Alternating Least-Squares-Based Microphone Array Parameter Estimation for A Single-Source Reverberant and Noisy Acoustic Scenario

Li, Changheng; Hendriks, Richard C.

DOI
10.1109/TASLP.2023.3306713

Publication date
2023

Document Version
Final published version

Published in
IEEE Trans. Audio, speech and Lang. Proc.

Alternating Least-Squares-Based Microphone Array Parameter Estimation for a Single-Source Reverberant and Noisy Acoustic Scenario

Changheng Li, *Graduate Student Member, IEEE*, and Richard C. Hendriks

***Abstract*—Acoustic-scene-related parameters such as relative transfer functions (RTFs) and power spectral densities (PSDs) of the target source, late reverberation and ambient noise are essential for microphone array signal processing and are challenging to estimate. Existing methods typically only estimate a subset of the parameters by assuming the other parameters are known. This can lead to unmatched scenarios and reduced estimation performance on the parameters of interest. Moreover, many methods process time frames independently, despite they share common information such as the same RTF. In this work, we consider a noisy scenario by modelling the noise component as a spatially homogeneous sound field with a time-invariant spatial coherence matrix and time-varying PSD. We first modify an existing alternating least squares (ALS) method to obtain more accurate estimates using a single time frame. Then, we extend the method to use multiple time frames that share the same RTF. Furthermore, we propose more robust constraints on the PSDs to avoid large estimation errors. We compare our proposed methods to the state-of-the-art simultaneously confirmatory factor analysis (SCFA) method, a joint maximum likelihood estimation (JMLE) method and an existing ALS-based method. The experimental results in terms of estimation accuracy, noise reduction performance, predicted speech quality, and predicted speech intelligibility demonstrate that our proposed methods achieve similar performance compared to the state-of-the-art SCFA method, which outperforms the existing ALS method in all scenarios and outperforms the JMLE method particularly in low SNR scenarios. Moreover, our proposed methods have significantly lower computational complexity than SCFA.**

***Index Terms*—Dereverberation, noise reduction, microphone array signal processing, RTF estimation, PSD estimation.**

I. Introduction

HANDS-FREE speech communication applications like mobile phones and hearing aids are commonly used nowadays. Equipped with microphone arrays, these devices can record and analyze the speech signal for various applications. Unavoidably, the microphone signals are corrupted by reverberation and ambient noise, which can degrade the speech quality and intelligibility [1], [2]. Hence, techniques like spatial filtering are used to extract the target signal from the noisy microphone signals. Typically, these spatial filters depend on acoustic-scene-related parameters such as relative transfer functions (RTFs) and power spectral densities (PSDs) of the source, the late reverberation and the ambient noise. In practice, these parameters are typically unknown. Therefore, an essential problem with hands-free speech communication applications is to estimate the aforementioned parameters. Note that there are non-parametric techniques such as blind beamforming or blind source separation [3], [4] that can extract the target signal without estimating the parameters. However, in this work we only focus on parametric beamformers where the estimated parameters can be used as a prior information on the acoustic scene.

Due to the non-stationarity of the speech signal, the PSDs of the target source and the late reverberation are time-varying. The PSDs of the ambient noise can be time-varying as well, depending on the working environment of the microphone arrays. The RTFs can change over time as well depending on whether the source is moving relative to the array. The facts that these parameters can be time-varying and corruptions caused by reverberation and ambient noise are present, make the estimation of these parameters rather challenging.

In recent years, many methods have been proposed to estimate these parameters, see e.g., [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. Many of these methods only estimate a subset of the parameters by making some strict assumptions about the acoustic scenarios and the knowledge of the remaining parameters. For example, in [5], [9], [12], the RTFs of the target source are assumed to be known such that the speech PSD, late reverberation PSD and noise PSD can be estimated. In [6], the PSD of the late reverberation is assumed to be known and the RTF of the target source is estimated. In [7], the RTFs and the PSDs of all sources and the noise covariance matrix are estimated. However, it is assumed that the late reverberation component is stationary and only a single source is active per time frequency tile. In [8], the noise covariance matrix is assumed known and the late reverberation PSD is estimated. In [13], [14], the noiseless scenario is assumed, neglecting the estimation of the ambient noise PSD.

Manuscript received 12 April 2023; revised 13 July 2023; accepted 6 August 2023. Date of publication 24 August 2023; date of current version 20 October 2023. This work was 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. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ito. *(Corresponding author: Changheng Li.)*

The authors are with the Faculty of Electrical Engineering, Mathematics and Computer Science, 2628CD Delft, The Netherlands (e-mail: l.li-7@tudelft.nl; r.c.hendriks@tudelft.nl).

Digital Object Identifier 10.1109/TASLP.2023.3306713

From the above overview, we see that existing methods for parameter estimation from the acoustic scene all assume a subset of parameters to be known. However, erroneously assuming a subset of the parameters to be known can lead to unmatched scenarios, and thus to reduced noise reduction performance. This emphasizes the importance of accurate joint parameter estimation. A second important point is the fact that, apart from a few exceptions, e.g., [11], [14], many of these methods process the time frames independently, despite the fact that they may share some common information. For instance, the RTFs corresponding to some adjacent time frames are the same if the sound source is static during these time frames. In such cases, we could use these time frames jointly to obtain better estimates of the RTFs [11], [14].

The joint estimation of parameters using multiple time frames is realized in [11] in a reverberant and noisy environment, using the simultaneous confirmatory factor analysis (SCFA) method. As expected, SCFA has much better estimation performance compared to methods using each time frame independently, especially for the RTF estimation [11]. Nevertheless, SCFA has a rather high computational cost. Therefore, we recently proposed some alternative methods that can achieve a nearly similar performance as SCFA, but at a much lower complexity [14].

In [14], we considered a single reverberant source scenario and proposed a joint maximum likelihood estimator (JMLE) for the parameters of interest. In the current work, we extend the signal model from [14] to the noisy case. Specifically, we model the noise component as a spatially homogeneous sound field characterized by a time-invariant spatial coherence matrix with a time-varying PSD. We can assume the spatial coherence matrix is known, as assumed in [9]. Further, we consider the use of multiple time frames to jointly form a segment. The RTF is considered constant across the segment, while the PSDs of the target's early reflections, the PSDs of the late reverberation and the ambient noise PSD are allowed to change from frame-to-frame. The focus herein is to jointly estimate the source's RTF, and the PSDs of the early reflections, the late reverberation and the ambient noise at low complexity. We will use the least squares (LS) error as a cost function, i.e., minimizing the Frobenious norm of model error matrices. Note that the LS cost function has been considered in [10] as well to estimate these parameters and the LS minimization was solved by an alternating least squares (ALS) method. However, we will show in this work that the ALS based method from [10] can suffer from a parameter identifiability issue and thus needs to be modified to obtain more accurate estimates. Note also that the ALS method from [10] uses each time frame separately. Hence, we will extend the modified ALS method such that it uses multiple time frames jointly to improve the estimation performance. In addition, we propose constraints on the estimated PSDs that are more robust than the ones used in [10] to avoid large estimation errors. Note that minimizing the least squares cost function for multiple time frames jointly can be seen as a special case of the joint diagonalization problems modeled in [15], [16], [17], except that the problem proposed in our work has additional constraints on some of the parameters and the single target source is disturbed by both the late reverberation and the ambient noise.

The remaining parts of the article are structured as follows. In Section II, we introduce the notation used in this article, present the signal model and formulate the problem discussed in this article. In Section III, we will present the existing ALS method, propose a modified ALS method and extend it to a method using multiple time frames. After that, we will compare our proposed methods to some state-of-the-art reference methods in various simulated acoustic experiments in Section IV. Finally, we will draw the conclusions in Section V.

II. Preliminaries

A. Notation

In this article, we use lower-case letters to denote scalars, bold-face lower-case letters for vectors and bold-face upper-case letters for matrices. Matrix notation with subscripts using two lower-case letters (e.g. $\mathbf{P}_{\mathbf{y}_{i,j}}$) denotes the element of the matrix. Matrix notation with superscripts $T, *, H$ denotes taking the transpose, the conjugate and the conjugate transpose of the matrix, respectively. $\Re(x)$ and $\Im(x)$ represent the real part and the imaginary part of a complex-valued variable x, respectively. Further, $\mathrm{E}[\cdot]$ refers to the expectation operator, $\mathrm{tr}(\cdot)$ refers to taking the trace of a matrix, and if not further specified, $|\cdot|$ denotes taking the determinant of a matrix. Finally, $\mathrm{diag}\,[a_1, \ldots, a_M]$ denotes a diagonal matrix with diagonal elements $a_1, \ldots, a_M$ and $\|\cdot\|_F$ denotes taking the Frobenius norm of a matrix.

B. Signal Model

We consider a reverberant and noisy environment, in which a single acoustic point source is recorded by an array of M microphones with an arbitrary geometric structure. The microphone signal received at the m_{th} microphone in the short-time Fourier transform (STFT) domain is given by

$$y_m(l,k) = e_m(l,k) + r_m(l,k) + v_m(l,k), \tag{1}$$

where l is the time-frame index and k is the frequency bin index, $e_m(l,k)$ is the sum of the direct sound and the early reflections, $r_m(l,k)$ is the sum of all the late reflections in time frame l and frequency bin k, and $v_m(l,k)$ contains the ambient noise and microphone self-noise. Since the direct components and early reflections are beneficial for speech intelligibility [18], the combination of these components forms our target signal,

$$e_m(l,k) = a_m(l,k)s(l,k), \tag{2}$$

where $s(l,k)$ contains the direct and early speech component recorded by the reference microphone and $a_m(l,k)$ is the relative transfer function (RTF) between the reference microphone and the m_{th} microphone. By selecting the first microphone as the reference microphone, we have the prior information that $a_1 = 1$. Note that, we use the multiplicative transfer function (MTF) approximation in (2) for ease of analyzing, instead of the convolutive transfer function (CTF) approximation [19], [20]. Stacking the M microphone STFT coefficients into a column vector, we have

$$\mathbf{y}(l,k) = \mathbf{a}(l,k)s(l,k) + \mathbf{r}(l,k) + \mathbf{v}(l,k) \in \mathbb{C}^{M\times 1}, \tag{3}$$

where $\mathbf{y}(l,k) = [y_1(l,k), \ldots, y_M(l,k)]^T$ and the other vectors are defined in the same way.

C. Cross Power Spectral Density Matrices

By processing in short time frames, we can assume the three components in (3) to be stationary and mutually uncorrelated within a time frame. The PSD matrix of the noisy microphone recordings can therefore be expressed as

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

where $\mathbf{P}_{\mathbf{e}}$ is given by

$$\mathbf{P}_{\mathbf{e}}(l,k) = \phi_s(l,k)\mathbf{a}(l,k)\mathbf{a}^H(l,k), \quad (5)$$

and $\phi_s(l,k) = \mathrm{E}[|s(l,k)|^2]$ is the PSD of the target source at the reference microphone with $|\cdot|$ taking the absolute value. However, notice that across frames, s and r might be correlated.

The CPSD matrix of the late reverberation component is commonly modelled as [5], [21]

$$\mathbf{P}_{\mathbf{r}}(l,k) = \phi_\gamma(l,k)\mathbf{\Gamma}(k), \quad (6)$$

which is a spatially homogeneous and isotropic sound field with a time varying PSD $\phi_\gamma(l,k)$. The spatial coherence matrix $\mathbf{\Gamma}(k)$ is time-invariant. Hence, $\mathbf{\Gamma}(k)$ can be estimated in advance using the information on the microphone array geometry [22], [23], [24]. We assume a spherically isotropic noise field [25] and model the $\{i,j\}$-th element of $\mathbf{\Gamma}(k)$ as

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

where $\mathrm{sinc}(x) = \frac{\sin x}{x}$, $d_{i,j}$ is the inter-distance between microphones i and j, f_s is the sampling frequency, c denotes the speed of sound and K is the number of frequency bins.

Lastly, we assume that the residual noise component has a similar CPSD matrix formulation as the late reverberation, i.e.,

$$\mathbf{P}_{\mathbf{v}}(l,k) = \phi_v(l,k)\mathbf{\Psi}(k), \quad (8)$$

where $\mathbf{\Psi}(k)$ is the known spatial coherence matrix and $\phi_v(l,k)$ is unknown PSD. We assume that $\mathbf{\Psi}(k)$ is non-singular and linearly independent with $\mathbf{\Gamma}(k)$ (i.e. $\mathbf{\Psi}(k)$ is not a scaled version of $\mathbf{\Gamma}(k)$). Note that when considering the microphone self noise only, we have $\mathbf{\Psi}(k) = \mathbf{I}$.

D. Problem Formulation

Based on the assumptions made in the previous subsection and (5), (6) and (8), we can rewrite the noisy CPSD matrix for each time frame l as

$$\mathbf{P}_{\mathbf{y}}(l) = \phi_s(l)\mathbf{a}(l)\mathbf{a}^H(l) + \phi_\gamma(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}. \quad (9)$$

Note that we omit the frequency bin index k in (9) and hereafter for legibility since the signals will be processed for each k independently. By making the RTF vector $\mathbf{a}$ dependent on the time-frame index l, we implicitly assume that the relative source position or room acoustics can change from time frame to time frame. However, we consider in this work a semi-static source scenario by assuming the RTF $\mathbf{a}$ does not change for N (a finite number) time frames (N ranges from 1 to 8 in our experiments, corresponding to a duration of approximately 0.5 s to 5 s). We denote the set of N time frames sharing a single RTF by a time segment with index β. The noisy CPSD matrix then becomes

$$\mathbf{P}_{\mathbf{y}}(l) = \phi_s(l)\mathbf{a}(\beta)\mathbf{a}^H(\beta) + \phi_\gamma(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}, \quad (10)$$

with $\beta = \lfloor \frac{l-1}{N} \rfloor + 1$.

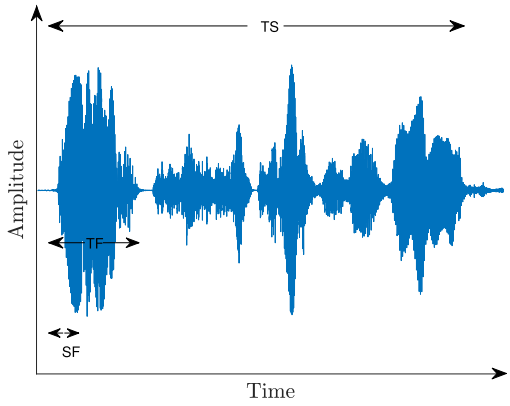


Fig. 1. Visualisation of the definition of time segment (TS), time frames (TF) and sub frames (SF).

Further, we define sub frames indexed by t_s, where T_{sf} overlapping sub frames form a time frame. See Fig. 1 for a visual interpretation of time segment, time frame and sub frame. Since the noisy signal is assumed to be stationary within a time frame, we can estimate the CPSD matrix per time frame i based on a sampled covariance matrix using the sub-time frames, that is,

$$\hat{\mathbf{P}}_{\mathbf{y}}(l) = \frac{1}{T_{sf}} \sum_{t_s=1+(l-1)T_{sf}}^{lT_{sf}} \mathbf{y}(t_s)\,\mathbf{y}(t_s)^H, \quad (11)$$

where $\mathbf{y}(t_s)$ denotes the STFT coefficients vector.

Accurate estimation of the parameters from the signal model in (10) is very important for speech enhancement and intelligibility improvement algorithms. However, this is also very challenging when the source is only stationary for a short time and microphone and source positions are time varying. The main goal of this article therefore is to estimate the RTF vector, the PSD of the source, the PSD of the late reverberation and the PSD of self-noise simultaneously using N sequentially estimated CPSD matrices $\hat{\mathbf{P}}_{\mathbf{y}}(l)$ for one time segment β, i.e., for N time frames, while the source is only stationary within a time frame and the RTF changes from segment-to-segment.

III. ALS-Based Joint Estimation

To jointly estimate the parameters of interest, we consider the use of alternating least squares (ALS) based methods. Note that a two-step ALS method has been proposed before in this context [10]. In Section III-A, we will first introduce the method proposed in [10]. Then in Section III-B we will propose a modified version of the ALS method based on two improvements

over the original method to overcome parameter identifiability issues and potential numerical issues due to matrix singularities. Note that in [10] each time frame is utilized separately. However, if we assume the CPSD matrices for multiple time frames in a single time segment share the same RTF vector, we can use these time frames jointly to estimate RTF $\mathbf{a}$ with improved accuracy. Therefore, we will extend the modified ALS method to the case using the PSD matrices for multiple time frames in Section III-C.

A. ALS for a Single Time Frame

In [10], for each single time frame, the estimates of the RTF vector $\mathbf{a}$ and the PSD vector $\boldsymbol{\phi} = [\phi_s, \phi_\gamma, \phi_v]^T$ are obtained by minimizing the Frobenius norm of a model mismatch error matrix, i.e.,

$$\arg\min_{\mathbf{a},\boldsymbol{\phi}} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \phi_s \mathbf{a}\mathbf{a}^H - \phi_\gamma \hat{\boldsymbol{\Gamma}} - \phi_v \hat{\boldsymbol{\Psi}} \right\|_F^2, \tag{12}$$

where $\hat{\mathbf{A}}$ means the estimated $\mathbf{A}$. Note that the cost function in (12) is non-convex. To solve (12), a two-step ALS method is used by assuming that for either $\mathbf{a}$ or $\boldsymbol{\phi}$, an estimate is given and then estimating the other parameter vector.

More specifically, by assuming the RTF vector $\mathbf{a}$ is known or already estimated, the estimate of $\boldsymbol{\phi}$ can be obtained by solving

$$\arg\min_{\boldsymbol{\phi}} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \phi_s \hat{\mathbf{a}}\hat{\mathbf{a}}^H - \phi_\gamma \hat{\boldsymbol{\Gamma}} - \phi_v \hat{\boldsymbol{\Psi}} \right\|_F^2, \tag{13}$$

which has the following closed form solution

$$\hat{\boldsymbol{\phi}} = \boldsymbol{\Phi}_{\mathbf{a}}^{-1} \mathbf{b}, \tag{14}$$

where

$$\boldsymbol{\Phi}_{\mathbf{a}} = \begin{bmatrix} \left(\hat{\mathbf{a}}^H \hat{\mathbf{a}}\right)^2 & \hat{\mathbf{a}}^H \hat{\boldsymbol{\Gamma}} \hat{\mathbf{a}} & \hat{\mathbf{a}}^H \hat{\boldsymbol{\Psi}} \hat{\mathbf{a}} \\ \hat{\mathbf{a}}^H \hat{\boldsymbol{\Gamma}} \hat{\mathbf{a}} & \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H \hat{\boldsymbol{\Gamma}}\right\} & \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H \hat{\boldsymbol{\Psi}}\right\} \\ \hat{\mathbf{a}}^H \hat{\boldsymbol{\Psi}} \hat{\mathbf{a}} & \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H \hat{\boldsymbol{\Psi}}\right\} & \mathrm{tr}\left\{\hat{\boldsymbol{\Psi}}^H \hat{\boldsymbol{\Psi}}\right\} \end{bmatrix}, \tag{15}$$

and

$$\mathbf{b} = \begin{bmatrix} \hat{\mathbf{a}}^H \hat{\mathbf{P}}_{\mathbf{y}} \hat{\mathbf{a}} \\ \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H \hat{\mathbf{P}}_{\mathbf{y}}\right\} \\ \mathrm{tr}\left\{\hat{\boldsymbol{\Psi}}^H \hat{\mathbf{P}}_{\mathbf{y}}\right\} \end{bmatrix}. \tag{16}$$

When assuming the PSD vector $\hat{\boldsymbol{\phi}}$ is already estimated, the RTF vector $\mathbf{a}$ can be estimated by minimizing the cost function with respect to $\mathbf{a}$, that is

$$\arg\min_{\mathbf{a}} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_s \mathbf{a}\mathbf{a}^H - \hat{\phi}_\gamma \hat{\boldsymbol{\Gamma}} - \hat{\phi}_v \hat{\boldsymbol{\Psi}} \right\|_F^2, \tag{17}$$

which also has a closed form solution [26] given by the scaled principal eigenvector of the matrix $\hat{\mathbf{P}}_{\mathbf{x}} = \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_\gamma \hat{\boldsymbol{\Gamma}} - \hat{\phi}_v \hat{\boldsymbol{\Psi}}$, which is

$$\hat{\mathbf{a}} = \sqrt{\frac{\lambda}{\hat{\phi}_s}} \boldsymbol{\nu}, \tag{18}$$

where λ and ν are the principal eigenvalue and eigenvector of $\hat{\mathbf{P}}_{\mathbf{x}}$. The two steps are performed iteratively.

For the first step, the method in [10] finds an initial estimate of the RTF vector $\mathbf{a}$ by taking a random value or using a coarse estimate of the direction of arrival of the target source. For the second step, the PSD vector $\boldsymbol{\phi}$ is estimated via (14) with $\boldsymbol{\Phi}_{\mathbf{a}}$ and $\mathbf{b}$ calculated using the initial estimate $\hat{\mathbf{a}}$. Using the estimate of $\boldsymbol{\phi}$, matrix $\hat{\mathbf{P}}_{\mathbf{x}}$ is calculated in the second step and the RTF vector $\mathbf{a}$ can be estimated again via (18). For the next iterations, the two steps are repeated and the estimates of $\mathbf{a}$ and $\boldsymbol{\phi}$ are updated in an alternating fashion until a given convergence criterion is achieved or a certain number of iterations I are executed. Note that since each step reduces the cost function value, this method can converge to a local minimum even though the global minimum is not guaranteed. The ALS method is summarized in Algorithm 1.

Algorithm 1: ALS Method.

Input: $\hat{\mathbf{P}}_{\mathbf{y}}$, $\hat{\boldsymbol{\Gamma}}$,$\hat{\boldsymbol{\Psi}}$,init.$\hat{\mathbf{a}}$, I
Output: $\mathbf{a}$, $\boldsymbol{\phi}$
for *all* k, l **do**
  for *iter=1:I* **do**
    Compute $\boldsymbol{\Phi}_{\mathbf{a}}$ using (15) and $\mathbf{b}$ using (16).
    Estimate $\boldsymbol{\phi}$ using (14).
    Constrain the estimates of PSDs using (19) and (20).
    Calculate $\hat{\mathbf{P}}_{\mathbf{x}} = \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_\gamma \hat{\boldsymbol{\Gamma}} - \hat{\phi}_v \hat{\boldsymbol{\Psi}}$.
    Take EVD of $\hat{\mathbf{P}}_{\mathbf{x}}$ to find its principal eigenvalue and eigenvector.
    Estimate $\mathbf{a}$ using (18).
  for next time frame: use $\mathbf{a} = \mathbf{a}/a_1$ as the initial estimate.

Since PSDs should be positive by definition, all the estimated PSDs need to be lower bounded. In [10], the estimates of the PSDs are updated in the following way:

$$\{\phi_s, \phi_\gamma, \phi_v\} = \max\left(\{\phi_s, \phi_\gamma, \phi_v\}, \epsilon\right), \tag{19}$$

and

$$\{\phi_s, \phi_\gamma, \phi_v\} = \min\left(\{\phi_s, \phi_\gamma, \phi_v\}, \frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\right)}{M}\right), \tag{20}$$

where ϵ is the machine precision.

B. Modified-ALS for a Single Time Frame

An important condition for parameter estimation is the fact that the estimation problem itself needs to be identifiable [27]. Specifically, in the problem of jointly estimating the RTF vector $\mathbf{a}$ and the PSDs, the following condition should be satisfied for any two sets of parameters $\{\mathbf{a}, \phi_s, \phi_\gamma, \phi_v\}$ and $\{\bar{\mathbf{a}}, \bar{\phi}_s, \bar{\phi}_\gamma, \bar{\phi}_v\}$:

$$\begin{gathered} \phi_s \mathbf{a}\mathbf{a}^H + \phi_\gamma \boldsymbol{\Gamma} + \phi_v \boldsymbol{\Psi} = \bar{\phi}_s \bar{\mathbf{a}}\bar{\mathbf{a}}^H + \bar{\phi}_\gamma \boldsymbol{\Gamma} + \bar{\phi}_v \boldsymbol{\Psi} \\ \Leftrightarrow \\ \phi_s = \bar{\phi}_s, \mathbf{a} = \bar{\mathbf{a}}, \phi_\gamma = \bar{\phi}_\gamma, \phi_v = \bar{\phi}_v \end{gathered} \tag{21}$$

In the ALS method [10], however, (21) does not hold. To see this, let $\bar{\phi}_s = 4\phi_s$ and $\bar{\mathbf{a}} = \frac{\mathbf{a}}{2}$, we have $\phi_s \mathbf{a}\mathbf{a}^H = \bar{\phi}_s \bar{\mathbf{a}}\bar{\mathbf{a}}^H$ but $\bar{\phi}_s \neq \phi_s$ and $\bar{\mathbf{a}} \neq \mathbf{a}$. Therefore, any proper scaling of $\mathbf{a}$ and ϕ_s can be a solution as well. To solve this issue, we use the prior

information that $a_1 = 1$. In the final iteration, after estimating $\mathbf{a}$ using (18), we add a normalization step for both $\mathbf{a}$ and ϕ_s using the constant $c = \hat{a}_1$:

$$\hat{\mathbf{a}} \leftarrow \frac{\hat{\mathbf{a}}}{c} \tag{22}$$

and

$$\hat{\phi}_s \leftarrow \hat{\phi}_s \left|c\right|^2. \tag{23}$$

Notice also that in each iteration of the ALS method, if the estimated ϕ_s has an unusually small value (e.g. eps), the elements of the estimate of $\mathbf{a}$ in (18) will have rather large values. This will lead to large values of the first column and the first row of the matrix $\mathbf{\Phi_a}$ in (15), which means $\mathbf{\Phi_a}$ is close to being singular or badly scaled. To solve this issue, we can constrain the norm of the estimate of the scaled RTF vector to 1 by simply using the principal eigenvector instead of the scaled one in (18), i.e. $\hat{\mathbf{a}} = \boldsymbol{\nu}$. Note that, estimating the scaled $\mathbf{a}$ and ϕ_s is allowable because we will normalize them using (22) and (23) eventually in the last step.

The modified alternating least squares (MALS) method aims at minimizing the following cost function

$$\underset{\tilde{\mathbf{a}},\tilde{\phi}_s,\phi_\gamma,\phi_v}{\arg\min} \left\| \hat{\mathbf{P}}_\mathbf{y} - \tilde{\phi}_s \tilde{\mathbf{a}}\tilde{\mathbf{a}}^H - \phi_\gamma \hat{\mathbf{\Gamma}} - \phi_v \hat{\mathbf{\Psi}} \right\|_F^2, \tag{24}$$

where $\tilde{\mathbf{a}} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{a}}}$ and $\tilde{\phi}_s = \phi_s \mathbf{a}^H\mathbf{a}$. Since $\tilde{\phi}_s\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H = \phi_s \mathbf{a}\mathbf{a}^H$, the solution to (24) will also be the solution to (12). Once the estimates $\tilde{\mathbf{a}}$ and $\tilde{\phi}_s$ are obtained, the estimates of the RTF vector and the PSD of the source are given by

$$\mathbf{a} \leftarrow \frac{\tilde{\mathbf{a}}}{\tilde{a}_1}, \tag{25}$$

and

$$\phi_s \leftarrow \tilde{\phi}_s |\tilde{a}_1|^2. \tag{26}$$

Similarly as in [10] and as described in Section III-A, The optimization problem in (24) can be solved in an alternating fashion. Assuming $\tilde{\mathbf{a}}$ is already available (from a previous iteration or initialization), $\tilde{\boldsymbol{\phi}} = [\tilde{\phi}_s, \phi_\gamma, \phi_v]$ is estimated by the least squares estimate

$$\hat{\tilde{\boldsymbol{\phi}}} = \mathbf{\Phi}_{\tilde{\mathbf{a}}}^{-1} \tilde{\mathbf{b}}, \tag{27}$$

where

$$\mathbf{\Phi}_{\tilde{\mathbf{a}}} = \begin{bmatrix} 1 & \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Gamma}} \hat{\tilde{\mathbf{a}}} & \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Psi}} \hat{\tilde{\mathbf{a}}} \\ \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Gamma}} \hat{\tilde{\mathbf{a}}} & \mathrm{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Gamma}}\right\} & \mathrm{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Psi}}\right\} \\ \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Psi}} \hat{\tilde{\mathbf{a}}} & \mathrm{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Psi}}\right\} & \mathrm{tr}\left\{\hat{\mathbf{\Psi}}^H \hat{\mathbf{\Psi}}\right\} \end{bmatrix}, \tag{28}$$

and

$$\tilde{\mathbf{b}} = \begin{bmatrix} \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{P}}_\mathbf{y} \hat{\tilde{\mathbf{a}}} \\ \mathrm{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{P}}_\mathbf{y}\right\} \\ \mathrm{tr}\left\{\hat{\mathbf{\Psi}}^H \hat{\mathbf{P}}_\mathbf{y}\right\} \end{bmatrix}. \tag{29}$$

When an estimate of $\hat{\tilde{\boldsymbol{\phi}}}$ is known from the previous iteration, we calculate the matrix $\hat{\mathbf{P}}_\mathbf{x} = \hat{\mathbf{P}}_\mathbf{y} - \hat{\phi}_\gamma \hat{\mathbf{\Gamma}} - \hat{\phi}_v \hat{\mathbf{\Psi}}$ and obtain the estimate of $\tilde{\mathbf{a}}$ by

$$\hat{\tilde{\mathbf{a}}} = \boldsymbol{\nu}, \tag{30}$$

where $\boldsymbol{\nu}$ is the principal eigenvector of $\hat{\mathbf{P}}_\mathbf{x}$. After a sufficient number of iterations, $\mathbf{a}$ and $\boldsymbol{\phi}$ are obtained using (25) and (26).

The MALS method is summarized in Algorithm 2.

Algorithm 2: MALS Method.

Input: $\hat{\mathbf{P}}_\mathbf{y}$, $\hat{\mathbf{\Gamma}}$,$\hat{\mathbf{\Psi}}$,init.$\hat{\tilde{\mathbf{a}}}$, I
Output: $\mathbf{a}$, $\boldsymbol{\phi}$
for *all* k, l **do**
  for *iter=1:I* **do**
    Compute $\mathbf{\Phi}_{\tilde{\mathbf{a}}}$ using (28) and $\tilde{\mathbf{b}}$ using (29).
    Estimate $\tilde{\boldsymbol{\phi}}$ using (27).
    Calculate $\hat{\mathbf{P}}_\mathbf{x} = \hat{\mathbf{P}}_\mathbf{y} - \hat{\gamma}\mathbf{\Gamma} - \hat{\phi}_v \hat{\mathbf{\Psi}}$.
    Take EVD of $\hat{\mathbf{P}}_\mathbf{x}$ to find its principal eigenvector.
    Estimate $\tilde{\mathbf{a}}$ using (30).
  Estimate $\mathbf{a}$ and ϕ_s using (25) and (26).

C. ALS for Multiple Time Frames

In the previous subsections, the joint estimation of the RTF vector $\mathbf{a}$ and the PSD vector $\boldsymbol{\phi}$ is performed for a single time frame based on the ALS approach. However, in many cases, $\mathbf{a}$ can be assumed to be constant across multiple frames in a time segment. With this prior information, we consider in this subsection the joint estimation of $\mathbf{a}$, and the PSD vector $\boldsymbol{\phi} = [\boldsymbol{\phi}(1 + (\beta - 1)N)^T, \ldots, \boldsymbol{\phi}(\beta N)^T]^T$ using all time-frames in a segment, where $\boldsymbol{\phi}(l) = [\phi_s(l), \phi_\gamma(l), \phi_v(l)]^T$ for $l = 1 + (\beta - 1)N, \ldots, \beta N$.

The alternating least squares method using multiple time frames jointly (JALS) aims at minimizing the sum of the Frobenius norms of the model mismatch error matrices for all time frames l that fall in the same segment β, i.e.,

$$\underset{\mathbf{a},\boldsymbol{\phi}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_\mathbf{y}(l) - \phi_s(l)\mathbf{a}\mathbf{a}^H - \phi_\gamma(l)\hat{\mathbf{\Gamma}} - \phi_v(l)\hat{\mathbf{\Psi}} \right\|_F^2. \tag{31}$$

Like the MALS method, we reparameterize $\mathbf{a}$ and $\phi_s(l)$ for $l = 1 + (\beta - 1)N, \ldots, \beta N$ by $\tilde{\mathbf{a}} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{a}}}$ and $\tilde{\phi}_s(l) = \phi_s(l)\mathbf{a}^H\mathbf{a}$, which gives us the following cost function

$$\underset{\tilde{\mathbf{a}},\tilde{\boldsymbol{\phi}}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_\mathbf{y}(l) - \tilde{\phi}_s(l)\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H - \phi_\gamma(l)\hat{\mathbf{\Gamma}} - \phi_v(l)\hat{\mathbf{\Psi}} \right\|_F^2. \tag{32}$$

To solve (32), we also use a two-step ALS method by either assuming $\tilde{\mathbf{a}}$ is given and estimating $\tilde{\boldsymbol{\phi}}$ or assuming $\tilde{\boldsymbol{\phi}}$ is estimated and estimating $\tilde{\mathbf{a}}$.

When an estimate of $\tilde{\mathbf{a}}$ is already given, the minimization with respect to $\tilde{\boldsymbol{\phi}}$ is

$$\arg\min_{\tilde{\boldsymbol{\phi}}} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(l) - \tilde{\phi}_s(l)\hat{\tilde{\mathbf{a}}}\hat{\tilde{\mathbf{a}}}^H - \phi_\gamma(l)\hat{\boldsymbol{\Gamma}} - \phi_v(l)\hat{\boldsymbol{\Psi}} \right\|_F^2, \quad (33)$$

which is equivalent to minimizing the cost function for each time frame l separately, i.e.

$$\arg\min_{\substack{\tilde{\boldsymbol{\phi}}(l),\\ \forall l\in 1+(\beta-1)N,\ldots,\beta N}} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(l) - \tilde{\phi}_s(l)\hat{\tilde{\mathbf{a}}}\hat{\tilde{\mathbf{a}}}^H - \phi_\gamma(l)\hat{\boldsymbol{\Gamma}} - \phi_v(l)\hat{\boldsymbol{\Psi}} \right\|_F^2, \quad (34)$$

as $\tilde{\boldsymbol{\phi}}(l)$ is defined per time frame. For each time frame l, (34) has a closed form solution

$$\tilde{\boldsymbol{\phi}}(l) = \tilde{\boldsymbol{\Phi}}_{\tilde{\mathbf{a}}}^{-1}\tilde{\mathbf{b}}(l), \quad (35)$$

where

$$\tilde{\boldsymbol{\Phi}}_{\tilde{\mathbf{a}}} = \begin{bmatrix} 1 & \hat{\tilde{\mathbf{a}}}^H\hat{\boldsymbol{\Gamma}}\hat{\tilde{\mathbf{a}}} & \hat{\tilde{\mathbf{a}}}^H\hat{\boldsymbol{\Psi}}\hat{\tilde{\mathbf{a}}} \\ \hat{\tilde{\mathbf{a}}}^H\hat{\boldsymbol{\Gamma}}\hat{\tilde{\mathbf{a}}} & \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\boldsymbol{\Gamma}}\right\} & \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\boldsymbol{\Psi}}\right\} \\ \hat{\tilde{\mathbf{a}}}^H\hat{\boldsymbol{\Psi}}\hat{\tilde{\mathbf{a}}} & \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\boldsymbol{\Psi}}\right\} & \mathrm{tr}\left\{\hat{\boldsymbol{\Psi}}^H\hat{\boldsymbol{\Psi}}\right\} \end{bmatrix}, \quad (36)$$

and

$$\tilde{\mathbf{b}}(l) = \begin{bmatrix} \hat{\tilde{\mathbf{a}}}^H\hat{\mathbf{P}}_{\mathbf{y}}(l)\hat{\tilde{\mathbf{a}}} \\ \mathrm{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\mathbf{P}}_{\mathbf{y}}(l)\right\} \\ \mathrm{tr}\left\{\hat{\boldsymbol{\Psi}}^H\hat{\mathbf{P}}_{\mathbf{y}}(l)\right\} \end{bmatrix}. \quad (37)$$

When an estimate of $\tilde{\boldsymbol{\phi}}$ is given, $\tilde{\mathbf{a}}$ can be obtained for a segment β by minimizing

$$\arg\min_{\tilde{\mathbf{a}}} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\tilde{\phi}}_s(l)\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H - \hat{\gamma}(l)\hat{\boldsymbol{\Gamma}} - \hat{\phi}_v(l)\hat{\boldsymbol{\Psi}} \right\|_F^2. \quad (38)$$

The solution for $\tilde{\mathbf{a}}$ is the principal eigenvector of $\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_s(l)\left[\hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\phi}_\gamma(l)\hat{\boldsymbol{\Gamma}} - \hat{\phi}_v(l)\hat{\boldsymbol{\Psi}}\right]$ (See Appendix A).

The alternating least squares method using multiple time frames jointly (JALS) is summarized in Algorithm 3.

Algorithm 3: JALS Method.

Input: $\hat{\mathbf{P}}_{\mathbf{y}}$, $\hat{\boldsymbol{\Gamma}}$,$\hat{\boldsymbol{\Psi}}$,init.$\hat{\tilde{\mathbf{a}}}$, I
Output: $\mathbf{a}, \phi$
for *all* k, β **do**
  for *iter=1:I* **do**
    Calculate $\boldsymbol{\Phi}_{\tilde{\mathbf{a}}}$ using (36) and $\tilde{\mathbf{b}}(l)$ using (37).
    Estimate $\tilde{\boldsymbol{\phi}}(l)$ using (35) for each l.
    Calculate $\hat{\mathbf{P}}_{\mathbf{x}}(l) = \hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\phi}_\gamma(l)\hat{\boldsymbol{\Gamma}} - \hat{\phi}_v(l)\hat{\boldsymbol{\Psi}}$.
    Estimate $\tilde{\mathbf{a}}$ using the principal eigenvector of $\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_s(l)\hat{\mathbf{P}}_{\mathbf{x}}(l)$.
  Estimate $\mathbf{a}$ and $\phi_s(l)$ using (25) and (26).

D. Robust PSDs Constraints

In [11], it has been shown that linear inequality constraints on the parameters of interest can be used to improve the robustness of the estimation. In [10], the PSD of the source, the PSD of the late reverberation and the PSD of the ambient noise are constrained by (19) and (20). In this section, we introduce more robust constraints on the PSDs to avoid large underestimation and overestimation errors.

1) Upper Bounds: To avoid large overestimation errors, we can use upper bounds for the PSDs. For the diagonal elements of $\mathbf{P}_y$, it holds that

$$\mathbf{P}_{y_{m,m}}(l) = \tilde{\phi}_s(l)\left|\tilde{a}_m\right|^2 + \phi_\gamma(l)\boldsymbol{\Gamma}_{m,m} + \phi_v(l)\boldsymbol{\Psi}_{m,m}. \quad (39)$$

Since the three additive terms in (39) are positive, we have

$$\left\{\tilde{\phi}_s(l)\left|\tilde{a}_m\right|^2, \phi_\gamma(l)\boldsymbol{\Gamma}_{m,m}, \phi_v(l)\boldsymbol{\Psi}_{m,m}\right\} \le \mathbf{P}_{y_{m,m}}(l), \quad (40)$$

for all m. Hence, the upper bound for the PSDs of the target source is

$$\tilde{\phi}_s(l) \le \min_m \left\{ \frac{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}}(l)}{\left|\hat{\tilde{a}}_m\right|^2} \right\}. \quad (41)$$

Similarly, the upper bounds for the PSDs of the late reverberation and the ambient noise are

$$\phi_\gamma(l) \le \min_m \left\{ \frac{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}}(l)}{\hat{\boldsymbol{\Gamma}}_{m,m}} \right\}, \quad (42)$$

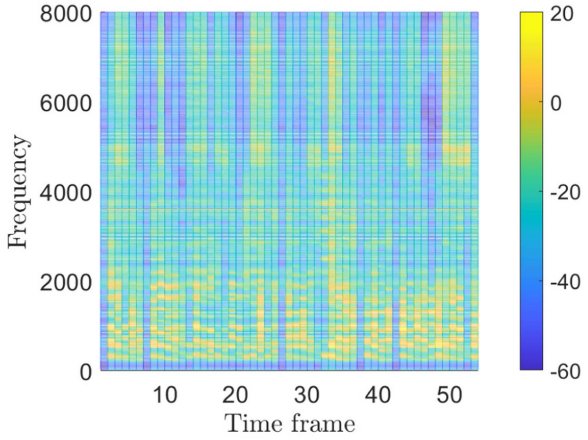
$$\phi_v(l) \le \min_m \left\{ \frac{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}}(l)}{\hat{\boldsymbol{\Psi}}_{m,m}} \right\}. \quad (43)$$

Note that $\hat{\boldsymbol{\Gamma}}_{m,m} = 1$ in (7) and that $\hat{\boldsymbol{\Psi}}_{m,m} = 1$ when considering only self-noise and each microphone has the same self-noise PSD. In that case we thus have

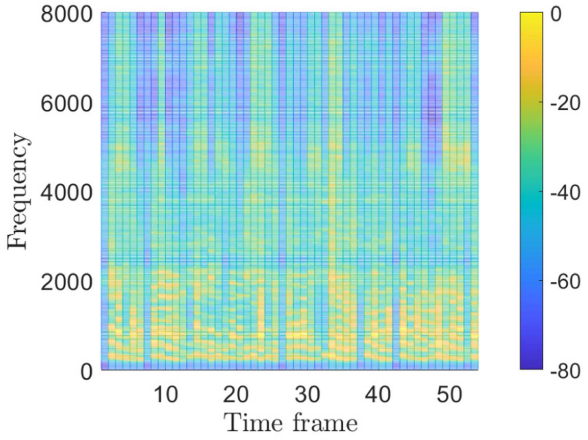
$$\{\phi_\gamma(l), \phi_v(l)\} \le \min_m \left\{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}}(l)\right\} \le \frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\right)}{M}, \quad (44)$$

which is tighter than the bound in (20) as used in [10]. Hence, by using (42) and (43), the overestimation errors for the PSDs of the late reverberation and the ambient noise are smaller than the errors using (20), resulting in better speech intelligibility performance [28], [29].

2) Lower Bounds: To avoid large underestimation errors, we need lower bounds for the PSDs as well. In both [10] and [11], the prior information was used that the PSDs should be positive, setting the lower bounds for all PSDs to ϵ. That is, when obtaining negative incorrect estimates of the PSDs, these are replaced by the minimum value ϵ. However, this will lead to very large under estimation errors. Therefore, we propose the use of tighter lower bounds derived from other prior information on the PSDs.



(a) Normalized PSD of the target source $\tilde{\phi}_s$ (in dB).



(b) PSD of the late reverberation ϕ_γ (in dB).

Fig. 2. Time frame and frequency distribution of the target source PSD and the late reverberation PSD.

For the normalized PSD of the source $\tilde{\phi}_s$ and the PSD of the late reverberation ϕ_γ, we can see that they have a similar distribution on the time-frequency domain as illustrated in Fig. 2.

Based on this, we make the assumption that the ratio between the normalized PSD of the source and the PSD of the late reverberation is bounded on both sides, i.e.

$$C_1 \leq \frac{\tilde{\phi}_s(l)}{\phi_\gamma(l)} \leq \frac{1}{C_2}, \tag{45}$$

or

$$C_1 \phi_\gamma(l) \leq \tilde{\phi}_s(l), \tag{46}$$

and

$$C_2 \tilde{\phi}_s(l) \leq \phi_\gamma(l), \tag{47}$$

for all (l, k) pairs. Note that this assumption is weaker than the one made in [30], where it is assumed that the ratio between the sound source PSD and the late reverberation PSD is a constant. Using (46) and (47), we can constrain the estimated PSDs of the source and the PSDs of the late reverberation in the following way. We first initialize C_1 and C_2 by an initial value like $C_1 = C_2 = 1$ for the first time frame $l = 1$. For the l-th time frame, we update C_1 and C_2 while making $\hat{\tilde{\phi}}_s(l)$ and $\hat{\phi}_\gamma(l)$ positive in the way shown in Fig. 3 and Appendix B, where $\hat{\tilde{\phi}}_s^{\min}(l)$ is calculated by

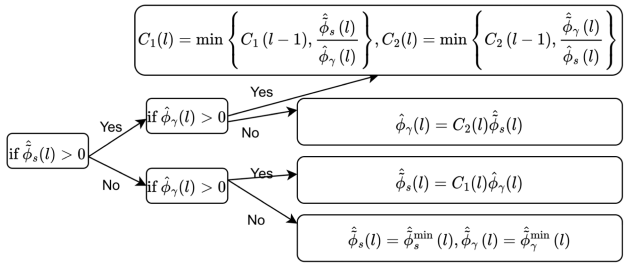


Fig. 3. Decision flow for updating $C_1, C_2, \hat{\tilde{\phi}}_s$ and $\hat{\phi}_\gamma$.

$$\hat{\tilde{\phi}}_s^{\min} = \min\left\{\left|\frac{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}} - \hat{\mathbf{P}}_{\mathbf{y}_{m+1,m+1}}}{\left|\hat{\tilde{\mathbf{a}}}_m\right|^2 - \left|\hat{\tilde{\mathbf{a}}}_{m+1}\right|^2}\right|\right\}_{m=1}^{M-1}, \tag{48}$$

and

$$\hat{\phi}_\gamma^{\min} = \min\left\{\left|\hat{\mathbf{P}}_{\mathbf{y}_{m,m}} - \hat{\tilde{\phi}}_s^{\min}\left|\hat{\tilde{\mathbf{a}}}_m\right|^2 - \hat{\phi}_v\right|\right\}_{m=1}^{M}. \tag{49}$$

For the PSD of the ambient noise, the lower bounds depend on the stochastic property of the noise component. We use the following way to constrain $\phi_v(l)$. First, we give the lower bound C_3 an initial small value ϵ. Then, we update C_3 as

$$C_3(l) = \begin{cases} \frac{C_3(l-1)+\hat{\phi}_v(l)}{2} & \text{if} \hat{\phi}_v(l) > 0 \\ C_3(l-1) & \text{else} \end{cases}. \tag{50}$$

With C_3, we estimate $\phi_v(l)$ by

$$\hat{\phi}_v(l) = \begin{cases} \hat{\phi}_v(l) & \text{if} \hat{\phi}_v(l) > 0 \\ C_3(l-1) & \text{else} \end{cases}, \tag{51}$$

Note that the above procedure dealing with non-positive estimates of the PSDs might give us values larger than the upper bounds we derived before in (41) to (43). Therefore, we first execute the above procedure and then upper bound all the estimates.

IV. Experiments

In this section, we will evaluate our proposed ALS-based methods in various scenarios. In addition to the ALS method proposed in [10], we introduce in Section IV-A two more reference methods, namely JMLE [14] and SCFA [11]. In Section IV-B, we present the evaluation metrics for all methods. We compare the performance of all methods in various scenarios in Sections IV-C and IV-D.

A. Reference Methods

The two reference methods introduced here are both based on the maximum likelihood (ML) cost function:

$$\min \sum_{l=1}^{N} \log |\mathbf{P}_{\mathbf{y}}(l)| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}(l)\mathbf{P}_{\mathbf{y}}^{-1}(l)\right). \tag{52}$$

1) JMLE: In our recent work [14], we assumed a noiseless scenario and proposed a joint maximum likelihood estimator (JMLE) to estimate the RTF of the target source, the PSDs of the target source and the PSDs of the late reverberation jointly. The JMLE method performances well and has low computational complexity. However, the performance of JMLE is not robust for low SNR scenarios due to the noiseless signal model assumed in [14], which is

$$\mathbf{P}_{\mathbf{y}}(l) = \phi_s(l)\mathbf{a}(\beta)\mathbf{a}^H(\beta) + \phi_\gamma(l)\mathbf{\Gamma}. \tag{53}$$

2) SCFA: The last reference method we use for comparison is the simultaneous confirmatory factor analysis (SCFA) method [11]. SCFA performs well in reverberant and noisy environments. However, SCFA comes with a high computational cost due to solving the following non-convex optimization problem

$$\begin{aligned} &\arg\min_{\substack{\phi_s(l),\,\mathbf{a}(\beta)\\ \phi_\gamma(l),\,\phi_v}} \sum_{l=1}^{N} \log |\mathbf{P}_{\mathbf{y}}(l)| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}(l)\mathbf{P}_{\mathbf{y}}^{-1}(l)\right) \\ &\text{s.t. } \mathbf{P}_{\mathbf{y}}(l) = \phi_s(l)\mathbf{a}(\beta)\mathbf{a}^H(\beta) + \phi_\gamma(l)\mathbf{\Gamma} + \phi_v\mathbf{I}, \\ &\quad a_1(\beta) = 1, \phi_s(l) \geq 0, \phi_\gamma(l) \geq 0, \phi_v \geq 0, \end{aligned} \tag{54}$$

where $\phi_v\mathbf{I}$ corresponds to the microphone self noise, which is assumed to be white Gaussian noise. In [11], the above optimization problem is computed iteratively. At each iteration, the parameters are updated and the cost function value is reduced by solving a non-linear constrained optimization problem. The updating procedure is terminated when meeting a local minimum. Note that due to the non-convexity of the optimization problem, the number of iterations needed is large. Hence the computational cost of this method is relatively high.

B. Evaluation Metrics

1) Estimation Errors: The first evaluation metric is the estimation error of the parameters of interest. For the RTF vector, we calculate the Hermitian angles between the estimated RTFs and the true RTFs and average them over different frequency bins and time segments, that is,

$$E_{\mathbf{a}} = \frac{\sum_{\beta=1}^{B}\sum_{k=1}^{K/2+1} \operatorname{acos}\left(\frac{|\mathbf{a}(\beta,k)^H\hat{\mathbf{a}}(\beta,k)|}{\|\mathbf{a}(\beta,k)\|_2\|\hat{\mathbf{a}}(\beta,k)\|_2}\right)}{\mathrm{B}\,(K/2+1)}. \tag{55}$$

Note that this metric evaluates the alignment of the estimated RTF with the ground-truth RTF, but cannot reflect scaling errors. For all types of PSDs, we use the symmetric log-error distortion measure [31]

$$E_i = \frac{10\sum_{\beta=1}^{B}\sum_{l=1+(\beta-1)N}^{\beta N}\sum_{k=1}^{K/2+1}\left|\log\left(\frac{\phi_i(l,k)}{\hat{\phi}_i(l,k)}\right)\right|}{\mathrm{B}N\,(K/2+1)}, \tag{56}$$

with $i \in \{s, \gamma, v\}$. In the following experiments, we will also show the detailed PSD estimation performance by using the overestimating errors (denoted as $E_{\phi_i}^{\mathrm{ov}}$) and the underestimation errors (denoted as $E_{\phi_i}^{\mathrm{un}}$) as used in [28]

$$E_i^{\mathrm{ov}} = \frac{10\sum_{\beta=1}^{B}\sum_{l=1+(\beta-1)N}^{\beta N}\sum_{k=1}^{K/2+1}\left|\min\left\{0, \log\left(\frac{\phi_i(l,k)}{\hat{\phi}_i(l,k)}\right)\right\}\right|}{\mathrm{B}N\,(K/2+1)}, \tag{57}$$

and

$$E_i^{\mathrm{un}} = \frac{10\sum_{\beta=1}^{B}\sum_{l=1+(\beta-1)N}^{\beta N}\sum_{k=1}^{K/2+1}\max\left\{0, \log\left(\frac{\phi_i(l,k)}{\hat{\phi}_i(l,k)}\right)\right\}}{\mathrm{B}N\,(K/2+1)}. \tag{58}$$

Note that, typically, large underestimation errors in the source PSDs and large overestimation errors in the noise PSDs can cause large target source distortions when applying the estimates in a noise reduction framework. Also, large underestimation errors in the noise PSD are likely to cause musical noise [28]. We therefore also quantify the performance in terms of predicted quality and intelligibility when used in combination with a noise reduction algorithm, as explained below.

2) Predicted Quality and Intelligibility: We can construct the following multi-channel Wiener filter (MWF) [32] based on the estimated parameters to extract the target signal,

$$\hat{\mathbf{w}}(l) = \frac{\hat{\phi}_s(l)\hat{\mathbf{w}}_{\mathrm{MVDR}}(l)}{\hat{\phi}_s(l) + \hat{\mathbf{w}}_{\mathrm{MVDR}}(l)^H\hat{\mathbf{R}}_{nn}(l)\hat{\mathbf{w}}_{\mathrm{MVDR}}(l)}, \tag{59}$$

where $\mathbf{w}_{\mathrm{MVDR}}(l)$ is the minimum variance distortionless response (MVDR) beamformer [33]

$$\hat{\mathbf{w}}_{\mathrm{MVDR}}(l) = \frac{\hat{\mathbf{R}}_{nn}^{-1}(l)\hat{\mathbf{a}}(l)}{\hat{\mathbf{a}}(l)^H\hat{\mathbf{R}}_{nn}^{-1}(l)\hat{\mathbf{a}}(l)}, \tag{60}$$

and

$$\hat{\mathbf{R}}_{nn}(l) = \hat{\phi}_\gamma(l)\hat{\mathbf{\Gamma}} + \hat{\phi}_v(l)\hat{\mathbf{\Psi}}, \tag{61}$$

and where $\hat{\mathbf{w}}(l)$ is used as $\hat{s}(l) = \hat{\mathbf{w}}(l)^H\mathbf{y}(l)$. After estimating $\hat{s}(l)$, the time domain signal is reconstructed by calculating the IFFT followed by an overlap-add procedure. Note that for the JMLE method, $\hat{\mathbf{R}}_{nn}(l) = \hat{\phi}_\gamma(l)\hat{\mathbf{\Gamma}}$ due to its noiseless signal model.

After applying the MWF filter to the noisy signal, we obtain the estimated target signal and evaluate the noise reduction performance using the segmental signal-to-noise-ratio (SSNR) [34], the speech intelligibility performance using the speech intelligibility in bits (SIIB) measure [35], [36] and the speech quality performance using the perceptual evaluation of speech quality (PESQ) measure [37].

3) Computation Time: The last evaluation metric is the computational time comparison between our proposed methods and the reference methods.

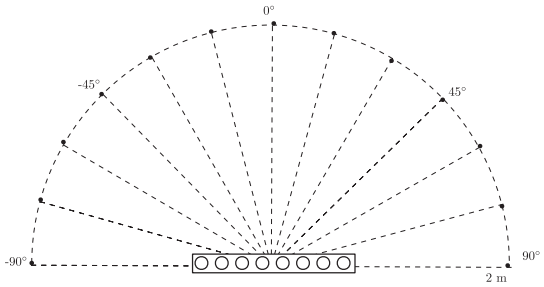


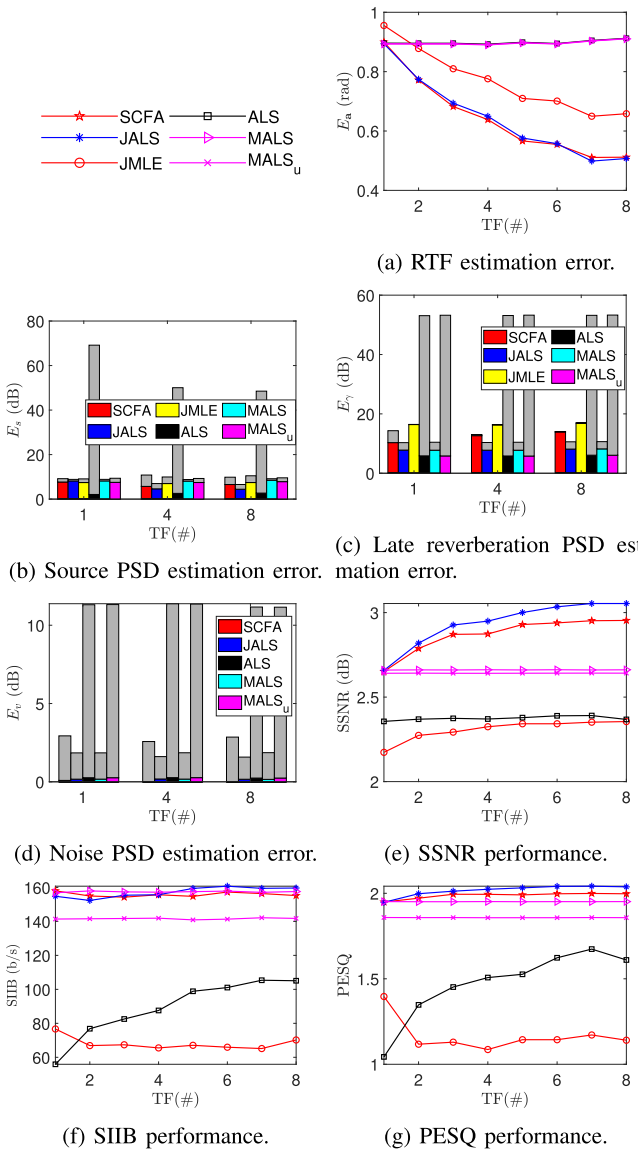
Fig. 4. Geometric setup for the real RIRs.

C. Experiment 1

1) Setup: We use speech signals originating from the TIMIT database [38] and recorded RIRs to simulate realistic acoustic scenarios. The RIRs are downloaded from the database in [39], which were recorded in a room with size $6 \times 6 \times 2.4$ m. The geometric setup for the recording is shown in Fig. 4. The sound source was placed 2 m away from the center of the uniform linear microphone array at 0°. This array has 8 microphones and 8 cm inter-distances. At each microphone, we synthesize the reverberant signal by convolving the speech source (with a duration of 35 s) with the corresponding RIR. Subsequently, we add noise components to the reverberant signals simulating the microphone noise at specified signal-to-noise ratios (SNRs) to synthesize the microphone signals. In the following, we will consider white Gaussian noise to simulate microphone selfnoise with variance σ_v^2 calculated from given SNR values for each microphone. Since the signal is non-stationary, we calculate the SNR by averaging the target signal-to-noise ratio over the whole time duration.

In this experiment, we used the following parameters setting: The sampling rate is $f_s = 16$ kHz. The sampled noisy microphone signals are processed by the STFT for each sub-time frame. As analysis and synthesis window we use the square-root Hann window with a length of 32 ms with 50% overlap between adjacent sub-time frames. Note that each time frame consists of $T_{sf} = 40$ overlapping sub-time frames and has a duration of 0.64 s. The FFT length is 512. The speed of sound is set to 344 m/s. Note that the first 512 samples of the RIRs are used to calculate the true RTFs as these parts of the RIRs fall within each current sub-time frame and the remaining parts are considered as the late reverberation. Note that for ALS-based methods, we use the same random vector as an initial estimate of the RTF for the first time frame in a time segment (ALS and MALS) or for a time segment (JALS).

2) Results: In Fig. 5, we compare our proposed methods with all the other reference methods as a function of the number of time frames in a time segment varying from 1 to 8. The reverberation time is 0.61 s and the SNR is fixed at 0 dB. To evaluate how the robust constraints of the PSDs proposed in Section III-D help the estimation of the parameters of interest, we also included the modified ALS method without using the robust constraints in Section III-D but using (19) and (20), referred to as MALS$_u$. When using only a single time frame in each time segment, the RTF estimation errors for the ALS-based



(a) RTF estimation error.

(b) Source PSD estimation error.

(c) Late reverberation PSD estimation error.

(d) Noise PSD estimation error.

(e) SSNR performance.

(f) SIIB performance.

(g) PESQ performance.

Fig. 5. Performance vs the number of time frames. In Figs (b), (c) and (d), the gray bars indicate the underestimation errors, the colored bars indicate overestimation errors and the methods from left to right are SCFA, JALS, JMLE(in Figs (b) and (c)), ALS, MALS and MALS$_u$.

methods and the SCFA method have similar values which are much lower than the JMLE method as shown in Fig. 5(a). The reason is that JMLE was derived from a noiseless signal model [40]. The signal model mismatch error is thus large for the JMLE in a 0 dB environment. When increasing the number of time frames in each time segment, the RTF estimation errors for ALS, MALS$_u$ and MALS (the three ALS-based methods using a single time frame) do not vary much; while the RTF estimation errors for JALS, JMLE and SCFA (methods using multiple time frames) become much lower. The RTF estimation errors $E_{\mathbf{a}}$ for methods using a single time frame fluctuate slightly because the first time frame of a time segment use random initial estimate of the RTF. The other time frames use the estimate in the previous time frame as the initial estimate. $E_{\mathbf{a}}$ for ALS and MALS$_u$ are close to MALS due to the normalization

process in the Hermitian angle metric in (55). The drawback of the Hermitian angle metric is that any scaled estimate will have the same value. The bad scaling of ALS can be reflected in the target source PSD estimation errors, where ALS has much larger errors compared to the other methods as shown in Fig. 5(b). In Fig. 5(c) and (d), we can see that MALS$_u$ has similar performance with ALS, which both use the PSDs constraints in (19) and (20). While, MALS using the robust constraints of the PSDs proposed in Section III-D has much lower errors compared to ALS and MALS$_u$. As expected, the PSDs estimation errors do not change much as a function of the number of frames in a segment since the PSDs are time frame variant parameters. In Fig. 5(b)–(d), we show the underestimation error and overestimation error for the PSDs. Our proposed methods (MALS and JALS) have improved performance compared to ALS and similar performance compared to SCFA. As shown, ALS has the worst underestimation errors for all the PSDs. This is due to the lack of a normalization step and using the value ϵ to replace negative values in the ALS method. JMLE has the largest overestimation errors for PSDs of the late reverberation. This is due to the noiseless signal model that is assumed with JMLE. In a low SNR environment, the JMLE method considers the ambient noise as late reverberation and gives larger values when estimating the PSDs of the late reverberation. For noise reduction performance evaluated by SSNR in Fig. 5(e), our proposed JALS has the best performance, which is slightly better than SCFA but much better than the other methods. For the speech intelligibility performance evaluated by SIIB and the speech quality performance evaluated by PESQ in Fig. 5(f) and (g), the proposed JALS, MALS and the reference method SCFA outperform the other methods.

In Fig. 6, we compare all the methods while changing the variance of the ambient noise component such that the SNR increases from 0 dB to 40 dB. The reverberation time is 0.36 s and each time segment contains 8 time frames. As shown in Fig. 6(a), the RTF estimation errors become lower for all methods when the SNR becomes larger. SCFA and our proposed method JALS have the best overall performance, which is much better than methods using a single time frame (ALS and MALS). For low SNR, JMLE is worse than JALS, but when increasing the SNR, JMLE improves the fastest as its model mismatch error is smaller and has a smaller RTF estimation error than JALS for 40 dB SNR. We can see that when the signal model mismatch error is neglectable, the MLE-based methods (SCFA and JMLE) perform better than the ALS-based methods (JALS). For the PSDs estimation errors in Fig. 6(b) to (d), SCFA has the best performance with only JMLE reaching a similar performance for high SNR scenarios. Our proposed ALS-based methods (MALS and JALS) perform much better than ALS. For noise reduction and speech intelligibility performance in Fig. 6(e) to (g), MALS and JALS have similar performance with SCFA and much better performance than ALS. When increasing the SNR, JMLE has the most significant improvement and gets close to the performance of MALS, JALS and SCFA for 40 dB SNR.

We also evaluate the computation time for all methods and average these over all cases shown in Fig. 6. Then, we averaged and normalized the run time for all methods with respect to the

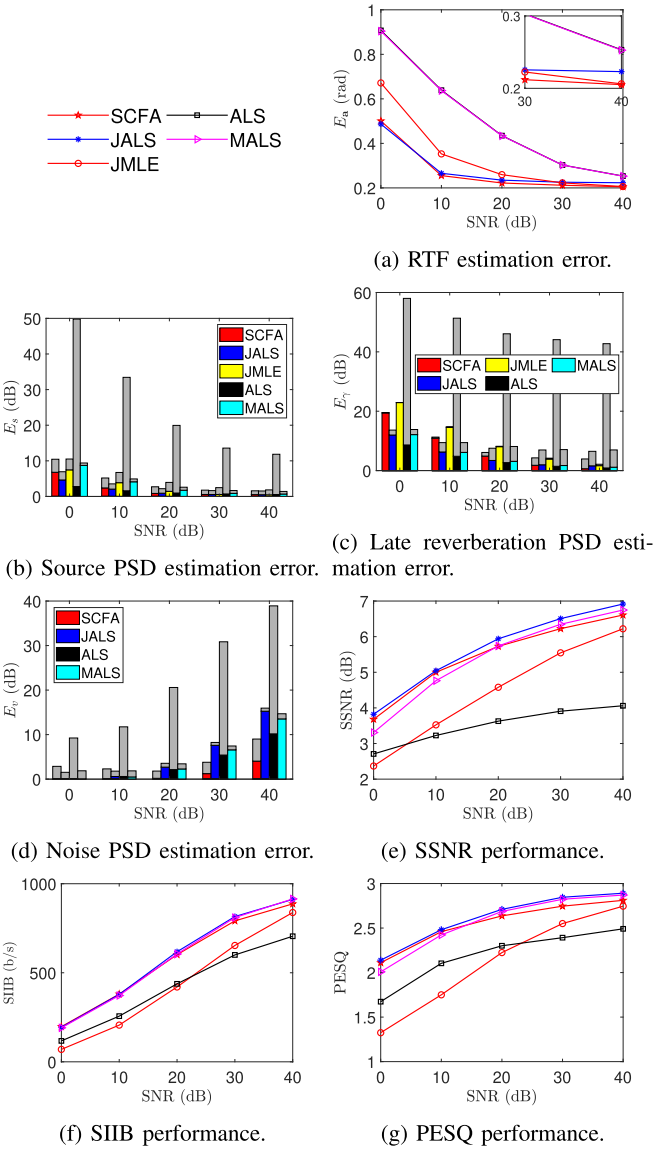


Fig. 6. Performance vs SNR. In Figs (b), (c) and (d), the gray bars indicate the underestimation errors and the colored bars indicate overestimation errors.

TABLE I
COMPUTATION TIME COMPARISON

method	SCFA	ALS	MALS	JMLE	JALS
Normalized run time	154.65	6.27	5.7	1.66	1

run time for JALS as shown in Table I. We sort the run time for all the methods in descending order from left to right. As expected, SCFA is the most time-consuming method. JALS and JMLE are the two fastest methods. The computational cost mainly comes from the inversion of a 3×3 matrix (complexity of order 3^3) and the eigenvalue decomposition of an $M \times M$ matrix (complexity of order M^3) for the ALS-based methods (ALS, MALS and JALS). For the case that each time segment has N time frames, ALS and MALS process each time frame separately and execute I iterations N times. Hence, they have a complexity of order $I \times N \times (3^3 + M^3)$. For JALS, we only need to calculate the

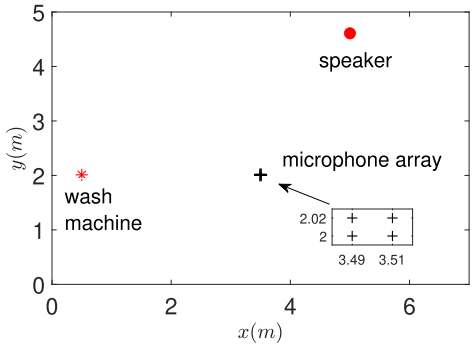


Fig. 7. Top view of the acoustic scene with a zoom-in of microphones.

eigenvalue decomposition I times. Hence, its complexity order is $I \times M^3 + I \times N \times 3^3$. The complexity order of JMLE is $(N + I) \times M^3$ [40]. In this experiment, we have $M = 8, N = 8$ and $I = 10$. Therefore, the time cost ratio among ALS/MALS, JMLE and JALS is $I \times N \times (3^3 + M^3) : (N + I) \times M^3 : I \times M^3 + I \times N \times 3^3 \approx 5.92 : 1.27 : 1$, which is approximately similar to the real averaged run time ratio in Table I.

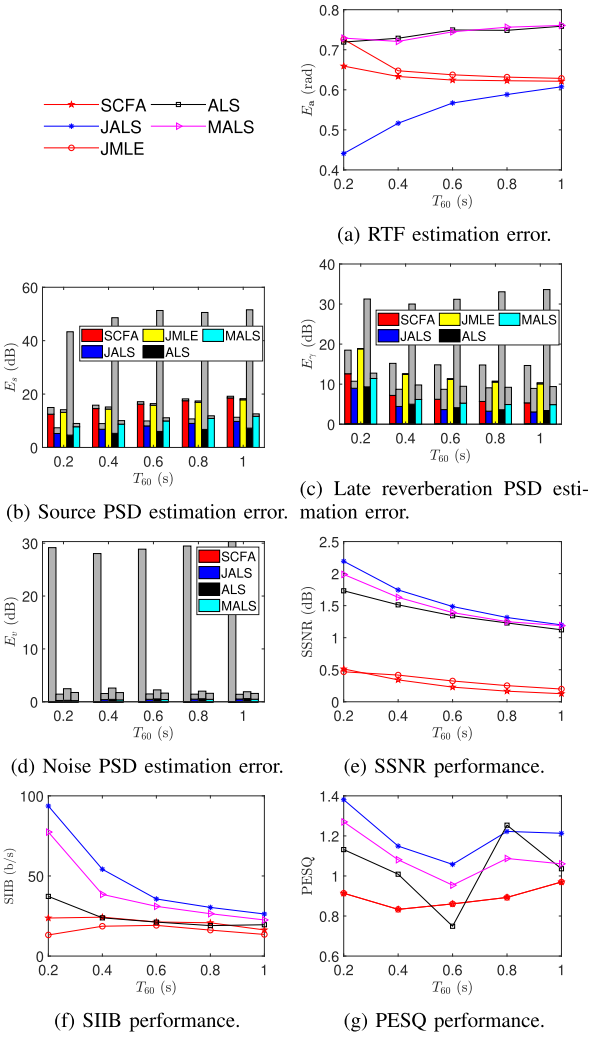
D. Experiment 2

1) Setup: In this experiment, we generate the RIRs using the image source method [41]. The dimension of the room is $7 \times 5 \times 4\ m$. In this simulated room, we have a single speaker, four microphones and a recorded wash machine noise from the ESC-50 database [42] as shown in Fig. 7. Note that we also added a white Gaussian noise to each microphone signal to simulate the microphone selfnoise at a SNR of 50 dB. The other settings are the same as those of Experiment 1. For ALS-based methods, we assume an ideal voice activity detector is used and the spatial coherence matrix of the ambient noise is calculated using the noise only time frame with the following equation

$$\mathbf{\Psi}_{i,j}(k) = \frac{\sum_{t_n=1}^{T_n} y_i(t_n,k)\, y_j(t_n,k)^*}{\sqrt{\left(\sum_{t_n=1}^{T_n} |y_i(t_n,k)|^2\right)\left(\sum_{t_n=1}^{T_n} |y_j(t_n,k)|^2\right)}}, \tag{62}$$

with $|x|$ the absolute value of x and $\{i, j\}$ the microphone indices. For SCFA, the spatial coherence matrix of the ambient noise is modeled as the identity matrix in [11]. For JMLE, the ambient noise is not considered. Hence, these two methods will have sever model mismatch errors in this experiment. Note that SCFA can be extended to handle spatial coherence matrices different from the identity matrix. However, it takes some effort to calculate the gradient and the Hessian matrix of the cost function and will not be addressed in this work.

2) Results: In Fig. 8, we compare all the methods while changing the reverberation time T_{60} of the RIRs from 0.2 s to 1 s. Each time segment contains 8 time frames. We can see that our proposed JALS method has the best performance in all the metrics evaluated. For the RTF estimation error in Fig. 8(a), the ALS-based methods ALS, MALS and JALS have degraded performance as T_{60} increases. However, SCFA and JMLE have improved performance. This is due to the model mismatch caused by the ambient noise component. When increasing T_{60},



(a) RTF estimation error.
(b) Source PSD estimation error.
(c) Late reverberation PSD estimation error.
(d) Noise PSD estimation error.
(e) SSNR performance.
(f) SIIB performance.
(g) PESQ performance.

Fig. 8. Performance vs T_{60}. In Figs (b), (c) and (d), the gray bars indicate the underestimation errors and the colored bars indicate overestimation errors.

the ratio between the correctly modeled late reverberation component and the incorrectly modeled ambient noise component becomes larger. For the PSDs estimation errors of the target source and the late reverberation in Fig. 8(b) and (c), SCFA and JMLE have large over estimation errors due to considering the ambient noise as the target source and the late reverberation. The ALS method still has the worst performance in Fig. 8(b) and (c). While, for the noise PSD estimation error in Fig. 8(d), SCFA has the worst performance due to erroneous spatial coherence matrix used. In Fig. 8(e) to (g), our proposed multiple time frames method JALS has improved performance over our single time frame method MALS, which both outperform all the other reference methods.

V. Concluding Remarks

We proposed alternating least square (ALS) based methods to estimate the RTFs, the PSDs of the source, the PSDs of the late reverberation, and the PSDs of the ambient noise jointly for a single reverberant and noisy scenario. We first modified an existing ALS method to obtain more accurate estimates using a

single time frame. Then, we extend the method to use multiple time frames that share the same RTF jointly. Furthermore, we imposed more robust constraints on the estimated PSDs. Experimental results demonstrated that the proposed methods achieve similar performance compared to the SCFA method in terms of estimation accuracy, noise reduction performance, speech quality, and speech intelligibility. The proposed methods outperform the existing ALS-based method and the JMLE method assuming a noiseless signal model, especially in low SNR scenarios.

Further studies can be conducted to extend the proposed methods to handle more complicated scenarios, such as multi-source signals.

Appendix A
Solution to (38)

We define $\hat{\mathbf{P}}_{\mathbf{x}}(l) = \hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\phi}_{\gamma}(l)\hat{\mathbf{\Gamma}} - \hat{\phi}_{v}(l)\hat{\mathbf{\Psi}}$ and reformulate (38) as

$$\begin{aligned}
&\arg\min_{\tilde{\mathbf{a}}} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_{\mathbf{x}}(l) - \hat{\tilde{\phi}}_{s}(l)\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H} \right\|_{F}^{2} \\
&= \arg\min_{\tilde{\mathbf{a}}} \sum_{l=1+(\beta-1)N}^{\beta N} \left[\left(\hat{\phi}_{s}(l)\tilde{\mathbf{a}}^{H}\tilde{\mathbf{a}} \right)^{2} - 2\hat{\tilde{\phi}}_{s}(l)\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{x}}(l)\tilde{\mathbf{a}} \right] \\
&= \arg\min_{\tilde{\mathbf{a}}} -2\tilde{\mathbf{a}}^{H} \left(\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_{s}(l)\hat{\mathbf{P}}_{\mathbf{x}}(l) \right) \tilde{\mathbf{a}} \\
&= \arg\max_{\tilde{\mathbf{a}}} \tilde{\mathbf{a}}^{H} \left(\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_{s}(l)\hat{\mathbf{P}}_{\mathbf{x}}(l) \right) \tilde{\mathbf{a}}, \qquad (63)
\end{aligned}$$

where we have used the fact that $\tilde{\mathbf{a}}^{H}\tilde{\mathbf{a}} = 1$. The solution for $\tilde{\mathbf{a}}$ is the principal eigenvector of $\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_{s}(l)\hat{\mathbf{P}}_{\mathbf{x}}(l)$.

Appendix B
Decision Flow for Updating $C_1, C_2, \hat{\tilde{\phi}}_s$ and $\hat{\phi}_{\gamma}$

We first update $C_1(l)$ and $C_2(l)$ by

$$C_1(l) = \begin{cases} \min\left\{ C_1(l-1), \frac{\hat{\tilde{\phi}}_{s}(l)}{\hat{\phi}_{\gamma}(l)} \right\} & \text{if} \hat{\tilde{\phi}}_{s}(l) > 0, \hat{\phi}_{\gamma}(l) > 0 \\ C_1(l-1) & \text{else.} \end{cases}, \qquad (64)$$

and

$$C_2(l) = \begin{cases} \min\left\{ C_2(l-1), \frac{\hat{\phi}_{\gamma}(l,k)}{\hat{\tilde{\phi}}_{s}(l)} \right\} & \text{if} \hat{\tilde{\phi}}_{s}(l) > 0, \hat{\phi}_{\gamma}(l) > 0 \\ C_2(l-1) & \text{else.} \end{cases} \qquad (65)$$

With $C_1(l)$ and $C_2(l)$, we update $\hat{\tilde{\phi}}_{s}(l)$ by

$$\hat{\tilde{\phi}}_{s}(l) = \begin{cases} \hat{\tilde{\phi}}_{s}(l) & \text{if} \hat{\tilde{\phi}}_{s}(l) > 0 \\ C_1(l)\hat{\phi}_{\gamma}(l) & \text{if} \hat{\phi}_{\gamma}(l) > 0, \hat{\tilde{\phi}}_{s}(l) \le 0 \\ \hat{\tilde{\phi}}_{s}^{\min}(l) & \text{if} \hat{\phi}_{\gamma}(l) \le 0, \hat{\tilde{\phi}}_{s}(l) \le 0 \end{cases} \qquad (66)$$

where $\hat{\tilde{\phi}}_{s}^{\min}(l)$ is calculated by

$$\hat{\tilde{\phi}}_{s}^{\min} = \min\left\{ \left| \frac{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}} - \hat{\mathbf{P}}_{\mathbf{y}_{m+1,m+1}}}{\left|\hat{\tilde{a}}_{m}\right|^{2} - \left|\hat{\tilde{a}}_{m+1}\right|^{2}} \right| \right\}_{m=1}^{M-1}, \qquad (67)$$

where we used the fact that $\mathbf{P}_{\mathbf{y}_{m,m}} = \tilde{\phi}_{s}|\tilde{a}_{m}|^{2} + \phi_{\gamma} + \phi_{v}$ for $m = 1, \ldots, M$ and $\mathbf{P}_{\mathbf{y}_{m,m}} - \mathbf{P}_{\mathbf{y}_{m+1,m+1}} = \tilde{\phi}_{s}(|\tilde{a}_{m}|^{2} - |\tilde{a}_{m+1}|^{2})$. Then, we update $\hat{\phi}_{\gamma}(l)$ by

$$\hat{\phi}_{\gamma}(l) = \begin{cases} \hat{\phi}_{\gamma}(l) & \text{if} \hat{\tilde{\phi}}_{\gamma}(l) > 0 \\ C_2(l)\hat{\tilde{\phi}}_{s}(l) & \text{if} \hat{\phi}_{s}(l) > 0, \hat{\tilde{\phi}}_{\gamma}(l) \le 0 \\ \hat{\phi}_{\gamma}^{\min}(l) & \text{if} \hat{\phi}_{s}(l) \le 0, \hat{\phi}_{\gamma}(l) \le 0 \end{cases} \qquad (68)$$

where $\hat{\phi}_{\gamma}^{\min}(l)$ is calculated by

$$\hat{\phi}_{\gamma}^{\min} = \min\left\{ \left| \hat{\mathbf{P}}_{\mathbf{y}_{m,m}} - \hat{\tilde{\phi}}_{s}^{\min}\left|\hat{\tilde{a}}_{m}\right|^{2} - \hat{\phi}_{v} \right| \right\}_{m=1}^{M}. \qquad (69)$$

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] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Cambridge, MA, USA: Academic Press, 2010.

[4] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, May 2014.

[5] 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.

[6] 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.

[7] 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, Lang. Process.*, vol. 25, no. 11, pp. 2209–2222, Nov. 2017.

[8] 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.

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

[10] 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.

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

[12] 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, Lang. Process.*, vol. 28, pp. 61–76, 2020.

[13] 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.

[14] 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, Lang. Process.*, vol. 31, pp. 695–705, 2023.
[15] A.-J. Van Der Veen, "Joint diagonalization via subspace fitting techniques," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2001, pp. 2773–2776.
[16] K. Rahbar and J. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 832–844, Sep. 2005.
[17] S. Degerine and E. Kane, "A comparative study of approximate joint diagonalization algorithms for blind source separation in presence of additive noise," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 3022–3031, Jun. 2007.
[18] 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.
[19] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.
[20] Y. Avargel and I. Cohen, "System identification in the Short-Time Fourier Transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.
[21] 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.
[22] 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.
[23] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
[24] H. Kuttruff, *Room Acoustics*. Boca Raton, FL, USA: CRC Press, 2016.
[25] 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.
[26] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep.. 1936.
[27] T. Söderström and P. Stoica, *System Identification*. Hoboken, NJ, USA: Prentice-Hall Int., 1989.
[28] 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.
[29] 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.
[30] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
[31] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement," in *Proc. Interspeech*, 2007, pp. 830–833.
[32] H. L. V. Trees, *Optimum Array Processing*. Hoboken, NJ, USA: Wiley, Mar. 2002.
[33] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2013.
[34] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.
[35] 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. 2017.
[36] 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.
[37] 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.
[38] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *Darpa Timit Acoustic-Phonetic Continuous Speech Corpus CD-ROM TIMIT*, Gaithersburg, MD, USA: NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Feb. 1, 1993.
[39] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, 2014, pp. 313–317.
[40] 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.
[41] 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.
[42] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd Annu. ACM Conf. Multimedia*, ACM Press, 2015, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390