Noise PSD Insensitive RTF Estimation in a Reverberant and Noisy Environment

Li, Changheng; Hendriks, Richard

DOI
10.1109/ICASSP49357.2023.10094840

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094840

NOISE PSD INSENSITIVE RTF ESTIMATION IN A REVERBERANT AND NOISY ENVIRONMENT

Changheng Li and Richard C. Hendriks

Signal Processing Systems (SPS) Group, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Spatial filtering techniques typically rely on estimates of the target relative transfer function (RTF). However, the target speech signal is typically corrupted by late reverberation and ambient noise, which complicates RTF estimation. Existing methods subtract the noise covariance matrix to obtain the target plus late reverberation covariance matrix, from where the RTF is estimated. However, the noise covariance matrix is typically unknown. More specifically, the noise power spectral density (PSD) is typically unknown, while the spatial coherence matrix can be assumed known as it might remain time-invariant for a longer time. Using the spatial coherence matrices we simplify the signal model such that the off-diagonal elements are not affected by the PSDs of the late reverberation and the ambient noise. Then we use these elements to estimate the target covariance matrix, from where the RTF can be obtained. Hence, the resulting estimate of the RTF is insensitive to the noise PSD. Experiments demonstrate the estimation performance of our proposed method.

***Index Terms*—** RTF estimation, spatial filter, Eigenvalue Decomposition

1. INTRODUCTION

Microphone arrays are widely used for hands-free speech communication applications such as mobile phones and hearing aids. Spatial filtering techniques like the minimum variance distortionless response (MVDR) beamformer [1, 2] and the multichannel Wiener filter (MWF) [2, 3] are often used to extract target signals from the noisy microphone recordings typically corrupted by reverberation and ambient noise. However, these filters critically rely on knowing the relative transfer functions (RTFs) from source to microphones. In this work we therefore address the RTF estimation problem of a single source in a reverberant and noisy environment.

Several RTF estimation methods have been proposed in recent years [4–11], including the covariance subtraction (CS) method [7–9] and the covariance whitening (CW) method [8–10]. In reverberant and noisy environments, these methods require the noise and late reverberation covariance matrices to be known. The CW method subtracts the noise covariance matrix from the noisy covariance matrix prior to whitening by the late reverberation covariance matrix. However, the noise covariance matrix is usually unknown. In this paper, we model the noise covariance matrix as a time-varying noise PSD multiplied by a time-invariant spatial coherence matrix. In that case, the noise PSD is assumed unknown, but the spatial coherence matrix can be assumed known as it might remain time-invariant for a longer time. Under this relaxed assumption, we propose a method to estimate the RTF in a reverberant and noisy environment, which avoids using the noise PSD and is insensitive to noise PSD estimation errors.

Changheng Li is supported by the China Scholarship Council.

2. PRELIMINARIES

2.1. Signal model

We consider the problem of estimating the RTFs of a single acoustic source in a reverberant and noisy environment using an array of M microphones with an arbitrary configuration. In the short-time Fourier transform (STFT) domain, the signal received at the m-th microphone is given by

$$y_m(l,k) = x_m(l,k) + r_m(l,k) + v_m(l,k), \quad (1)$$

with l the time-frame index, k the frequency bin index, and m the microphone index. Let x_m denote the speech including the direct and early reflections of the source. Let r_m denote the late reverberation including all the late reflections of the source, which can be considered diffuse. Further, v_m denotes the ambient noise component and microphone self-noise. The early speech component can be modelled as

$$x_m(l,k) = a_m(l,k)\, s(l,k), \quad (2)$$

with $a_m(l,k)$ the RTF of the source from the reference microphone to the m-th microphone. Without loss of generality, we select in this work the first microphone as the reference microphone, which means $a_1 = 1$. Stacking all M microphone signals $\{y_m\}_{m=1}^{M}$ into a vector, we have

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

Assuming the three components in Eq. (3) to be mutually uncorrelated, the noisy covariance matrix is given by

$$\mathbf{\Phi}_{\mathbf{y}}(l,k) \triangleq \mathbf{\Phi}_{\mathbf{x}}(l,k) + \mathbf{\Phi}_{\mathbf{r}}(l,k) + \mathbf{\Phi}_{\mathbf{v}}(l,k), \quad (4)$$

where $\mathbf{\Phi_p} \triangleq E\left\{\mathbf{p}\mathbf{p}^H\right\}$ for $\mathbf{p} = \mathbf{y}, \mathbf{x}, \mathbf{r}$ or $\mathbf{v}$ with $E\{\cdot\}$ the expectation. From Eq. (2), we have

$$\mathbf{\Phi_x}(l,k) = \phi_s(l,k)\,\mathbf{a}(l,k)\,\mathbf{a}^H(l,k), \tag{5}$$

with $\phi_s(l,k)$ the PSD of the source at the reference microphone. For the late reverberation, we adopt the commonly used model from [12]

$$\mathbf{\Phi_r}(l,k) = \phi_\gamma(l,k)\,\mathbf{\Gamma}(k), \tag{6}$$

where $\phi_\gamma(l,k)$ is the unknown PSD of the late reverberation and $\mathbf{\Gamma}(k)$ is the non-singular and known spatial coherence matrix which can be calculated using the microphone array geometry [13]. For the residual noise, we assume its covariance matrix has a similar form, i.e.,

$$\mathbf{\Phi_v}(l,k) = \phi_v(l,k)\,\mathbf{\Psi}(k), \tag{7}$$

where $\phi_v(l,k)$ is the unknown PSD and $\mathbf{\Psi}(k)$ is the known spatial coherence matrix.

2.2. Problem formulation

Using Eqs. (5) to (7), we can formulate the noisy covariance matrix as

$$\mathbf{\Phi_y}(l,k) = \phi_s(l,k)\,\mathbf{a}(l,k)\,\mathbf{a}^H(l,k) + \phi_\gamma(l,k)\mathbf{\Gamma}(k) + \phi_v(l,k)\mathbf{\Psi}(k). \tag{8}$$

We assume the microphone signals to be stationary over a frame consisting of L_s sub-time frames, indexed by l_s, and estimate $\mathbf{\Phi_y}(t,k)$ for one frame using the sample covariance matrix $\hat{\mathbf{\Phi}}_\mathbf{y}(l,k) = 1/L_s \sum_{l_s=1}^{L_s} \mathbf{y}(l_s,k)\,\mathbf{y}^H(l_s,k)$.

The aim of this work is to estimate the RTF vector $\mathbf{a}(l,k)$ using the estimated covariance matrix $\hat{\mathbf{\Phi}}_\mathbf{y}(l,k)$ and the known spatial coherence matrices $\mathbf{\Gamma}(k)$ and $\mathbf{\Psi}(k)$, while the PSDs $\phi_s(l,k), \phi_\gamma(l,k)$, and $\phi_v(l,k)$ are all unknown. Prior to presenting our proposed method in Section 4, we summarize in Section 3 the CW method from [10] that is meant to estimate $\mathbf{a}(l,k)$ assuming the complete $\mathbf{\Phi_v}(l,k)$ is known instead of only $\mathbf{\Psi}(k)$. For notational simplicity, we omit the frequency and time indices as all processing will be done per time-frequency bin independently.

3. STATE OF THE ART AND MOTIVATION

Existing methods for RTF estimation include the covariance subtraction (CS) method and the covariance whitening (CW) method. The CW method has been shown to outperform the CS method [8,9]. Therefore, we introduce here only the CW method. To use the CW method, we need to assume the covariance matrix of the noise $\mathbf{\Phi_v}$ is given, and subtract it from the noisy covariance matrix $\mathbf{\Phi_y}$, that is

$$\mathbf{\Phi_{x+r}} = \mathbf{\Phi_y} - \mathbf{\Phi_v} = \phi_s\mathbf{a}\mathbf{a}^H + \phi_\gamma\mathbf{\Gamma}. \tag{9}$$

With the signal model from Eq. (9), the CW method can estimate the RTF vector in three steps. First, it whitens the noisy signal using $\mathbf{\Gamma}^{\frac{1}{2}}$, which is the principal square-root of the spatial coherence matrix $\mathbf{\Gamma}$ satisfying $\mathbf{\Gamma} = \mathbf{\Gamma}^{\frac{1}{2}}\mathbf{\Gamma}^{\frac{H}{2}}$ with $\mathbf{\Gamma}^{\frac{H}{2}}$ the Hermitian transpose of $\mathbf{\Gamma}^{\frac{1}{2}}$. Note that the square-root is not unique and in this work, we use the Cholesky decomposition. The covariance matrix after whitening has the form

$$\mathbf{\Phi}_w = \mathbf{\Gamma}^{-\frac{1}{2}}\mathbf{\Phi_{x+r}}\mathbf{\Gamma}^{-\frac{H}{2}} = \phi_s\mathbf{a}_w\mathbf{a}_w{}^H + \phi_\gamma\mathbf{I}, \tag{10}$$

where $\mathbf{a}_w = \mathbf{\Gamma}^{-\frac{1}{2}}\mathbf{a}$ is a scaled version of the principal eigenvector of $\mathbf{\Phi}_w$. Hence the second step is to take the eigenvalue decomposition of $\hat{\mathbf{\Phi}}_w$ and find its principal eigenvector $\mathbf{u}$. The last step is to estimate the RTF vector by

$$\hat{\mathbf{a}} = \frac{\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{u}}{\mathbf{e}^T\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{u}}, \tag{11}$$

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

A weakness of the CW method is that it needs to assume the covariance matrix of the ambient noise is known and subtracted. Subtracting an estimated noise covariance matrix $\hat{\mathbf{\Phi}}_\mathbf{v} = \hat{\phi}_v\mathbf{\Psi}$, the covariance matrix after whitening becomes

$$\mathbf{\Phi}_w = \phi_s\mathbf{a}_w\mathbf{a}_w{}^H + \phi_\gamma\mathbf{I} + \Delta\phi_v\mathbf{\Gamma}^{-\frac{1}{2}}\mathbf{\Psi}\mathbf{\Gamma}^{-\frac{H}{2}}, \tag{12}$$

with $\Delta\phi_v$ the noise PSD estimation error. Here, $\mathbf{a}_w$ is no longer a scaled principal eigenvector of $\mathbf{\Phi}_w$. Hence, inaccuracies in $\hat{\phi}_v$ will lead to significant estimation errors in $\hat{\mathbf{a}}$.

4. PROPOSED METHOD

For the case that not the complete covariance matrix of the ambient noise is known, but only $\mathbf{\Psi}$, we propose an alternative way to estimate the RTF vector by using the off-diagonal elements of a simplified covariance matrix. The proposed method will be less sensitive to estimation errors due to variations in the noise PSD ϕ_v. Note that, the technique using only off-diagonal elements of a matrix was used before in [14] for the PSDs estimation and in [15] for radio telescope arrays.

4.1. Parameter Identifiability

Before using any estimation methods, the identifiability condition that the number of equations is equal or larger than the number of unknowns should be satisfied [16]. Since $\mathbf{\Phi}_y$ is a Hermitian matrix, in Eq. (8) there are M^2 knowns (taking Hermitian symmetry and complex values of the data into account). Since $a_1 = 1$, there are $2(M-1)$ unknowns due to the complex-valued $\mathbf{a}$ and there are 3 unknown real-valued PSDs. Therefore, we have altogether the necessary condition

$$M^2 \geq 2(M-1) + 3, \tag{13}$$

which means $M \geq \sqrt{2} + 1$. Noticing that M should be an integer value, we have $M \geq 3$.

4.2. Simplification

In Eq. (8), since the spatial coherence matrices $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ are assumed to be known, we can simplify the signal model by using the square-root decomposition (e.g. the Cholesky decomposition) of $\mathbf{\Psi} = \mathbf{\Psi}^{\frac{1}{2}}\mathbf{\Psi}^{\frac{H}{2}}$

$$\tilde{\mathbf{\Phi}}_{\mathbf{y}} = \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Phi}_{\mathbf{y}}\mathbf{\Psi}^{-\frac{H}{2}} = \phi_s\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \phi_\gamma\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Psi}^{-\frac{H}{2}} + \phi_v\mathbf{I}, \tag{14}$$

and the eigenvalue decomposition (EVD) of $\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Psi}^{-\frac{H}{2}} = \mathbf{U}\mathbf{\Lambda}_\gamma\mathbf{U}^H$, such that

$$\bar{\mathbf{\Phi}}_{\mathbf{y}} = \mathbf{U}^H\tilde{\mathbf{\Phi}}\mathbf{U} = \underbrace{\phi_s\bar{\mathbf{a}}\bar{\mathbf{a}}^H}_{\bar{\mathbf{\Phi}}_{\mathbf{x}}} + \phi_\gamma\mathbf{\Lambda}_\gamma + \phi_v\mathbf{I}, \tag{15}$$

where $\bar{\mathbf{a}} = \mathbf{U}^H\tilde{\mathbf{a}} = \mathbf{U}^H\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{a}$.

4.3. covariance matrix reconstruction

The simplified covariance matrix in Eq. (15) is now a summation of a rank-1 matrix $\bar{\mathbf{\Phi}}_{\mathbf{x}}$ and a diagonal matrix $\phi_\gamma\mathbf{\Lambda}_\gamma + \phi_v\mathbf{I}$. Hence, the elements of $\bar{\mathbf{\Phi}}_{\mathbf{y}}$ have the form

$$\bar{\mathbf{\Phi}}_{\mathbf{y}\{i,j\}} = \begin{cases} \phi_s|\bar{a}_m|^2 + \phi_\gamma\lambda_m + \phi_v & i = j = m \\ \phi_s\bar{a}_i\bar{a}_j^* & i \neq j \end{cases}, \tag{16}$$

where λ_m is the $\{m,m\}$-th element of $\mathbf{\Lambda}_\gamma$. From Eq. (16), we know that the off-diagonal elements of $\bar{\mathbf{\Phi}}_{\mathbf{x}}$ are equal to the corresponding off-diagonal elements of $\bar{\mathbf{\Phi}}_{\mathbf{y}}$, i.e.,

$$\bar{\mathbf{\Phi}}_{\mathbf{x}\{i,j\}} = \bar{\mathbf{\Phi}}_{\mathbf{y}\{i,j\}} \text{ for } i \neq j. \tag{17}$$

Therefore, in order to estimate $\bar{\mathbf{\Phi}}_{\mathbf{x}}$ by $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{x}}$ prior to calculating $\mathbf{a}$, we first have to estimate the diagonal elements of $\bar{\mathbf{\Phi}}_{\mathbf{x}}$ as the off diagonal elements are already known from $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}}$. From now on we will use the estimated covariance matrix $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}}$ and show that we can use the off-diagonal elements of $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}}$ to estimate the diagonal elements of $\bar{\mathbf{\Phi}}_{\mathbf{x}}$.

For the m_p-th diagonal element, we can select any 2 other microphones m_q, m_r from the remaining $M-1$ microphones and obtain the following estimates

$$\widehat{\phi_s|\bar{a}_{m_p}|^2} \approx \frac{\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}\{m_p,m_q\}}\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}\{m_r,m_p\}}}{\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}\{m_r,m_q\}}} = \frac{\widehat{\phi_s\bar{a}_{m_p}\bar{a}_{m_q}^*}\,\widehat{\phi_s\bar{a}_{m_r}\bar{a}_{m_p}^*}}{\widehat{\phi_s\bar{a}_{m_r}\bar{a}_{m_q}^*}}, \tag{18}$$

or

$$\widehat{\phi_s|\bar{a}_{m_p}|^2} \approx \frac{\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}\{m_p,m_r\}}\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}\{m_q,m_p\}}}{\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}\{m_q,m_r\}}} = \frac{\widehat{\phi_s\bar{a}_{m_p}\bar{a}_{m_r}^*}\,\widehat{\phi_s\bar{a}_{m_q}\bar{a}_{m_p}^*}}{\widehat{\phi_s\bar{a}_{m_q}\bar{a}_{m_r}^*}}. \tag{19}$$

Since $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}}$ is Hermitian, Eq. (19) is the conjugate of Eq. (18). By taking the average of Eq. (19) and Eq. (18), one can insure a real valued estimate of $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{x}\{m_p,m_p\}}$.

The choice of m_q and m_r should satisfy that $m_q \neq m_r \neq m_p$ and $1 \leq m_q, m_r \leq M$. Therefore, there are $(M-1)(M-2)$ different estimates of $\phi_s|\bar{a}_{m_p}|^2$, say the set $\mathbb{L}$. We find all the estimates and take their mean value as the final estimate of $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{x}\{m_p,m_p\}}$, that is,

$$\hat{\bar{\mathbf{\Phi}}}_{\mathbf{x}\{m_p,m_p\}} = \frac{1}{(M-1)(M-2)} \sum_{\widehat{\forall\phi_s|\bar{a}_{m_p}|^2}\in\mathbb{L}} \widehat{\phi_s|\bar{a}_{m_p}|^2} \tag{20}$$

4.4. RTF estimation

Since $\bar{\mathbf{\Phi}}_{\mathbf{x}} = \phi_s\bar{\mathbf{a}}\bar{\mathbf{a}}^H$, we can estimate a scaled version of $\bar{\mathbf{a}}$ by the principal eigenvector of $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{x}}$ denoted as $\mathbf{u}$. From $\bar{\mathbf{a}} = \mathbf{U}^H\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{a}$ and $a_1 = 1$, we can estimate the RTF by

$$\hat{\mathbf{a}} = \frac{\mathbf{\Psi}^{\frac{1}{2}}\mathbf{U}\mathbf{u}}{\mathbf{e}^T\mathbf{\Psi}^{\frac{1}{2}}\mathbf{U}\mathbf{u}}. \tag{21}$$

5. EXPERIMENTS

To verify the performance of our proposed method, we simulate a room with dimension $7 \times 5 \times 4$ m and place a speech source as well as 10 microphones in the room forming a line array, as depicted in Fig. 1. Note that for some experiments, only the first a few microphones are used from left to right. The signal received at each microphone is a

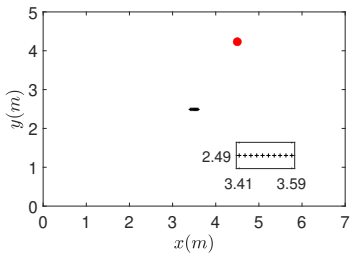
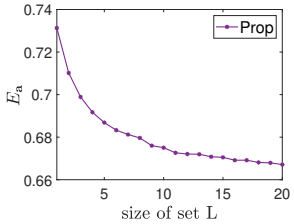


Fig. 1: Top view of the acoustic scene with a zoom-in of microphones. The source is denoted by the red circle.

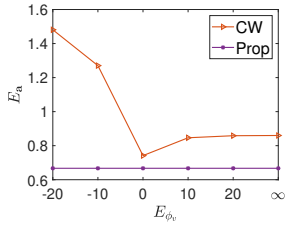
convolution between the speech source and the corresponding room impulse response. The room impulse responses are simulated by the image source method [17]. Moreover, we calculate the spatial coherence matrix of the late reverberation by assuming a spherically diffuse sound field, i.e., $\mathbf{\Gamma}_{i,j}(k) = \text{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{i,j}}{c}\right)$, with $\text{sinc}(x) = \sin x/x$, $d_{i,j}$ the inter-distance between microphones i and j, f_s the sampling frequency, c the speed of sound and K the number of frequency bins. The spatial coherence matrix of the ambient noise is set to the identity matrix, i.e. $\mathbf{\Psi} = \mathbf{I}$ simulating microphone self-noise by a zero-mean uncorrelated Gaussian process with the same variance for each microphone. The noisy microphone signals are sampled at a frequency of $f_s = 16$ kHz and processed by the STFT procedure including windowing and FFT. We use a square-root Hann window with a duration of 12.5 ms and an overlap of 75% between two adjacent time frames. The FFT length is 256. The true RTF is calculated by 256-length FFT of the first 200 samples of the room impulse responses. The RTF estimation error is

evaluated by the Hermitian angle measure (in rad) [6]

$$E_{\mathbf{a}} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K/2+1} \operatorname{acos}\left(\frac{\left|\mathbf{a}^H(l,k)\hat{\mathbf{a}}(l,k)\right|}{\left\|\mathbf{a}^H(l,k)\right\|_2\|\hat{\mathbf{a}}(l,k)\|_2}\right)}{L\,(K/2+1)}\,(\text{rad})\,. \tag{22}$$



(a) Performance of the proposed method as a function of the size of $\mathbb{L}$.

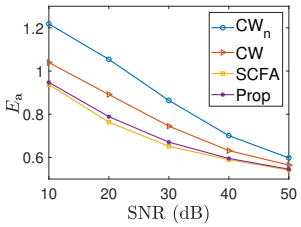


(b) Performance in terms of $E_{\mathbf{a}}$ as a function of the noise PSD estimation errors.

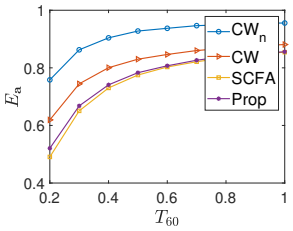
Fig. 2: Evaluation of the proposed and CW method.

For the results shown in Fig. 2, we use 6 microphones with reverberation time $T_{60} = 0.3$ s and signal-to-noise ratio (SNR) of 30 dB. Hence, we will have $(M-1)\,(M-2) = 20$ different estimates of each of the diagonal elements of $\bar{\mathbf{\Phi}}_{\mathbf{x}}$. As shown in Fig. 2a, the more estimates we average, the smaller the RTF estimation error becomes. Therefore, in the following experiments, we will average all different estimates in our proposed method. In Fig. 2b, the estimation performance of the CW method and our proposed method (referred to as 'Prop') are compared as a function of the noise PSD estimation error in dB, i.e., $E_{\phi_v} = 10\log_{10}\left(\phi_v/\hat{\phi}_v\right)$. Note that ϕ_v is the mean of the trace of the noise covariance matrix. E_{ϕ_v} ranges from an overestimation error of -20 dB to an underestimation error of ∞ dB (i.e. not subtracting anything before whitening) in Fig. 2b. Since the proposed method is independent of the noise PSD, the proposed method is not affected by E_{ϕ_v} and is presented as a horizontal line in Fig. 2b. Note that even at 0 dB, the proposed method outperforms 'CW', because the true noise spatial coherence matrix is not identical to, although close to, the identity matrix in the experiments.

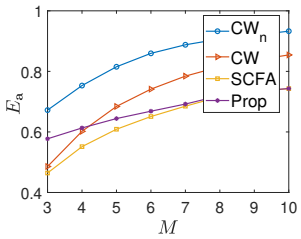
In Fig. 3, the simultaneous confirmatory factor analysis method (SCFA) [5] is also included for comparison, which minimizes the maximum likelihood cost function using the 'fmincon' MATLAB procedure after calculating the gradient and Hessian matrix at each updating step. Note that 'CW_n' refers to CW without subtracting the noise covariance matrix, i.e., $E_{\phi_v} = \infty$ dB, while 'CW' refers to $E_{\phi_v} = 0$ dB. In Fig. 3a, we use 6 microphones and fix the reverberation time to 0.3 s, and only change the SNR from 10 dB to 50 dB. In Fig. 3b, we use 6 microphones, fix the SNR to 30 dB, and only change T_{60} from 0.2 s to 1 s. From these results, it follows that our proposed method and the SCFA method have a similar performance and both outperform the CW method in most scenarios. As the SNR increases or the T_{60} decreases, all methods improve. However, the proposed method has better performance compared to 'CW' for low SNR or small T_{60}, as



(a) RTF estimation error vs SNR.



(b) RTF estimation error vs T_{60}.



(c) RTF estimation error vs M.

Fig. 3: Performance comparison of the proposed method, the CW method and the SCFA method.

the reverberation-to-noise ratio is small in both cases resulting in relatively large impact from the noise component.

In Fig. 3c, we fix the reverberation time to 0.3 s, the SNR to 30 dB, and only change the number of microphones from 3 to 10. The estimation performance of the proposed method is shown to be less good for a small number of microphones, but improves very fast when using more microphones and reaches almost the same performance as the SCFA method for large M. The reason is that we use only the off-diagonal elements of the simplified covariance matrix $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}}$ in the proposed method. The percentage of the number of elements in $\hat{\bar{\mathbf{\Phi}}}_{\mathbf{y}}$ we omit is $M/M^2 = 1/M$, which decreases as the number of microphones increases. In Table 1, we average and normalize the computation time over all scenarios per method. The runtime for Prop is close to CW, but much lower than for SCFA.

Table 1: Computation time comparison.

methods	SCFA	Prop	CW
run time	286.97	1	0.67

6. CONCLUSIONS

We considered the problem of estimating the RTF for a single source in a reverberant and noisy environment. We proposed a method that uses only off-diagonal elements of the simplified covariance matrix which are not affected by the late reverberation and the noise PSDs. Experiments show that the RTF estimation performance of the proposed method is insensitive to the noise PSD errors and reaches the performance of the SCFA method while using much less computation time. Both the proposed method and the SCFA method outperform the CW method, in most scenarios, especially for low SNR, low reverberation time and a large number of microphones.

7. REFERENCES

[1] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[2] 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.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[4] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 795–799.

[5] 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.

[6] 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.

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

[8] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 544–548.

[9] 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.

[10] 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.

[11] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multimicrophone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.

[12] 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.

[13] 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.

[14] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2101–2110, 2011.

[15] A.-J. Boonstra and A.-J. van der Veen, "Gain calibration methods for radio telescope arrays," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 25–38, 2003.

[16] S. A. Mulaik, *Foundations of factor analysis*. CRC press, 2009.

[17] 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.