This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario

Changheng Li, Jorge Martinez and Richard C. Hendriks

***Abstract*—Estimation of the acoustic-scene related parameters such as relative transfer functions (RTFs) from source to microphones, source power spectral densities (PSDs) and PSDs of the late reverberation is essential and also challenging. Existing maximum likelihood estimators typically consider only subsets of these parameters and use each time frame separately. In this paper we explicitly focus on the single source scenario and first propose a joint maximum likelihood estimator (MLE) to estimate all parameters jointly using a single time frame. Since the RTFs are typically invariant for a number of consecutive time frames we also propose a joint maximum likelihood estimator (MLE) using multiple time frames which has similar estimation performance compared to a recently proposed reference algorithm called simultaneously confirmatory factor analysis (SCFA), but at a much lower complexity. Moreover, we present experimental results which demonstrate that the estimation accuracy, together with the performance of noise reduction, speech quality and speech intelligibility, of our proposed joint MLE outperform those of existing MLE based approaches that use only a single time frame.**

***Index Terms*—Maximum likelihood estimation, dereverberation, microphone array signal processing, RTF estimation, PSD estimation.**

I. Introduction

Microphone array signal processing has ubiquitous applications like source dereverberation [1]–[4], noise reduction [5]–[8], source separation [9]–[11] and source localization [12]. These applications heavily depend on acoustic-scene related parameters such as relative transfer functions (RTFs), power spectral densities (PSDs) of the source, PSDs of the late reverberation and PSDs of the microphone self noise. These parameters are typically unknown in practical scenarios. Therefore, estimation of these parameters is an essential problem for microphone array signal processing applications.

As speech sources are typically non-stationary, their PSD changes over time. Moreover, the source might be moving, resulting in changes in the RTF as well. The estimation of the RTF and the PSDs of the source and the late reverberation is therefore rather challenging, especially when considering to estimate them simultaneously at low complexity. To get a full understanding of the problem, we constrain ourselves in this paper to the single source reverberant scenario and focus on the joint estimation of the source's RTF, PSD of the early reflections and the PSD of the late reverberation. In future work, we will extend this towards the multi-source scenario.

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

There are many existing methods that consider maximum likelihood estimation of these parameters [1], [13]–[16]. However, most of these methods do not estimate the parameters in a joint manner. In [1], [13], the RTFs are assumed to be known and the MLE for the PSDs of the source and the late reverberation is proposed. In [2], the estimate of the late reverberation is obtained without estimating the RTFs or the PSDs of the source. In [14], the RTFs are estimated given that the PSDs of the late reverberation are assumed to be known or have been estimated. In [15], by assuming the late reverberation is stationary, the expectation maximization (EM) method [17] was used to estimate the RTFs and the PSD of the source. However, in practice, the late reverberation is non-stationary and the PSDs of the late reverberation can change from time-frame to time-frame, which limits the scenarios to which the method in [15] can be applied.

Apart from the fact that most reference methods only estimate a subset of these parameters, all these methods, i.e., [1], [13]–[16], use each time frame separately. However, in most practical scenes, the RTFs change slower than the PSDs of the source and the late reverberation, and can be assumed invariant for a number of consecutive time frames. Therefore, better estimates of these parameters can be obtained by using the time frames that share the same RTFs jointly. A recently proposed method referred to as the simultaneous confirmatory factor analysis (SCFA) method considers the joint estimation of these parameters using multiple time frames [18] and has a much better estimation performance compared to methods using each time frame separately. However, since the problem formulated in [18] is non-convex, this method suffers from a rather high computational cost, which makes it difficult to be applied when dealing with practical problems.

To estimate all the aforementioned parameters of interest jointly and accurately with low computational complexity, we first propose a joint maximum likelihood estimator (MLE) using a single time frame. This has a closed form solution and can be solved efficiently. Note that recently the joint MLE using a single time frame is also proposed in [16], but we provide an alternative proof. More importantly, we propose an extension, which is a joint MLE using multiple time frames. This extension uses the rough estimates obtained by the MLE for a single time frame as initialisation and estimates all the parameters in an iterative manner. Since the computational cost for each step in the proposed method mainly comes from an eigenvalue decomposition, it has similar computational complexity as the MLE approach for a single time frame.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

Experimental results demonstrate that our proposed MLE for multiple time frames has similar estimation performance compared to the recently proposed SCFA method from [18], but, at a much lower computational complexity. Moreover, both the proposed and SCFA method outperform two other reference methods that consist of combining several existing state-of-the-art methods.

In the current work we thus focus on the single source scenario. In our recent work published in [19], we proposed a method that can jointly estimate the RTFs as well as the source PSDs for multiple simultaneously present sources using multiple time frames. The method proposed in [19] also has a much lower computational complexity compared to SCFA, while maintaining similar estimation performance. However, in [19], a non-reverberant environment is assumed. In the current work we therefore also consider the late reverberation components constrained to the single source scenario. In future work, we will consider the joint estimation of these parameters for a combination of the two scenarios (i.e., multiple simultaneously present sources in a reverberant environment).

The remaining parts of the paper are structured as follows. We present the notation, the signal model and the main goal of this paper in Section II. In Section III, we propose the joint maximum likelihood estimator using a single time frame in Section III-A and using multiple time frames in Section III-B. In Section IV, we first introduce some reference methods and compare them to our proposed joint MLE in different acoustic scenarios. In the last section, Section V, conclusions will be drawn.

II. Preliminaries

A. Notation

In this paper, we denote scalars using lower-case letters, vectors using bold-face lower-case letters and matrices using bold-face upper-case letters (in some cases with subscripts using bold-face lower-case letters, e.g. $\mathbf{P}_{\mathbf{y}}$). Matrix notation with subscripts using two lower-case letters (e.g. $\mathbf{P}_{\mathbf{y}\,i,j}$) denotes the element of the matrix. $\Re\left(\cdot\right)$ and $\Im\left(\cdot\right)$ represents the real part and the imaginary part of a complex-valued variable, respectively. Further, $\mathrm{E}\left(\cdot\right)$ denotes the expected value of a random variable, $\mathrm{tr}\left(\cdot\right)$ denotes the trace of a matrix, and if not further specified, $|\cdot|$ denotes the determinant of a matrix. Finally, $\mathrm{diag}\left[a_1,\cdots,a_M\right]$ denotes a diagonal matrix with diagonal elements $a_1,\cdots,a_M$ and $\|\cdot\|_2$ denotes the Frobenius norm of a matrix.

B. Signal model

We consider a single acoustic point source observed by a microphone array consisting of M microphones with an arbitrary geometric structure in a reverberant and noisy environment. Decomposing the signal into its direct component with its early reflections, and the late reverberant components, we can write the signal received at the m_{th} microphone in the short-time Fourier transform (STFT) domain as

$$y_m\left(i,k\right)=e_m\left(i,k\right)+l_m\left(i,k\right)+v_m\left(i,k\right), \tag{1}$$

where i is the time-frame index and k is the frequency bin index, $e_m\left(i,k\right)$ is the sum of the direct components and the early reflections, $l_m\left(i,k\right)$ is the sum of all late reflections and $v_m\left(i,k\right)$ is the microphone self-noise. The direct components and early reflections are beneficial for speech intelligibility [20]. The combination of these components, denoted by $e_m\left(i,k\right)$ in Eq. (1), forms our target signal. In this work, we differentiate between time segments (indexed by β) and time frames (indexed by i). Each time segment consists of N time frames, i.e., for each β, $i=\left(\beta-1\right)N+1,\cdots,\beta N$. The target signal at the m_{th} microphone is given by

$$e_m\left(i,k\right)=a_m\left(\beta,k\right)s\left(i,k\right), \tag{2}$$

where $a_m\left(\beta,k\right)$ is the relative transfer function (RTF) for source s from the reference location to the m_{th} microphone in time segment β and s is the target source including direct and early reflections at the reference microphone. Note that, for ease of analyzing, we use the multiplicative transfer function (MTF) approximation instead of the convolutive transfer function (CTF) approximation in Eq. (2). CTF can be more accurate than MTF but has a more complicated signal model [21], [22]. We assume that the RTFs are constant during a time segment (thus during multiple time frames that fall in one segment) and $a_1=1$, which means that the first microphone is selected as the reference microphone. Stacking the M microphone STFT coefficients into a column vector, we have

$$\mathbf{y}\left(i,k\right)=\mathbf{a}\left(\beta,k\right)s\left(i,k\right)+\mathbf{l}\left(i,k\right)+\mathbf{v}\left(i,k\right)\in\mathbb{C}^{M\times1}. \tag{3}$$

C. Cross Power Spectral Density Matrices

We assume the STFT coefficients of the microphone signal have a circularly-symmetric complex Gaussian distribution[1], i.e.: $\mathbf{y}\left(i,k\right)\sim\mathcal{N}_C\left(\mathbf{0},\mathbf{P}_{\mathbf{y}}\left(i,k\right)\right)$, where $\mathbf{P}_{\mathbf{y}}\left(i,k\right)$ is the noisy cross power spectral density (CPSD) matrix, expressing the covariance across microphones. Assuming that all components in Eq. (3) are mutually uncorrelated, we have

$$\mathbf{P}_{\mathbf{y}}\left(i,k\right)=\mathbf{P}_{\mathbf{e}}\left(i,k\right)+\mathbf{P}_{\mathbf{l}}\left(i,k\right)+\mathbf{P}_{\mathbf{v}}\left(i,k\right)\in\mathbb{C}^{M\times M}, \tag{4}$$

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

$$\mathbf{P}_{\mathbf{e}}\left(i,k\right)=p\left(i,k\right)\mathbf{a}\left(\beta,k\right)\mathbf{a}^H\left(\beta,k\right), \tag{5}$$

and where $p\left(i,k\right)=\mathrm{E}\left[\left|s\left(i,k\right)\right|^2\right]$ is the power spectral density (PSD) of the source at the reference microphone with $|\cdot|$ the absolute value. Note that although the mutual uncorrelation assumption is commonly used, these components are not perfectly uncorrelated in practice.

The CPSD matrix of the late reverberation component is commonly modelled as [1], [27]

$$\mathbf{P}_{\mathbf{l}}\left(i,k\right)=\gamma\left(i,k\right)\mathbf{\Gamma}\left(k\right), \tag{6}$$

where the time-varying coefficient $\gamma\left(i,k\right)$ is the PSD of the late reverberation and the time-invariant matrix $\mathbf{\Gamma}\left(k\right)$ is the

[1]Although a super-Gaussian distribution can better model the coefficients [23]–[25], the estimators based on it are much more cumbersome than that based on the Gaussian distribution [26] and hence are not considered in this paper.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

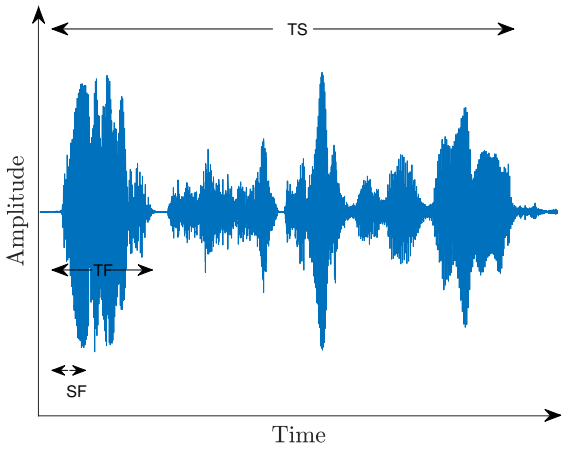

Figure 1: Time segment (TS), time frames (TF) and sub frames (SF).

spatial coherence matrix of the late reverberation. $\mathbf{\Gamma}(k)$ is assumed to be non-singular and known in this paper. Several methods have been proposed to measure $\mathbf{\Gamma}(k)$ by using pre-calculated room impulse responses [28] or by using knowledge on the microphone array geometry [29], [30]. We use the latter one and model the coherence matrix as a spherically isotropic noise field [31]

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

where $\operatorname{sinc}(x) = \frac{\sin x}{x}$, $d_{i,j}$ 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, the microphone self-noise component is assumed to have slow varying statistics and its CPSD matrix $\mathbf{P_v}(i,k)$ can be modelled as a time-invariant diagonal matrix with its M diagonal elements being the PSD of the self noise corresponding to the M microphones

$$\mathbf{P_v}(k) = \operatorname{diag}\left[n_1(k), \cdots, n_M(k)\right]. \tag{8}$$

Due to its time-invariant property, a voice activity detector (VAD) can be used to detect the noise-only segments of the signal such that the covariance matrix of the noise can be estimated [32]. Moreover, the power of the microphone self-noise is usually very small compared to the other components. Therefore, we assume in this paper that $\mathbf{P_v}(k)$ is neglectable or can be subtracted from the noisy covariance matrix.

D. Problem Formulation

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

$$\mathbf{P_y}(i,k) = p(i,k)\mathbf{a}(\beta,k)\mathbf{a}^H(\beta,k) + \gamma(i,k)\mathbf{\Gamma}(k). \tag{9}$$

Each time frame i consists of T_{sf} overlapping sub frames indexed by t_s, each with equal length N_s. For a visual interpretation of time segments, frames and sub frames see Figure 1. Assuming the noisy signal is 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,

$$\hat{\mathbf{P}}_{\mathbf{y}}(i,k) = \frac{1}{T_{sf}}\sum_{t_s=1}^{T_{sf}}\mathbf{y}(t_s,k)\,\mathbf{y}(t_s,k)^H, \tag{10}$$

where $\mathbf{y}(t_s,k)$ denotes the STFT coefficients vector and the FFT length is selected as $2^{\lceil \log_2 N_s \rceil}$, where $\lceil\cdot\rceil$ denotes taking the next highest integer. Note that each time frame contains multiple sub-time frames as illustrated in Figure 1 and these sub-time frames are used to estimate the covariance matrix of a single time frame. Notice that across the time frames of one time segment, the RTF vector is assumed to be constant and the PSDs of the source and late reverberation power $\gamma(i,k)$ are assumed to be time-variant.

Accurate estimation of the parameters from the signal model in Eq. (9) 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 paper therefore is to estimate the RTF vector, the PSD of the source and the PSD of the late reverberation simultaneously using N estimated CPSD matrices $\hat{\mathbf{P}}_{\mathbf{y}}(i,k)$ for $i = 1, \cdots, N$, while the source is only stationary within a time frame and the RTF changes from segment to segment. Since we process the signal for each frequency bin independently, we omit the frequency bin index k in the following sections for notational convenience.

III. Joint MLE

In this work, we present a novel maximum likelihood estimator (MLE) to jointly estimate the parameters from the signal model in Eq. (9). Note that MLEs have been proposed before in this context [1], [13], [15], but typically they assume that the RTF vector $\mathbf{a}$ is known and only determine the MLEs of $p(i)$ and $\gamma(i)$ for each time frame i separately. We will first in Section III-A propose the joint MLE estimator of $p(i), \mathbf{a}$ and $\gamma(i)$ using the estimated noisy CPSD matrix for a single time frame. Since the CPSD matrices for multiple time frames in a single time segment share the same RTF vector, we can use these matrices jointly to obtain a better estimate of $\mathbf{a}$. Therefore, we will also propose in Section III-B the joint MLE estimator of $p(i), \mathbf{a}$ and $\gamma(i)$ using the CPSD matrices for multiple time frames.

A. Joint MLE for a single time frame

Assuming that the T_{sf} sub-time frames in a single time frame i per frequency band k are independent and identically distributed (i.i.d.), we can write the joint PDF $f(\mathbf{y}(1,k), \cdots, \mathbf{y}(T_{sf},k))$ as

$$f(\mathbf{y}(1,k), \cdots, \mathbf{y}(T_{sf},k)) = \left(\frac{\exp\left[-\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right)\right]}{\pi^M\left|\mathbf{P}_{\mathbf{y}}\right|}\right)^{T_{sf}}, \tag{11}$$

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

where $\hat{\mathbf{P}}_{\mathbf{y}}$ is given in Eq. (10) and $\mathbf{P}_{\mathbf{y}}$ in Eq. (9). The negative log-likelihood function with respect to (w.r.t.) $p, \mathbf{a}$ and γ is given by

$$-L\left(p, \mathbf{a}, \gamma\right) = T_{sf}\left[\log\left|\mathbf{P}_{\mathbf{y}}\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right)\right], \tag{12}$$

where the additive constant term $T_{sf}M\log\pi$ has been omitted as it is irrelevant for the parameters of interest. The MLEs of $p, \mathbf{a}$ and γ are given by minimizing the cost function in Eq. (12), i.e.,

$$\underset{p,\mathbf{a},\gamma}{\arg\min}\log\left|\mathbf{P}_{\mathbf{y}}\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right). \tag{13}$$

To solve this problem, we reparameterize the signal model in Eq. (9) as

$$\begin{aligned}\mathbf{P}_{\mathbf{y}} &= p\mathbf{a}\mathbf{a}^{H} + \gamma\mathbf{\Gamma}\\ &= \mathbf{L}\left(p\mathbf{L}^{-1}\mathbf{a}\mathbf{a}^{H}\mathbf{L}^{-H} + \gamma\mathbf{I}\right)\mathbf{L}^{H}\\ &= \mathbf{L}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H} + \gamma\mathbf{I}\right)\mathbf{L}^{H},\end{aligned} \tag{14}$$

where $\mathbf{L}$ is the Cholesky factor of $\mathbf{\Gamma}$ (i.e. $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^{\mathbf{H}}$), $\tilde{\mathbf{a}} = \frac{\mathbf{L}^{-1}\mathbf{a}}{\sqrt{\mathbf{a}^{H}\mathbf{\Gamma}^{-1}\mathbf{a}}}$ and $\tilde{p} = p\mathbf{a}^{H}\mathbf{\Gamma}^{-1}\mathbf{a}$. Therefore, the optimization problem in Eq. (13) can be cast as

$$\underset{\tilde{p},\tilde{\mathbf{a}},\gamma}{\arg\min}\log\left|\mathbf{P}_{\mathbf{y}}\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right). \tag{15}$$

By using this reparameterization, we can make the estimation of $\tilde{\mathbf{a}}$ independent of the estimation of $\tilde{p}$ and γ. Therefore, the joint estimation of these parameters can be decomposed into two simpler estimation steps, as we will show below.

The first term in Eq. (15) can be rewritten as

$$\begin{aligned}\log\left|\mathbf{P}_{\mathbf{y}}\right| &= \log\left|\mathbf{L}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H} + \gamma\mathbf{I}\right)\mathbf{L}^{H}\right|\\ &= \log\left(\left|\mathbf{L}\right|\left(\tilde{p}\tilde{\mathbf{a}}^{H}\tilde{\mathbf{a}} + \gamma\right)\gamma^{M-1}\left|\mathbf{L}^{H}\right|\right)\\ &= \log\left(\left|\mathbf{\Gamma}\right|\right) + \log\left(\tilde{p} + \gamma\right) + \left(M-1\right)\log\left(\gamma\right),\end{aligned} \tag{16}$$

where we have used the fact that $\tilde{\mathbf{a}}^{H}\tilde{\mathbf{a}} = 1$. The second term in Eq. (15) can be rewritten as

$$\begin{aligned}\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right) &= \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left[\mathbf{L}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H} + \gamma\mathbf{I}\right)\mathbf{L}^{H}\right]^{-1}\right)\\ &= \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H} + \gamma\mathbf{I}\right)^{-1}\right)\\ &= \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(\gamma^{-1}\mathbf{I} - \frac{\gamma^{-2}\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H}}{1+\gamma^{-1}\tilde{p}\tilde{\mathbf{a}}^{H}\tilde{\mathbf{a}}}\right)\right)\\ &= \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_{\mathbf{w}}\right) - \operatorname{tr}\left(\frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\hat{\mathbf{P}}_{\mathbf{w}}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H}\right)\\ &= \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_{\mathbf{w}}\right) - \frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\tilde{\mathbf{a}},\end{aligned} \tag{17}$$

where $\hat{\mathbf{P}}_{\mathbf{w}} = \mathbf{L}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{L}^{-H}$ and the Sherman–Morrison formula [33] is used to calculate $\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^{H} + \gamma\mathbf{I}\right)^{-1}$.

Substituting Eq. (16) and Eq. (17) in Eq. (15) and omitting the constant irrelevant term $\log\left(\left|\mathbf{\Gamma}\right|\right)$, the cost function from Eq. (13) can eventually thus be expressed in the following useful form,

$$\begin{aligned}\underset{\tilde{p},\tilde{\mathbf{a}},\gamma}{\arg\min}&\log\left(\tilde{p} + \gamma\right)\left(\gamma^{M-1}\right) + \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_{\mathbf{w}}\right)\\ &- \frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\tilde{\mathbf{a}}.\end{aligned} \tag{18}$$

Since only the last term in Eq. (18) depends on $\tilde{\mathbf{a}}$ and $\frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}} > 0$, the estimate of $\tilde{\mathbf{a}}$ can be obtained by solving

$$\underset{\tilde{\mathbf{a}}}{\arg\max}\,\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\tilde{\mathbf{a}}. \tag{19}$$

The solution of Eq. (19) is known as the principal eigenvector of $\hat{\mathbf{P}}_{\mathbf{w}}$ and the optimum value of $\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\tilde{\mathbf{a}}$ is the principal eigenvalue $\lambda_{\max}$ of $\hat{\mathbf{P}}_{\mathbf{w}}$.

Substituting the optimal $\tilde{\mathbf{a}}$ from Eq. (19) in Eq. (18), we can find the estimates of $\tilde{p}$ and γ by solving

$$\begin{aligned}\underset{\tilde{p},\gamma}{\arg\min}\, f =& \log\left[\left(\tilde{p} + \gamma\right)\gamma^{M-1}\right] + \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_{\mathbf{w}}\right)\\ &- \frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\lambda_{\max}.\end{aligned} \tag{20}$$

Taking the partial derivatives of the cost function in Eq. (20) w.r.t. $\tilde{p}$ and γ and setting them equal to zero, respectively, we obtain

$$\begin{aligned}\frac{\partial f}{\partial\gamma} =& \frac{1}{\tilde{p}+\gamma} + \frac{M-1}{\gamma} - \frac{\operatorname{tr}\left(\mathbf{L}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{L}^{-H}\right)}{\gamma^{2}}\\ &+ \frac{\tilde{p}\left(2\gamma+\tilde{p}\right)}{\left(\gamma^{2}+\gamma\tilde{p}\right)^{2}}\lambda_{\max} = 0\end{aligned} \tag{21}$$

and

$$\frac{\partial f}{\partial\tilde{p}} = \frac{1}{\tilde{p}+\gamma} - \frac{\lambda_{\max}}{\left(\gamma+\tilde{p}\right)^{2}} = 0. \tag{22}$$

Solving Eq. (21) and Eq. (22) for $\tilde{p}$ and γ, we obtain

$$\hat{\tilde{p}} = \frac{M\lambda_{\max} - \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)}{M-1}, \tag{23}$$

$$\hat{\gamma} = \frac{\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right) - \lambda_{\max}}{M-1}. \tag{24}$$

To show that $\left(\hat{\tilde{p}}, \hat{\gamma}\right)$ is the minimum point of function f, we derive its second order derivatives

$$\begin{aligned}\frac{\partial^{2} f}{\partial\gamma^{2}} =& -\frac{1}{\left(\tilde{p}+\gamma\right)^{2}} - \frac{M-1}{\gamma^{2}} + \frac{2\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)}{\gamma^{3}}\\ &+ \frac{2\lambda_{\max}\left(-3\gamma^{2}\tilde{p} - 3\gamma\tilde{p}^{2} - \tilde{p}^{3}\right)}{\gamma^{3}\left(\gamma+\tilde{p}\right)^{3}},\end{aligned} \tag{25}$$

$$\frac{\partial^{2} f}{\partial\gamma\partial\tilde{p}} = -\frac{1}{\left(\tilde{p}+\gamma\right)^{2}} + \frac{2\lambda_{\max}}{\left(\gamma+\tilde{p}\right)^{3}}, \tag{26}$$

$$\frac{\partial^{2} f}{\partial\tilde{p}^{2}} = -\frac{1}{\left(\tilde{p}+\gamma\right)^{2}} + \frac{2\lambda_{\max}}{\left(\gamma+\tilde{p}\right)^{3}}. \tag{27}$$

At point $\left(\hat{\tilde{p}}, \hat{\gamma}\right)$, we have

$$\left.\frac{\partial^{2} f}{\partial\gamma^{2}}\right|_{\gamma=\hat{\gamma}} = \frac{\left(M-1\right)^{3}}{\left(\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right) - \lambda_{\max}\right)^{2}} + \frac{1}{\left(\lambda_{\max}\right)^{2}} > 0, \tag{28}$$

$$\left.\frac{\partial^{2} f}{\partial\tilde{p}^{2}}\right|_{\tilde{p}=\hat{\tilde{p}}} = \frac{1}{\left(\lambda_{\max}\right)^{2}} > 0, \tag{29}$$

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

$$\frac{\partial^2 f}{\partial \gamma^2}\frac{\partial^2 f}{\partial \tilde{p}^2}-\left(\frac{\partial^2 f}{\partial \gamma \partial \tilde{p}}\right)^2\Bigg|_{\substack{\gamma=\hat{\gamma}\\ \tilde{p}=\hat{\tilde{p}}}}=\frac{(M-1)^3/(\lambda_{\max})^2}{\left(\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)-\lambda_{\max}\right)^2}>0. \tag{30}$$

Furthermore, we can show that $\hat{\tilde{p}}, \hat{\gamma}$ are both positive such that they can be used as the estimates of $\tilde{p}$ and γ. Since $\hat{\mathbf{P}}_{\mathbf{w}}$ is a positive definite matrix, we have $\frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)}{M}<\lambda_{\max}<\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)$. Hence from Eqs. (23) and (24) it follows that $\hat{\tilde{p}}>0$ and $\hat{\gamma}>0$. Note that this examination of the Hessian matrix and $\hat{\tilde{p}}, \hat{\gamma}$ being positive is absent in [16].

Finally, we obtain the optimal estimates of p and $\mathbf{a}$ using the estimated $\hat{\tilde{p}}$ and $\hat{\tilde{\mathbf{a}}}$ by setting

$$\hat{\mathbf{a}}=\mathrm{No}\left(\mathbf{L}\hat{\tilde{\mathbf{a}}}\right) \tag{31}$$

and

$$\hat{p}=\frac{\hat{\tilde{p}}}{\hat{\mathbf{a}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}}, \tag{32}$$

where $\mathrm{No}\left(\mathbf{x}\right)$ means taking normalization w.r.t. the first element of $\mathbf{x}$.

As mentioned in [16], the estimation of $\mathbf{a}$ is consistent with the covariance whitening method [5], [14], while we provided an alternative proof that this estimate equals the MLE of $\mathbf{a}$. More specifically, the proof in [16] with respect to the estimation of the PSDs does not include the examination of the Hessian matrix and the estimates of the PSDs being positive. This examination of the Hessian matrix being positive definite is necessary, since setting the partial derivative to zero does not give us the optimal estimate when the Hessian matrix is not positive definite. Also, the examination of estimates of the PSDs being positive is necessary, since the PSDs should always be positive. Moreover, The proof in [16] is based on the proportion of the likelihood function, which makes it difficult to analyze the cost function for multiple time frames. While, in this work, our proof is based on the likelihood function itself and the extension to multiple time frames is straightforward.

B. Joint MLE for multiple time frames

In the previous subsection we considered the joint MLE for p, γ and $\mathbf{a}$ given a single time frame. As $\mathbf{a}$ is assumed to stay fixed across multiple frames in a segment, we consider in this subsection the joint ML optimal estimates of $p\left(i\right)$, $\gamma\left(i\right)$ for $i=1,\cdots,N$ and $\mathbf{a}$ using all time-frames in a segment.

Assuming that the N time frames are independent, we can write the negative log likelihood function of the STFT coefficients as

$$L=-\sum_{i=1}^{N}T_{sf}\left[\log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right|+\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right)\right], \tag{33}$$

where non-essential constant terms have been omitted. The joint MLEs for $p\left(i\right)$, $\gamma\left(i\right)\forall i=1,\cdots,N$ and $\mathbf{a}$ are the solution to the optimization problem

$$\underset{p(i),\mathbf{a},\gamma(i)}{\arg\min}\sum_{i=1}^{N}\log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right|+\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right). \tag{34}$$

By reparameterizing the signal model in a similar way as in the previous subsection, i.e., using $\tilde{\mathbf{a}}=\frac{\mathbf{L}^{-1}\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}}}$ and $\tilde{p}=p\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}$, the CPSD matrix for each time frame i has the form

$$\mathbf{P}_{\mathbf{y}}\left(i\right)=\mathbf{L}\left(\tilde{p}\left(i\right)\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H+\gamma\left(i\right)\mathbf{I}\right)\mathbf{L}^H, \tag{35}$$

and the optimization problem in Eq. (34) can be cast as

$$\underset{\tilde{p}(i),\tilde{\mathbf{a}},\gamma(i)}{\arg\min}\sum_{i=1}^{N}\log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right|+\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right). \tag{36}$$

Substituting Eq. (16) and Eq. (17) in Eq. (36) and omitting the irrelevant constant terms, the cost function can be expressed as

$$\begin{aligned}\underset{\tilde{p}(i),\tilde{\mathbf{a}},\gamma(i)}{\arg\min}\sum_{i=1}^{N}&\log\left[\left(\tilde{p}\left(i\right)+\gamma\left(i\right)\right)\left(\gamma\left(i\right)^{M-1}\right)\right]\\&+\mathrm{tr}\left(\gamma\left(i\right)^{-1}\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\right)\\&-\frac{\gamma\left(i\right)^{-2}\tilde{p}\left(i\right)}{1+\gamma\left(i\right)^{-1}\tilde{p}\left(i\right)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\tilde{\mathbf{a}},\end{aligned} \tag{37}$$

where similar manipulations have been carried out as in Eq. (18).

To estimate $\tilde{\mathbf{a}}$, we can focus on the last term of Eq. (37). Hence, the estimation of $\tilde{\mathbf{a}}$ is the solution of the following optimization problem

$$\underset{\tilde{\mathbf{a}}}{\arg\max}\sum_{i=1}^{N}\left(\frac{\tilde{p}\left(i\right)}{\gamma\left(i\right)+\tilde{p}\left(i\right)}\frac{1}{\gamma\left(i\right)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\tilde{\mathbf{a}}\right), \tag{38}$$

which is the principal eigenvector of the matrix

$$\sum_{i=1}^{N}\frac{\tilde{p}\left(i\right)}{\gamma\left(i\right)+\tilde{p}\left(i\right)}\frac{1}{\gamma\left(i\right)}\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right). \tag{39}$$

Note that unlike the estimation of $\tilde{\mathbf{a}}$ in a single time frame case where the estimate is the principal eigenvector of $\hat{\mathbf{P}}_{\mathbf{w}}$, the estimate is now the principal eigenvector of a weighted sum of the whitened CPSD matrices for all time frames and the weights depend on the estimation of $\tilde{p}\left(i\right)$ and $\gamma\left(i\right)$ for $i=1,\cdots,N$. Therefore, a closed form solution to Eq. (38) does not exist and we propose a recursive estimation approach.

For the first step, we estimate the parameters for each time frame independently using the method proposed in Section III-A. In this case, we will obtain N different estimates of the RTF vector, say, $\hat{\tilde{\mathbf{a}}}\left(i\right)$, which is the principal eigenvector of $\mathbf{L}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{L}^{-H}$ per frame i. Given $\hat{\tilde{\mathbf{a}}}\left(i\right)$ for a single frame i, the estimates of $\tilde{p}\left(i\right)$ and $\gamma\left(i\right)$ are obviously identical to expressions in Eq. (23) and Eq. (24), that is,

$$\hat{\tilde{p}}\left(i\right)=\frac{M\lambda_{\max}\left(i\right)-\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\right)}{M-1}, \tag{40}$$

$$\hat{\gamma}\left(i\right)=\frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\right)-\lambda_{\max}\left(i\right)}{M-1}, \tag{41}$$

where $\lambda_{\max}\left(i\right)$ is the principal eigenvalue of $\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)$.

For the second step, we use the initial estimates of $\tilde{p}\left(i\right)$ and $\gamma\left(i\right)$ to calculate the matrix in Eq. (39) and then use its principal eigenvector as the estimate of the RTF vector $\hat{\tilde{\mathbf{a}}}$. Next, we use the estimated $\hat{\tilde{\mathbf{a}}}$ in Eq. (37) and find new update estimates of $\tilde{p}\left(i\right)$ and $\gamma\left(i\right)$ based on the estimate $\hat{\tilde{\mathbf{a}}}$ which was

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

found using the joint information across all time frames in a segment. That is,

$$\hat{\tilde{p}}\left(i\right)=\frac{M\hat{\tilde{\mathbf{a}}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\hat{\tilde{\mathbf{a}}}-\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\right)}{M-1} \tag{42}$$

and

$$\hat{\gamma}\left(i\right)=\frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\right)-\hat{\tilde{\mathbf{a}}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\hat{\tilde{\mathbf{a}}}}{M-1}. \tag{43}$$

Note that $\hat{\tilde{\mathbf{a}}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\hat{\tilde{\mathbf{a}}}\leq\lambda_{\max}\left(i\right)<\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\right)$, hence $\hat{\gamma}\left(i\right)>0$. But $\hat{\tilde{p}}\left(i\right)$ in Eq. (42) can become negative. We replace these negative values using the initial estimates from Eq. (40) and store their corresponding time frame indices as index set G, which will not be included when calculating the weighted sum in Eq. (39) to estimate the RTF vector in the next step.

In the remaining steps, we repeat the second step until the relative change of $\hat{\tilde{\mathbf{a}}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}\left(i\right)\hat{\tilde{\mathbf{a}}}$ between the current iteration and the last iteration does not exceed a certain number ϵ, or a certain number of iterations has been executed.

C. Robust parameter estimation

In [18], it has been shown that linear inequality constraints on the parameters of interest can be used to improve the robustness of the estimation. Herein, we introduce these constraints on the RTF, the PSD of source and the PSD of the late reverberation. Note that, after obtaining estimates in each step of our proposed method, we can project the estimates into the constraint intervals introduced below. These constraints can effectively avoid large underestimation or overestimation errors and therefore can improve the robustness of our proposed joint MLE for multiple time frames.

1) Constraints for the RTFs: Considering only the direct path component, the anechoic acoustic transfer function (ATF) has the following equation [34]

$$\bar{a}_{i}=\frac{1}{4\pi d_{i}}\exp\left(-\frac{j2\pi kd_{i}}{Kc}\right), \tag{44}$$

where c denotes the sound speed, K is the FFT length and d_i is the distance between the source and the i_{th} microphone ($d_i>0$). The RTF in the k_{th} frequency bin is then given by (with the first microphone selected as the reference microphone)

$$a_{i}\left(k\right)=\frac{d_{1}}{d_{i}}\exp\left(-\frac{j2\pi k(d_{i}-d_{1})}{Kc}\right). \tag{45}$$

Using Eq. (45), for any frequency bin, a tight bound for both the real and imaginary parts of a_i is given by

$$-\frac{d_{1}}{d_{i}}\leq\Re\left(a_{i}\right),\Im\left(a_{i}\right)\leq\frac{d_{1}}{d_{i}}. \tag{46}$$

When not only the direct path component but also the early reflections are considered, the RTF value might exceed the tight bound above and we need to use a looser bound. Observing $d_1\leq d_{1,i}+d_i$ ($d_{1,i}$ is the distance between the first microphone and the i_{th} microphone) and assuming $d_i\geq d_{\max}$ (i.e. the distance between the source and each microphone is not smaller than a given small value $d_{\max}$), a looser bound for RTFs is

$$-\frac{d_{1,i}+d_{\max}}{d_{\max}}\leq\Re\left(a_{i}\right),\Im\left(a_{i}\right)\leq\frac{d_{1,i}+d_{\max}}{d_{\max}}. \tag{47}$$

Note that after obtaining $\hat{\tilde{\mathbf{a}}}$ at each step in our proposed method, we first normalize it with its first element to estimate the RTF vector $\hat{\mathbf{a}}$ and then project the estimated RTF vector into the interval $\left[-\frac{d_{1,i}+d_{\max}}{d_{\max}},\frac{d_{1,i}+d_{\max}}{d_{\max}}\right]$. Finally, we calculate the reparameterized vector using $\hat{\tilde{\mathbf{a}}}=\frac{\mathbf{L}^{-1}\hat{\mathbf{a}}}{\sqrt{\hat{\mathbf{a}}^{H}\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}}}$.

2) Constraints for the source PSDs: In Eq. (9), using the fact that $a_1=1$ and $\mathbf{\Gamma}_{1,1}=1$, we have

$$\mathbf{P}_{\mathbf{y}_{1,1}}\left(i\right)=p\left(i\right)+\gamma\left(i\right). \tag{48}$$

Hence, an upper bound for $p\left(i\right)$, by using a prefixed constant δ (with $\delta\geq1$), is found as

$$p\left(i\right)\leq\delta\mathbf{P}_{\mathbf{y}_{1,1}}\left(i\right)-\gamma\left(i\right), \tag{49}$$

and the upper bound for the reparametrized parameter $\tilde{p}\left(i,k\right)$ is

$$\tilde{p}\left(i\right)\leq\left[\delta\mathbf{P}_{\mathbf{y}_{1,1}}\left(i\right)-\gamma\left(i\right)\right]\mathbf{a}^{H}\mathbf{\Gamma}^{-1}\mathbf{a}. \tag{50}$$

3) Constraints for the late reverberation PSDs: As shown in [18], the following constraints can be applied to ensure better speech intelligibility performance by reducing overestimation errors on the PSD of the late reverberation [3], [35]

$$\gamma\leq\min\left[\mathrm{diag}\left(\mathbf{P}_{\mathbf{y}}\left(i\right)\right)\right]. \tag{51}$$

Since $\mathbf{\Gamma}_{m,m}=1$ for $m=1,\cdots,M$, we have

$$\mathbf{P}_{\mathbf{y}_{m,m}}\left(i\right)=p\left(i\right)a_{m}a_{m}^{H}+\gamma\left(i\right), \tag{52}$$

where $p\left(i\right)a_{m}{a_{m}}^{H}$ is positive. Hence we have $\mathbf{P}_{\mathbf{y}_{m,m}}\left(i\right)\geq\gamma\left(i\right)$ for all m and Eq. (51) holds.

IV. Experiments

In this section, we evaluate the estimation performance of the proposed methods as well as the performance on noise reduction, speech quality and speech intelligibility. We will first introduce the reference methods in Section IV-A and the evaluation metrics in Section IV-B. Then, in Section IV-C, we consider a static source scenario and use the simulated room impulse responses (RIRs) to construct the microphone signals. At last, in Section IV-D, we consider both the static source scenario and the source-moving scenario and use the RIRs recorded in real life from [36].

A. Reference methods

1) Combination of existing methods: The first reference method we consider utilizes several existing methods [2], [13], [14] to estimate the PSD of the late reverberation, the RTF vector and the PSD of the source successively. First, by assuming a noiseless or high SNR scenario, we use the eigenvalue decomposition-based method proposed in [2] to estimate the PSD of the late reverberation. With this estimate, we use the covariance whitening method in [14] to estimate the RTF vector. Finally, we use the method proposed in [13]

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

to estimate the PSD of the source. Note that although this reference method is a combination of existing state-of-the-art methods, this combination has the same estimation steps as the joint MLE estimator for a single time frame presented in Section III-A. Note also that this reference method only considers using the CPSD matrix for a single time frame. Therefore, when dealing with multiple time frames in one time segment, we can either use it to estimate parameters for all time frames independently or averaging the CPSD matrices for all time frames in a time segment and use it to estimate parameters with this averaged CPSD matrix. For convenience, we refer to this first case as 'Ref1' and the second case as 'Ref2' in each figure.

2) Simultaneous confirmatory factor analysis: The recently published method in [18] is also used for comparison in all the experiments. This method is based on confirmatory factor analysis (CFA) and non-orthogonal joint diagonalization principles and, hence, is called the simultaneous confirmatory factor analysis (SCFA) method. Note that the SCFA method is very accurate and can estimate the RTF matrix, the PSDs of the early components of the sources, the PSD of the late reverberation, and the PSDs of the microphone-self noise jointly, but, also has high computational complexity. With the SCFA method, the parameters estimation problem is modelled as the following optimization problem

$$\begin{aligned}\hat{p}\left(i\right),\hat{\mathbf{a}}\\ \hat{\gamma}\left(i\right),\hat{\mathbf{P}}_{\mathbf{v}}\end{aligned} = \begin{aligned}&\arg\min_{\substack{p\left(i\right),\mathbf{a}\\ \gamma\left(i\right),\mathbf{P}_{\mathbf{v}}}} \sum_{i=1}^{N}\log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right|+\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right)\\ &\text{s.t. }\mathbf{P}_{\mathbf{y}}\left(i\right)=\mathbf{P}_{\mathbf{e}}\left(i\right)+\mathbf{P}_{\mathbf{l}}\left(i\right)+\mathbf{P}_{\mathbf{v}}\end{aligned} \tag{53}$$

where $\mathbf{P}_{\mathbf{e}}\left(i\right)$, $\mathbf{P}_{\mathbf{l}}\left(i\right)$ and $\mathbf{P}_{\mathbf{v}}$ are defined in Eqs. (5), (6) and (8), respectively. This problem is not a convex problem and the computational complexity is high. In [18], the problem is solved iteratively and the *fmincon* procedure in the standard MATLAB optimization toolbox is used to decrease the value of the cost function in Eq. (53) for each iteration. The iteration terminates if a given estimation accuracy is achieved or the iteration number exceeds a certain number.

Although the SCFA method can estimate the RTF matrix and the PSDs jointly, it is computationally not efficient and sometimes may have a wrong estimate because it deals with a non-convex problem and does not assure a global optimal solution. Therefore, a set of "box constraints" is proposed in [18] to improve the robustness of the SCFA method. In our experiments, we used the same constraints as in Eqs. $(27), (38), (39)$ and (40) in [18].

B. Evaluation metrics

In all the experiments, three types of performance comparison between the proposed method and the reference methods are presented. We first compare the estimation error of the parameters of interest. For the RTF vector, we use the Hermitian angle measure (in rad) [37] which is averaged over all frequency bins and time segments

$$E_{\mathbf{a}}=\frac{\sum_{\beta=1}^{B}\sum_{k=1}^{K/2+1}\mathrm{acos}\left(\frac{\left|\mathbf{a}(\beta,k)^{H}\hat{\mathbf{a}}(\beta,k)\right|}{\|\mathbf{a}(\beta,k)\|_2\|\hat{\mathbf{a}}(\beta,k)\|_2}\right)}{B\left(K/2+1\right)}. \tag{54}$$

For the PSDs of the source and the late reverberation, we use the averaged error (in dB)

$$E_s=\frac{10\sum_{\beta=1}^{B}\sum_{i=1}^{N}\sum_{k=1}^{K/2+1}\left|\log\left(\frac{p(i,k)}{\hat{p}(i,k)}\right)\right|}{BN\left(K/2+1\right)} \tag{55}$$

and

$$E_{\gamma}=\frac{10\sum_{\beta=1}^{B}\sum_{i=1}^{N}\sum_{k=1}^{K/2+1}\left|\log\left(\frac{\gamma(i,k)}{\hat{\gamma}(i,k)}\right)\right|}{BN\left(K/2+1\right)}, \tag{56}$$

where $|\cdot|$ denotes taking the absolute value in Eqs. (54) to (56).

Then, we provide the speech intelligibility and quality comparison among the estimated sources constructed using parameters that are obtained by different methods. That is, we use estimated parameters to calculate the following multi-channel Wiener filter (MWF)

$$\hat{\mathbf{w}}=\frac{\hat{p}}{\hat{p}+\hat{\mathbf{w}}_{\mathrm{MVDR}}^{H}\hat{\mathbf{R}}_{nn}\hat{\mathbf{w}}_{\mathrm{MVDR}}}\hat{\mathbf{w}}_{\mathrm{MVDR}}, \tag{57}$$

where $\mathbf{w}_{\mathrm{MVDR}}$ is the minimum variance distortionless response (MVDR) beamformer [38]

$$\hat{\mathbf{w}}_{\mathrm{MVDR}}=\frac{\hat{\mathbf{R}}_{nn}^{-1}\hat{\mathbf{a}}}{\hat{\mathbf{a}}^{H}\hat{\mathbf{R}}_{nn}^{-1}\hat{\mathbf{a}}}, \tag{58}$$

and

$$\hat{\mathbf{R}}_{nn}=\hat{\gamma}\hat{\mathbf{\Gamma}}. \tag{59}$$

Note that $\hat{\mathbf{\Gamma}}$ is calculated by Eq. (7) for all methods by assuming the distance between each microphone pair is known. For the SCFA method, we set $\hat{\mathbf{R}}_{nn}=\hat{\gamma}\hat{\mathbf{\Gamma}}+\hat{\mathbf{P}}_{v}$, since SCFA can provide an estimate of the PSD of the microphone self noise.

After reconstructing the estimated sources, we use the segmental signal-to-noise-ratio (SSNR) [39] to measure the noise reduction performance. In addition, we compare the speech intelligibility performance using the speech intelligibility in bits (SIIB) measure [40], [41]. The speech-to-reverberation modulation energy ratio (SRMR) measure [42] is also calculated in each scenario to demonstrate the speech quality and intelligibility of all reconstructed sources.

Finally, we compare the computation time between our proposed method and the reference methods.

C. Experiments with Simulated RIRs

1) Setup: To simulate room impulse responses from source to microphones, we use the image source method [34]. The four microphone signals are then constructed by convolving the speech source (with a duration of 35 s) with each of the four room impulse responses corresponding to each microphone. The positions of four microphones and the position of the source are shown in Fig. 2, and the dimensions of the simulated room are set to $7\times5\times4$ m. Since we used the SCFA method as a reference method, the parameters used in

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

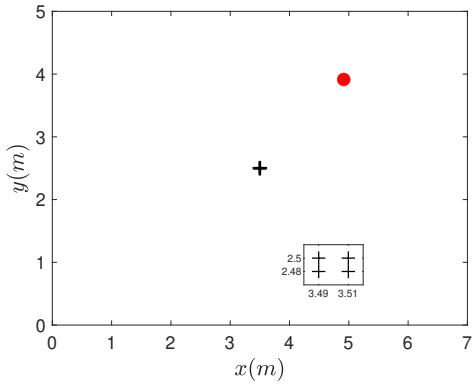

Figure 2: Top view of the acoustic scene. The red circle denotes the source. The cross in the center denotes the set of microphones. A zoom-in of that set of four microphones is provided in the little square.

the experiments are similar to those used in [18]. Subsequently, microphone self-noise is simulated by adding realizations of a zero-mean uncorrelated Gaussian process with variance σ_v^2, such that the SNR per microphone due to the self-noise is equal to the values as specified in each figure. Note that since we consider only the microphone self-noise, the noise energy is relatively low resulting in large SNR values of about 50 dB. The sampling frequency is $f_s = 16$ kHz. Per sub-time frame, the sampled noisy microphone signals are converted to the frequency domain using the STFT procedure, where the sub-time frames are windowed with a square-root Hann window with a length of 512 samples (i.e. 32 ms) and an overlap of 50% between sub-time frames. The true RTF is set to the early reflections of the room impulse response, which is set here as the 512-length FFT of the first 512 samples of the room impulse responses, as this equals the early part (first 32 ms) of the impulse response that falls within a single sub-frame. Each time frame consists of $N_s = 40$ overlapped sub frames. The prefixed parameters are $\delta = 1.1$ and $d_{\max} = 0.02$ (i.e. the distance between each microphone and the source is larger than 0.02 m).

2) Results: In Fig. 3, we fix the reverberation time T_{60} at 1 s and obtain noisy speech with the SNR fixed at 50 dB. We change the number of time frames in a time segment from 1 to 8. The CPSD matrix of the microphone self noise is subtracted from the noisy CPSD matrix for JMLE, Ref1 and Ref2 in this scenario. The performance comparison among JMLE and the other three reference methods is shown in Fig. 3 as the number of time frames used in each time segment changes from 1 to 8. When using only one time frame, JMLE, Ref1 and Ref2 have exactly the same estimates of the RTF and the PSDs of the source and the late reverberation as expected and their estimation performance is better than SCFA. When the number of time frames in a time segment increases, the RTF estimation performance for Ref1 nearly does not change since this method always uses each time frame independently and does not use the prior information that the RTF is constant for all time frames in a time segment. However, for JMLE, SCFA and Ref2, the estimation error of the RTF decreases

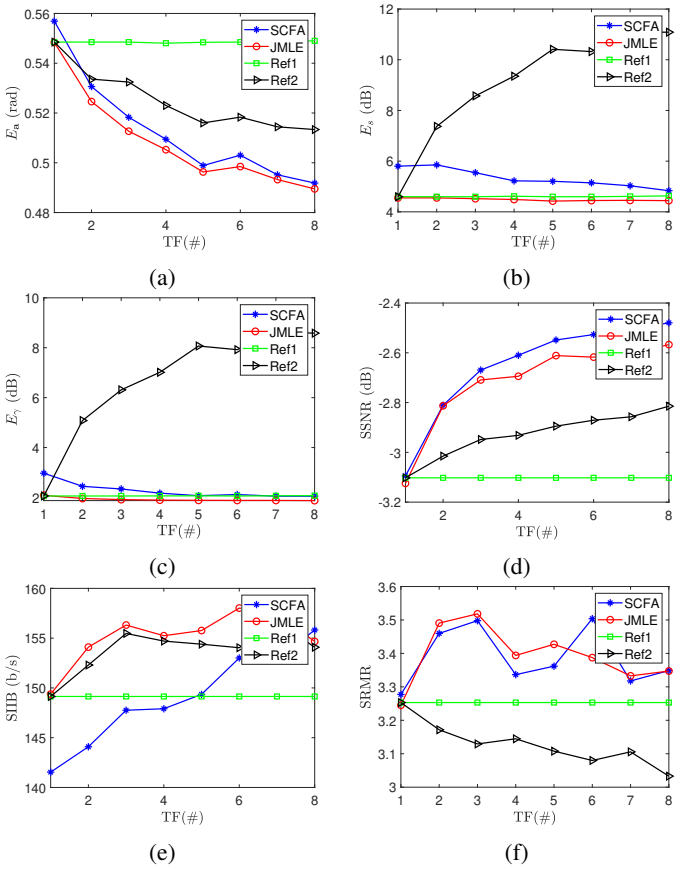

Figure 3: Performance vs the number of time frames.

with the increase of the number of time frames in a time segment. For a larger number of time frames, i.e. a longer segment, among these three methods, JMLE and SCFA have similar performance, and both notably outperform Ref2. The PSD estimation performance for JMLE, SCFA and Ref1 does not change much since the PSDs can differ time-frame by time-frame. However, the PSD estimation performance for Ref2 decreases when the number of time frames increases because Ref2 assumes the source is stationary during a time segment, which is mostly not true in a practical scene. For the noise reduction performance and the speech quality and intelligibility performance, we can see that JMLE and SCFA have larger SSNR, SIIB and SRMR values compared to the other two reference methods in most cases.

D. Experiments with Recorded RIRs

The performance of all methods is now compared using recorded room impulse responses from [36]. The reverberation time of the RIRs include 0.36 s and 0.61 s. The positions of the microphones and the position of the source used to record the impulse responses are shown in Fig. 4. The source is placed at a distance of 2 m from the center of the uniform linear microphone array which has inter-distances of 8 cm. Although the angles of the source include $\{-90°, -75°, \cdots, 90°\}$ in [36], we use only $\{0°, 15°, 30°, 45°, 60°\}$ in this work. We will first consider a static source scenario and evaluate the performance for various SNR values. Then, we will show the

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

influence on the estimation performance of all methods when the source position changes at specific moments.

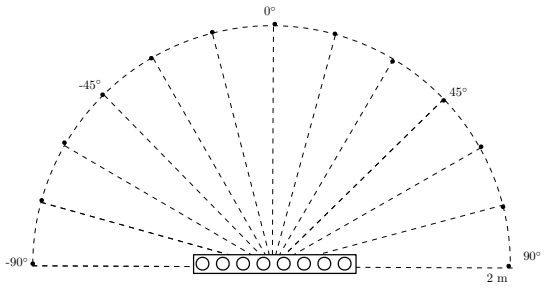

Figure 4: Setup for the real RIRs.

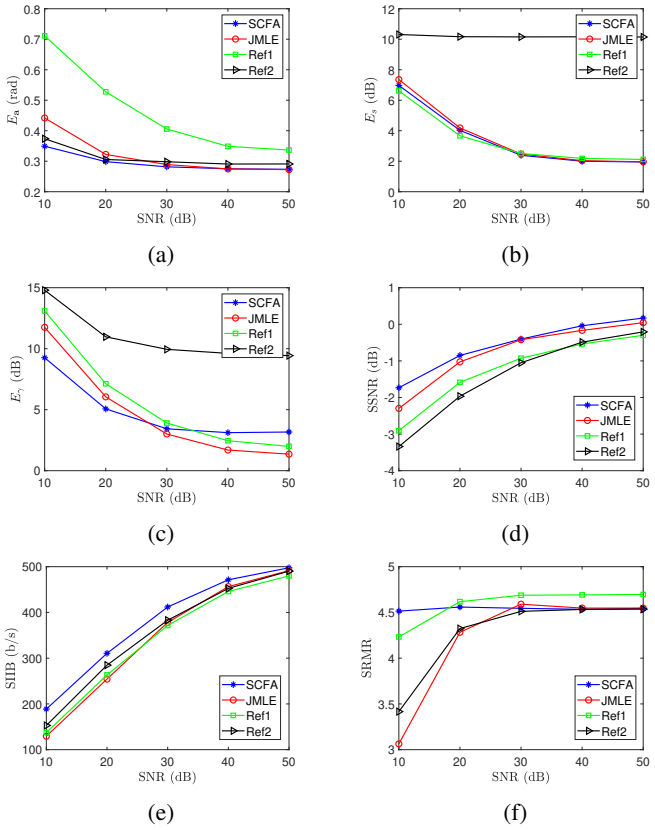

Figure 5: Performance vs SNR.

1) Static source: For the static source scenario, we use the RIRs for the source position fixed at $0°$ and the reverberation time of 0.61 s. We obtain noisy speech with the SNR simulating the microphone self noise ranging from 10 dB to 50 dB. Notice that realistic values for microphone self noise are in the order of 40 to 50 dB. Each time segment contains 8 time frames. Note that in this scenario, the prior information of the microphone self noise is used by none of the methods and for JMLE, Ref1 and Ref2, we simply ignore the microphone self noise and use the CPSD matrix of the noisy signal directly.

The performance comparison among JMLE and the other three reference methods is shown in Fig. 5 as the SNR increases from 10 dB to 50 dB. As shown in Fig. 5, JMLE and SCFA outperform Ref1 in the RTF estimation performance and outperform Ref2 in the PSDs estimation performance (of the source and the late reverberation). As the SNR becomes larger, all methods have both better RTF and PSD estimation performance. However, JMLE shows the most significant improvement compared to the other methods. For the noise reduction performance and the speech quality and intelligibility performance, JMLE and SCFA still outperform the other two reference methods.

TABLE I: Computation time comparison.

method	SCFA	JMLE	Ref1	Ref2
Normalized run time	1310	19	6	1

In Table I we show the normalized computation time comparison among all methods, where we have averaged the run time over all cases for each method. As expected, SCFA needs significantly more time compared to the other three methods. The computational cost of the proposed method using multiple time frames mainly comes from the calculation of the eigenvalue decomposition of an $M \times M$ matrix in each iteration, which has a complexity of order M^3. The total complexity order is thus $(N+N_i)M^3$ with one initial step and N_i iterative steps. Similarly, for Ref1, its complexity order is NM^3 with N the number of time frames in a time segment. For Ref2, its complexity order is M^3. Therefore, the time cost ratio among JMLE, Ref1 and Ref2 is $N_i + N : N : 1 = 18 : 8 : 1$, which is similar to the real averaged run time ratio in Table I. Note that the proposed method using multiple time frames can be initialized by either Ref1 or Ref2. In this work, we present only using Ref1 as the initialization step. If the Ref2 is used as the initialization, the complexity order of JMLE will be $(N_i + 1)M^3$.

2) Moving source: For the moving source scenario, we place the source at $0°$ and change the position to $60°$ in steps of $15°$ every 7 s. Since each time frame contains 40 sub-time frames of 32 ms taken with 50% overlap and each time segment contains 8 time frames, the time segment duration is about 5.12 s. The 35 s speech is divided into 6 complete time segments (the last incomplete time segment is not used). Only the microphone signals during the first and the fourth time segments are received from a single source position. In all other segments, the source position changes during the segment. We evaluate the estimation performance of all methods for per time segment.

In Fig. 6, the reverberation time is 0.36 s. For comparison, we show the estimation performance of all methods when the source position is fixed at $0°$ in Figs. 6a, 6c and 6e. As shown, the estimation performance of all methods does not change much for different time segments, except the poor PSDs estimation performance of the Ref2 method. In Figs. 6b, 6d and 6f, we show the estimation performance of all methods when the source position is moved from $0°$ to $60°$ by $15°$ every 7 s. The vertical dashed lines in these figures denote the time point when the source position is changed. As shown, the estimation performance during the first and the fourth time segments is best among others for the methods using multi-time frames in their estimation as during these time segments, the source position is fixed while during

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

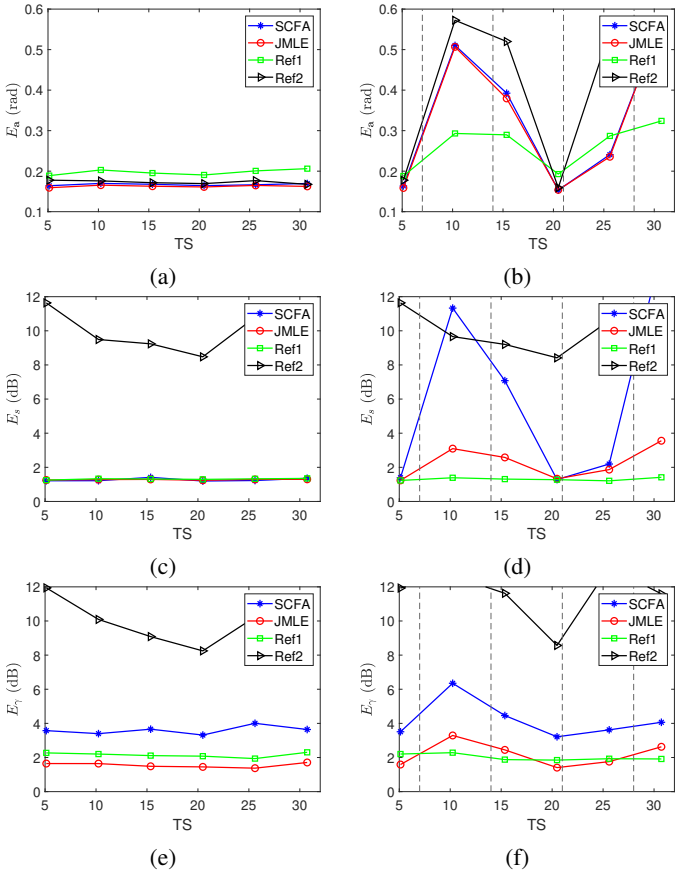

Figure 6: Performance vs time segments (TS) with the reverberation time 0.36 s.

other time segments the source position is changed. The RTF estimation performance is influenced the most while the late reverberation PSD estimation performance is influenced the least by source position change. The reason is that the RTF contains information on the source position, while the late reverberation can be considered as a diffuse noise field. For the Ref1 method, its estimation performance is not affected much since it estimates the parameters frame by frame instead of segment by segment and only four time frames are affected by source position change.

V. Concluding Remarks

We considered the problem of estimating the RTFs, the PSDs of the source and the PSDs of the late reverberation jointly for a single source scenario. We first proposed a joint maximum likelihood estimator (JMLE) using a single time frame, which has a closed form solution and can be solved efficiently. Then, we proposed a joint MLE using multiple time frames that share the same RTF and achieved similar estimation accuracy, together with the performance of noise reduction, speech quality and speech intelligibility, compared to the SCFA method, which both outperform the other reference methods combining several existing state-of-the-art methods. Moreover, it is also shown that the proposed JMLE for multiple time frames has a much lower computational complexity than that of the SCFA method.

To constrain the scope of this work, the focus of this work was on a single source in a reverberant environment. Understanding the single source scenario in future work, we will extend this work in combination with recent results [19] towards the multi-source formulation.

References

[1] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1599–1612, 2016.

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

[3] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, 2018.

[4] 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 *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1123–1127.

[5] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[6] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 137–152, 2017.

[7] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone subset selection for MVDR beamformer based noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 550–563, 2018.

[8] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks With Arbitrary Topology," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1434–1448, 2018.

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

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

[11] 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 Journal on Advances in Signal Processing*, vol. 2006, pp. 1–13, 2006.

[12] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[13] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 61–65.

[14] 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 *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 544–548.

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

[16] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6119–6123.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE/ACM 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/TASLP.2022.3231706

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

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

[19] C. Li, J. Martinez, and R. C. Hendriks, "Low complex accurate multi-source RTF estimation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4953–4957.

[20] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, p. 3233, 2003.

[21] R. Talmon, I. Cohen, and S. Gannot, "Relative Transfer Function Identification Using Convolutive Transfer Function Approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[22] Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[23] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, jul 2003.

[24] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 845–856, 2005.

[25] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, "A study of the distribution of time-domain speech samples and discrete Fourier coefficients," in *Proc. SPS-DARTS*, vol. 1, 2005, pp. 155–158.

[26] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[27] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *21st European Signal Processing Conference (EUSIPCO 2013)*, 2013, pp. 1–5.

[28] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–14, 2006.

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

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

[31] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1732–1736, 1962.

[32] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[33] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," nov 2012, version 20121115. [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html

[34] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

[36] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[37] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 11–15.

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

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

[40] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2017.

[41] ——, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[42] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.