ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747170

LOW COMPLEX ACCURATE MULTI-SOURCE RTF ESTIMATION

Changheng Li, Jorge Martinez and Richard C. Hendriks

Circuits and Systems (CAS) Group, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Many multi-microphone algorithms depend on knowing the relative acoustic transfer functions (RTFs) of the individual sound sources in the acoustic scene. However, accurate joint RTF estimation for multiple sources is a challenging problem. Existing methods to jointly estimate the RTF for multiple sources have either no satisfying performance, or, suffer from a very large computational complexity. In this paper, we propose a method for robust estimation of the individual RTFs in a multi-source acoustic scenario. The presented algorithm is based on linear algebraic concepts and therefore of lower computational complexity compared to a recently presented state-of-the-art algorithm, while having a similar performance. Experimental results are presented to demonstrate the RTF estimation performance as well as the noise reduction performance when combining the estimated RTFs with a beamformer.

Index Terms— Joint diagonalization, microphone array signal processing, source separation, RTF estimation, speech enhancement

1. INTRODUCTION

Microphone arrays are ubiquitous these days and can be used for applications like source separation [1–3], dereverberation [4–6], noise reduction [7–10] and sources localization [11]. These applications have in common that they heavily rely on acoustic-scene dependent parameters like relative acoustic transfer functions (RTFs), power spectral densities (PSDs) of the sources, PSDs of the late reverberation and PSDs of the microphone self-noise. In particular the RTF plays a very important role in beamforming applications. knowing and having an accurate estimate of the RTF per source is very important, for example, to steer a beamformer in the right direction [12], or preserve the spatial cues in binaural noise reduction algorithms [8]. However, accurate RTF estimation is also rather challenging. In this paper we therefore specifically focus on estimating the RTFs and present an algorithm to jointly estimate the individual RTFs of the sources in the acoustic scene.

RTF estimation for a single point source in noise is a problem that has been addressed before in several papers, e.g. [13–15]. In this work, we consider the more general and more challenging case of simultaneously RTF estimation for multiple sources. A few methods have been proposed for multiple source RTF estimation in recent years, e.g., [16–18]. In [16], the RTFs are estimated by updating the initial estimate of the RTFs in an iterative fashion. However in reality, the a priori information of the RTFs might be unknown. In [17], the expectation maximization (EM) method is used to estimate the RTFs by assuming that, in each time-frequency bin, only a single source is active, which thus puts limitations on the acoustic scenarios. In [18], a simultaneous confirmatory factor analysis (SCFA) method was proposed to estimate the RTFs and also the PSDs of sources, late reverberation and the microphone self-noise jointly. However, due to the non-convexity of the problem formulation, the SCFA method in [18] has a rather high computational cost and is therefore currently less applicable for real-time applications.

To accurately estimate the RTFs jointly for multiple sources, our starting point is the algorithm proposed in [1]. This algorithm was developed for blind source separation and is based on linear algebraic concepts. We start with presenting the method from [1], but from a different perspective, such that our proposed algorithm can be better understood. Next, we propose a more robust method, which is also based on linear algebraic concepts and has relatively low computational complexity. The simulations demonstrate that our method is more accurate compared to the reference algorithm [1] and of much lower complexity compared to the state-of-the-art SCFA method from [18], while having a comparable performance.

2. PRELIMINARIES

2.1. Signal model

We consider R acoustic point sources observed by a microphone array consisting of M microphones with an arbitrary geometric structure under the assumption that the signal-to-noise ratio (SNR), i.e., the SNR due to the diffuse noise, is relatively high, the late reverberation is neglectable and the number of microphones is larger than the number of the sources (i.e., $M > R$). In the short-time Fourier transform (STFT) domain, the signal received at the m-th microphone can be modelled as

$$y_m(i,k) = \sum_{r=1}^{R} a_{mr}(\beta,k)\, s_r(i,k), \tag{1}$$

where i is the time-frame index, k is the frequency bin index and $a_{mr}(\beta,k)$ is the m-th element of the RTF vector $\mathbf{a}_r(\beta,k)$ corresponding to source s_r in time segment β at microphone m. In this work, we differentiate between time segments (indexed by β) and time frames (indexed by i). Each time segment consists of multiple time frames. We assume that the RTF vector is constant during a time segment (thus during multiple time frames that fall within one segment) and $a_{1r} = 1$ for $r = 1, ..., R$, which means that the first microphone is selected as the reference microphone. Stacking the M microphone STFT coefficients into a vector, we have

$$\mathbf{y}(i,k) = \sum_{r=1}^{R} \mathbf{a}_r(\beta,k)\, s_r(i,k) \in \mathbb{C}^{M\times 1}. \tag{2}$$

We assume that all the sources are mutually uncorrelated for each frame of a time segment, which leads to the following second-order statistical signal model

$$\mathbf{P}_{\mathbf{y}}(i,k) = \sum_{r=1}^{R} p_r(i,k)\mathbf{a}_r(\beta,k)\mathbf{a}_r^H(\beta,k) \in \mathbb{C}^{M\times M}, \tag{3}$$

Changheng Li is supported by the China Scholarship Council.

where $p_r(i,k) = E\left[|s_r(i,k)|^2\right]$ is the power spectral density (PSD) of the r-th source at the reference microphone. The covariance matrix can be rewritten in the following matrix form

$$\mathbf{P_y}(i,k) = \mathbf{A}(\beta,k)\,\mathbf{P}(i,k)\,\mathbf{A}^H(\beta,k), \tag{4}$$

where the RTF matrix is given by

$$\mathbf{A}(\beta,k) = [\mathbf{a}_1(\beta,k), \cdots, \mathbf{a}_R(\beta,k)] \tag{5}$$

and the PSD matrix is given by

$$\mathbf{P}(i,k) = \text{diag}[p_1(i,k), \cdots, p_R(i,k)]. \tag{6}$$

The main goal of this paper is to estimate the RTF matrix $\mathbf{A}$ using estimated covariance matrices $\{\mathbf{P_y}(i,k)\}$ with $i = 1, \cdots, N$, where N is the number of time frames in a time segment.

2.2. Covariance Matrix Estimation

In addition to time frames and time segments, we now also define sub time-frames. Each time frame consists of N_s overlapping sub-frames indexed by n_s with equal length T_s, where the sub-frame length is much smaller than the time frame length such that N_s is a large integer. Assuming the signal is stationary across a time frame, we can estimate the covariance matrix per time frame i based on the sample covariance matrix using the sub-frames' samples, i.e.,

$$\hat{\mathbf{P}}_\mathbf{y}(i,k) = \frac{1}{N_s}\sum_{n_s=1}^{N_s}\mathbf{y}(n_s,k)\,\mathbf{y}(n_s,k)^H, \tag{7}$$

where $\mathbf{y}(n_s,k)$ is the STFT coefficient vector. Notice that within the time frames of one time segment, the RTF matrix is a constant matrix and the PSDs of the sources are assumed to be non-stationary, which means that the signal powers can change over the frames.

3. RTF ESTIMATION

In Section 3.2, we propose an improved algorithm to estimate the RTF matrix. The starting point is the method presented in [1], which is originally meant for blind source separation. Since the RTF is defined per frequency, from now on, frequency indices are neglected for ease of notation.

We first write the covariance matrices $\mathbf{P_y}(i)$ into the form

$$\mathbf{P_y}(i) = \tilde{\mathbf{A}}(i)\,\tilde{\mathbf{A}}^H(i), \text{for } i = 1, \cdots, N, \tag{8}$$

where $\tilde{\mathbf{A}}(i) = \mathbf{A}\sqrt{\mathbf{P}(i)}$ and the diagonal matrix $\sqrt{\mathbf{P}(i)}$ is the unique non-negative square root of $\mathbf{P}(i)$. Note that $\mathbf{A}$ equals the normalized version of matrix $\tilde{\mathbf{A}}(i)$ where the columns of $\mathbf{A}(i)$ are normalized with respect to their first element, which is the square root of the PSD of each corresponding source. Hence, estimation of $\mathbf{A}$ and $\mathbf{P}(i)$ can be converted into the estimation of $\tilde{\mathbf{A}}(i)$ for any time frame i. With this conversion, the covariance matrices for all the other time frames in the same segment can be represented by $\tilde{\mathbf{A}}(i)$. That is

$$\begin{aligned}\mathbf{P_y}(j) &= \mathbf{A}\mathbf{P}(j)\,\mathbf{A}^H \\ &= \mathbf{A}\sqrt{\mathbf{P}(i)}\sqrt{\mathbf{P}^{-1}(i)}\mathbf{P}(j)\sqrt{\mathbf{P}^{-1}(i)}\sqrt{\mathbf{P}(i)}\mathbf{A}^H \\ &= \tilde{\mathbf{A}}(i)\,\tilde{\mathbf{P}}(j)\,\tilde{\mathbf{A}}^H(i), \text{for } j = 1, \cdots, N,\end{aligned} \tag{9}$$

where $\tilde{\mathbf{P}}(j) = \sqrt{\mathbf{P^{-1}}(i)}\mathbf{P}(j)\sqrt{\mathbf{P^{-1}}(i)}$ is a diagonal matrix.

3.1. Joint Diagonalization Method

We first summarize in this section the joint diagonalization method from [1] to put our work in perspective. This method was originally proposed for blind source separation and used in, e.g., [1, 19], to estimate the mixing matrix instead of the RTF matrix. Therefore, although the estimation steps are the same as in [1], we summarize this method when used in a different context to better understand our proposed method that we present in Section 3.2.

The method in [1] focuses on estimating $\tilde{\mathbf{A}}(1)$. Then, matrices $\mathbf{P_y}(i)$ in the segment can be represented by $\tilde{\mathbf{A}}(1)$ using

$$\mathbf{P_y}(i) = \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{P}}(i)\,\tilde{\mathbf{A}}^H(1), \text{for } i = 2, \cdots, N, \tag{10}$$

where

$$\tilde{\mathbf{P}}(i) = \sqrt{\mathbf{P^{-1}}(1)}\mathbf{P}(i)\sqrt{\mathbf{P^{-1}}(1)} \tag{11}$$

is diagonal. Notice that $\mathbf{P_y}(1) = \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{A}}^H(1)$.

Consider the singular value decomposition (SVD) of $\tilde{\mathbf{A}}(1)$, i.e.,

$$\tilde{\mathbf{A}}(1) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{12}$$

where $\mathbf{U}$ is an $M \times R$ complex sub-unitary matrix (i.e., $\mathbf{U}^H\mathbf{U} = \mathbf{I}$), $\mathbf{\Sigma}$ is a $R \times R$ diagonal matrix and $\mathbf{V}$ is a complex valued $R \times R$ unitary matrix. The estimation of $\tilde{\mathbf{A}}$ is decomposed into the estimation of the three matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.

The estimates of $\mathbf{U}$ and $\mathbf{\Sigma}$ can be obtained from $\mathbf{P_y}(1)$. Using the SVD of $\tilde{\mathbf{A}}(1)$ in (8), $\mathbf{P_y}(1)$ can be expressed as:

$$\begin{aligned}\mathbf{P_y}(1) &= \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{A}}^H(1) \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{V}\mathbf{\Sigma}\mathbf{U}^H \\ &= \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^H.\end{aligned} \tag{13}$$

Since $\mathbf{U}$ is a sub-unitary matrix and $\mathbf{\Sigma^2}$ is a diagonal matrix, (13) is an eigenvalue decomposition of the matrix $\mathbf{P_y}(1)$. Hence we can calculate $\mathbf{U}$ and $\mathbf{\Sigma}$ by taking the EVD of $\mathbf{P_y}(1)$.

The estimation of $\mathbf{V}$ can be solved by using estimated $\mathbf{U}$, $\mathbf{\Sigma}$, and the covariance matrices for all other time frames in the same segment. Taking the SVD of $\tilde{\mathbf{A}}(1)$ in (10), $\mathbf{P_y}(i)$ for $i = 2, \cdots, N$ can be expressed as

$$\begin{aligned}\mathbf{P_y}(i) &= \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{P}}(i)\,\tilde{\mathbf{A}}^H(1) \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\tilde{\mathbf{P}}(i)\,\mathbf{V}\mathbf{\Sigma}\mathbf{U}^H.\end{aligned} \tag{14}$$

Now we construct a new set of matrices $\mathbf{P_w}(i)$ using $\mathbf{U}$ and $\mathbf{\Sigma}$

$$\begin{aligned}\mathbf{P_w}(i) &= \mathbf{\Sigma}^{-1}\mathbf{U}^H\mathbf{P_y}(i)\,\mathbf{U}\mathbf{\Sigma}^{-1} \\ &= \mathbf{V}^H\tilde{\mathbf{P}}(i)\,\mathbf{V}.\end{aligned} \tag{15}$$

As $\mathbf{V}$ is an orthogonal matrix, it can be obtained by computing the eigenvectors of the matrices $\{\mathbf{P_w}(i)\}$, with $i = 2, \cdots, N$.

If $N = 2$, we can estimate $\mathbf{V}$ by taking the EVD of $\mathbf{P_w}(2)$. In case of equal eigenvalues, the corresponding eigenvectors are not unique. Hence, in order to obtain the correct estimate of $\mathbf{V}$, we need to assume that the diagonal matrix $\tilde{\mathbf{P}}(i)$ has distinct diagonal elements, which means that the following inequalities should be satisfied for every two sources r_1 and r_2,

$$\frac{p_{r_1}(2)}{p_{r_1}(1)} \neq \frac{p_{r_2}(2)}{p_{r_2}(1)}, \tag{16}$$

where $p_r(i)$ denotes the PSD of the r_{th} source in the i_{th} time frame.

If $N > 2$, the estimation of $\mathbf{V}$ becomes a joint diagonalization problem: find a unitary matrix $\mathbf{V}$ such that $\left\{\mathbf{V}\mathbf{P}_{\mathbf{w}}(i)\mathbf{V}^H\right\}$ with $i = 2, \cdots, N$ is a set of diagonal matrices or has minimal off-diagonal elements. The Jacobi-like algorithm proposed in [20] can be used to solve this joint diagonalization problem, which reduces the original joint diagonalization problem into finite sub-problems having closed-form solutions (see [20] for more details). To make sure the joint diagonalization problem has a satisfying solution, we also need to make an assumption on the PSDs of the sources: for every source r_1, there exists one time frame i_0 such that the following inequality holds for any other source r_2:

$$\frac{p_{r_2}(i_0)}{p_{r_2}(1)} \neq \frac{p_{r_1}(i_0)}{p_{r_1}(1)} \text{ for any } r_2 \neq r_1. \tag{17}$$

Finally, we estimate $\tilde{\mathbf{A}}(1)$ by multiplying the estimated $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ as in (12). Normalizing $\tilde{\mathbf{A}}(1)$, we eventually obtain the estimate of the RTF matrix $\mathbf{A}$. Note that having estimated $\tilde{\mathbf{A}}(1)$, we can also estimate the individual source PSDs. To do so, we use the diagonal elements of $\mathbf{V}\mathbf{P}_{\mathbf{w}}(i)\mathbf{V}^H$ to estimate $\tilde{\mathbf{P}}(i)$. Using the definition $\tilde{\mathbf{A}}(1) = \mathbf{A}\sqrt{\mathbf{P}(1)}$ in combination with (11) we obtain the PSDs of all sources for all time frames in the segment.

This algorithm is summarized in Algorithm description 1.

Algorithm 1: Joint Diagonalization Method (JOINT)

Input: Estimated $\hat{\mathbf{P}}_{\mathbf{y}}(i)$, for $i = 1, \cdots, N$,
Output: $\mathbf{A}$ and $\mathbf{P}(i)$ for $i = 1, \cdots, N$,

1 Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\hat{\mathbf{P}}_{\mathbf{y}}(1)$.
2 Construct new matrices $\mathbf{P}_{\mathbf{w}}(i)$ for $i = 2, \cdots, N$.
3 Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(i)$ for $i = 2, \cdots, N$ using the Jacobi-like algorithm [20].
4 Estimate $\tilde{\mathbf{A}}(1)$ by multiplying $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.
5 Estimate $\mathbf{A}$ by normalizing $\tilde{\mathbf{A}}(1)$ with its first row.
6 Estimate $\mathbf{P}(i)$ for $i = 1, \cdots, N$ using the first row of $\tilde{\mathbf{A}}(1)$ and $\tilde{\mathbf{P}}(i)$ for $i = 2, \cdots, N$.

3.2. Robust Joint Diagonalization

The algorithm introduced in Section 3.1 focuses on estimating the RTF matrix $\mathbf{A}$ using the estimated covariance matrices $\hat{\mathbf{P}}(i)$ for $i = 1, \cdots, N$. However, instead of using the individual matrices $\hat{\mathbf{P}}(i)$ as done in the first step in Algorithm 1, we can also choose to estimate the RTF matrix from any linear combination of estimated covariance matrices in segment β. By using an average of estimated covariance matrices instead of a single estimated $\mathbf{P}_{\mathbf{y}}(1)$ in step 1 from Algorithm 1, we are able to significantly reduce the estimation error on estimating $\mathbf{A}$ if we are able to also select the best estimated covariance matrices to form this average. To see this, let us first look at the error on the estimated covariance matrix $\Delta\mathbf{P}_{\mathbf{y}}(i)$. This error can be decomposed into:

$$\Delta\mathbf{P}_{\mathbf{y}}(i) = \mathbf{A}\left(\mathbf{P}(i) - \hat{\mathbf{P}}(i)\right)\mathbf{A}^H - \mathbf{E}(i), \tag{18}$$

where the first part $\Delta\mathbf{P}(i) = \left(\mathbf{P}(i) - \hat{\mathbf{P}}(i)\right)$ is indeed the estimation error between the sampled covariance matrix and the true covariance matrix of sources, and the second part $\mathbf{E}(i)$ is due to the late reverberation component and the microphone self noise component, which can be assumed to be positive definite.

It is well known that the estimation error between a sampled covariance matrix and the true covariance matrix can be reduced by increasing the number of samples. Hence, to decrease $\Delta\mathbf{P}(i)$, we can average covariance matrices for as many time frames as possible in a time segment. However, the second error matrix $\mathbf{E}(i)$ might increase when using more time frames. The question now is, which estimated $\mathbf{P}_{\mathbf{y}}(i)$ for $i = 1, \cdots, N$, should we average to replace $\hat{\mathbf{P}}_{\mathbf{y}}(1)$ in step 1 from Algorithm 1 to reduce the estimation error. Notice that the rank of the true covariance matrix $\mathbf{P}_{\mathbf{y}}(i)$ is R, the rank of the estimated covariance matrix $\hat{\mathbf{P}}_{\mathbf{y}}(i)$ is M and we have assumed that $M > R$. Therefore the $R+1$ largest eigenvalue $\lambda_{R+1}(i)$ of $\hat{\mathbf{P}}_{\mathbf{y}}(i)$ can be used to evaluate how large the error matrix $\mathbf{E}(i)$ is.

Based on the analysis of the estimation error of covariance matrices, the next steps of the robust joint diagonalization algorithm are as follows: Take the EVD for the N estimated covariance matrices $\hat{\mathbf{P}}_{\mathbf{y}}(i)$ from a segment and reorder the time frame index such that $\lambda_{R+1}(i)$ is in an ascending order. Use the first estimated covariance matrix (i.e., the one with the smallest error $\mathbf{E}$) to do Algorithm 1 and obtain the first estimates of the RTF matrix $\hat{\mathbf{A}}_{\mathbf{1}}$ and PSDs of the R sources $\left\{\hat{\mathbf{P}}_{\mathbf{1}}(i)\right\}$. Use these estimates to calculate the following weighted cost function:

$$C(1) = \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_{R+1}^2(i)} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(i) - \hat{\mathbf{A}}_{\mathbf{1}}\hat{\mathbf{P}}_{\mathbf{1}}(i)\hat{\mathbf{A}}_{\mathbf{1}}^H \right\|_2, \tag{19}$$

where $\|\cdot\|_2$ denotes the matrix 2-norm. Next, average the first two estimated covariance matrices from the ordered sequence and use this in combination with Algorithm 1 to obtain the second estimates of the RTF matrix and PSDs of the R sources, and calculate the cost function:

$$C(2) = \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_{R+1}^2(i)} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(i) - \hat{\mathbf{A}}_{\mathbf{2}}\hat{\mathbf{P}}_{\mathbf{2}}(i)\hat{\mathbf{A}}_{\mathbf{2}}^H \right\|_2, \tag{20}$$

In each next iteration, we include an additional covariance matrix from the ordered sequence in the average and use the averaged covariance matrix in combination with Algorithm 1 to estimate the RTF matrix and PSDs of R sources until all the N covariance matrices are averaged and N cost function values are calculated. We then select the minimum cost function value with respect to iteration q, and use the estimate of the RTF matrix in the q_{th} iteration as the final estimate of the RTF matrix.

The algorithm steps are given in algorithm two. Since both the joint diagonalization method and our proposed method are based on linear algebra, computational costs of both algorithms are relatively low. Note that the computational cost of the proposed algorithm is about N times higher than for Algorithm 1.

4. EXPERIMENTS

The performance of the proposed methods is evaluated in the context of noise reduction with four microphones and three sources each with a duration of 25 s. The acoustic setup is depicted in Fig. 1. Each speech signal is convolved with a room impulse response in the time domain. The room impulse responses are generated using the image method [21]. To simulate a nearly non-reverberant noisy signal, we set the reflection coefficients of the six walls as $[0.5, -0.25, 0.1, -0.5, 0.25, -0.1]$ in the first scenario (the reverberation time is about 0.04 s). Besides, we also evaluate the performance of our proposed methods in a second scenario where the reverberation time of the room impulse response is 0.2 s. The sampling frequency is $f_s = 16$ kHz. The microphone self-noise is a

Algorithm 2: Robust Joint Diagonalization (PROP)

Input: Estimated $\mathbf{P_y}(i)$, for $i = 1, \cdots, N$,
Output: $\mathbf{A}$

Estimate λ_{R+1} from EVD of $\mathbf{P_y}(i)$, for $i = 1, \cdots, N$.
Reorder time frame index such that λ_{R+1} is ascending.
for $q = 1 : N$ **do**
    Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\sum_{i=1}^{q} \frac{1}{q} \hat{\mathbf{P}}_{\mathbf{y}}(i) = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^H$.
    Construct new matrices $\mathbf{P_w}(i)$ for $i = 2, \cdots, N$.
    Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(i)$ for $i = 2, \cdots, N$ using the Jacobi-like algorithm [20].
    Estimate $\tilde{\mathbf{A}}(1)$ by multiplying $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.
    Estimate $\mathbf{A}$ by normalizing $\tilde{\mathbf{A}}$ with its first row.
    Estimate $\mathbf{P}(i)$ for $i = 1, \cdots, N$ using the first row of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(i)$ for $i = 2, \cdots, N$.
    Use the estimate to calculate the cost function Eq. (19)
Find the minimum cost function value with respect to the q_{th} estimate of $\mathbf{A}$ and use it as the final estimate of the RTF matrix.

zero-mean uncorrelated Gaussian process with variance σ_v^2, such that the SNR due to the self-noise is equal to the values as specified in Fig. 2 per microphone. The noisy speech signal is converted into the STFT domain using a square-root Hann window with a length of 800 samples (i.e. 50 ms) and an overlap of 50%. The FFT length is 1024. Note that the true RTF matrix is calculated using the 1024-length FFT coefficients of the first 800 samples of the room impulse responses. Each time segment consists of $N = 8$ time frames and each time frame consists of $N_s = 40$ sub frames. For comparison, we used the SCFA method from [18] and the original joint diagonalization method from [1] as a reference as SCFA and JOINT, respectively. The proposed method will be referred to as PROP.

The RTF estimation error is evaluated by the Hermitian angle [22].

$$\frac{\sum_{r=1}^{R}\sum_{\beta}^{\mathrm{B}}\sum_{k=1}^{K/2+1} \operatorname{acos}\left(\frac{\left|\mathbf{a}_r^H(\beta,k)\hat{\mathbf{a}}_r(\beta,k)\right|}{\left\|\mathbf{a}_r^H(\beta,k)\right\|_2\left\|\hat{\mathbf{a}}_r(\beta,k)\right\|_2}\right)}{R\mathrm{B}\left(K/2+1\right)}\,(\mathrm{rad}), \qquad (21)$$

where K and B are the number of frequency bins and time segments, respectively. In Fig. 2(a), we show the estimation performance in the nearly no reverberation case (with subscript 'nr'), and $T_{60} = 0.2s$ (with subscript 'r'). For both scenarios, PROP and SCFA have a similar and much better performance compared to JOINT. For the nearly no reverberation case, SCFA has a somewhat better performance than PROP, because SCFA can model microphone self-noise and can better reduce the model mismatch error caused mainly by the diffuse noise. However, for the $T_{60} = 0.2s$ and high SNR case, PROP has a slightly better estimation performance than SCFA, because the model mismatch error now is mainly caused by the late reverberation component, which is not considered in the referenced version of SCFA. For the $T_{60} = 0.2s$ case, we also evaluated the noise reduction performance in combination with three minimum variance distortionless response (MVDR) beamformers [23], where we use each time one of the three estimated RTFs as the target and the remaining two sources as interferers. We then calculate the segmental-signal-to-noise-ratio (SSNR) and average this over the three sources. Note that for the SSNR calculation, we omit the sub frames in which the signal energy is zero. In addition to the methods PROP, JOINT and SCFA we also show the performance when using the true RTF.

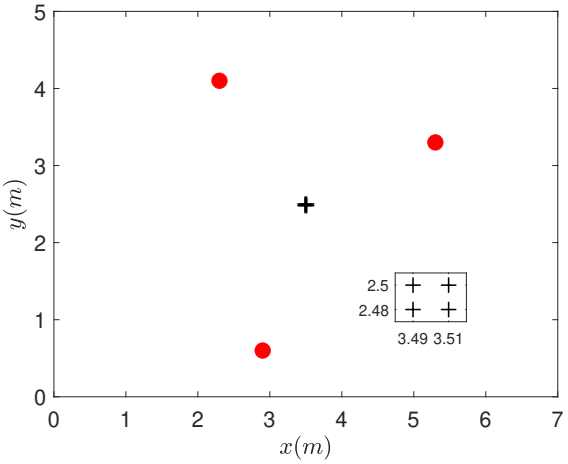

Fig. 1: Acoustic scene. The three red circles denote the sources. The cross in the center denotes the set of microphones. A zoom-in of that set of four microphones is provided in the little square.

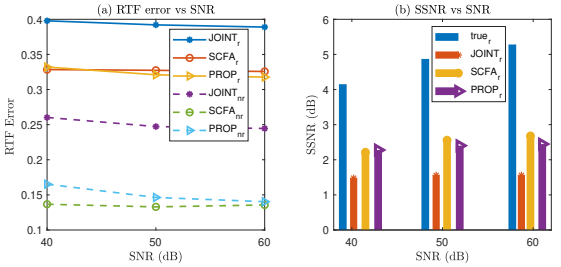

Fig. 2: RTF estimation error and SSNR vs SNR

As shown in Fig. 2(b), the SSNR for each method increases as the SNR increases. PROP has an almost similar performance compared to SCFA, while both PROP and SCFA improve over JOINT with slightly less than 1 dB in terms of SSNR. In Table 1 we show the

Table 1: Computation time comparison.

method	SCFA	PROP	JOINT
Normalized run time	1	0.0163	0.0024

normalized computation time for all methods after averaging the run time over all scenarios. As expected, the runtime for PROP is about $N = 8$ times larger than for JOINT, but PROP is significantly less complex than SCFA.

5. CONCLUSIONS

We considered the problem of estimating the RTF for multiple sources jointly. We proposed a robust method which averages covariance matrices for as many time frames as possible without suffering too much from model mismatch errors caused by late reverberation and microphone self noise. Experiments show that the RTF estimation performance of the proposed method is similar to the SCFA method, but at a significantly lower complexity, and much better than the joint diagonalization method from [1]. Note that SCFA can also be used to estimate the RTF matrix for larger reverberation times, which we will address in future research.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[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] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2499–2503.

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

[17] 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.

[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] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[20] J.-F. Cardoso and A. Souloumiac, "Jacobi Angles For Simultaneous Diagonalization." *SIAM Journal on Matrix Analysis and Applications*, vol. 17, 01 1996.

[21] 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.

[22] 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.

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