev}}$ problem, since it does not estimate γ. A less tight box constraint that can be used for both $\text{SCFA}_{\text{no-rev}}$, SCFA_{rev} problems is

$$0 \leq ||\text{diag}\left(\mathbf{P}\right)||_1 \leq \delta_2 \hat{p}^{\mathbf{y}}_{\rho\rho}. \tag{28}$$

One may see the inequality in (28) as a sparsity constraint, natural in audio and speech processing as the number of the active sound sources is small (typically much smaller than the maximum number of sources, r, existing in the acoustic scene) for a single time-frequency tile. In this case, δ_2 controls the sparsity. A low δ_2 implies large sparsity, while a large δ_2 implies low sparsity. The sparsity is over frequency and time.

B. Box Constraints for the Early RATFs

Extra robustness can be achieved if the elements of the early RATFs are box-constrained as follows:

$$l_{ij,R} \leq \Re(a_{ij}) \leq u_{ij,R}, l_{ij,I} \leq \Im(a_{ij}) \leq u_{ij,I}, \tag{29}$$

where $u_{ij,R}$ and $l_{ij,R}$ are some upper and lower bounds, respectively of $\Re(a_{ij})$, while $u_{ij,I}$, and $l_{ij,I}$ are some upper and lower bounds, respectively of $\Im(a_{ij})$.[2] We select the values of $u_{ij,R}, l_{ij,R}, u_{ij,I}$, and $l_{ij,I}$ based on relative Green functions. Let us denote with $\mathbf{f}_j \in \mathbb{R}^{3\times 1}$ the location of the j-th source, with $\mathbf{m}_i$ the location of the i-th microphone, and with $d_{ij} = ||\mathbf{f}_j - \mathbf{m}_i||_2$ the distance between the j-th source and i-th microphone. The anechoic ATF (direct path only) at the frequency-bin k between the j-th source i-th microphone is given by [43]

$$\tilde{a}_{ij}(k) = \frac{1}{4\pi d_{ij}} \text{exp}\left(-\frac{j2\pi k}{K}\frac{d_{ij}}{c}\right), \tag{30}$$

where K is the FFT length, c is the speed of sound, and d_{ij}/c is the time of arrival (TOA) of the j-th source to the i-th microphone. The corresponding anechoic relative ATF with respect to the reference microphone ρ is given by

$$a_{ij}(k) = \frac{\tilde{a}_{ij}(k)}{\tilde{a}_{\rho j}(k)} = \frac{d_{\rho j}}{d_{ij}} \text{exp}\left(-\frac{j2\pi k}{K}\frac{(d_{ij} - d_{\rho j})}{c}\right), \tag{31}$$

where $(d_{ij} - d_{\rho j})/c$ is the time difference of arrival (TDOA) of the j-th source between microphones i and ρ. What becomes clear from (31) is that the anechoic relative ATF depends only on the two unknown parameters $d_{ij}, d_{\rho j}$. The upper and lower bounds of the real and imaginary parts of (31) can be written compactly using the following box inequality

$$-\frac{d_{\rho j}}{d_{ij}} \leq \Re\left(a_{ij}(k)\right), \Im\left(a_{ij}(k)\right) \leq \frac{d_{\rho j}}{d_{ij}}. \tag{32}$$

Among all the points on the circle with any constant radius and center the middle point between microphones with indices i and ρ, the inequality in (32) becomes maximally relaxed for the maximum possible $d_{\rho j}$ and minimum possible d_{ij}, i.e., when the ratio $d_{\rho j}/d_{ij}$ becomes maximum. This happens when the j-th source is in the endfire direction of the two microphones and closest to i-th microphone. In this case we have $d_{\rho j} = d_{\rho i} + d_{ij}$ and, therefore, (32) becomes

$$-\frac{d_{\rho i} + d_{ij}}{d_{ij}} \leq \Re\left(a_{ij}(k)\right), \Im\left(a_{ij}(k)\right) \leq \frac{d_{\rho i} + d_{ij}}{d_{ij}}. \tag{33}$$

In the inequality in (33), the parameters $d_{\rho i}, d_{ij}$ are unknown. Now, we try to relax this inequality and find ways that are independent of these unknown parameters.

Note that the quantity $|d_{ij} - d_{\rho j}|/c$ (in seconds) should not be allowed to be greater than the sub-frame length in seconds, i.e., N/f_s, where N is the sub-frame length in samples. If it is greater than N/f_s, the signal model given in (7) is invalid, i.e., the CPSDM of the j-th point source cannot be written as a rank-1 matrix, because it will not be fully correlated between microphones i, ρ. Therefore, we have

$$\frac{|d_{ij} - d_{\rho j}|}{c} \leq \frac{N}{f_s} \iff |d_{ij} - d_{\rho j}| \leq \frac{Nc}{f_s}. \tag{34}$$

Note that the inequality in (34) should also hold in the endfire direction of the two microphones, which means

$$d_{\rho i} \leq \frac{Nc}{f_s}. \tag{35}$$

The inequality in (33) is maximally relaxed for the maximum possible $d_{\rho i}$ and the minimum possible d_{ij}. The maximum allowable $d_{\rho i}$ is given by (35). Moreover, another practical observation is that the sources cannot be in the same location as the microphones. Therefore, we have

$$d_{ij} \geq \lambda, \tag{36}$$

where λ is a very small distance (e.g., 0.01 m). Therefore, the maximum range of the real and imaginary parts of the relative anechoic ATF is given by

$$-\frac{\frac{Nc}{f_s} + \lambda}{\lambda} \leq \Re\left(a_{ij}(k)\right), \Im\left(a_{ij}(k)\right) \leq \frac{\frac{Nc}{f_s} + \lambda}{\lambda}. \tag{37}$$

The above inequality is based on anechoic free-field RATFs. In practice, we have early RATFs which include early echoes and/or directivity patterns which means that we might want to make the box constraint in (37) less tight.

C. Tight Box Constraints for the Early RATFs Based on $\hat{\mathbf{D}}$

In Sec. V-B we proposed the box constraints in (37) based on practical considerations without knowing the distance between sensors or between sources and sensors. In this section we assume that we have an estimate of the distance matrix (see Sec. II-C), $\hat{\mathbf{D}}$. Consequently we know $\hat{d}_{\rho i}$ and, therefore, we can make the box constraint in (37) even tighter. That is,

$$-\frac{\hat{d}_{\rho i} + \lambda}{\lambda} \leq \Re\left(a_{ij}(k)\right), \Im\left(a_{ij}(k)\right) \leq \frac{\hat{d}_{\rho i} + \lambda}{\lambda}. \tag{38}$$

D. Box Constraints for the Late Reverberation PSD

In this section, we take into consideration the late reverberation. We can be almost certain that the following box constraint is satisfied:

$$0 \leq \gamma(t,k)\text{min}\left(\text{diag}(\hat{\mathbf{\Phi}})\right) \leq \text{min}\left[\text{diag}\left(\hat{\mathbf{P}}_{\mathbf{y}}(t,k)\right)\right]. \tag{39}$$

[2] An alternative method would be to constrain $||a_{ij}||$ with lower and upper bounds but that would lead to a non-linear inequality constraint and, thus, a more complicated implementation.

This box constraint is only applicable in the SCFA_{rev} problem. The box-constraint in (39) prevents large overestimation errors which may result in speech intelligibility reduction in noise reduction applications [18], [44].

E. All Microphones Have the Same Microphone-Self Noise PSD

Here we examine the special case where $\mathbf{P}_{\mathbf{v}}(k) = q(k)\mathbf{I}$, i.e., all microphones have the same self-noise PSD. Moreover, since the microphone self-noise is stationary, we can be almost certain that the following box-constraint holds

$$0 \leq q(k) \leq \min_{\forall t \in \mathcal{B}_\beta} \left(\min \left[\text{diag} \left(\hat{\mathbf{P}}_{\mathbf{y}}(t) \right) \right] \right). \tag{40}$$

Similar to the constraint in (39), the constraint in (40) avoids large overestimation errors.

By having a common self-noise PSD for all microphones, the number of parameters are reduced by $M-1$, since we have only one microphone-self noise PSD for all microphones. Hence, the first identifiability condition for the $\text{SCFA}_{\text{no-rev}}$ problem is now given by

$$|\mathcal{B}_\beta| M(M+1) \geq 2Mr + |\mathcal{B}_\beta| r - 2r + 1. \tag{41}$$

Similarly, the first identifiability condition for the SCFA_{rev} problem is now given by

$$|\mathcal{B}_\beta| M(M+1) \geq 2Mr + |\mathcal{B}_\beta| r - 2r + |\mathcal{B}_\beta| + 1. \tag{42}$$

VI. Practical Considerations

In this section, we discuss practical problems regarding the choice of several parameters of the proposed methods and implementation aspects. Although, we have already explained the problem of over-determination in Sec. IV-B, in Sec VI-A, we discuss additional ways of achieving over-determination. In Sec. VI-B, we discuss about some limitations of the proposed methods. Finally, in Secs. VI-C and VI-D, we discuss how to implement the proposed methods.

A. Over-Determination Considerations

Increasing the ratio of the number of equations over the number of unknowns obviously fits better the CPSDM model to the measurements under the assumption that the model is accurate enough and the early RATFs do not change within a time-segment. There are two main approaches to increase the ratio of the number of equations over the number of unknowns. The first approach is to reduce the number of the parameters to be estimated while fixing the number of equations as already explained in Sec. IV-B. In addition to the explanation provided in IV-B, we could also reduce the number of parameters by source counting per time-frequency tile and adapt r. However, this is out of the scope of the present paper and here we assume that we have a constant r in the entire time-frequency horizon which is the maximum possible r. The second approach is to increase the number of time-frames $|\mathcal{B}_\beta|$ in a time-segment and/or the number of microphones M. Increasing $|\mathcal{B}_\beta|$ is not practical, because typically, the acoustic sources are moving. Thus, $|\mathcal{B}_\beta|$ should not be too small but also not too large. Note that $|\mathcal{B}_\beta|$ is also effected by the time-frame length denoted by $\mathcal{T}$. If $\mathcal{T}$ is small we can use a larger $|\mathcal{B}_\beta|$, while if $\mathcal{T}$ is large, we should use a small $|\mathcal{B}_\beta|$ in order to be able to also track moving sources. However, if we select $\mathcal{T}$ to be very small, the number of sub-frames will be smaller and consequently the estimation error in $\hat{\mathbf{P}}_{\mathbf{y}}$ will be large and will cause performance degradation.

B. Limitations of the Proposed Methods

From the identifiability conditions in (23), (25), (41) and (42) for fixed $|\mathcal{B}_\beta|$ and r, we can obtain the minimum number of microphones needed to satisfy these inequalities. Alternatively, for a fixed M and r we can obtain the minimum number of time-frames $|\mathcal{B}_\beta|$ needed to satisfy these inequalities. Finally, for a fixed M and $|\mathcal{B}_\beta|$ we can find the maximum number of sources r for which we can identify their parameters (early RATFs and PSDs). Let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ and $\mathcal{M}_4$ the minimum number of microphones satisfying the identifiability conditions in (23), (25), (41) and (42), respectively. Moreover, let $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3$ and $\mathcal{J}_4$ the minimum number of time-frames satisfying the identifiability conditions in (23), (25), (41) and (42), respectively. In addition, let $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ and $\mathcal{R}_4$ the maximum number of sources satisfying the identifiability conditions in (23), (25), (41) and (42), respectively. The following inequalities can be easily proved:

$$\begin{aligned}
&\mathcal{M}_3 \leq \mathcal{M}_4, \quad \mathcal{M}_1 \leq \mathcal{M}_2, \quad \mathcal{M}_4 \leq \mathcal{M}_2, \quad \mathcal{M}_3 \leq \mathcal{M}_1, \\
&\mathcal{J}_3 \leq \mathcal{J}_4, \quad \mathcal{J}_1 \leq \mathcal{J}_2, \quad \mathcal{J}_4 \leq \mathcal{J}_2, \quad \mathcal{J}_3 \leq \mathcal{J}_1, \\
&\mathcal{R}_3 \geq \mathcal{R}_4, \quad \mathcal{R}_1 \geq \mathcal{R}_2, \quad \mathcal{R}_4 \geq \mathcal{R}_2, \quad \mathcal{R}_3 \geq \mathcal{R}_1.
\end{aligned}$$

C. Online Implementation Using Warm-Start

The estimation of the parameters is carried out for all time-frames within one time-segment. Subsequently, in order to have low latency, we shift the time-segment one time-frame. For the $|\mathcal{B}_\beta| - 1$ time-frames in the current time-segment that overlap with the time-frames in the previous time-segment, the parameters are initialized using the estimates from the corresponding $|\mathcal{B}_\beta| - 1$ time-frames in the previous time-segment. The parameters of the most recent time-frame are initialized by selecting a value that is drawn from a uniform distribution with boundaries corresponding to the lower and upper bound of the corresponding box constraint. Only for the first time-segment, the early RATFs are initialized with the r most dominant relative eigenvectors from the averaged CPSDM over all time-frames of the first time-segment.

D. Solver

For a single frequency-bin of the β-th time segment, the variables of our proposed optimization problem are stacked in a real-valued vector $\boldsymbol{\xi} \in \mathbb{R}^{(r(2M-2)+r|\mathcal{B}_\beta|+|\mathcal{B}_\beta|+M)\times 1}$. The vector $\boldsymbol{\xi}$ can be mapped to the matrix $\mathbf{P}_{\mathbf{y}}$ through the function $f(\boldsymbol{\xi}) = \mathbf{P}_{\mathbf{y}}$.

The optimal $\hat{\boldsymbol{\xi}}$ is computed iteratively. Specifically, in the $k+1$-th iteration, the vector $\boldsymbol{\xi}$ is computed as

$$\boldsymbol{\xi}^{(k+1)} = U\left(\boldsymbol{\xi}^{(k)}, \nabla_{\boldsymbol{\xi}} F\left(\hat{\mathbf{P}}_{\mathbf{y}}, f(\boldsymbol{\xi}^{(k)}) \right), \nabla^2_{\boldsymbol{\xi}} F(\hat{\mathbf{P}}_{\mathbf{y}}, f(\boldsymbol{\xi}^{(k)})) \right),$$

where $\nabla_{\boldsymbol{\xi}} F(\hat{\mathbf{P}}_{\mathbf{y}}, f(\boldsymbol{\xi}^{(k)}))$, $\nabla^2_{\boldsymbol{\xi}} F(\hat{\mathbf{P}}_{\mathbf{y}}, f(\boldsymbol{\xi}^{(k)}))$ are the gradient and Hessian matrix of $F(\cdot)$ with respect to $\boldsymbol{\xi}$ at $\boldsymbol{\xi}^{(k)}$, respectively, and $U(\cdot)$ is the update procedure. The update procedure updates $\boldsymbol{\xi}$ such that all the inequality constraints that we have proposed are satisfied and

$$F\left(\hat{\mathbf{P}}_{\mathbf{y}}, f(\boldsymbol{\xi}^{(k+1)})\right) \leq F\left(\hat{\mathbf{P}}_{\mathbf{y}}, f(\boldsymbol{\xi}^{(k)})\right). \tag{43}$$

The update procedure terminates when a local minimum is found. The first-order derivatives of the cost functions in (11) with respect to most parameters have been obtained already in [4], [34]–[36] without taking into account the late reverberation PSD. Thus, here we provide only the first-order derivatives with respect to the late reverberation PSD parameter. We have

$$\text{ML: } \frac{\partial F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P}_{\mathbf{y}})}{\partial \gamma} = \text{tr}\left(\mathbf{P}_{\mathbf{y}}^{-1}\left(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}\right)\mathbf{P}_{\mathbf{y}}^{-1}\hat{\boldsymbol{\Phi}}\right),$$

$$\text{LS: } \frac{\partial F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P}_{\mathbf{y}})}{\partial \gamma} = \text{tr}\left(\left(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}\right)\hat{\boldsymbol{\Phi}}\right),$$

$$\text{GLS: } \frac{\partial F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P}_{\mathbf{y}})}{\partial \gamma} = \text{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}^{-1}\left(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}\right)\hat{\mathbf{P}}_{\mathbf{y}}^{-1}\hat{\boldsymbol{\Phi}}\right).$$

In this paper, we computed the Hessian using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximation [36]. The non-convex optimization problems that we proposed can be solved with various existing update procedures within the literature such as [45]–[48]. In this paper, we used the update procedure *fmincon* from the standard MATLAB optimization toolbox to solve the optimization problems which implements a combination of the iterative methods in [46]–[48].

VII. Experiments

In this section, we evaluate the performance of the proposed methods in the context of two multi-microphone applications. The first application is dereverberation of a single point source ($r = 1$). The second application is source separation combined with dereverberation examined in an acoustic scene with $r = 3$ point sources. In this paper, we use the perfect permutation matrix for all compared methods in the source separation experiments. For these experiments we selected the maximum-likelihood (ML) cost function in (11). The values of the parameters that we selected for both applications are summarized in Table II. All methods based on the diagonal SCFA methodology are implemented using the online implementation explained in Sec. VI-C. The acoustic scene we consider for the source separation example is depicted in Fig. 2. The acoustic scene we consider for the dereverberation example is similar with the only difference that the music signal and male talker sources (see Fig. 2) are not present. The room dimensions are $7 \times 5 \times 4$ m. The reverberation time for the dereverberation application is selected $T_{60} = 1$ s, while for the source separation, $T_{60} = 0.2$ and 0.6 s. The microphone signals have a duration of 50 s and the duration of the impulse responses used to construct the microphone signals is 0.5 s. The microphone signals were constructed using the image method [43]. The microphone array is circular with a consecutive microphone distance of 2 cm. The reference

TABLE II
Summary of Parameters Used in the Experiments

Parameter	Definition	Value
M	number of microphones	4
K	FFT length	256
$\mathcal{T}$	time-frame length	2000 samples (0.125 s)
N	sub-frame length	200 samples (0.0125 s)
ov_N	overlapping of sub-frames	75%
$\hat{\boldsymbol{\Phi}}$	spatial coherence matrix	spherical isotropic model
ρ	reference microphone index	1
δ_1	controls overestimation underestimation	1.2
δ_2	controls sparsity	1
c	speed of sound	343m/s
λ	minimum possible source-microphone distance	1 cm
f_s	sampling frequency	16 kHz
q	mic. self noise PSD	$9 * 10^{-6}$

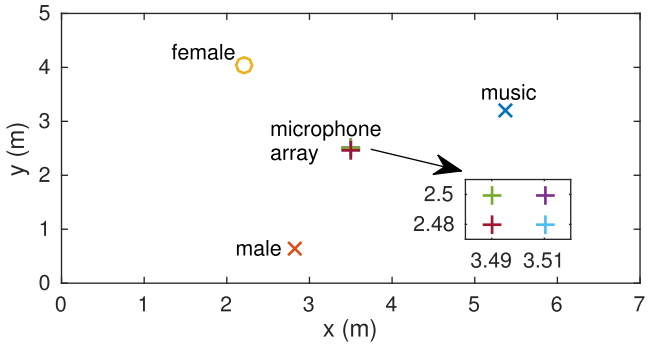


Fig. 2. Acoustic scene with $r = 3$ sources and $M = 4$ microphones.

microphone is the right-top microphone in Fig. 2. Moreover, we assume that the microphone-self noise has the same PSD at all microphones. Finally, it is worth mentioning that the early part of a room impulse response (see Sec. II-B) is of the same length as the sub-frame length.

A. Performance Evaluation

We will perform two types of performance evaluations in both applications. The first one measures the error of the estimated parameters, while the second one measures the performance using the estimated parameters in a source estimation algorithm and measure instrumental intelligibility and sound quality of the estimated source waveforms. We measure the average PSD errors of the sources, the average PSD error of the late reverberation, and the average PSD error of the microphone-self noise using the following three measures [49]:

$$E_s = \frac{10}{C(K/2+1)r}\sum_{t=1}^{C}\sum_{k=1}^{K/2+1}\sum_{j=1}^{r}\left|\log\frac{p_j(t,k)}{\hat{p}_j(t,k)}\right| \text{(dB)}, \tag{44}$$

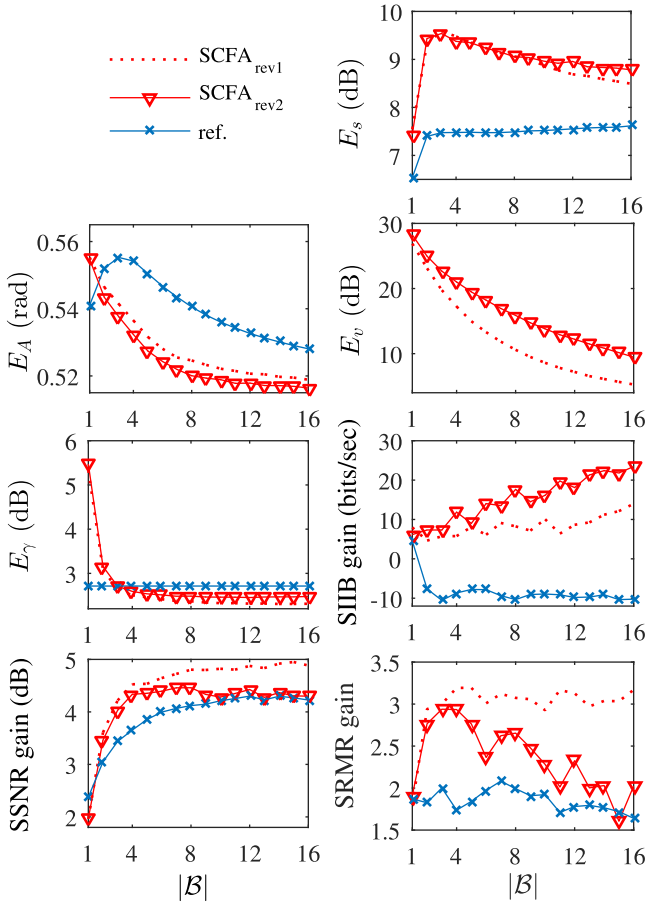


Fig. 3. Dereverberation results: The proposed methods are denoted by $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$. The ref. is the reference method reviewed in Sec. VII-B.

$$E_\gamma = \frac{10}{C(K/2+1)} \sum_{t=1}^{C} \sum_{k=1}^{K/2+1} \left| \log \frac{\gamma(t,k)}{\hat{\gamma}(t,k)} \right| \text{(dB)}, \tag{45}$$

$$E_v = \frac{10}{C(K/2+1)} \sum_{t=1}^{C} \sum_{k=1}^{K/2+1} \left| \log \frac{q(t,k)}{\hat{q}(t,k)} \right| \text{(dB)}, \tag{46}$$

where C is the number of time-frames in the microphone recordings. We also compute the underestimates (denoted as above with superscript un) and overestimates (denoted as above with superscript ov) of the above averages as in [44] since a large overestimation error in the noise PSDs and a large underestimation error in the target PSD typically results in large target source distortions in the context of a noise reduction framework [44]. On the other hand, a large underestimation error in the noise PSDs may result in musical noise [44]. We also evaluate the average early RATF estimation error using the Hermitian angle measure [50] given by

$$E_A = \frac{\sum_{j=1}^{r} \sum_{\beta=1}^{V} \sum_{k=1}^{K/2+1} \text{acos}\left(\frac{|\mathbf{a}_j^H(\beta,k)\hat{\mathbf{a}}_j(\beta,k)|}{||\mathbf{a}_j^H(\beta,k)||_2 ||\hat{\mathbf{a}}_j(\beta,k)||_2} \right)}{V(K/2+1)r} \text{(rad)}, \tag{47}$$

where V is the number of time-segments in the microphone recordings. Since, we use the warm-start procedure in Sec. VI-C in which the consecutive time-segments highly overlap, V is approximately equal to C. If the PSD of a source in a frequency-bin is negligible for all time-frames within a time-segment, the estimated PSD and RATF of this source at that time-frequency tile are skipped from the above averages.

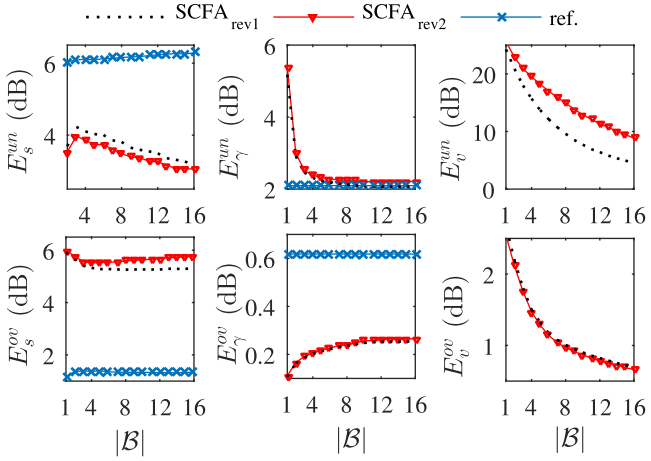


Fig. 4. Underestimates (with superscript un) and overestimates (with superscript ov): The proposed methods are denoted by $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$. The ref. is the reference method described in Sec. VII-B.

To evaluate the intelligibility and quality of the j-th target source signal, the estimated parameters are used to construct a multi-channel Wiener filter (MWF) as a concatenation of a single-channel Wiener filter (SWF) and a minimum variance distortionless response (MVDR) beamformer [1]. That is,

$$\hat{\mathbf{w}}_j = \frac{\hat{p}_j}{\hat{p}_j + \hat{\mathbf{w}}_{j,\text{MVDR}}^H \hat{\mathbf{P}}_{j,\mathbf{n}} \hat{\mathbf{w}}_{j,\text{MVDR}}} \hat{\mathbf{w}}_{j,\text{MVDR}}, \tag{48}$$

and

$$\hat{\mathbf{w}}_{j,\text{MVDR}} = \frac{\hat{\mathbf{P}}_{j,\mathbf{n}}^{-1} \hat{\mathbf{a}}_j}{\hat{\mathbf{a}}_j^H \hat{\mathbf{P}}_{j,\mathbf{n}}^{-1} \hat{\mathbf{a}}_j}, \tag{49}$$

where

$$\hat{\mathbf{P}}_{j,\mathbf{n}} = \sum_{\forall i \neq j} \hat{p}_i \hat{\mathbf{a}}_i \hat{\mathbf{a}}_i^H + \hat{\gamma} \mathbf{\Phi} + \hat{q} \mathbf{I}. \tag{50}$$

The noise reduction of the j-th source is evaluated using the segmental-signal-to-noise-ratio (SSNR) [51] for the j-th source only in sub-frames where the j-th source is active after which we average the SSNRs over all sources. Moreover, for speech sources, we measure the predicted intelligibility with the speech intelligibility in bits (SIIB) measure [52], [53], as it has been shown in [53] that SIIB correlates reasonably well with speech intelligibility in scenarios of processed reverberated noisy speech. Subsequently, we average SIIB over all speech sources. Finally, for the dereverberation application we measure the predicted quality and intelligibility with the speech to reverberation modulation energy ratio (SRMR) measure [54].

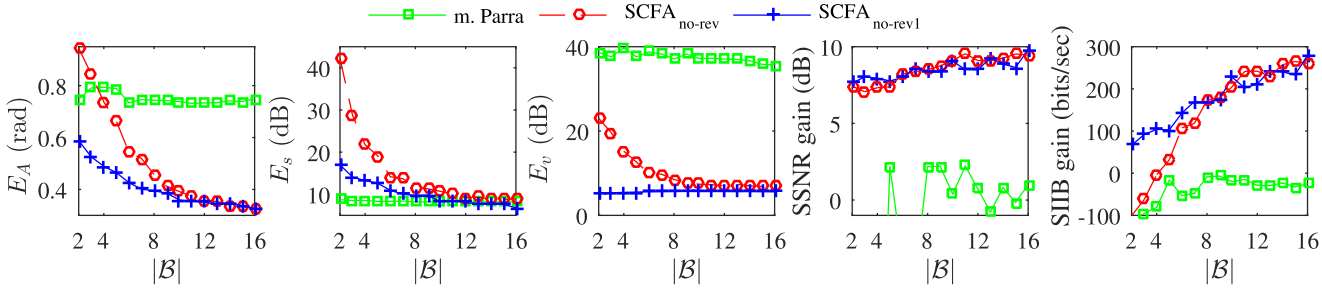


Fig. 5. Source separation results for $T_{60} = 0.2$ s: Comparison of m. Parra method and the proposed blind methods SCFA$_{\text{no-rev}}$ and SCFA$_{\text{no-rev1}}$.

B. Reference State-of-the-Art Dereverberation and Parameter-Estimation Methods

The reference method that we use in our comparison is a combination of the methods in [8], [15], [19], [26], [28], [55]. Specifically, we first estimate the PSD of the late reverberation using the method proposed in [19], [26]. That is, we first compute the Cholesky decomposition $\hat{\mathbf{\Phi}} = \mathbf{L}_{\mathbf{\Phi}}\mathbf{L}_{\mathbf{\Phi}}^H$ after which we compute the whitened estimated noisy CPSDM as

$$\mathbf{P}_{\mathbf{w1}} = \mathbf{L}_{\mathbf{\Phi}}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}(\mathbf{L}_{\mathbf{\Phi}}^H)^{-1}. \tag{51}$$

Next, we compute the eigenvalue decomposition $\mathbf{P}_{\mathbf{w1}} = \mathbf{V}\mathbf{R}\mathbf{V}^H$, where the diagonal entries of $\mathbf{R}$ are sorted in descending order. The PSD of the late reverberation is then computed as

$$\hat{\gamma} = \frac{1}{M-1}\sum_{i=2}^{M}\mathbf{R}_{ii}. \tag{52}$$

Having an estimate of the late reverberation, we compute the noise CPSDM matrix as $\hat{\mathbf{P}}_{\mathbf{n}} = \hat{\gamma}\hat{\mathbf{\Phi}} + \mathbf{P}_{\mathbf{v}}$ and use it to estimate the early RATF and PSD of the target in the sequel.

We estimate the early RATF of the target using the method proposed in [8], [55]. We first compute the Cholesky decomposition $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{L}_{\mathbf{n}}\mathbf{L}_{\mathbf{n}}^H$. We then compute the whitened estimated noisy CPSDM as $\mathbf{P}_{\mathbf{w2}} = \mathbf{L}_{\mathbf{n}}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}(\mathbf{L}_{\mathbf{n}}^H)^{-1}$. Next, we compute the eigenvalue decomposition $\mathbf{P}_{\mathbf{w2}} = \mathbf{V}\mathbf{R}\mathbf{V}^H$, where the diagonal entries of $\mathbf{R}$ are sorted in descending order. We compute the early RATF as

$$\hat{\mathbf{a}} = \frac{\mathbf{L}_{\mathbf{n}}\mathbf{V}_1}{\mathbf{e}_1^T\mathbf{L}_{\mathbf{n}}\mathbf{V}_1}, \tag{53}$$

with $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$. We improve even further the accuracy of the estimated RATF by estimating the RATFs of all time-frames within each time-segment and then use the average of these as the RATF estimate. Finally, the target PSD is estimated as proposed in [15], [28], i.e.,

$$\hat{p} = \hat{\mathbf{w}}_{\text{MVDR}}^H\left(\hat{\mathbf{P}}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{n}}\right)\hat{\mathbf{w}}_{\text{MVDR}}, \tag{54}$$

where $\hat{\mathbf{w}}_{\text{MVDR}}$ is given in (49).

C. Dereverberation

We compare two different versions of the proposed SCFA$_{\text{rev}}$ problem in (24) referred to as SCFA$_{\text{rev1}}$ and SCFA$_{\text{rev2}}$. Unlike the SCFA$_{\text{no-rev}}$ problem (see Sec. IV-A), the SCFA$_{\text{rev}}$ problem also estimates the late reverberation PSD and thus is more appropriate in the context of dereverberation. Both versions use the box constraint for the γ parameter in (39) and the box constraint of the early RATF in (38). Moreover, since we assume that the microphones-self noise PSDs are all equal, both versions will use the box constraint in (40). Both methods use the true distance matrix $\hat{\mathbf{D}} = \mathbf{D}$. The SCFA$_{\text{rev1}}$ uses the linear inequality in (27), while the SCFA$_{\text{rev2}}$ does not use a constraint for the sum of PSDs. We also include in the comparisons the state-of-the-art approach described in Sec. VII-B (denoted as ref.). The reference method does not estimate the microphone-self noise PSD and we assume for the reference method that we have a perfect estimate, i.e., $\mathbf{P}_{\mathbf{v}} = q\mathbf{I}$. We consider a single target source without interfering signals so that the signal model in (7) reduces to

$$\mathbf{P}_{\mathbf{y}} = p_1\mathbf{a}_1\mathbf{a}_1^H + \underbrace{\gamma\mathbf{\Phi} + q\mathbf{I}}_{\mathbf{P}_{\mathbf{n}}}. \tag{55}$$

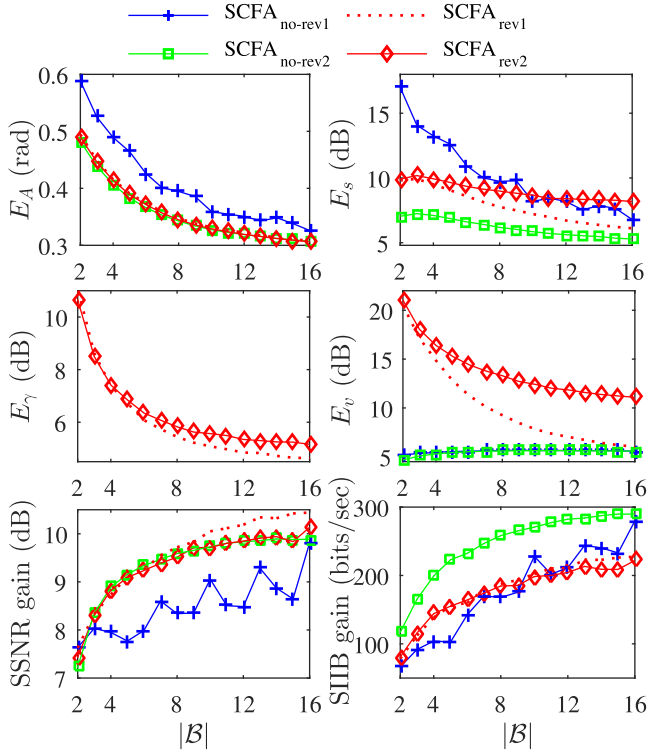


Fig. 6. Source separation results for $T_{60} = 0.2$ s: Comparison of the proposed SCFA$_{\text{no-rev2}}$, SCFA$_{\text{rev1}}$ and SCFA$_{\text{rev2}}$ methods, which assume knowledge of $\mathbf{D}$, and the proposed blind method denoted by SCFA$_{\text{no-rev1}}$.

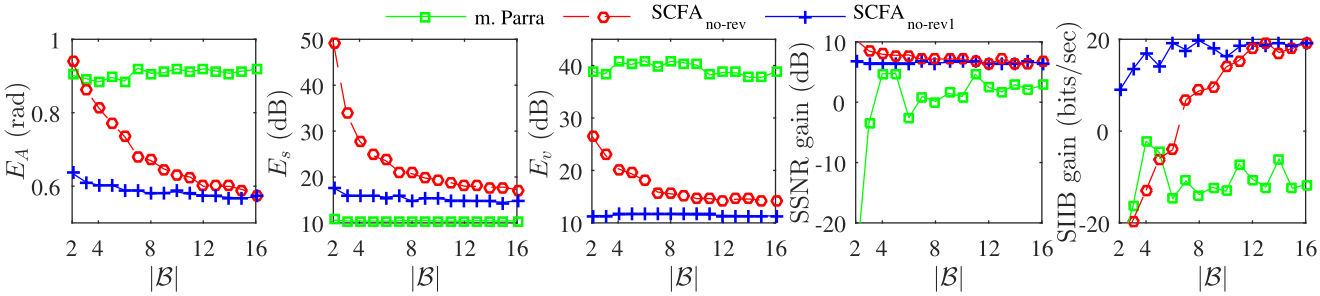


Fig. 7. Source separation results for $T_{60} = 0.6$ s: Comparison of m. Parra method and the proposed blind methods SCFA$_{\text{no-rev}}$ and SCFA$_{\text{no-rev1}}$.

After having estimated all the model parameters for the proposed and reference methods, the estimated parameters are used within the MWF given in (48), which is applied to the reverberant target source in order to enhance it.

Fig. 3 shows the results of the compared methods. The input SSNR of the female talker at the reference microphone is -9.73 dB. It is clear that in almost all evaluation criteria both proposed methods are significantly outperforming the reference method, except for the overall source PSD error E_s. However, the proposed methods have all larger intelligibility gain and better noise reduction performance compared to the reference method for $|\mathcal{B}_\beta| \geq 2$. Fig. 4 shows the underestimates and overestimates for the PSDs. It is clear that although the overall PSD error E_s is lower for the reference method, the proposed method has a lower underestimation error for the target, E_s^{un}, and a lower overestimation for the noise, E_γ^{ov}, which means less distortions to the target signal and therefore increased intelligibility.

D. Source Separation

We consider $r = 3$ source signals. In this acoustic scenario, the signal model is given by

$$\mathbf{P}_{\mathbf{y}} = \mathbf{P}_{\mathbf{e}} + \gamma\mathbf{\Phi} + q\mathbf{I}. \tag{56}$$

First we estimate the signal model parameters. We examine the performance of the proposed SCFA$_{\text{no-rev}}$ method (see Sec. IV-A) and the proposed methods SCFA$_{\text{no-rev1}}$, SCFA$_{\text{no-rev2}}$, SCFA$_{\text{rev1}}$, SCFA$_{\text{rev2}}$. Unlike the methods SCFA$_{\text{rev1}}$, SCFA$_{\text{rev2}}$, the methods SCFA$_{\text{no-rev1}}$ and SCFA$_{\text{no-rev2}}$ are based on the SCFA$_{\text{no-rev}}$ problem. The SCFA$_{\text{no-rev2}}$ method uses the box constraints in (28), (38) (which assumes full knowledge of $\hat{\mathbf{D}} = \mathbf{D}$), and (40). We also use the method SCFA$_{\text{no-rev1}}$ where the only difference with SCFA$_{\text{no-rev2}}$ is that SCFA$_{\text{no-rev1}}$ uses the RATF box constraint in (37) which does not depend on $\hat{\mathbf{D}}$. For the reference method, we use the method proposed in [4] (denoted as m. Parra), modified such that is as much aligned as possible with the proposed methods. Specifically, we solved the optimization problem of the reference method differently compared to [4]. Unlike [4] which uses the constraints $a_{ii} = 1$, we set the reference microphone row of $\mathbf{A}$ equal to the all-ones vector, as we did in all proposed methods. In addition, instead of the LS cost function used in [4], we used the ML cost function as with the proposed methods. We also used the same solver (see Sec. VI-D) for all compared methods. Note that the authors in [4] have solved the iterative problem using first-order derivatives only, while here we also use an approximation of the Hessian. Finally, the extracted parameters for both the reference and proposed methods are combined with the MWF in (48) where for each different source signal we use a different MWF $\hat{\mathbf{w}}_i$.

1) Low reverberation time. $T_{60} = 0.2s$: The input SSNR of the female talker, male talker and music signal at the reference microphone are -7.98, -10.35, and -4.71 dB, respectively. In order to have a clear visualization of the performance differences, we group the comparisons in two figures. Fig. 5 compares all blind methods that do not depend on $\hat{\mathbf{D}}$ or $\hat{\mathbf{\Phi}}$, i.e., SCFA$_{\text{no-rev}}$, SCFA$_{\text{no-rev1}}$ and the reference method (referred to as m. Parra). Recall that the only difference between the SCFA$_{\text{no-rev}}$ method and the m. Parra is the positivity constraints for the PSDs. It is clear that using these positivity constraints improves performance significantly. In contrast, the m. Parra often obtains negative PSD estimates which lead to an unpredicted noise reduction gain. This problem becomes more profound when $|\mathcal{B}_\beta|$ is small. For instance, for $|\mathcal{B}_\beta| = 2$, 1.8% and 97.2% of the sources' PSD estimates and microphone self noise PSD estimates, respectively, are negative, while for $|\mathcal{B}_\beta| = 16$, 0.5% and 94.1% of the sources' PSD estimates and microphone self noise PSD estimates are negative. Finally, note that the usage of extra inequality constraints from SCFA$_{\text{no-rev1}}$ is beneficial for improving the performance even more significantly.

In Fig. 6, we compare the best-performing SCFA$_{\text{no-rev1}}$ method of Fig. 5 with SCFA$_{\text{no-rev2}}$, SCFA$_{\text{rev1}}$ and SCFA$_{\text{rev2}}$. The problems that estimate the late reverberation parameter γ have worse estimation accuracy for the PSD of the sources and microphone-self noise and worse predicted intelligibility improvement compared to the SCFA$_{\text{no-rev2}}$ method. This is mainly due to the low reverberation time ($T_{60} = 0.2$ s) and the large number of parameters of SCFA$_{\text{rev1}}$ and SCFA$_{\text{rev2}}$ as argued in Sec. IV-B. However, SCFA$_{\text{rev1}}$ achieve a better noise reduction performance than the other methods when $|\mathcal{B}_\beta|$ becomes large.

2) Large reverberation time. $T_{60} = 0.6s$: In Figs. 7 and 8, we compare the same methods as in Fig. 5 and 6, respectively, but with $T_{60} = 0.6$. Here, the input SSNR of the female talker, male talker, and music signal at the reference microphone are -13.84, -16.56, and -11.53 dB, respectively. Here we observe that the methods which estimate γ become more accurate in RATF estimation, since now the contribution of late reverberation is significant (see the explanation in Sec. IV-B). Moreover, when the number of time-frames per time-segment $|\mathcal{B}_\beta|$ increases significantly the methods SCFA$_{\text{rev1}}$ and SCFA$_{\text{rev2}}$ have approximately the same predicted intelligibility improvement

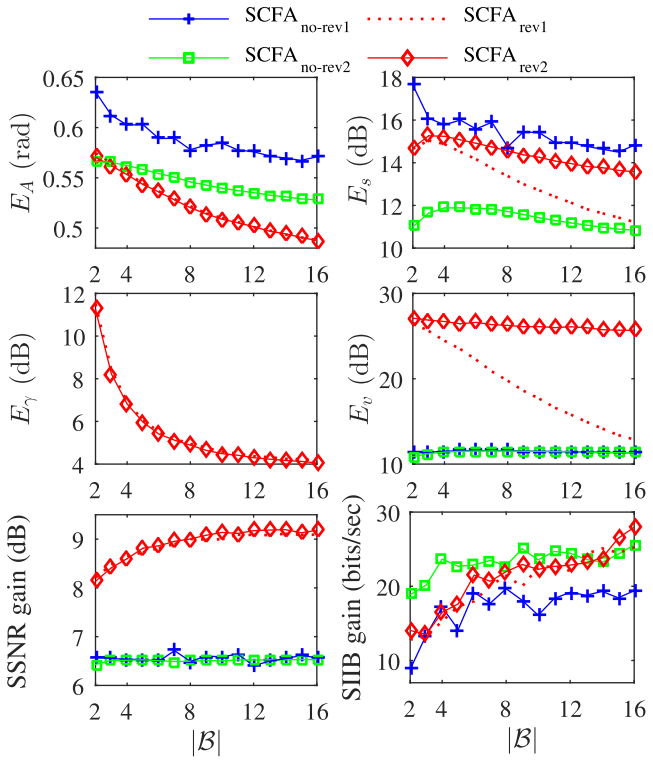


Fig. 8. Source separation results for $T_{60} = 0.6$ s: Comparison of the proposed $\text{SCFA}_{\text{no-rev2}}$, $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$ methods, which assume knowledge of $\mathbf{D}$, and the proposed blind method denoted by $\text{SCFA}_{\text{no-rev1}}$.

compared to the $\text{SCFA}_{\text{no-rev2}}$ method but have a much better noise reduction performance. Finally, the $\text{SCFA}_{\text{no-rev2}}$ method outperforms the $\text{SCFA}_{\text{no-rev1}}$ method in most cases. This is due to the tight box constraint in (38) used in $\text{SCFA}_{\text{no-rev2}}$.

In conclusion, we observe that in both applications the proposed approaches have shown remarkable robustness in highly reverberant environments. The box constraints that we used indeed provided estimates that are useful in both examined applications. Specifically, the box constraints avoided large overestimation errors in the late reverberation and microphone-self noise PSDs and large underestimation errors for the point sources PSDs. As a result the sources were not distorted significantly and combined with the good noise reduction performance we achieved large predicted intelligibility gains compared to the reference methods.

VIII. Conclusion

In this paper, we proposed several methods based on the combination of confirmatory factor analysis (CFA) and non-orthogonal joint diagonalization principles for estimating jointly several parameters of the multi-microphone signal model. The proposed methods achieved, in most cases, a better parameter estimation accuracy and a better performance in the context of dereverberation and source separation compared to existing state-of-the-art approaches. The inequality constraints introduced to limit the feasibility set in the proposed methods resulted in increased robustness in highly reverberant environments in both applications.

For future research it will be interesting to extend the proposed CFA problems to more general acoustic environments with for instance an additional diffuse noise component such as vehicle cabin noise. In this case, the signal model in (7) will become

$$\mathbf{P_y} = \mathbf{P_e} + \gamma\mathbf{\Phi} + \omega\mathbf{\Omega} + \mathbf{P_v}, \tag{57}$$

where the additional terms ω and $\mathbf{\Omega}$ are the PSD and spatial coherence matrix of the new diffuse noise component. Moreover, is interesting to examine whether the matrices $\mathbf{\Phi}$ and $\mathbf{\Omega}$ can be estimated as well via the proposed CFA problems. Such extended signal models might be more accurate in some acoustical scenarios, but at the price of increased number of parameters to be estimated.

References

[1] M. Brandstein and D. Ward (Eds.), *Microphone Arrays: Signal Processing Techniques and Applications*. New York, NY, USA: Springer, 2001.

[2] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[3] J. F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[5] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP J. Appl. Signal Process.*, vol. 2006, no. 1, pp. 1–13, 2006.

[6] D. Nion, K. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive parafac-based blind separation of convolutive speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1193–1207, Aug. 2010.

[7] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2006, no. 1, pp. 1–14, Dec. 2006.

[8] 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.

[9] 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.

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

[11] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 137–152, Jan. 2017.

[12] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1434–1448, Aug. 2018.

[13] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone subset selection for MVDR beamformer based noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 550–563, Mar. 2018.

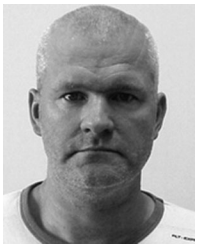
[14] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. EURASIP Europ. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

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

[16] S. Braun and E. A. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, 2015, Art. no. 34.
[17] 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.
[18] 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.
[19] 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. 1106–1118, Jun. 2018.
[20] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.
[21] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 106–110.
[22] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *ELSEVIER Signal Process.*, vol. 107, pp. 54–67, 2015.
[23] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 611–623, Mar. 2017.
[24] F. Antonacci *et al.*"Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.
[25] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Nat. Acad. Sci.*, vol. 110, no. 30, pp. 12 186–12 191, 2013.
[26] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on eigenvalue decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 611–615.
[27] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. EURASIP Eur. Signal Process. Conf.*, Aug. 2012, pp. 295–299.
[28] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 5728–5732.
[29] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.
[30] B. Schwartz, S. Gannot, and E. A. P. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2209–2222, Nov. 2017.
[31] A. Kuklasinski and J. Jensen, "Multichannel Wiener filters in binaural and bilateral hearing aids: Speech intelligibility improvement and robustness to DoA errors," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 8–16, 2017.
[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.
[33] D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method*. London, U.K.: Butterworths, 1963.
[34] K. G. Jöreskog, "A general approach to confirmatory maximum likelihood factor analysis," *Psychometrika*, vol. 34, no. 2, pp. 183–202, 1969.
[35] K. G. Jöreskog, "Simultaneous factor analysis in several populations," *Psychometrika*, vol. 36, no. 4, pp. 409–426, 1971.
[36] S. A. Mulaik, *Foundations of Factor Analysis*. Boca Raton, FL, USA: CRC Press, 2009.
[37] H. Kuttruff, *Room Acoustics*. Boca Raton, FL, USA: CRC Press, 1973.
[38] K. G. Jöreskog, "Factoring the multitest-multioccasion correlation matrix," *ETS Res. Bull. Ser.*, vol. 1969, no. 2, 1969, pages i-32, doi: 10.1002/j.2333-8504.1969.tb00740.x.
[39] K. G. Jöreskog, "Factor analysis by generalized least squares," *Psychometrika*, vol. 37, no. 3, pp. 243–260, 1972.
[40] K. G. Jöreskog and D. N. Lawley, "New methods in maximum likelihood factor analysis," *Br. J. Math. Statist. Psycol.*, vol. 21, pp. 85–96, 1968.
[41] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions with application to arithmetic complexity and statistics," *Linear Alg. Appl.*, vol. 18, no. 2, pp. 95–138, 1977.
[42] L. D. Lathauwer, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1096–1105, Mar. 2008.
[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[44] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[45] D. P. Bertsekas, "Projected newton methods for optimization problems with simple constraints," *SIAM J. Control Optim.*, vol. 20, no. 2, pp. 221–246, 1982.
[46] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, 1999.
[47] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Math. Program.*, vol. 89, no. 1, pp. 149–185, 2000.
[48] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.
[49] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement," in *Proc. ISCA Interspeech*, 2007, pp. 830–833.
[50] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Int. Workshop Hands-Free Speech Commun.*, 2017, pp. 11–15.
[51] P. C. Loizou, *Speech Enhancement: Theory Practice*. Boca Raton, FL, USA: CRC press, 2007.
[52] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.
[53] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2153–2166, Nov. 2018.
[54] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.
[55] S. Markovich 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.*, 2015, pp. 544–548.

Andreas I. Koutrouvelis received the B.Sc. degree in computer science from the University of Crete, Heraklion, Greece, in 2011, the M.Sc. degree in electrical engineering from Delft University of Technology (TU-Delft), Delft, The Netherlands, in 2014, and the Ph.D. degree from TU-Delft, in 2018. He is currently a Post-Doctoral Researcher with the Circuits and Systems Group (CAS), Faculty of Electrical Engineering, Mathematics and Computer Science, TU-Delft. His research interests include array signal processing and speech enhancement.

Richard C. Hendriks was born in Schiedam, The Netherlands. He received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively. He is currently an Associate Professor with the Circuits and Systems (CAS) Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. His main research interests include biomedical signal processing, and audio and speech processing, including speech enhancement, speech intelligibility improvement, and intelligibility modeling. In March 2010, he received the prestigious VENI grant for his proposal Intelligibility Enhancement for Speech Communication Systems. He was the recipient of the several best paper awards, among which the IEEE Signal Processing Society best paper award in 2016. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the *EURASIP Journal on Advances in Signal Processing*.



Richard Heusdens received the M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively. Since 2002, he has been an Associate Professor with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. In the spring of 1992, he joined the digital signal processing group with the Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio/speech signal processing activities within the ICT group. He held visiting positions at KTH (Royal Institute of Technology, Sweden), in 2002 and 2008, and is a Part-Time Professor with Aalborg University, Aalborg, Denmark. He is involved in research projects that cover subjects such as audio and acoustic signal processing, speech enhancement, and distributed signal processing for sensor networks.

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was the Ph.D. student and Assistant Research Professor with the Center for Person Kommunikation, Aalborg University. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. He is currently a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Signal and Information Processing, Department of Electronic Systems, Aalborg University. His research interests include acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.