This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3533374

Multimicrophone Signal Parameter Estimation in A Multi-Source Noisy Reverberant Scenario

Changheng Li and Richard C. Hendriks

***Abstract*—Estimation of acoustic parameters is of great interest but very challenging in the multichannel microphone signal processing area. Existing methods either assume simple, but less realistic scenarios, or suffer from very high computational costs. In this work, we consider the more general scenario where multiple sources, late reverberation and noise exist concurrently. The parameters of interest include the relative transfer functions (RTFs) of the point sources (both target and interferers) and individual power spectral densities (PSDs) of the sources and the late reverberation. We first propose a robust late reverberation PSD estimator using an iterative compensation scheme. Then, based on an analysis of the variance of the sample covariance matrices, we propose a robust and joint estimator for the sources RTFs and PSDs using multiple time frames that share the same RTFs. We compare the proposed method with the state-of-the-art simultaneously confirmatory factor analysis (SCFA) method and the second order blind identification (SOBI) method. Experiments show that our proposed method reaches the estimation performance of SCFA, which significantly outperforms SOBI, but uses much less computational costs compared to SCFA.**

***Index Terms*—Source separation, dereverberation, noise reduction, microphone array signal processing, RTF estimation, PSD estimation.**

I. Introduction

Microphone arrays are widely used in various devices, such as mobile phones, ear/headphones, hearing aids and all sorts of speech recognition applications. Typically, signals recorded by the microphones include not only the direct sounds from one or more point sources, but also reflections and ambient noise. In particular, the late reflections, known as late reverberation, are, next to the direct sound of interfering point sources, harmful to the speech quality and intelligibility [1], [2], even if these late reflections originate from the target source. Therefore, to achieve satisfying speech communication performance, microphone signals are processed by multi-microphone or single-microphone noise reduction and dereverberation algorithms [3], [4]. Multi-microphone noise reduction algorithms typically perform significantly better than their single-microphone counterparts [5] and typically depend on the relative transfer functions (RTFs) of the sources, the power spectral densities (PSDs) of the sources, the late reverberation and the ambient noise. However, in practice these parameters are unknown and their estimation is thus an essential problem for microphone array signal processing.

Many methods have been proposed in recent years to estimate these acoustic parameters [6]–[19]. However, when considering multiple sources and the coexistence of late reverberation and ambient noise, the estimation of the aforementioned parameters can be very challenging. Therefore, many of these works consider simplified signal models [7]–[12], [14], [15], [17]–[19], where either simplifying assumptions are used, or a subset of the parameters is assumed known. For instance, in [9], it is assumed that there is only a single active source in each time-frequency bin. In [8]–[10], [17], [18], either the late reverberation or the noise component is not considered in the model. Note that these methods based on simplified signal models have been widely used in practice, due to their simplicity and the properties of speech signals such as sparsity.

Some works considered a more general signal model, but have some other strict assumptions. For example, in [15], only the direct sound is considered as the target signal. The RTFs are assumed to only depend on the direction of arrival (DOA) of the source position and the microphone array geometry. By further assuming the DOA is known, the RTFs are considered known. However, the early reflections, which are beneficial to speech intelligibility [20], are sometimes included in the target sound. The number of unknown real parameters in each RTF vector is $2(M-1)$ with M the number of microphones. Some methods use prior knowledge (like hearing aids that assume a target in front). When considering both the direct sound and the early reflections as the target signal without prior knowledge of the scene, it is very challenging to estimate the RTFs. In [21], [22], the sound sources are assumed to be active successively and in [23], the interferers are assumed to be active earlier than the target sound source, which means that these methods cannot be used if two or more sources become active simultaneously.

The joint estimation of all the parameters considering multiple sources, late reverberation and ambient noise is achieved in [13] using the simultaneous confirmatory factor analysis (SCFA) method. Although this method is very effective, it comes with a very high computational cost. The goal of this paper is therefore to develop a method that can estimate the signal parameters (RTFs and PSDs of multiple sources, as well as the late reverberation PSD) at high accuracy and low complexity.

An important aspect of this problem formulation is the estimation of the late reverberation PSD. In [24], a comparison between many state-of-the-art late reverberation PSD estimators was published. All methods in this comparison considered only a single source and the RTF was assumed to be known for the spatial coherence-based methods. In this work, as part of the joint estimation of all unknown parameters, we propose a late reverberation PSD estimator that

This work is supported in part by the China Scholarship Council under Grant 202006340031 and in part by the Signal Processing Systems Group, Delft University of Technology, Delft, The Netherlands.

does not require knowledge on the RTFs. This can be seen as an extension of the method in [10] from a single-source to the multi-source scenarios.

In [25], a low complex blind source separation method was proposed based on a joint diagonalization of a set of covariance matrices. In [26], we modified this method to estimate the RTFs of multiple sources in a nearly non-reverberant and noiseless environment. In the current work, we extend the methods from [25], [26] to jointly estimate not only the RTFs in a noise-free and non-reverberant environment as in [26], but to estimate both the RTFs and the PSDs of the sources in a reverberant and noisy environment. Note that eventually, the noise component in this work refers to microphone self-noise. Although not strictly necessary for the proposed method, this is often modelled as spatially white Gaussian noise. Given a set of covariance matrices corresponding to a sequence of time-windows, [25] exploits the covariance matrix of the first time window and [26] exploits an average of a subset of these covariance matrices to jointly diagonalize the complete set and then estimate the RTFs. We show in this paper that any proper linear combination (e.g., a random combination or their average) of these matrices can be used and propose the optimal linear combination that minimizes the variances of the error matrix of the sample covariance matrix.

This paper is structured in the following way. Section II presents the signal model, statistical assumptions and problem formulation of this work. In Section III, we will first propose our late reverberation PSD estimator. Then in Section III-B, we modify the second order blind identification (SOBI) method from [25] to our estimation problem. After that, we will analyze the variance of the sample covariance matrices and propose our minimum variance joint diagonalization (MVJD) method to estimate the RTFs and the PSDs of the sources. In Section IV, experiments in different scenarios will be presented to compare our proposed method to some state-of-the-art reference methods. Finally, Section V concludes the paper.

II. Signal model

We consider the presence of R acoustic point sources recorded by a microphone array of M microphones in a reverberant and noisy environment. The number of sources R is assumed known in this work. (In practice, it can be estimated using some existing methods such as [27], [28].) The microphones can be placed compactly with various geometric structures (e.g., linear, circular or spherical). Each microphone records the signals generated from sound sources via both a direct propagation path and (infinite) reflections of surrounding objects (e.g. walls). These signals can be modeled as the convolution between the sound sources and the room impulse response (RIR). In the short-time Fourier transform (STFT) domain, the signal received at the m-th microphone can then be modeled as

$$y_m(l,k) = \underbrace{\sum_{r=1}^{R} x_{mr}(l,k)}_{x_m(l,k)} + \underbrace{\sum_{r=1}^{R} d_{mr}(l,k)}_{d_m(l,k)} + v_m(l,k) \ , \quad (1)$$

where l is the time index of the STFT window, which we will refer to as a sub-time frame, and k is the frequency-bin index. In addition to sub-time frames indexed by l, we will later also define time frames and time segments. The source reflections are typically labeled as direct component, early reflections (typically the first 50 ms), and late reflections. When considering the target source, these early reflections are actually beneficial for the speech intelligibility [20]. For a source r, we will therefore consider the direct component and early reflections combined, denoted by $x_{mr}(l,k)$, and differentiate these from the late reflections, denoted by $d_{mr}(l,k)$. The additive noise component is denoted by $v_m(l,k)$. In addition to potential interfering sources, both the late reverberation and additive noise are detrimental to speech intelligibility and quality.

As multiplication in the STFT domain can approximate the convolution in the time domain [29], we can model the r-th source at the m-th microphone as

$$x_{mr}(l,k) = a_{mr}(l,k)\, s_r(l,k) \ , \quad (2)$$

where $s_r(l,k)$ contains the direct sound and early reflections at the reference microphone and $a_{mr}(l,k)$ is the relative transfer function (RTF) [29] of the r-th source between the m-th microphone and the reference microphone. Without any limitation, we use the first microphone as our reference (i.e., $a_{1r} = 1$). For the duration that the sources are static relative to the microphone array, we can assume that the RTFs are constant. We refer to this duration as a time segment (TS) indexed by β. In vector form, the multi-microphone signal model is then given by

$$\mathbf{y}(l,k) = \underbrace{\sum_{r=1}^{R} \mathbf{a}_r(\beta,k)\, s_r(l,k)}_{\mathbf{x}(l,k)} + \mathbf{d}(l,k) + \mathbf{v}(l,k) \in \mathbb{C}^{M\times 1} \ , \quad (3)$$

where each column vector is stacked with M elements such as $\mathbf{y}(l,k) = [y_1(l,k), \cdots, y_M(l,k)]^T$.

Although speech-related signals $s_r(l,k)$ and $\mathbf{d}(l,k)$ are realizations of non-stationary processes, they can be assumed stationary for a short duration of a time frame (TF). The duration of a TF is much longer than that of the STFT window, which we already denoted as a sub-time frame (SF). Hence, we assume the t-th TF contains T consecutive SFs indexed by l from $l = 1 + (t-1)T$ to $l = tT$. In addition, we assume in this work that all sources are static for N consecutive TFs (e.g. $N = 8$ for approximately 2.5 s in our experiments), which means that the β-th TS contains N TFs indexed by t from $t = 1 + (\beta-1)N$ to $t = \beta N$. The relation between TS, TF and SF is visualized in Fig. 1. In the situation that sources are not static for the duration of a TS, we can use an adaptive time-segmentation e.g. as proposed in [30].

Within the t-th TF, the STFT coefficients vector $\mathbf{y}(l,k)$, with sub-frame index $l = 1 + (t-1)T, \cdots, tT$, is assumed to follow a circularly-symmetric complex Gaussian distribution with zero mean and cross power spectral density (CPSD) matrix $\mathbf{P}_{\mathbf{y}}(t,k) \in \mathbb{C}^{M\times M}$. Since $\mathbf{x}(l,k)$, $\mathbf{d}(l,k)$ and $\mathbf{v}(l,k)$ are commonly assumed to be mutually uncorrelated (even

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3533374

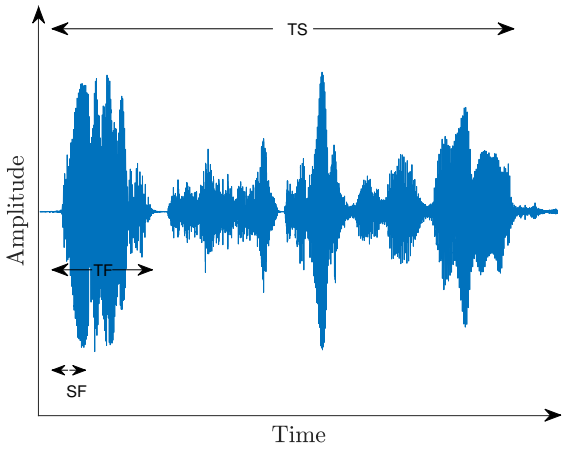


Figure 1: Visualisation of the definition of time segment (TS), time frames (TF) and sub frames (SF).

though strictly speaking $\mathbf{x}$ and $\mathbf{d}$ are weakly correlated), we can decompose $\mathbf{P}_{\mathbf{y}}(t,k)$ into

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(t,k) &= \mathrm{E}\left[\mathbf{y}(l,k)\,\mathbf{y}^{H}(l,k)\right] \\ &= \mathbf{P}_{\mathbf{x}}(t,k) + \mathbf{P}_{\mathbf{l}}(t,k) + \mathbf{P}_{\mathbf{v}}(t,k) \in \mathbb{C}^{M\times M}\ .\end{aligned} \tag{4}$$

For the source component $\mathbf{x}(l,k)$, containing the direct and early reflections for all sources, the CPSD matrix $\mathbf{P}_{\mathbf{x}}(t,k)$ is given by

$$\begin{aligned}\mathbf{P}_{\mathbf{x}}(t,k) &= \sum_{r=1}^{R} \phi_r(t,k)\,\mathbf{a}_r(\beta,k)\,\mathbf{a}_r^{H}(\beta,k) \\ &= \mathbf{A}(\beta,k)\,\mathbf{P}(t,k)\,\mathbf{A}^{H}(\beta,k)\end{aligned} \tag{5}$$

with $\mathbf{A}(\beta,k) = [\mathbf{a}_1(\beta,k),\cdots,\mathbf{a}_R(\beta,k)]$, $\mathbf{P}(t,k) = \mathrm{diag}\left[\phi_1(t,k),\cdots,\phi_R(t,k)\right]$ and $\phi_r(t,k) = \mathrm{E}\left[|s_r(l,k)|^2\right]$ the power spectral density (PSD) of the r-th source at the reference microphone with $|\cdot|$ denoting the absolute value. Note that in Eq. (5), we used the assumption that all sources are mutually uncorrelated and made explicit that the RTFs $\mathbf{a}_r(\beta,k)$ are constant over a time segment β.

For the late reverberation component, $\mathbf{P}_{\mathbf{l}}(t,k)$ is commonly assumed to be the product of a time-invariant full rank spatial coherence matrix $\mathbf{\Gamma}(k)$ and a time-varying PSD $\phi_\gamma(t,k)$ [7], [31], that is,

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

Here, $\mathbf{\Gamma}(k)$ is assumed to be measured or calculated *a priori* since it is time-invariant and independent of the microphone array position [32]–[34]. For instance, if a spherically isotropic noise field is assumed [35] and inter-microphone distances are assumed known, $\mathbf{\Gamma}(k)$ can be calculated to be

$$\mathbf{\Gamma}_{i,j}(k) = \mathrm{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{i,j}}{c}\right)\ , \tag{7}$$

with $\mathrm{sinc}(x) = \frac{\sin x}{x}$, $d_{i,j}$ the inter-distance between microphones i and j, f_s the sampling frequency, c the speed of sound and K the total frequency bin number. Also note that when a room has ceilings and floors that are more absorbing than the walls, the cylindrical isotropic noise field is a more realistic model. Note that with Eqs. (6) and (7), $\mathbf{P}_{\mathbf{l}}$ could also model other isotropic noise sources, i.e., noise sources that are not due to the late reverberation.

The noise component $\mathbf{v}$ is usually a summation of the microphone self-noise and other non-point noise sources that are approximately spatially uncorrelated. For this kind of noise, we assume that it has a time-invariant covariance matrix $\mathbf{P}_{\mathbf{v}}(k)$ for each frequency. Therefore, we can also measure $\mathbf{P}_{\mathbf{v}}(k)$ *a priori* by assuming a noise-only segment is available. In this work, we consider only the microphone self-noise to be present with each microphone having the same spatially white Gaussian noise distribution, which means $\mathbf{P}_{\mathbf{v}}(k) = \phi_v\mathbf{I}$. However, notice that we can always introduce a whitening step to guarantee $\mathbf{P}_{\mathbf{v}}(k)$ is spatially white.

With these assumptions, we can now write the covariance matrix of $\mathbf{y}(l)$ as

$$\mathbf{P}_{\mathbf{y}}(t) = \mathbf{A}(\beta)\,\mathbf{P}(t)\,\mathbf{A}^{H}(\beta) + \phi_\gamma(t)\,\mathbf{\Gamma} + \phi_v\mathbf{I}\ , \tag{8}$$

Note that we omitted the frequency indices for legibility in Eq. (8) and will do so for all the following equations since the estimators proposed in this work are independent across frequency. Based on the previously discussed stationarity of the signal, we can estimate $\mathbf{P}_{\mathbf{y}}(t)$ using the sample covariance matrix

$$\hat{\mathbf{P}}_{\mathbf{y}}(t) = \frac{1}{T}\sum_{l=1+(t-1)T}^{tT} \mathbf{y}(l)\,\mathbf{y}(l)^{H}\ . \tag{9}$$

Note that, to compute an STFT with a meaningful frequency resolution at 16kHz, the subframe duration or STFT window length cannot be too small. Meanwhile, to estimate the second-order statistics in practice, each time frame is composed of many subframes. Therefore, the time frame here is longer than the commonly assumed duration for stationarity. This will lead to an average of the PSDs but will maintain the RTF matrix or the spatial coherence matrix of the signal components [33]. Within the β-th TS, the *a priori* known or estimated parameters from the signal model given in Eq. (8) now include N sample covariance matrices $\left\{\hat{\mathbf{P}}_{\mathbf{y}}(t)\right\}_{t=1+(\beta-1)N}^{\beta N}$ (i.e., for the N time frames in segment β), the estimated spatial coherence matrix of the late reverberation $\hat{\mathbf{\Gamma}}$ and the estimated noise PSD $\hat{\phi}_v$. Note that as analyzing the errors of estimated $\hat{\mathbf{\Gamma}}$ is outside of the scope of this work, we assume that $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$ in the next section. The main goal of this paper is to develop an algorithm that can estimate the RTF matrix $\mathbf{A}(\beta)$, the diagonal PSD matrices of the sources $\{\mathbf{P}(t)\}_{t=1+(\beta-1)N}^{\beta N}$ and the PSDs of the late reverberation $\{\phi_\gamma(t)\}_{t=1+(\beta-1)N}^{\beta N}$ for each segment β.

III. Parameter Estimation

In this section, we propose our joint estimator based on a joint diagonalization scheme. We first introduce the estimator of the late reverberation PSDs in Section III-A. Then, we use the estimated late reverberation PSDs and the other *a priori*

given parameters to estimate the RTF matrix and the source PSDs in Section III-B.

A. Estimator of the Late Reverberation PSDs

We assume here that the late reverberation PSDs $\phi_\gamma(t)$ across time frames are unrelated and will estimate these per time frame t using $\hat{\mathbf{P}}_\mathbf{y}(t)$ for the t-th time frame only. Hence, for legibility, we will omit the time frame index in this subsection. Subtracting the true noise covariance matrix $\mathbf{P}_\mathbf{v}$ from $\mathbf{P}_\mathbf{y}$, we get

$$\mathbf{P}_\gamma = \mathbf{P}_\mathbf{y} - \mathbf{P}_\mathbf{v} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \phi_\gamma \mathbf{\Gamma} \ . \tag{10}$$

Taking the square-root decomposition such as the Cholesky decomposition of the full rank matrix $\mathbf{\Gamma}$, we have $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H$. Using $\mathbf{L}$, we can whiten matrix $\mathbf{P}_\gamma$ by calculating

$$\bar{\mathbf{P}}_\gamma = \mathbf{L}^{-1}\mathbf{P}_\gamma\mathbf{L}^{-H} = \left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H + \phi_\gamma\mathbf{I} \ . \tag{11}$$

Since the rank of $\left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H$ is R, we can see that, after whitening, the $M-R$ smallest eigenvalues of $\bar{\mathbf{P}}_\gamma$ should be equal to ϕ_γ. To see this, we can take the eigenvalue decomposition (EVD) of $\left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ with $\mathbf{U}$ unitary and $\mathbf{\Lambda}$ diagonal. The $M-R$ smallest diagonal elements of $\mathbf{\Lambda}$ are all zero. Taking the EVD of $\bar{\mathbf{P}}_\gamma$ using $\mathbf{U}$ we get

$$\mathbf{U}^H\bar{\mathbf{P}}_\gamma\mathbf{U} = \mathbf{\Lambda} + \phi_\gamma\mathbf{I} \ , \tag{12}$$

which shows that the $M-R$ smallest eigenvalues of $\bar{\mathbf{P}}_\gamma$ equal ϕ_γ. Because $\mathbf{P}_\mathbf{y}$ is estimated from limited data, the $M-R$ smallest eigenvalues will have some distribution around ϕ_γ. Therefore, we take their mean value as our estimate of ϕ_γ. That is,

$$\hat{\phi}_\gamma = \sum_{i=R+1}^{M} \frac{\lambda_{\gamma i}}{M-R} \ . \tag{13}$$

Note that, we assume all the eigenvalues in this work are ordered in descending order, i.e., λ_1 is the largest eigenvalue. The error of the estimator in Eq. (13) is analyzed for the special case with $R = 1$ in [10]. Eq. (13) is indeed a biased estimate of ϕ_γ (underestimation) due to using the subset of the ordered eigenvalues. Although Eq. (13) is not the optimal estimate of the late reverberation PSD, we choose this estimator since it does not need any RTF information.

Note that Eq. (13) can be seen as an extension of the method proposed in [10] where a single source scenario (i.e., $R = 1$) was assumed. However, we work with estimates of $\mathbf{P}_\mathbf{y}, \mathbf{\Gamma}$ and $\mathbf{P}_\mathbf{v}$. Therefore, similar to other spatial coherence-based methods as evaluated in [24], this method can have overestimation errors or underestimation errors when the late reverberation PSD is relatively small compared to the noise PSD (e.g. under low reverberant signal-to-noise ratios (RSNRs) in [24]). Even when the true covariance matrix $\mathbf{P}_\mathbf{y}$ is used, we can only obtain an estimated noise PSD, implying a residual noise PSD error will remain. Hence, we have

$$\hat{\mathbf{P}}_\gamma = \mathbf{P}_\mathbf{y} - \hat{\phi}_v\mathbf{I} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \phi_\gamma\mathbf{\Gamma} + \underbrace{\left(\phi_v - \hat{\phi}_v\right)\mathbf{I}}_{\text{residual noise}} \ . \tag{14}$$

The whitened matrix is then given by

$$\begin{aligned}\hat{\bar{\mathbf{P}}}_\gamma &= \mathbf{L}^{-1}\left(\mathbf{P}_\mathbf{y} - \hat{\phi}_v\mathbf{I}\right)\mathbf{L}^{-H} \\ &= \left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H + \phi_\gamma\mathbf{I} + \left(\phi_v - \hat{\phi}_v\right)\mathbf{\Gamma}^{-1} \ .\end{aligned} \tag{15}$$

If $\left(\phi_v - \hat{\phi}_v\right) \gg \phi_\gamma$, the $M-R$ smallest eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$ can be much larger than ϕ_γ resulting in large overestimation errors of ϕ_γ. If $-\left(\phi_v - \hat{\phi}_v\right) \gg \phi_\gamma$, the eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$ can be negative. A common way to deal with negative PSD estimates is to replace the negative estimates with ϵ as done in [12]. However, this will result in very large underestimation errors. To avoid large overestimation errors and underestimation errors, we propose the following estimation procedure for ϕ_v and ϕ_γ.

First of all, notice that $\mathbf{P}_\mathbf{x} = \mathbf{P}_\mathbf{y} - \phi_v\mathbf{I} - \phi_\gamma\mathbf{\Gamma} = \mathbf{A}\mathbf{P}\mathbf{A}^H$ is positive semi-definite with rank R. In practice, we have the estimated matrix

$$\hat{\mathbf{P}}_\mathbf{x} = \hat{\mathbf{P}}_\mathbf{y} - \hat{\phi}_v\mathbf{I} - \hat{\phi}_\gamma\mathbf{\Gamma} \ , \tag{16}$$

which can have negative eigenvalues even when we know the actual values of the PSDs ϕ_v and ϕ_γ, since we only have an estimated $\hat{\mathbf{P}}_\mathbf{y}$. Therefore, instead of adjusting $\hat{\phi}_v$ and $\hat{\phi}_\gamma$ to make $\hat{\mathbf{P}}_\mathbf{x}$ positive semi-definite with a rank R, we only constrain the estimated matrix $\hat{\mathbf{P}}_\mathbf{x}$ to have no less than R positive eigenvalues to overcome adjustments that will lead to overestimation of ϕ_γ. We now consider three cases in which this constraint is violated due to large overestimation errors of $\hat{\phi}_v$ and $\hat{\phi}_\gamma$.

1) If the given initial estimate $\hat{\phi}_v$ (estimated from speech absence frames) is larger than λ_{yR}, with λ_{yR} the R-th largest eigenvalue of $\hat{\mathbf{P}}_\mathbf{y}$, for any non-negative $\hat{\phi}_\gamma$, we have

$$\begin{aligned}\hat{\mathbf{P}}_\mathbf{x} &= \hat{\mathbf{P}}_\mathbf{y} - \hat{\phi}_v\mathbf{I} - \hat{\phi}_\gamma\mathbf{\Gamma} \\ &\preceq \hat{\mathbf{P}}_\mathbf{y} - \lambda_{yR}\mathbf{I} - \hat{\phi}_\gamma\mathbf{\Gamma} \\ &\preceq \hat{\mathbf{P}}_\mathbf{y} - \lambda_{yR}\mathbf{I} \ ,\end{aligned} \tag{17}$$

where the matrix inequality $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semi-definite. Since $\hat{\mathbf{P}}_\mathbf{y} - \lambda_{yR}\mathbf{I}$ has at most $R-1$ positive eigenvalues, $\hat{\mathbf{P}}_\mathbf{x}$ has less than R positive eigenvalues. Therefore, to make sure $\hat{\mathbf{P}}_\mathbf{x}$ has no less than R positive eigenvalues, we need $\hat{\phi}_v < \lambda_{yR}$. In this work, we update $\hat{\phi}_v$ by

$$\hat{\phi}_v \leftarrow \min\left\{\hat{\phi}_v, \frac{\sum_{i=R}^{M}\lambda_{yi}}{M-R+1}\right\} \ , \tag{18}$$

such that $\hat{\phi}_v \leq \frac{\sum_{i=R}^{M}\lambda_{yi}}{M-R+1} \leq \lambda_{yR}$, where, for the second inequality, the equality holds only when $\lambda_{yR} = \lambda_{yR+1} = \cdots = \lambda_{yM}$.

2) Next, $\hat{\phi}_v$ can still be largely overestimated such that the eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$ in Eq. (15) are too small to get a positive $\hat{\phi}_\gamma$ using Eq. (13). Therefore, we iteratively update $\hat{\phi}_v$ by $\hat{\phi}_v \leftarrow c_v\hat{\phi}_v$ with $0 < c_v < 1$ a constant value such as $c_v = 0.9$ and estimate $\hat{\phi}_\gamma$ using Eq. (13) again until a positive $\hat{\phi}_\gamma$ is obtained. Note that this

procedure has at most $\left\lceil \log_{c_v}\left(\frac{\lambda_{yM}}{\hat{\phi}_v}\right)\right\rceil + 1$ iterations since after these iterations, we have

$$\hat{\phi}_v c_v^{\left\lceil \log_{c_v}\left(\frac{\lambda_{yM}}{\hat{\phi}_v}\right)\right\rceil+1} < \hat{\phi}_v c_v^{\log_{c_v}\left(\frac{\lambda_{yM}}{\hat{\phi}_v}\right)} = \lambda_{yM} \ , \quad (19)$$

and $\hat{\tilde{\mathbf{P}}}_\gamma$ will be positive definite. This results in positive eigenvalues of $\hat{\tilde{\mathbf{P}}}_\gamma$, and hence, a positive $\hat{\phi}_\gamma$.

3) Finally, $\hat{\phi}_\gamma$ can be overestimated such that $\hat{\mathbf{P}}_\mathbf{x}$ has less than R positive eigenvalues. Therefore, we iteratively update $\hat{\phi}_\gamma$ by $\hat{\phi}_\gamma \leftarrow c_\gamma \hat{\phi}_\gamma$ with $0 < c_\gamma < 1$ a constant value such as $c_\gamma = 0.1$ until $\hat{\mathbf{P}}_\mathbf{x}$ has R positive eigenvalues. Since we have updated $\hat{\phi}_v$ by Eq. (18), we have $\hat{\phi}_v \leq \lambda_{yR}$. Hence, in the worst case that we need many iterations, $\hat{\phi}_\gamma$ approaches zero and $\hat{\mathbf{P}}_\mathbf{x} \approx \hat{\mathbf{P}}_\mathbf{y} - \hat{\phi}_v \mathbf{I}$ can have R positive eigenvalues.

The late reverberation PSD estimator is summarized in Algorithm 1.

Algorithm 1: ϕ_γ estimator

Input: Estimated $\mathbf{P}_\mathbf{y}$, $\mathbf{\Gamma}$,init.$\hat{\phi}_v$, $IterN$
Output: $\hat{\phi}_\gamma$,$\hat{\mathbf{P}}_\mathbf{x}$

```
for all k, l do
    Calculate the EVD of P_y and update φ̂_v using Eq. (18).
    Use φ̂_v and Γ to do subtraction and whitening using Eq. (14) and Eq. (15).
    Calculate the EVD of P̂̃_γ.
    Calculate φ̂_γ using Eq. (13).
    while φ̂_γ < 0 do
        Update φ̂_v by φ̂_v ← c_v φ̂_v
        Calculate the EVD of P̂̃_γ.
        Calculate φ̂_γ using Eq. (13).
    Calculate P̂_x using Eq. (16).
    Calculate the R-th largest eigenvalue of P̂_x, λ_xR.
    while λ_xR < 0 do
        Update φ̂_γ ← c_γ φ̂_γ. Calculate P̂_x using Eq. (16).
        Calculate the R-th largest eigenvalue of P̂_x, λ_xR.
```

B. Estimator of the RTF matrix and the source PSDs

Without loss of generality, we consider the estimator of the RTF matrix and the source PSDs for the first time segment (i.e., $\beta = 1$) and neglect the index β for notational convenience. Since all time frames in a time segment are assumed to share the same RTFs, we can estimate the RTF matrix with improved accuracy using all time frames jointly, similar to the recently proposed methods in [13], [26]. Having estimated $\hat{\phi}_\gamma$ and $\hat{\phi}_v$, we can subtract both the late reverberation and noise components from $\hat{\mathbf{P}}_\mathbf{y}(t)$ for $t = 1, \cdots, N$, and we get

$$\hat{\mathbf{P}}_\mathbf{x}(t) = \hat{\mathbf{P}}_\mathbf{y}(t) - \hat{\phi}_v \mathbf{I} - \hat{\phi}_\gamma(t)\mathbf{\Gamma} = \widehat{\mathbf{A}\mathbf{P}(t)\mathbf{A}^H} \ , \quad (20)$$

for $t = 1, \cdots, N$.

1) parameter identifiability: Before estimating the RTF matrix and the source PSDs, we need to analyze the parameter identifiability to avoid biased estimates. In general, the parameters are said to be identifiable meaning that if two matrices have the form as in Eq. (20) (i.e., $\mathbf{P}_{\mathbf{x}1}(t) = \mathbf{A}_1 \mathbf{P}_1(t) {\mathbf{A}_1}^H$ and $\mathbf{P}_{\mathbf{x}2}(t) = \mathbf{A}_2 \mathbf{P}_2(t) {\mathbf{A}_2}^H$) for $t = 1, \cdots, N$, then $\mathbf{P}_{\mathbf{x}1} = \mathbf{P}_{\mathbf{x}2}$ is equivalent to $\mathbf{A}_1 = \mathbf{A}_2$ and $\mathbf{P}_1 = \mathbf{P}_2$. Note that for a given matrix $\mathbf{P}_\mathbf{x} = \mathbf{A}\mathbf{P}\mathbf{A}^H$, we can find different solutions by simply permuting the columns of $\mathbf{A}$ and corresponding diagonal elements of $\mathbf{P}$. However, since this permutation ambiguity can be further solved by methods such as post-processing [36], we consider the parameters to be equal to their permuted versions in this work. Note that the first row of $\mathbf{A}$ are all ones and $\mathbf{P}$ is diagonal.

We now show that for multiple sources (i.e., $R > 1$) and a time-segment consisting of one time-frame (i.e., $N = 1$), the parameters are not identifiable. That means, for any matrix $\mathbf{A}_1$ with its first row all ones and $\mathbf{P}_1(1)$ diagonal, we can find $\mathbf{A}_2 \neq \mathbf{A}_1$ and $\mathbf{P}_2 \neq \mathbf{P}_1$ while $\mathbf{A}_1\mathbf{P}_1{\mathbf{A}_1}^H = \mathbf{A}_2\mathbf{P}_2{\mathbf{A}_2}^H$, where the first row of $\mathbf{A}_2$ are all ones and $\mathbf{P}_2$ is diagonal.

For any unitary matrix $\mathbf{Q} \in \mathbb{C}^{R\times R}$, we can construct

$$\mathbf{A}_2 = \mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \Big(\text{diag}\left(\mathbf{e}_1^H \mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \right) \Big)^{-1} \quad (21)$$

and

$$\mathbf{P}_2 = \Big(\text{diag}\left(\mathbf{e}_1^H \mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \right) \Big) \Big(\text{diag}\left(\mathbf{e}_1^H \mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \right) \Big)^H \quad (22)$$

with $\mathbf{e}_1 = [1, 0, \cdots, 0]^T \in \mathbb{C}^{M\times 1}$ (where the subscript in $\mathbf{e}_1$ indicates that the first microphone is the reference). The diagonal matrix $\mathbf{P} = \text{diag}\left(\mathbf{e}_1^H \mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \right)$ is used to make the first row of $\mathbf{A}_2$ all ones and $\mathbf{P}_2$ diagonal. We then have

$$\begin{aligned} \mathbf{P}_{\mathbf{x}2} &= \mathbf{A}_2 \mathbf{P}_2 {\mathbf{A}_2}^H \\ &= \left(\mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \mathbf{P}^{-1} \right) \left(\mathbf{P}\mathbf{P}^H \right) \left(\mathbf{A}_1 \mathbf{P}_1^{\frac{1}{2}} \mathbf{Q} \mathbf{P}^{-1} \right)^H \\ &= \mathbf{A}_1 \mathbf{P}_1 {\mathbf{A}_1}^H \\ &= \mathbf{P}_{\mathbf{x}1} \ , \end{aligned} \quad (23)$$

but $\mathbf{A}_2 \neq \mathbf{A}_1$ and $\mathbf{P}_2 \neq \mathbf{P}_1$ for the non-diagonal unitary $\mathbf{Q}$ and $R > 1$ ($\mathbf{Q}$ is a scalar when $R = 1$). Hence, if no other prior information is used, the parameters for a single time frame are not identifiable, and we need multiple time frames, i.e., $N \geq 2$, for each time segment to estimate the RTF matrix and the PSDs uniquely. Note that $N \geq 2$ is only a necessary condition for the identifiability of the parameters. For a sufficient condition, We need further assumptions on the PSDs of the sources, which we will introduce in Section III-B2.

2) SOBI: Although the SOBI method was proposed in [25] to estimate the mixing matrix and separate the source signals directly, we slightly modify this method and use it to estimate the RTF matrix and the PSDs. We therefore first introduce a modified SOBI as the reference method for RTF estimation in this subsection. Subsequently, in the next subsection, we propose a significantly improved method based on SOBI, referred to as the minimum variance joint diagonalization method (MVJD).

Given is a set of covariance matrices $\{\mathbf{P}_\mathbf{x}(t)\}_{t=1}^N$, with $N \geq 2$. To find $\mathbf{A}$ and $\mathbf{P}(t)$, such that $\mathbf{P}_\mathbf{x}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H$,

for $t = 1, \cdots, N$, we can make use of a joint diagonalization of the set $\{\mathbf{P_x}(t)\}_{t=1}^{N}$. That means, instead of estimating $\mathbf{A}$ and $\mathbf{P}(t)$ directly, we first estimate $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}(1)^{\frac{1}{2}}$ and $\tilde{\mathbf{P}}(t) = \mathbf{P}(1)^{-\frac{1}{2}}\mathbf{P}(t)\mathbf{P}(1)^{-\frac{H}{2}}$ via solving a joint diagonalization problem, as we will show later. Let us for now assume we know $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(t)$. In that case, we can estimate the RTF matrix and $\mathbf{P}(t)$ by

$$\mathbf{A} = \tilde{\mathbf{A}}\text{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)^{-1} , \tag{24}$$

$$\mathbf{P}(1) = \text{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)\text{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)^H , \tag{25}$$

and

$$\mathbf{P}(t) = \mathbf{P}(1)^{\frac{1}{2}}\tilde{\mathbf{P}}(t)\mathbf{P}(1)^{\frac{H}{2}} , \tag{26}$$

where $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$.

Now, we show how to estimate $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(t)$. Consider estimating the SVD components of $\tilde{\mathbf{A}} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^H$. We can reformulate $\mathbf{P_x}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H$ by

$$\begin{aligned}\mathbf{P_x}(t) &= \underbrace{\mathbf{A}\mathbf{P}(1)^{\frac{1}{2}}}_{\tilde{\mathbf{A}}}\underbrace{\mathbf{P}(1)^{-\frac{1}{2}}\mathbf{P}(t)\mathbf{P}(1)^{-\frac{H}{2}}}_{\tilde{\mathbf{P}}(t)}\mathbf{P}(1)^{\frac{H}{2}}\mathbf{A}^H \\ &= \tilde{\mathbf{A}}\tilde{\mathbf{P}}(t)\tilde{\mathbf{A}}^H \\ &= \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\underbrace{\mathbf{V}^H\tilde{\mathbf{P}}(t)\mathbf{V}}_{\mathbf{P_w}(t)}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{U}^H .\end{aligned} \tag{27}$$

For $t = 1$, we have $\mathbf{P_x}(1) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ as then $\tilde{\mathbf{P}}(1) = \mathbf{I}$ and $\mathbf{V}^H\mathbf{V} = \mathbf{I}$. Therefore, both $\mathbf{U}$ and $\mathbf{\Sigma}$ are known from the above EVD of $\mathbf{P_x}(1)$. Then we can use $\mathbf{U}$ and $\mathbf{\Sigma}$ to calculate

$$\begin{aligned}\mathbf{P_w}(t) &= \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{U}^H\mathbf{P_x}(t)\mathbf{U}\mathbf{\Sigma}^{-\frac{1}{2}} \\ &= \mathbf{V}^H\tilde{\mathbf{P}}(t)\mathbf{V} .\end{aligned} \tag{28}$$

Since $\mathbf{V}$ is unitary and $\tilde{\mathbf{P}}(t)$ is diagonal with its r-th diagonal element $\frac{\phi_r(t)}{\phi_r(1)}$, Eq. (28) is indeed the EVD of $\mathbf{P_w}(t)$ with $\left\{\frac{\phi_r(t)}{\phi_r(1)}\right\}_{r=1}^{R}$ the eigenvalues and $\mathbf{V}$ the joint eigenvector matrix for all t in the segment. To make sure we get a unique estimate of $\mathbf{V}$, we need to assume that for any r_1-th and r_2-th eigenvectors (i.e. the r_1-th and r_2-th columns of $\mathbf{V}$), there exist one time frame t_0 such that the r_1-th and r_2-th eigenvalues are distinct [25], i.e.,

$$\frac{\phi_{r_1}(t_0)}{\phi_{r_1}(1)} \neq \frac{\phi_{r_2}(t_0)}{\phi_{r_2}(1)} . \tag{29}$$

The joint eigenvector matrix $\mathbf{V}$ diagonalizes $\{\mathbf{P_w}(t)\}_{t=1}^{N}$ simultaneously, i.e.,

$$\mathbf{V}\mathbf{P_w}(t)\mathbf{V}^H = \tilde{\mathbf{P}}(t), \forall t \in \{1, \cdots, N\} . \tag{30}$$

However, such a joint diagonalization might not be achieved in practice since we only have the estimated $\mathbf{P_w}(t)$. Therefore, an approximate joint diagonalization was pursued in [25] by minimizing the off-diagonal elements of $\mathbf{V}\mathbf{P_w}(t)\mathbf{V}^H$, which is

$$\begin{aligned}&\min_{\mathbf{V}} \textstyle\sum_{t=2}^{N} \text{off}\left(\mathbf{V}\mathbf{P_w}(t)\mathbf{V}^H\right) \\ &\text{s.t. } \mathbf{V}^H\mathbf{V} = \mathbf{I} ,\end{aligned} \tag{31}$$

where $\text{off}(\mathbf{C}) = \sum_{1\leq i\neq j\leq M}|C_{i,j}|^2$ for a matrix $\mathbf{C} \in \mathcal{C}^{M\times M}$. Then, the algorithm proposed in [37] is used to solve Eq. (31), which is numerically very efficient. With $\mathbf{V}$ estimated, we use $\text{diag}\left(\mathbf{V}\mathbf{P_w}(t)\mathbf{V}^H\right)$ as the estimate of $\tilde{\mathbf{P}}(t)$.

The SOBI method is summarized in Algorithm 2.

Algorithm 2: SOBI

Input: Estimated $\hat{\mathbf{P}}_{\mathbf{x}}(t)$, for $t = 1, \cdots, N$,
Output: $\mathbf{A}$ and $\mathbf{P}(t)$ for $t = 1, \cdots, N$,
1 Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\hat{\mathbf{P}}_{\mathbf{x}}(1)$.
2 Construct new matrices $\mathbf{P_w}(t)$ for $t = 2, \cdots, N$ using Eq. (28).
3 Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(t)$ for $t = 2, \cdots, N$ using the Jacobi-like algorithm [37].
4 Estimate $\tilde{\mathbf{A}}$ with $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.
5 Estimate $\mathbf{A}$ and $\mathbf{P}(t)$ using Eqs. (24) to (26).

Note that, with this SOBI-based algorithm, the matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ in the SVD of $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}(1)^{\frac{1}{2}}$ are first estimated before estimating the RTF matrix and the PSDs. The estimation accuracy of $\mathbf{U}$ and $\mathbf{\Sigma}$ depends fully on the estimation accuracy of the first covariance matrix $\mathbf{P_x}(1) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, which can be hugely erroneous. For instance, when the late reverberation and noise have large energy during the first time frame. Instead of using $\mathbf{P_x}(1)$ to do the EVD at the first step, we can use any proper linear combination of all the covariance matrices $\sum_{t=1}^{N} c_t\mathbf{P_x}(t)$ with $c_t \geq 0$, such as the average of a subset of the covariance matrices as we proposed in [26]. The estimation accuracy of the RTF matrix and the PSDs can be improved by using values for c_t that minimize the error between $\sum_{t=1}^{N} c_t\mathbf{P_x}(t)$ and its estimated counterpart $\sum_{t=1}^{N} c_t\hat{\mathbf{P}}_{\mathbf{x}}(t)$.

3) MVJD: In this subsection, we first show our generalization of the SOBI method. Then we propose our minimum variance joint diagonalization method (MVJD) based on the analysis of the variance of the sample covariance matrices.

Instead of using the first covariance matrix $\mathbf{P_x}(1)$ to do the EVD at the first step of SOBI, we can use any proper linear combination of all the covariance matrices $\sum_{t=1}^{N} c_t\mathbf{P_x}(t) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ with $c_t \geq 0$. Therefore, $\mathbf{U}$ and $\mathbf{\Sigma}$ can be obtained from $\sum_{t=1}^{N} c_t\mathbf{P_x}(t)$ for

$$\tilde{\mathbf{A}} = \mathbf{A}\left(\sum_{t=1}^{N} c_t\mathbf{P}(t)\right)^{\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^H . \tag{32}$$

Then, using Eqs. (28) and (31), we can get $\mathbf{V}$ and $\tilde{\mathbf{P}}(t) = \left(\sum_{t=1}^{N} c_t\mathbf{P}(t)\right)^{-\frac{1}{2}}\mathbf{P}(t)\left(\sum_{t=1}^{N} c_t\mathbf{P}(t)\right)^{-\frac{H}{2}}$. To get a unique estimate of $\mathbf{V}$, we assume that for any r_1-th and r_2-th columns of $\mathbf{V}$, there exists one time frame t_0 such that

$$\frac{\phi_{r_1}(t_0)}{\sum_{t=1}^{N} c_t\phi_{r_1}(t)} \neq \frac{\phi_{r_2}(t_0)}{\sum_{t=1}^{N} c_t\phi_{r_2}(t)} . \tag{33}$$

With $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ estimated, we can estimate $\tilde{\mathbf{A}}$ using Eq. (32), which further gives us the RTF matrix $\mathbf{A} = \tilde{\mathbf{A}}\text{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)^{-1}$ and $\sum_{t=1}^{N} c_t\mathbf{P}(t) =$

$\text{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)\text{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)^H$. Finally, from $\sum_{t=1}^{N} c_t\mathbf{P}(t)$ and $\tilde{\mathbf{P}}(t)$, we can calculate the PSDs matrix for all time frames by $\mathbf{P}(t)=\tilde{\mathbf{P}}(t)\sum_{t=1}^{N} c_t\mathbf{P}(t)$.

Since the estimation errors for the estimated covariance matrices $\mathbf{P}_{\mathbf{x}}(t)$ are different for different t, using different coefficients c_t in step 1 will result in a different $\mathbf{U}$ and $\mathbf{\Sigma}$, and thus in different estimates of the RTF matrix and the PSDs.

We will now explain how we can optimally select the coefficients c_t such that the summation of the variances of the error matrix in the estimated $\mathbf{P}_{\mathbf{x}}(t)$ is minimized. Suppose we have the true PSDs of the late reverberation and noise. The estimated covariance matrix for $\mathbf{P}_{\mathbf{x}}(t)$ is then given by

$$\begin{aligned}\hat{\mathbf{P}}_{\mathbf{x}}(t)=&\hat{\mathbf{P}}_{\mathbf{y}}(t)-\phi_v\mathbf{I}-\phi_\gamma(t)\mathbf{\Gamma}\\ =&\sum_{l=1+(t-1)T}^{tT}\frac{\mathbf{y}(l)\mathbf{y}(l)^H}{T}-\phi_v\mathbf{I}-\phi_\gamma(t)\mathbf{\Gamma}\\ =&\sum_{l}\frac{(\mathbf{x}(l)+\mathbf{n}(l))(\mathbf{x}(l)^H+\mathbf{n}(l)^H)}{T}\\ &-\phi_v\mathbf{I}-\phi_\gamma(t)\mathbf{\Gamma}\\ =&\sum_{l}\frac{\mathbf{x}(l)\mathbf{x}(l)^H}{T}\\ &+\sum_{l}\frac{\mathbf{x}(l)\mathbf{n}(l)^H+\mathbf{n}(l)\mathbf{x}(l)^H}{T}\\ &+\sum_{l}\frac{\mathbf{n}(l)\mathbf{n}(l)^H}{T}-\phi_\gamma(t)\mathbf{\Gamma}-\phi_v\mathbf{I}\,,\end{aligned}\tag{34}$$

where $\mathbf{n}(l)=\mathbf{d}(l)+\mathbf{v}(l)$. Since we assumed that $\mathbf{x}$, $\mathbf{l}$ and $\mathbf{v}$ are uncorrelated, we omit the cross correlation terms in Eq. (34) and get

$$\begin{aligned}\hat{\mathbf{P}}_{\mathbf{x}}(t)\approx&\sum_{l}\frac{\mathbf{x}(l)\mathbf{x}(l)^H}{T}\\ &+\sum_{l}\frac{\mathbf{n}(l)\mathbf{n}(l)^H}{T}-\phi_\gamma(t)\mathbf{\Gamma}-\phi_v\mathbf{I}\,.\end{aligned}\tag{35}$$

Applying this to the weighted sum of the estimated covariance matrices used at the first step of our proposed method, we get

$$\begin{aligned}\sum_{t=1}^{N}c_t\hat{\mathbf{P}}_{\mathbf{x}}(t)\approx&\sum_{t=1}^{N}c_t\frac{\sum_{l=1+(t-1)T}^{tT}\mathbf{x}(l)\mathbf{x}(l)^H}{T}\\ &+\underbrace{\sum_{t=1}^{N}c_t\frac{\sum_{l=1+(t-1)T}^{tT}\mathbf{n}(l)\mathbf{n}(l)^H}{T}-\sum_{t=1}^{N}c_t\left(\phi_\gamma(t)\mathbf{\Gamma}+\phi_v\mathbf{I}\right)}_{\mathbf{W}}\,,\end{aligned}\tag{36}$$

where the first term is a weighted sum of the sample covariance matrices for the target sources and the remaining terms are unwanted errors that we will denote by matrix $\mathbf{W}$. Since $\mathbf{n}(l)$, for $l=1+(t-1)T,\cdots,tT$, is assumed to follow a circularly-symmetric complex Gaussian distribution with zero mean and covariance matrix $\mathbf{P}_{\mathbf{n}}(t)=\phi_\gamma(t)\mathbf{\Gamma}+\phi_v\mathbf{I}$, the random matrix $\mathbf{W}_t=\sum_{l=1+(t-1)T}^{tT}\mathbf{n}(l)\mathbf{n}(l)^H$ has a complex Wishart distribution $\sim\mathcal{W}_M^C(T,\mathbf{P}_{\mathbf{n}}(t))$ with T degrees of freedom [38]. The expectation of $\mathbf{W}_t$ is $T\mathbf{P}_{\mathbf{n}}(t)$ [39]. Hence the expectation of $\mathbf{W}$ is

$$\begin{aligned}\mathrm{E}\{\mathbf{W}\}=&\sum_{t=1}^{N}c_t\frac{\mathrm{E}\{\mathbf{W}_t\}}{T}-\sum_{t=1}^{N}c_t\mathbf{P}_{\mathbf{n}}(t)\\ =&\sum_{t=1}^{N}c_t\frac{T\mathbf{P}_{\mathbf{n}}(t)}{T}-\sum_{t=1}^{N}c_t\mathbf{P}_{\mathbf{n}}(t)\\ =&0\,.\end{aligned}\tag{37}$$

The variance of the $\{i,j\}$-th element of $\mathbf{W}_t$ is $\text{var}\{W_{t,i,j}\}=P_{n,i,i}P_{n,j,j}$ [39]. Hence the summation of the variances of all the elements of $\mathbf{W}_t$ is

$$\begin{aligned}\sum_{i,j=1}^{M}\text{var}\{W_{i,j}\}=&\sum_{i,j=1}^{M}\text{var}\left\{\sum_{t=1}^{N}c_t\frac{W_{t,i,j}}{T}\right\}\\ =&\sum_{i,j=1}^{M}\sum_{t=1}^{N}\frac{c_t^2}{T^2}\text{var}\{W_{t,i,j}\}\\ =&\sum_{i,j=1}^{M}\sum_{t=1}^{N}\frac{c_t^2}{T^2}P_{n,i,i}P_{n,j,j}\\ =&\sum_{t=1}^{N}\frac{c_t^2}{T^2}\left[\text{tr}\left(\mathbf{P}_{\mathbf{n}}(t)\right)\right]^2\\ \geq&\frac{1}{T^2}\left[\sum_{t=1}^{N}c_t\text{tr}\left(\mathbf{P}_{\mathbf{n}}(t)\right)\right]^2\,,\end{aligned}\tag{38}$$

where the equality holds when $c_1\text{tr}(\mathbf{P}_{\mathbf{n}}(1))=c_2\text{tr}(\mathbf{P}_{\mathbf{n}}(2))=\cdots=c_N\text{tr}(\mathbf{P}_{\mathbf{n}}(N))$. Since $\text{tr}(\mathbf{P}_{\mathbf{n}}(t))=M(\phi_\gamma(t)+\phi_v)$, we can choose $c_t=\frac{1}{\phi_\gamma(t)+\phi_v}$ to minimize the variances of the error matrix.

The MCJD method is summarized in Algorithm 3.

Algorithm 3: MVJD

Input: Estimated $\hat{\mathbf{P}}_{\mathbf{x}}(t)$ $\phi_\gamma(t)$ and ϕ_v,for $t=1,\cdots,N$,
Output: $\mathbf{A}$ and $\mathbf{P}(t)$ for $t=1,\cdots,N$,
1 Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\sum_{t=1}^{N}\frac{1}{\phi_\gamma(t)+\phi_v}\hat{\mathbf{P}}_{\mathbf{x}}(t)$.
2 Construct new matrices $\mathbf{P}_{\mathbf{w}}(t)$ for $t=1,\cdots,N$ using Eq. (28).
3 Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(t)$ for $t=1,\cdots,N$ using the Jacobi-like algorithm [37].
4 Estimate $\tilde{\mathbf{A}}$ with $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ using Eq. (32).
5 Estimate $\mathbf{A}$ and $\mathbf{P}(t)$ using $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(t)$.

IV. Experiments

In this section, we evaluate the estimation performance of our proposed method in various simulated acoustic scenarios using multiple microphones. We compare our method to both the SOBI based method introduced in Section III-B2 and the

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3533374

SCFA method [13] that we will introduce in Section IV-A. In Section IV-B, we evaluate the different methods using performance measures for the estimation accuracy, the predicted speech quality and the predicted speech intelligibility. Finally, the performances including the computational complexity of all methods are presented and discussed in Sections IV-C and IV-D.

A. Reference methods

In addition to the SOBI method introduced in Section III-B2, we include another state-of-the-art method for comparison, which is the simultaneous confirmatory factor analysis (SCFA) method [13]. The SCFA method is based on the maximum likelihood cost function:

$$\min \sum_{t=1}^{N} \log \left|\mathbf{P}_{\mathbf{y}}\left(t\right)\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(t\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(t\right)\right) . \tag{39}$$

Specifically, the following non-convex optimization problem is formalized in [13]

$$\begin{aligned} &\underset{\mathbf{P}\left(t\right),\mathbf{A}}{\arg\min} \sum_{t=1}^{N} \log \left|\mathbf{P}_{\mathbf{y}}\left(t\right)\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(t\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(t\right)\right) \\ &\phi_{\gamma}\left(t\right),\phi_{v} \\ &\quad \text{s.t. } \mathbf{P}_{\mathbf{y}}\left(t\right) = \mathbf{A}\mathbf{P}\left(t\right)\mathbf{A}^{H} + \phi_{\gamma}\left(t\right)\mathbf{\Gamma} + \phi_{v}\mathbf{I}, \\ &\qquad \mathbf{P}\left(t\right) = \operatorname{diag}\left[\phi_{1}\left(t\right),\cdots,\phi_{R}\left(t\right)\right], \\ &\qquad a_{1r} = 1, \phi_{r}\left(t\right) \geq 0, \phi_{\gamma}\left(t\right) \geq 0, \phi_{v} \geq 0, \\ &\qquad \text{for } t = 1,\cdots,N; r = 1,\cdots,R \ . \end{aligned} \tag{40}$$

Note that the signal model assumed here is the same as our proposed model in Eq. (8). According to [13], a local minimum for Eq. (40) can be found by iteratively reducing the cost function value. At each iteration, a non-linear constrained optimization problem needs to be solved to update the parameters. The number of required iterations is very large (e.g. in the order of 500) due to the non-convexity of the problem and the high dimension of the parameters. Therefore, the SCFA method has a relatively high computational cost.

B. Evaluation measures

1) Estimation accuracy: Since the main goal of this work is to find accurate estimates of the parameters of interest, we first introduce the estimation accuracy measures for the different parameters.

For the RTF matrix, to evaluate the alignment of the estimated RTF with the ground-truth RTF, we calculate the Hermitian angle by means of

$$E_{\mathbf{a}} = \frac{\sum_{\beta=1}^{B}\sum_{k=1}^{K/2+1}\sum_{r=1}^{R} \operatorname{acos}\left(\frac{\left|\mathbf{a}_r(\beta,k)^H \hat{\mathbf{a}}_r(\beta,k)\right|}{\|\mathbf{a}_r(\beta,k)\|_2 \|\hat{\mathbf{a}}_r(\beta,k)\|_2}\right)}{BR\left(K/2+1\right)} , \tag{41}$$

where the error has been averaged over different sources, frequency bins and the number of time segments B. Note that the ground-truth RTF is calculated using the first 32ms of the corresponding RIR. For the PSD of the r-th source ϕ_r and the PSD of the late reverberation ϕ_γ, we use the symmetric log-error distortion measure [40]

$$E_i = \frac{10\sum_{t,k\in\mathcal{Q}}\left|\log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right|}{|\mathcal{Q}|} , \tag{42}$$

for $i = r$ or γ, where the index set $\mathcal{Q}$ is used to discard zero PSDs, as used in [41] and $|\mathcal{Q}|$ is the cardinality of $\mathcal{Q}$. For the errors of the source PSDs, we use E_s to denote the average value of them, i.e., $E_s = \frac{\sum_{r=1}^{R} E_r}{R}$. Note that the error in Eq. (42) can be seen as the summation of the overestimation error and the underestimation error, which are

$$E_i^{\text{ov}} = \frac{10\sum_{t,k\in\mathcal{Q}}\left|\min\left\{0, \log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right\}\right|}{|\mathcal{Q}|} , \tag{43}$$

and

$$E_i^{\text{un}} = \frac{10\sum_{t,k\in\mathcal{Q}}\max\left\{0, \log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right\}}{|\mathcal{Q}|} , \tag{44}$$

respectively. The ground-truth PSDs of the r-th source and the late reverberation are calculated using

$$\phi_r\left(t\right) = \frac{1}{T}\sum_{l=1+(t-1)T}^{tT} \left|s_r\left(l\right)\right|^2. \tag{45}$$

and

$$\phi_\gamma\left(t\right) = \frac{1}{TM}\sum_{l=1+(t-1)T}^{tT} \mathbf{d}^H\left(l\right)\mathbf{d}(l). \tag{46}$$

In a noise reduction method, under or overestimates of target source PSDs or noise/interference PSDs have each its own effect. When the target source PSDs have large underestimation errors or the noise or interference PSD has large overestimation errors, the target source obtained by a noise reduction algorithm using these estimates typically has large distortions. On the other hand, if the estimate of the noise or interference PSD has a large underestimation error, the reconstructed signal often comes with musical noise [42]. Therefore, we will also present in detail the underestimation errors and overestimation errors in the experiments.

2) Predicted Quality and intelligibility: Since the estimated parameters are commonly used in noise reduction algorithms, we use the estimates in the well-known multi-channel Wiener filter (MWF) [43] and use the MWF outputs to reconstruct each point source signal. For estimating the r-th signal, the MWF can be expressed as a combination of a minimum variance distortionless response (MVDR) beamformer [44] and a single-channel Wiener filter, which is

$$\hat{\mathbf{w}}_r = \frac{\hat{\phi}_r}{\hat{\phi}_r + \hat{\mathbf{w}}_{r,\text{MVDR}}^H \hat{\mathbf{R}}_{r,nn} \hat{\mathbf{w}}_{r,\text{MVDR}}} \hat{\mathbf{w}}_{r,\text{MVDR}} , \tag{47}$$

where $\mathbf{w}_{r,\text{MVDR}}$ is the MVDR beamformer

$$\hat{\mathbf{w}}_{r,\text{MVDR}} = \frac{\hat{\mathbf{R}}_{r,nn}^{-1}\hat{\mathbf{a}}_r}{\hat{\mathbf{a}}_r^H \hat{\mathbf{R}}_{r,nn}^{-1}\hat{\mathbf{a}}_r} , \tag{48}$$

and

$$\hat{\mathbf{R}}_{r,nn} = \sum_{i=1, i\neq r}^{R} \hat{\phi}_i \hat{\mathbf{a}}_i \hat{\mathbf{a}}_i^H + \hat{\phi}_\gamma \hat{\mathbf{\Gamma}} + \hat{\phi}_v \mathbf{I} \ . \tag{49}$$

Note that the permutation ambiguity exists after estimating the RTF matrix and the sources PSDs (i.e., we cannot determine which column of $\mathbf{A}$ belongs to which source for different frequency bins). This problem is beyond the scope of this work and methods on this topic, to name a few, were investigated in [36], [45], [46]. In the experiments of this work, we use the oracle RTF matrix as guidance to permute the columns of the estimated RTF matrix per time-frequency tile.

The predicted speech quality of each reconstructed signal is evaluated by calculating the segmental-signal-to-noise-ratio (SSNR) [47] and the perceptual evaluation of speech quality (PESQ) measure [48]. The predicted speech intelligibility performance is evaluated by the speech intelligibility in bits (SIIB) measure [49], [50]. Alternately, we select one of the R sources as the target and the remaining $R-1$ sources as interferers. We than average all measures we used in the experiments over these R different setups.

C. Experiments with simulated RIRs

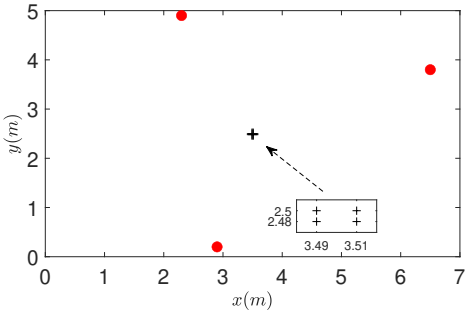
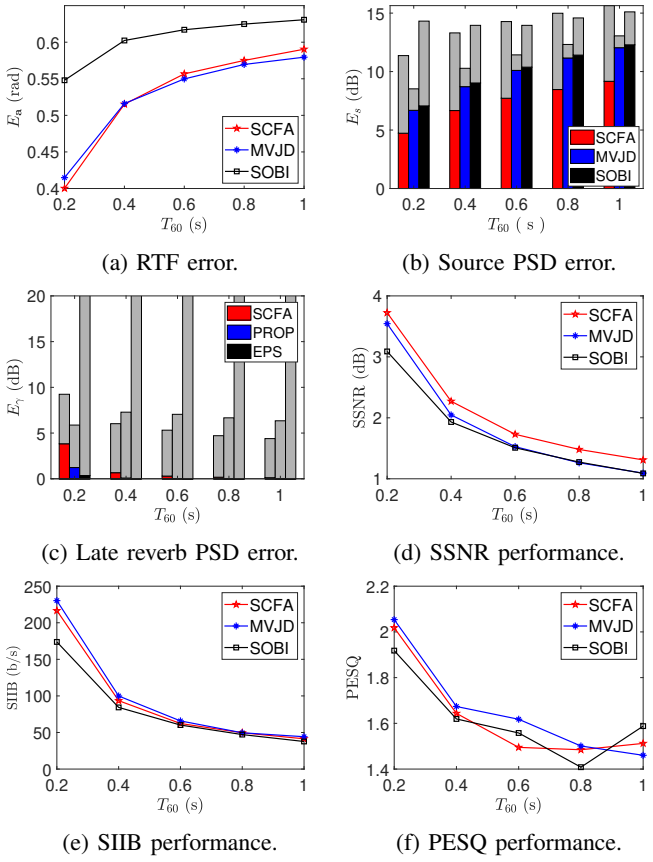


Figure 2: Top view of the acoustic scene with a zoom-in of microphones.

The acoustic scene of the first experiment is shown in Fig. 2, where four microphones and three sources are placed in the room with a dimension of $7 \times 5 \times 4$m. The speech signals are downloaded from the TIMIT database [51]. To simulate the reverberant signal recorded by each microphone, we convolve the speech signals (with a duration of 33 s) with the room impulse responses (RIRs) generated by the image source method [52]. The microphones are omnidirectional and the RIRs have a duration of 1 s. Then, to synthesize the noisy microphone signals, we add independently generated white Gaussian noise to each reverberant signal. The variance of the noise is fixed at a value calculated from given signal-to-noise ratios (SNRs). The SNR value is the ratio between the overall energy of the direct and early reflections of the first speech signal at position $(3.49, 2.5)$ and the energy of the noise component at the first microphone.

The microphone signals are sampled at a frequency of $f_s = 16$ kHz after which they are transformed to the frequency domain by the STFT procedure, in which the 50 % overlapping square-root Hann window with a length of 32 ms and the FFT length of 512 are used. Note that the window length is the same as the sub-time frame length and also equals the early part of the RIRs. Each time frame has $T = 20$ overlapping sub-time frames and thus a duration of 0.32 s. Note that this duration can be longer than the actual speech source stationary period and the PSDs can be seen as the averages of the PSDs over each time frame.



(a) RTF error. (b) Source PSD error.

(c) Late reverb PSD error. (d) SSNR performance.

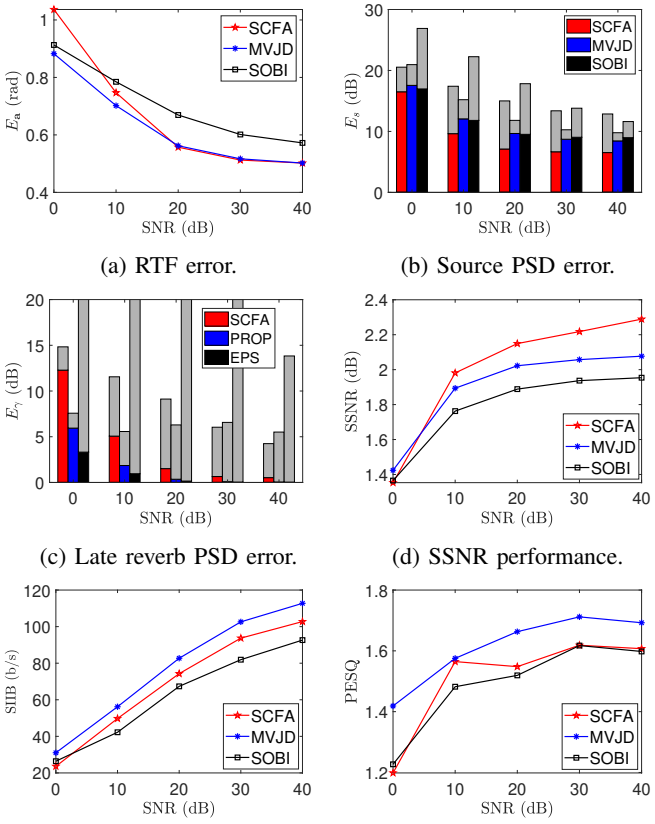
(e) SIIB performance. (f) PESQ performance.

Figure 3: Performance vs the late reverberation time. In Figs. b and c, the top gray bars indicate the underestimation errors, and the bottom colored bars indicate overestimation errors.

1) Performance comparison: In Fig. 3, we present the performance comparison among our proposed method and the two reference methods, where we adjust the reverberation time from 0.2 s to 1 s. The number of time frames per segment is 8 and the SNR is fixed at 30 dB. We first show the RTF estimation error calculated by Eq. (41) in Fig. 3a. The error for each method increases as the room becomes more reverberant. Our proposed MVJD method has similar performance compared to SCFA, which both outperform the SOBI method. In Figs. 3b and 3c, we show the PSDs estimation error calculated by Eq. (42). For each bar (each overall error), we also show the overestimation error using the bottom colored bar and the underestimation error using the top gray bar. In Fig. 3b, we show the source PSD estimation errors, where the errors also become larger when the reverberation time increases. Our proposed method has the smallest error compared to SOBI and a slightly larger overestimation error compared to SCFA. In particular, the underestimation error (gray bar) of our proposed

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3533374

method outperforms the other two methods. In Fig. 3c, the late reverberation PSD errors are presented. For visibility, parts of the bars over 20 dB are not shown. Note that we use our proposed late reverberation estimator in both SOBI and our proposed MVJD. The 'EPS' method in Fig. 3c refers to replacing the negative estimates from Eq. (12) with ϵ, the machine precision, as used in [12]. Our proposed estimator has similar errors compared to SCFA, both of which are much smaller than EPS. Note that the overestimation errors of our method are smaller than SCFA. In Figs. 3d to 3f, it is shown that our proposed method and SCFA outperform SOBI in general regarding to the predicted speech quality and speech intelligibility evaluated by SSNR, PESQ and SIIB. Note that our proposed method has better predicted intelligibility and predicted quality in terms of PESQ but a worse SSNR than SCFA.



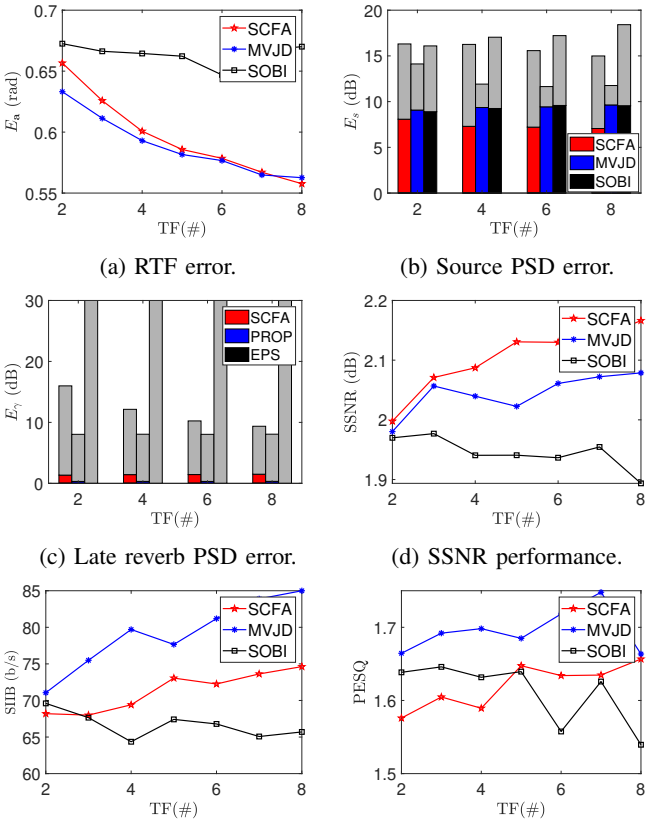
(a) RTF error. (b) Source PSD error.

(c) Late reverb PSD error. (d) SSNR performance.

(e) SIIB performance. (f) PESQ performance.

Figure 4: Performance vs SNR. In Figs. b and c, the top gray bars indicate the underestimation errors, the bottom colored bars indicate overestimation errors.

In Fig. 4, we compare all the methods while changing the noise level by increasing the SNR from 0 dB to 40 dB. The number of time frames per segment is again eight and the reverberation time is fixed at 0.4 s. The RTF estimation error is first shown in Fig. 4a. For all the methods, the RTF error is relatively small for high SNR. Our proposed MVJD method has the best performance, which outperforms SCFA at low SNR values and outperforms SOBI at high SNR values. In Fig. 4b, the source PSD estimation errors also reduces when the SNR increases. Our proposed method has the smallest underestimation errors, while the SCFA method has the smallest overestimation errors. In Fig. 4c, the late reverberation PSD errors are compared, where our proposed late reverberation estimator and SCFA have much smaller errors compared to using the ϵ procedure. For visibility, parts of the bars over 20 dB are again not shown. Note that the overestimation errors of our method are smaller than SCFA, both which decreases when the SNR increases. In Figs. 4d to 4f, it is also shown that our proposed method and SCFA outperform SOBI in general regarding to the predicted speech quality and speech intelligibility.



(a) RTF error. (b) Source PSD error.

(c) Late reverb PSD error. (d) SSNR performance.

(e) SIIB performance. (f) PESQ performance.

Figure 5: Performance vs the number of time frames per segment. In Figs. b and c, the top gray bars indicate the underestimation errors, the bottom colored bars indicate overestimation errors.

Fig. 5 shows the performance comparison for different time segment durations (i.e., different numbers of time frames per segment). For visibility, parts of the bars over 20 dB in Fig. 5c are not shown. Our proposed method and the SCFA method still outperform the SOBI method in estimation errors, speech quality and speech intelligibility.

TABLE I: Estimation errors using different steps

	E_a	E_s	E_γ	E_s^{un}	E_s^{ov}	E_γ^{un}	E_γ^{ov}
Step 1	0.52	10.19	43.50	1.31	8.89	43.48	0.02
Step 1+2	0.52	10.62	7.01	1.99	8.62	6.87	0.14
Step 1+2+3	0.52	10.32	7.40	1.60	8.72	7.33	0.07

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3533374

To evaluate the impact of the three robustification steps we proposed for the late reverberation estimator, we compared the estimation errors using different steps, where we fixed reverberation time at 0.4 s, SNR at 30 dB and the number of time frames at 8. We can see from the table that by adding step 2 after step 1, the late reverberation PSD error is reduced a lot. Adding step 3 after step 2 shows slight reduction on the error of the source PSDs. The reason is that without step 3, the covariance matrices for the sources might not have R positive eigenvalues, which will likely lead to negative source PSD estimates that will be replaced by small positive value like eps, resulting in a huge underestimation error of the source PSDs for that frequency bin. However, the overall improvement is not big as step 3 is only executed for some time-frequency bins. The average iteration number of the frequency bins executing step 3 in this experiment is 0.05 (with the total number of frequency bins 257).

D. Experiments with recorded RIRs

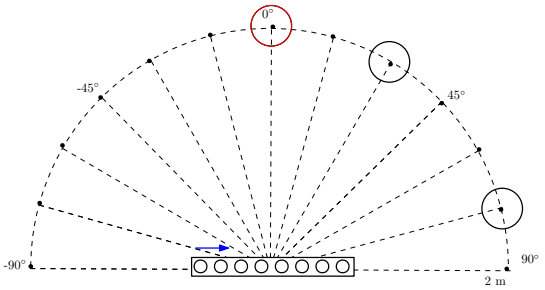
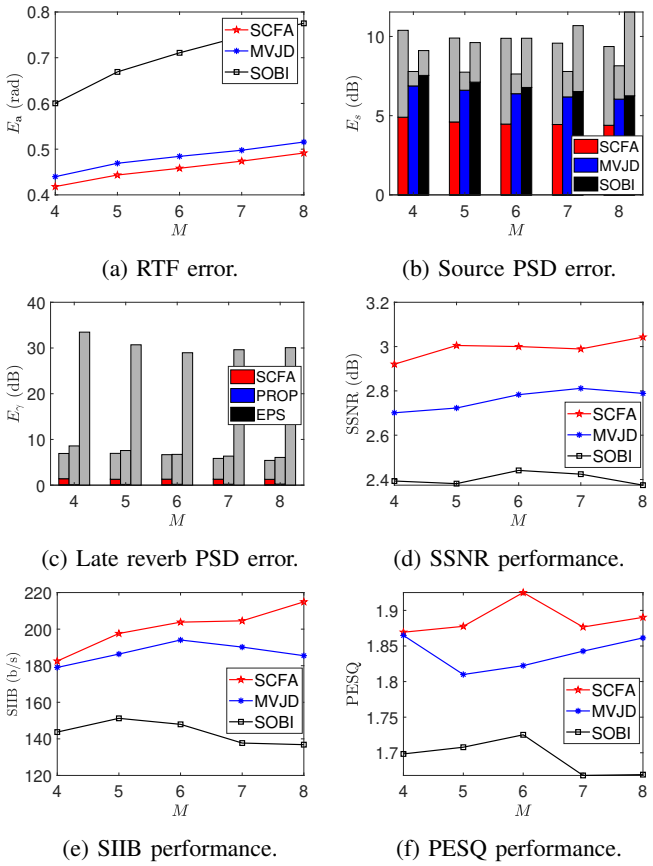


Figure 6: Geometric setup of the acoustic scene [53] with big red circles representing the positions of sources. From left to right, as shown by the blue arrow, the first M microphones are used with M changing from 4 to 8.

1) Setup: In this section, we use RIRs recorded in a real room with dimension $6 \times 6 \times 2.4$ m [53]. We consider two scenarios in this experiment. For the first scenario with three sources, the geometric positions of the sources and the microphones are shown in Fig. 6. The microphones form a uniform linear array with 8 cm interdistance. The data base in [53] contains RIRs measured at a 2 m distance from the microphone array center at different angles. We convolve the RIRs at 0°, 30° and 75° with different speech signals. We also add white Gaussian noise to simulate the microphone self-noise. For the second scenario with two sources, where one source is fixed at an angle of $-15°$ and the other source is placed at different angles ranging from 0° to 90° in steps of 15°. We use the same STFT procedure as we used in the first experiment to transfer the time domain signals to the frequency domain.

2) Performance comparison: In Fig. 7, we show the performance comparison for three sources for all methods as a function of the number of microphones. The SNR is 30 dB and the reverberation time is 0.36 s. Note that when using a larger number of microphones, the theoretical spatial coherence matrix calculated by Eq. (7) can be close to singular, particularly for low frequency bins as observed in our



(a) RTF error. (b) Source PSD error.

(c) Late reverb PSD error. (d) SSNR performance.

(e) SIIB performance. (f) PESQ performance.

Figure 7: Performance vs the number of microphones. In Figs. b and c, the top gray bars indicate the underestimation errors, the bottom colored bars indicate overestimation errors.

experiments. To avoid numerical issues, we regularize such matrices by $\mathbf{\Gamma} = \mathbf{\Gamma} + \mu\mathbf{I}$ with $\mu = 10^{-3}$ in this experiment. In terms of RTF errors in Fig. 7a, MVJD and SCFA show similar performance and both outperform SOBI. In terms of the source PSD errors in Fig. 7b, SCFA has a lower underestimation error than MVJD, but MVJD has lower overestimation errors than SCFA. For the late reverberation PSD errors, 'EPS' still shows the worst performance. In terms of the predicted speech quality and intelligibility performance as shown in Figs. 7d to 7f, our proposed method has performances close to the SCFA method, while both outperform the SOBI method.

For estimators of the late reverberation PSD, we extended another state-of-the-art method [54] from single-source to multi-source as an additional reference method. We estimate the PSDs of the sources and the late reverberation using the RTF matrix estimated by our proposed MVJD method. We minimize the following cost function:

$$\left\| \hat{\mathbf{P}}_{\mathbf{y}} - \left(\hat{\mathbf{A}}\mathbf{P}\hat{\mathbf{A}}^H + \phi_\gamma \mathbf{\Gamma} + \hat{\phi}_v \mathbf{I} \right) \right\|^2. \tag{50}$$

The solution of the PSDs $\mathbf{P} = \mathrm{diag}\,[\phi_1, \cdots, \phi_R]$ and ϕ_γ is

$$\hat{\phi} = \mathbf{\Phi}^{-1}\mathbf{b} \tag{51}$$

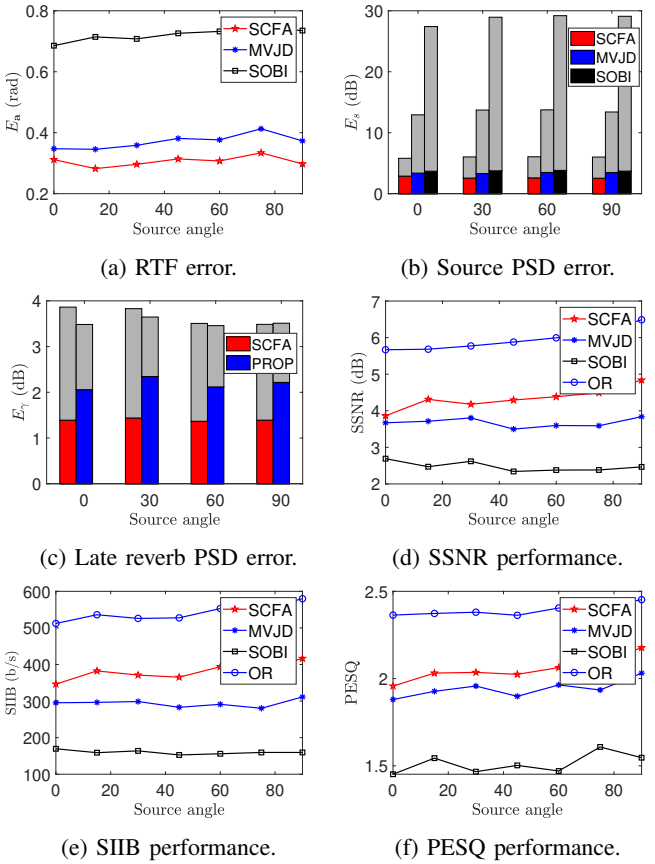
with

$$\mathbf{\Phi} = \begin{bmatrix} \left(\mathbf{a}_1^H \mathbf{a}_1\right)^2 & \cdots & \left|\mathbf{a}_1^H \mathbf{a}_R\right|^2 & \mathbf{a}_1^H \mathbf{\Gamma} \mathbf{a}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \left|\mathbf{a}_1^H \mathbf{a}_R\right|^2 & \cdots & \left(\mathbf{a}_R^H \mathbf{a}_R\right)^2 & \mathbf{a}_R^H \mathbf{\Gamma} \mathbf{a}_R \\ \mathbf{a}_1^H \mathbf{\Gamma} \mathbf{a}_1 & \cdots & \mathbf{a}_R^H \mathbf{\Gamma} \mathbf{a}_R & \text{trace}\left\{\mathbf{\Gamma}^H \mathbf{\Gamma}\right\} \end{bmatrix} \quad (52)$$

and

$$\mathbf{b} = \begin{bmatrix} \mathbf{a}_1^H \left(\hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_v \mathbf{I}\right) \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_R^H \left(\hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_v \mathbf{I}\right) \mathbf{a}_R \\ \text{trace}\left\{\mathbf{\Gamma}^H \left(\hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_v \mathbf{I}\right)\right\} \end{bmatrix}. \quad (53)$$

The last element of $\hat{\phi}$ is the estimated late reverberation PSD. For the case of using 8 microphones in Fig. 7, the least squares-based estimator has an error of 11.38 (overestimation error 1 + underestimation error 10.38), which is larger than our proposed estimator with an error of 6.39 (overestimation error 0.56 + underestimation error 5.83).



(a) RTF error. (b) Source PSD error.

(c) Late reverb PSD error. (d) SSNR performance.

(e) SIIB performance. (f) PESQ performance.

Figure 8: Performance vs source position. In Figs. b and c, the top gray bars indicate the underestimation errors, and the bottom colored bars indicate overestimation errors.

In Fig. 8, we show the performance comparison for two sources for all methods for different positions of the second source ranging from 0° to 90° by every 15°. It is shown that the performances, on both estimation errors and predicted speech quality and intelligibility, do not change much with different positions of the second source. Note that the 'EPS' method for late reverberation estimation has been left out in Figure 8c for better visibility of the other methods. MVJD shows comparable performance with SCFA, which both greatly outperform SOBI. We also show the predicted speech quality and intelligibility performance when using oracle parameters to calculate the MWF, which is referred to as 'OR' in Fig. 8.

In previous experiments, the maximum number R of sources per time segment and frequency is assumed known. In practice, it needs to be estimated using methods such as [27], [28]. The estimated $\hat{R}$ can be smaller, equal or larger than R. To evaluate this problem, we show in Table II the predicted speech quality and intelligibility performance of our proposed method for $\hat{R} - R$ being -1, 0 and 1. We considered two sources placed at 0° and 60°. It is shown that our proposed method with overestimated $\hat{R} = R + 1$ is similar to the case of $\hat{R} = R$. However, the performance with underestimated $\hat{R} = R - 1$ is much worse than the other cases.

TABLE II: Predicted speech quality and intelligibility comparison.

$\hat{R} - R$	-1	0	1
SSNR	0.92	3.93	3.80
SIIB	163.92	316.60	318.26
PESQ	1.00	2.04	2.01

Finally, we show the computation time using MATLAB for processing the microphone signals with a duration of 33 s using different methods and different number of microphones in Table III. We can see that the SCFA method needs the longest run time, which increases as the number of microphones increases. Our proposed method and the SOBI method have a similar run time, which is in the order of 700 times faster than SCFA. For our proposed method, The computation time of the late reverberation PSD estimator mainly comes from iterative steps with EVD. The average iteration number for each frequency bin is less than 1 as observed in our experiments. In practice, it depends on the accuracy of given noise PSDs. The computation cost of the RTF matrix and source PSDs estimator mainly comes from the joint diagonalization algorithm, which is in the order of R^3. Based on the analysis, if the overall iteration number for the late reverberation PSD estimator is I, the computational complexity of the proposed algorithm is in the order of $IM^3 + R^3$.

TABLE III: Computation time (in s) comparison.

M	4	5	6	7	8
SCFA	3523	3708	4362	5169	5768
MVJD	9	10	11	8	8
SOBI	8	11	10	9	9

E. Experiments with real recordings

1) Setup: In this section, we used signals recorded by four microphones mounted on a dummy head in the BRUDEX Database [55], i.e., including natural reverberation. We considered two sources, speaker 1 at 0° and speaker 2 at 60° with medium reverberant condition in [55]. The sampling

frequency is 48000 Hz in this experiment and the FFT length is 2048. The other settings for the STFT procedure are the same as the previous two experiments. Note that besides the real recordings, the RIRs were also measured in [55], with which we can simulate source components such as the late reverberation. For the spatial coherence matrix of the late reverberation, we calculate it using the simulated late reverberation component by

$$\mathbf{\Gamma}_{i,j}(k) = \frac{\sum_l d_i(l,k)\, d_j(l,k)^*}{\sqrt{\sum_l |d_i(l,k)|^2}\sqrt{\sum_l |d_j(l,k)|^2}}. \quad (54)$$

For the noise component, we assume a spatially white (spectrally non-white) model and use the first second recordings (speech absent duration) to measure the noise PSD for each frequency bin. Note that in this experiment, we added another reference method, ARMA-FastMNMF [56] as a comparison to a state-of-the-art speech enhancement method. For ARMA-FastMNMF, we used the following parameters: number of speech: 2, speech model: NMF, number of noise: 0, tap length of the MA model $L_{MA} = 8$, tap length of the AR model $L_{AR} = 4$, delay of the late reverberation $\Delta = 1$ and the Iterative Source Steering (ISS) algorithm was used. Note that all methods were run in a device with Intel(R) Core(TM) i7-10610U CPU @ 1.80GHz 2.30 GHz without using GPU. Notice that ARMA-FastMNMF does not estimate the underlying parametric model (as the proposed method and SCFA), but directly performs the source separation.

2) Performance comparison: In Fig. 9, we evaluate the predicted speech quality and intelligibility performance of all methods. As shown in the figures, our proposed method outperforms SOBI in all measures and outperforms ARMA-FastMNMF in PESQ and SIIB. We also show the computation time normalized by the time it takes for MVJD in Table IV. We can see that although SCFA has the best performance in this experiment, its computation time is again very high compared to MVJD. Also, MVJD is about 150 times faster than the ARMA-FastMNMF method.

TABLE IV: Computation time comparison.

Methods	SCFA	MVJD	SOBI	ARMA-FastMNMF
Normalized run time	832.56	1	0.89	154.41

V. Concluding Remarks

In this paper we considered the complex scenario where multiple sources, late reverberation and noise exist concurrently. For this scenario, we proposed a joint estimator of the parameters include the RTFs of the sources and the PSDs of the sources and the late reverberation. We first proposed a late reverberation PSD estimator that does not require the knowledge of the RTFs. Then we proposed the minimum variance joint diagonalization (MVJD) method to estimate the RTFs and the PSDs of the sources. The proposed MVJD method is more robust than the existing joint-diagonalization SOBI method, since we considered an optimal linear combination of a set of covariance matrices instead of only the first one as done with SOBI. The optimality is obtained by minimizing the variances of the error matrix of the linearly combined sample covariance matrices. Experiments demonstrated that our proposed method outperforms the SOBI method in terms of estimation errors, the predicted speech quality and the speech intelligibility. The results also show that our proposed method achieves similar performance compared to the state-of-the-art SCFA method but has a significantly lower computational complexity.

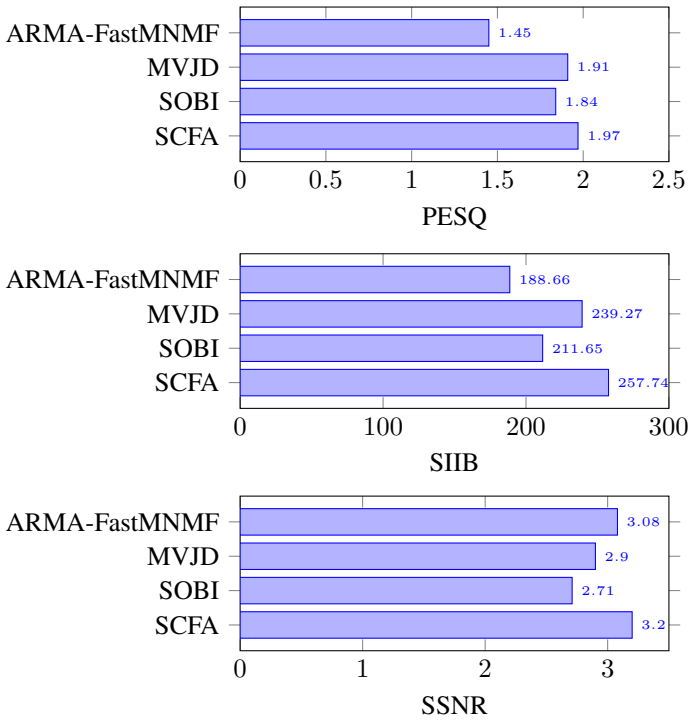


Figure 9: Predicted speech quality and intelligibility performance comparison.

References

[1] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994.

[2] J. Xia, B. Xu, S. Pentony, J. Xu, and J. Swaminathan, "Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 1523–1533, Mar. 2018.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[4] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement*. Springer Nature, 2022.

[5] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Process.*, vol. 87, no. 8, pp. 1933–1950, 2007.

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

[7] 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, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[8] 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.*, 2015, pp. 544–548.

[9] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[10] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[11] ——, "Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[12] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

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

[14] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1507–1519, 2019.

[15] Y. Laufer and S. Gannot, "Scoring-Based ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in a Spatially Homogeneous Noise Field," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 61–76, 2020.

[16] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square Root-Based Multi-Source Early PSD Estimation and Recursive RETF Update in Reverberant Environments by Means of the Orthogonal Procrustes Problem," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 755–769, 2020.

[17] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[18] C. Li, J. Martinez, and R. C. Hendriks, "Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 695–705, 2023.

[19] C. Li and R. C. Hendriks, "Alternating Least-Squares-Based Microphone Array Parameter Estimation for A Single-Source Reverberant and Noisy Acoustic Scenario," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2023.

[20] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[21] D. Cherkassky and S. Gannot, "Successive Relative Transfer Function Identification Using Blind Oblique Projection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 474–486, 2020.

[22] H. Gode and S. Doclo, "Covariance Blocking and Whitening Method for Successive Relative Transfer Function Vector Estimation in Multi-Speaker Scenarios," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, 2023, pp. 1–5.

[23] Y. Laufer, B. Laufer-Goldshtein, and S. Gannot, "ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in Rank-Deficient Noise Field," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 619–634, 2020.

[24] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

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

[26] C. Li, J. Martinez, and R. C. Hendriks, "Low Complex Accurate Multi-Source RTF Estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4953–4957.

[27] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6458–6473, 2018.

[28] H. Sun, P. Samarasinghe, and T. Abhayapala, "Blind source counting and separation with relative harmonic coefficients," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[29] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[30] C. Li and R. C. Hendriks, "Adaptive time segmentation for improved signal model parameter estimation for a single-source scenario," to appear in Proc. IEEE Asilomar Conf. Signals, Syst., Comput., 2023.

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

[32] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[33] 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, Language Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[34] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[35] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, 1962.

[36] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping Separated Frequency Components by Estimating Propagation Model Parameters in Frequency-Domain Blind Source Separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1592–1604, 2007.

[37] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Mat. Anal. Appl.*, vol. 17, no. 1, pp. 161–164, Jan. 1996.

[38] N. R. Goodman, "Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution (An Introduction)," *Ann. Math. Stat.*, vol. 34, no. 1, pp. 152–177, 1963. [Online]. Available: http://www.jstor.org/stable/2991290

[39] D. Maiwald and D. Kraus, "On moments of complex Wishart and complex inverse Wishart distributed matrices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 5, 1997, pp. 3817–3820 vol.5.

[40] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement," in *Proc. Interspeech*, 2007, pp. 830–833.

[41] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[42] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[43] H. L. V. Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., Mar. 2002.

[44] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2013.

[45] 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. Adv. Signal. Process.*, vol. 2006, pp. 1–13, 2006.

[46] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193–1207, 2009.

[47] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[48] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[49] 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, 2017.

[50] ——, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[51] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[52] 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, 1979.

[53] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, Sept. 2014.

[54] O. Schwartz, S. Gannot, and E. A. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm," in *Proc. EURASIP Eur. Signal Process. Conf.*, 2016, pp. 1123–1127.

[55] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX Database: Binaural Room Impulse Responses with Uniformly Distributed External Microphones," in *Speech Commun.; 15th ITG Conference*, 2023, pp. 126–130.

[56] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, "Autoregressive moving average jointly-diagonalizable spatial covariance analysis for joint source separation and dereverberation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2368–2382, 2022.