ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464924

SPATIAL COVARIANCE MATRIX RECONSTRUCTION FOR SPEECH ENHANCEMENT IN REVERBERANT MULTI-SOURCE ENVIRONMENTS

Wei Liu[1,4]*, Xueqin Luo*[2]*, Jilu Jin*[2]*, Gongping Huang*[1]*, Jingdong Chen*[2]*, Jacob Benesty*[3]*, Shoji Makino*[4]

[1]School of Electronic Information, Wuhan University, Wuhan, Hubei 430072, China
[2]CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi, China
[3]INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada
[4]Graduate School of Information, Production and Systems, Waseda University, 808-0135, Japan

ABSTRACT

Accurate estimation of the noise covariance matrix is critical yet challenging in multichannel speech enhancement. In this work, we propose a spatial covariance matrix (SCM) reconstruction method for speech enhancement using compact microphone arrays in reverberant, multi-source environments. At each time–frequency bin, the normalized SCM of the array observations is modeled as a linear combination of predefined coherence matrices representing individual sources, late reverberation, and ambient noise. The combination coefficients, termed variance ratios, are estimated by minimizing the Frobenius norm between the modeled and observed normalized SCMs, subject to nonnegativity and unity-sum constraints. An adaptive algorithm is introduced to efficiently estimate these ratios, and the reconstructed SCMs are subsequently used in the multichannel Wiener filter. Simulation and experimental results show that the proposed SCM estimation method enables the multichannel Wiener filter to achieve robust and effective speech enhancement.

***Index Terms*—** Microphone arrays, speech enhancement, spatial covariance matrix reconstruction, multi-source scenario.

1. INTRODUCTION

Microphone arrays are widely employed across numerous applications [1–3]. The array observations typically contain not only the direct-path component of the target source but also reflections, interference, and ambient noise, all of which can significantly degrade speech quality. To address these challenges, a range of signal enhancement methods have been developed, including fixed and adaptive beamforming [4–8] as well as the multichannel Wiener filter (MWF) [9, 10]. The effectiveness of these approaches, however, largely depends on the accurate and robust estimation of key acoustic parameters, particularly the spatial covariance matrices (SCMs) of the sources, noise, and reverberation.

A common approach to SCM estimation uses time–frequency masks predicted by neural networks [11–14]. In this approach, a neural separator first generates soft masks indicating the proportion of speech, noise, or other components in each time–frequency bin of the mixture. These masks are then applied as weights to average the spatial correlations of the microphone signals over time, producing separate SCMs for speech and noise. However, such methods are generally offline due to their model size and computational complexity [15]. More recently, lightweight alternatives have been proposed, including directional-gain-based methods [16, 17], which estimate noise SCMs using a few fixed beamformers; their resolution, however, remains limited by the array geometry.

In general, an SCM can be decomposed into a variance term and a coherence matrix that captures the spatial structure. The coherence matrix is typically assumed to be known or pre-estimated: for example, from direction-of-arrivals (DOAs) or relative transfer functions (RTFs) for sources, or using diffuse-field models for reverberation [18–20]. Consequently, much research has focused on estimating the variance component, including reverberation variance [18, 21, 22] and noise variance [23–25]. Furthermore, several studies have investigated joint estimation of source variances and RTFs to better capture spatial characteristics [26–29].

In this work, we propose an online method for SCM reconstruction. The observation covariance is first normalized by its trace and then modeled as a linear combination of predefined spatial coherence matrices representing multiple sources, late reverberation, and ambient noise. The corresponding combination weights, referred to as variance ratios, reflect the relative contributions of these components to the noisy observation. The ratios are estimated by minimizing the Frobenius norm between the modeled and observed normalized SCMs, subject to nonnegativity and unity-sum constraints. We further introduce an adaptive algorithm to efficiently estimate the variance ratios online. The reconstructed SCMs are subsequently used to derive a multichannel Wiener filter (MWF), which is evaluated using both simulations and experiments. Results demonstrate that the proposed method provides a practical solution for multichannel speech enhancement, making it suitable for real-time applications.

2. SIGNAL MODEL AND PROBLEM FORMULATION

Consider a compact planar microphone array with M elements in a reverberant and noisy environment containing I acoustic point sources. Without loss of generality, the first microphone is chosen as the reference. In the short-time Fourier transform (STFT) domain, the array observation vector is given by

$$\begin{aligned}\mathbf{y}\left(k,n\right) &= \left[\begin{array}{cccc} Y_1\left(k,n\right) & Y_2\left(k,n\right) & \cdots & Y_M\left(k,n\right)\end{array}\right]^T \\ &= \sum_{i=1}^{I}\mathbf{x}_i\left(k,n\right)+\mathbf{r}\left(k,n\right)+\mathbf{v}\left(k,n\right), \end{aligned} \tag{1}$$

where $Y_m(k,n)$ is the signal at the mth microphone, k and n denote the frequency-bin and time-frame indices, respectively, ^{T} is the transpose operator, $\mathbf{x}_i\left(k,n\right)=\mathbf{a}_i\left(k,n\right)S_i\left(k,n\right)$, $S_i\left(k,n\right)$ represents the early signal component from the source i at the reference

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62471340 and 62192713, and in part by the China Scholarship Council (CSC) under Grant 202506270089.

microphone, $\mathbf{a}_i(k,n)$ is the RTF vector of the source i, whose mth entry is the RTF from the reference microphone to the mth microphone, $\mathbf{r}(k,n)$ is the late reverberation vector, and $\mathbf{v}(k,n)$ is the additive noise vector. In general, $\mathbf{x}_i(k,n)$, $\mathbf{r}(k,n)$, and $\mathbf{v}(k,n)$ are zero-mean, mutually uncorrelated random vectors.

If the reverberation is modeled as a diffuse sound field and the noise is assumed to be spatially white, the SCM of the array observations can be expressed as

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) &= E\left[\mathbf{y}(k,n)\,\mathbf{y}^H(k,n)\right] \\ &= \sum_{i=1}^{I}\mathbf{\Phi}_i(k,n) + \mathbf{\Phi}_{\mathbf{r}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n), \end{aligned} \tag{2}$$

where the superscript ^{H} represents the conjugate-transpose operator,

$$\begin{aligned}\mathbf{\Phi}_i(k,n) &= E\left[\mathbf{x}_i(k,n)\,\mathbf{x}_i^H(k,n)\right] \\ &= \phi_i(k,n)\,\mathbf{\Gamma}_i(k,n)\end{aligned} \tag{3}$$

is the SCM of the source i, with $\phi_i(k,n) = E\left[|S_i(k,n)|^2\right]$ and

$$\mathbf{\Gamma}_i(k,n) = \mathbf{a}_i(k,n)\,\mathbf{a}_i^H(k,n) \tag{4}$$

being the rank-one normalized covariance matrix corresponding to the source i, $E(\cdot)$ represents the expectation operator, and

$$\mathbf{\Phi}_{\mathbf{r}}(k,n) = E\left[\mathbf{r}(k,n)\,\mathbf{r}^H(k,n)\right] = \phi_R(k,n)\,\mathbf{\Gamma}_{\mathrm{d}}(k), \tag{5}$$

$$\mathbf{\Phi}_{\mathbf{v}}(k,n) = E\left[\mathbf{v}(k,n)\,\mathbf{v}^H(k,n)\right] = \phi_V(k,n)\,\mathbf{I}_M \tag{6}$$

are, respectively, the SCM of the late reverberation and noise, with $\phi_R(k,n) = \frac{1}{M}\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{r}}(k,n)\right] = E\left[|R_1(k,n)|^2\right]$, $\phi_V(k,n) = \frac{1}{M}\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{v}}(k,n)\right] = E\left[|V_1(k,n)|^2\right]$, $R_1(k,n)$ and $V_1(k,n)$ being the late reverberation component and noise component received by the reference microphone, $\mathbf{I}_M$ being the identity matrix of size $M \times M$, and $\mathbf{\Gamma}_{\mathrm{d}}(k)$ being the normalized SCM of the late reverberation. Considering a spherically isotropic noise field, the (i,j)th element of $\mathbf{\Gamma}_{\mathrm{d}}(k)$ can be written as

$$\left[\mathbf{\Gamma}_{\mathrm{d}}(k)\right]_{ij} = \mathrm{sinc}\left(\frac{2\pi f_{\mathrm{s}} k \delta_{ij}}{Kc}\right), \ \forall\, i,j \in \{1,2,\ldots,M\}, \tag{7}$$

where $\mathrm{sinc}(x) = \frac{\sin x}{x}$, f_{s} is the sampling rate, and δ_{ij} represents the distance between microphones i and j.

The normalized SCM of the array observation can be written as

$$\mathbf{\Gamma}_{\mathbf{y}}(k,n) \triangleq \frac{\mathbf{\Phi}_{\mathbf{y}}(k,n)}{\phi_Y(k,n)}, \tag{8}$$

where $\phi_Y(k,n) \triangleq \frac{1}{M}\,\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{y}}(k,n)\right] = E\left[|Y_1(k,n)|^2\right]$. For the sake of simplicity, we will disregard the dependence on k in the following content when it is evident from the context.

Substituting (3), (5), (6), and (2) into (8) , we obtain

$$\mathbf{\Gamma}_{\mathbf{y}}(n) = \sum_{i=1}^{I}\psi_i(n)\,\mathbf{\Gamma}_i(n) + \psi_R(n)\,\mathbf{\Gamma}_{\mathrm{d}} + \psi_V(n)\,\mathbf{I}_M, \tag{9}$$

where $\psi_i(n) \triangleq \frac{\phi_i(n)}{\phi_Y(n)}$, $\psi_R(n) \triangleq \frac{\phi_R(n)}{\phi_Y(n)}$, and $\psi_V(n) \triangleq \frac{\phi_V(n)}{\phi_Y(n)}$ are referred to as variance ratios. As mentioned above, for a compact microphone array, $\psi_i(n)$, $\psi_R(n)$, $\psi_V(n)$, and $\psi_Y(n)$ correspond, respectively, to the variances of the early signal component from the source i, the late reverberation component, the noise component, and the noisy signal received at the reference microphone. Accordingly, we have $\psi_i(n) \geq 0$, $\psi_R(n) \geq 0$, $\psi_V(n) \geq 0$, and

$$\sum_{i=1}^{I}\psi_i(n) + \psi_R(n) + \psi_V(n) = 1. \tag{10}$$

According to (9), $\mathbf{\Gamma}_{\mathbf{y}}(n)$ can be decomposed into a linear combination of several matrices, namely $\mathbf{\Gamma}_i(n)$, $\mathbf{\Gamma}_{\mathrm{d}}$, and the identity matrix $\mathbf{I}_M$. The matrix $\mathbf{\Gamma}_i(n)$ can be obtained in two ways: (i) from (4) by estimating the RTF vector $\mathbf{a}_i(n)$, e.g., through calibration or data-driven approaches [26, 27]; or (ii) from a known DOA θ_i using $\mathbf{\Gamma}_i = \mathbf{d}(\theta_i)\,\mathbf{d}^H(\theta_i)$ when a suitable DOA estimator is available [30], with $\mathbf{d}(\theta_i)$ being the steering vector of the planar array [2]. In this manner, all covariance matrices can be pre-modeled, and the problem of SCM reconstruction reduces to estimating the variance ratios $\{\psi_i(n)\}_{i=1}^{I}$, $\psi_R(n)$, and $\psi_V(n)$.

3. SPATIAL COVARIANCE MATRIX RECONSTRUCTION

Given $\mathbf{\Gamma}_{\mathbf{y}}(n)$, $\{\mathbf{\Gamma}_i(n)\}_{i=1}^{I}$, $\mathbf{\Gamma}_{\mathrm{d}}$, and $\mathbf{I}_M$, we can estimate the variance ratios $\boldsymbol{\psi}(n) \triangleq \left\{\{\psi_i(n)\}_{i=1}^{I}, \psi_R(n), \psi_V(n)\right\}$ by solving the following optimization problem:

$$\min_{\boldsymbol{\psi}(n)} \ \mathcal{J}\left[\boldsymbol{\psi}(n)\right] \tag{11}$$

$$\text{s. t.} \ \sum_{i=1}^{I}\psi_i(n) + \psi_R(n) + \psi_V(n) = 1,$$
$$\{\psi_i(n) \geq 0\}_{i=1}^{I}, \quad \psi_R(n) \geq 0, \quad \psi_V(n) \geq 0, \tag{12}$$

where

$$\begin{aligned}&\mathcal{J}\left[\boldsymbol{\psi}(n)\right] \\ &= \left\|\mathbf{\Gamma}_{\mathbf{y}}(n) - \sum_{i=1}^{I}\psi_i(n)\,\mathbf{\Gamma}_i(n) - \psi_R(n)\,\mathbf{\Gamma}_{\mathrm{d}} - \psi_V(n)\,\mathbf{I}_M\right\|_{\mathrm{F}}^2,\end{aligned} \tag{13}$$

with $\|\cdot\|_{\mathrm{F}}$ denoting the Frobenius norm.

Vectorizing $\mathbf{\Gamma}_{\mathbf{y}}(n)$, $\mathbf{\Gamma}_i(n)$, $\mathbf{\Gamma}_{\mathrm{d}}$, and $\mathbf{I}_M$, one can rewrite the cost function (13) as

$$\mathcal{J}\left[\mathbf{h}(n)\right] = \left\|\mathbf{c}(n) - \mathbf{\Upsilon}\mathbf{h}(n)\right\|_2^2, \tag{14}$$

where $\|\cdot\|_2$ represents the ℓ_2-norm,

$$\mathbf{h}(n) = \left[\begin{array}{cccccc}\psi_1(n) & \cdots & \psi_I(n) & \psi_R(n) & \psi_V(n)\end{array}\right]^T, \tag{15}$$

$$\mathbf{\Upsilon}(n) = \left[\begin{array}{cccccc}\boldsymbol{\gamma}_1(n) & \cdots & \boldsymbol{\gamma}_I(n) & \boldsymbol{\gamma}_{\mathrm{d}} & \mathbf{i}\end{array}\right], \tag{16}$$

with $\mathbf{c}(n) = \mathrm{vec}\left[\mathbf{\Gamma}_{\mathbf{y}}(n)\right]$, $\boldsymbol{\gamma}_i(n) = \mathrm{vec}\left[\mathbf{\Gamma}_i(n)\right]$, $\boldsymbol{\gamma}_{\mathrm{d}} = \mathrm{vec}\left[\mathbf{\Gamma}_{\mathrm{d}}\right]$, and $\mathbf{i} = \mathrm{vec}\left[\mathbf{I}_M\right]$ with $\mathrm{vec}[\cdot]$ denoting the vectorization of a matrix. Then, the original optimization problem in (11) can be converted to

$$\min_{\mathbf{h}(n)} \ \mathcal{J}\left[\mathbf{h}(n)\right] \ \text{ s. t. } \ \mathbf{h}(n) \succeq 0, \ \|\mathbf{h}(n)\|_1 = 1. \tag{17}$$

In real-time processing, the a priori and a posteriori errors are defined respectively as

$$\mathbf{e}(n) = \mathbf{c}(n) - \mathbf{\Upsilon}(n)\,\mathbf{h}(n-1), \tag{18}$$

$$\boldsymbol{\varepsilon}(n) = \mathbf{c}(n) - \mathbf{\Upsilon}(n)\,\mathbf{h}(n). \tag{19}$$

To solve (17), we introduce the following Lagrangian function:

$$\mathcal{L}\left[\mathbf{h}\left(n\right)\right]=\mathcal{J}\left[\mathbf{h}\left(n\right)\right]+\lambda\mathcal{K}\left[\mathbf{h}\left(n\right)\right]+\mu\left[\left\|\mathbf{h}\left(n\right)\right\|_{1}-1\right], \quad (20)$$

where

$$\mathcal{K}\left[\mathbf{h}\left(n\right)\right]=\sum_{i=1}^{I}\psi_{i}\left(n\right)\ln\frac{\psi_{i}\left(n\right)}{\psi_{i}\left(n-1\right)}+\psi_{R}\left(n\right)\ln\frac{\psi_{R}\left(n\right)}{\psi_{R}\left(n-1\right)}+\psi_{V}\left(n\right)\ln\frac{\psi_{V}\left(n\right)}{\psi_{V}\left(n-1\right)} \quad (21)$$

is the Kullback-Leibler (KL) divergence between $\mathbf{h}\left(n\right)$ and $\mathbf{h}\left(n-1\right)$, minimized to control the update step size.

Since $\mathbf{h}\left(n\right)\succeq 0$, the derivative of $\left\|\mathbf{h}\left(n\right)\right\|_{1}-1$ with respect to $\mathbf{h}\left(n\right)$ is $\mathbf{1}$, a vector of ones of length $\left(I+2\right)$. Therefore, the derivative of $\mathcal{L}\left[\mathbf{h}\left(n\right)\right]$ with respect to $\mathbf{h}\left(n\right)$ is

$$\frac{d\mathcal{L}\left[\mathbf{h}\left(n\right)\right]}{d\mathbf{h}\left(n\right)}=\lambda\left[\ln\mathbf{h}\left(n\right)-\ln\mathbf{h}\left(n-1\right)+\mathbf{1}\right]-2\Re\left[\boldsymbol{\Upsilon}^{H}\left(n\right)\boldsymbol{\varepsilon}\left(n\right)\right]+\mu\mathbf{1}, \quad (22)$$

where $\Re[\cdot]$ denotes the real part. Setting this expression to zero gives

$$\ln\mathbf{h}\left(n\right)-\ln\mathbf{h}\left(n-1\right)=-\frac{\lambda+\mu}{\lambda}\mathbf{1}+\frac{2}{\lambda}\Re\left[\boldsymbol{\Upsilon}^{H}\left(n\right)\boldsymbol{\varepsilon}\left(n\right)\right]. \quad (23)$$

Converting it to exponential representation, we obtain

$$\mathbf{h}\left(n\right)=\mathbf{h}\left(n-1\right)\circ\mathbf{g}\left(n\right), \quad (24)$$

where $\circ$ denotes the Hadamard product and

$$\mathbf{g}\left(n\right)=\exp\left\{-\frac{\lambda+\mu}{\lambda}\mathbf{1}+\frac{2}{\lambda}\Re\left[\boldsymbol{\Upsilon}^{H}\left(n\right)\boldsymbol{\varepsilon}\left(n\right)\right]\right\}. \quad (25)$$

To enforce $\left\|\mathbf{h}\left(n\right)\right\|_{1}=1$, we make the following normalization:

$$\mathbf{h}\left(n\right)=\frac{\mathbf{h}\left(n-1\right)\circ\mathbf{r}\left(n\right)}{\mathbf{h}^{T}\left(n-1\right)\mathbf{r}\left(n\right)}, \quad (26)$$

where the multiplicative vector:

$$\begin{aligned}\mathbf{r}\left(n\right)&=\mathbf{g}\left(n\right)\circ\exp\left\{\frac{\lambda+\mu}{\lambda}\mathbf{1}\right\}\\&=\exp\left\{\frac{2}{\lambda}\Re\left[\boldsymbol{\Upsilon}^{H}\left(n\right)\boldsymbol{\varepsilon}\left(n\right)\right]\right\}\\&=\exp\left\{\eta\Re\left[\boldsymbol{\Upsilon}^{H}\left(n\right)\boldsymbol{\varepsilon}\left(n\right)\right]\right\},\end{aligned} \quad (27)$$

with a stepsize $\eta>0$. With the normalization, the dependence of $\mathbf{r}(n)$ on μ is eliminated, while λ remains but is incorporated into the coefficient η. Since the *a posteriori* estimate is not known before the update, we use the *a priori* estimate instead, so we have

$$\mathbf{r}\left(n\right)=\exp\left\{\eta\Re\left[\boldsymbol{\Upsilon}^{H}\left(n\right)\mathbf{e}\left(n\right)\right]\right\}. \quad (28)$$

For real-time applications, the SCM of the array observations can be updated recursively as

$$\hat{\boldsymbol{\Phi}}_{\mathbf{y}}\left(n\right)=\alpha\hat{\boldsymbol{\Phi}}_{\mathbf{y}}\left(n-1\right)+\left(1-\alpha\right)\mathbf{y}\left(n\right)\mathbf{y}^{H}\left(n\right), \quad (29)$$

where $\alpha\in(0,1)$ is a forgetting factor.

Next, we present the detailed pseudocode of the proposed SCM reconstruction method in Algorithm 1. The per time-frequency bin computational complexity of the algorithm is $\mathcal{O}\left(M^{2}\left(I+2\right)\right)$.

Algorithm 1

Input: $\boldsymbol{\Upsilon}\left(n\right)$, $\mathbf{y}\left(n\right)$, $\mathbf{h}\left(n-1\right)$, α, and η.

repeat {at each index n}
  Update the SCM of array observation $\hat{\boldsymbol{\Phi}}_{\mathbf{y}}\left(n\right)$ by (29).
  Normalize $\hat{\boldsymbol{\Phi}}_{\mathbf{y}}\left(n\right)$ by (8).
  Vectorize $\boldsymbol{\Gamma}_{\mathbf{y}}\left(n\right)$ via $\mathbf{c}\left(n\right)=\mathrm{vec}\left[\boldsymbol{\Gamma}_{\mathbf{y}}\left(n\right)\right]$.
  Compute the *priori* error $\mathbf{e}\left(n\right)$ by (18).
  Compute the multiplicative vector $\mathbf{r}\left(n\right)$ by (28).
  Update the variance ratio vector $\mathbf{h}\left(n\right)$ by (26).
until

4. EXPERIMENTS

In this section, we apply the proposed SCM reconstruction method to derive the MWF. Assuming the first source is the target one, the filter is expressed as

$$\mathbf{h}_{\mathrm{W},1}\left(n\right)=\psi_{1}\left(n\right)\boldsymbol{\Gamma}_{\mathbf{y}}^{-1}\left(n\right)\boldsymbol{\Gamma}_{1}\left(n\right)\mathbf{u}, \quad (30)$$

where $\mathbf{u}=\left[\begin{array}{cccc}1 & 0 & \cdots & 0\end{array}\right]^{T}$ and $\boldsymbol{\Gamma}_{\mathbf{y}}\left(n\right)$ is computed according to (9). We refer to this formulation as the SCM reconstruction-based MWF (R-MWF). The stepsize η in (28) is set to 0.1, and the forgetting factor α in (29) is set to 0.5. The DOAs of all sources are assumed to be known, such that $\boldsymbol{\Gamma}_{i}=\mathbf{d}\left(\theta_{i}\right)\mathbf{d}^{H}\left(\theta_{i}\right)$.

We compare the proposed R-MWF with two recent approaches: the directional-gain MVDR beamformer (DG-MVDR) [17] and the MWF with minimum variance joint diagonalization (MVJD-MWF) [29]. Unlike R-MWF and DG-MVDR, MVJD-MWF requires prior knowledge of the noise covariance but can estimate the RTFs of multiple sources. For fairness, we consider two variants: MVJD-MWF-I, which operates without source prior knowledge, and MVJD-MWF-II, which assumes known steering vectors for RTF estimation. In both cases, the noise covariance is computed from noise-only segments. Evaluation on simulated data is presented in Section 4.1, and evaluation on real recordings in Section 4.2.

4.1. Simulations with Generated RIRs

In this subsection, we assess the speech enhancement performance of the proposed method in a simulated acoustic environment. The simulation is conducted in a rectangular room of size $8\times6\times3$ m^3, with the origin of the 3D Cartesian coordinate system at the corner $(0,0,0)$ and the x-, y-, and z-axes aligned with the room's length, width, and height. A uniform linear array (ULA) of $M=4$ omni-directional microphones with an inter-element spacing of 2.0 cm, is placed at the room center. As shown in Fig. 1, three sound sources

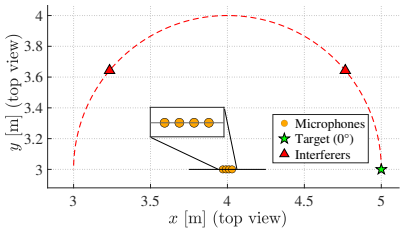

Fig. 1: Top view of the simulated acoustic scene.

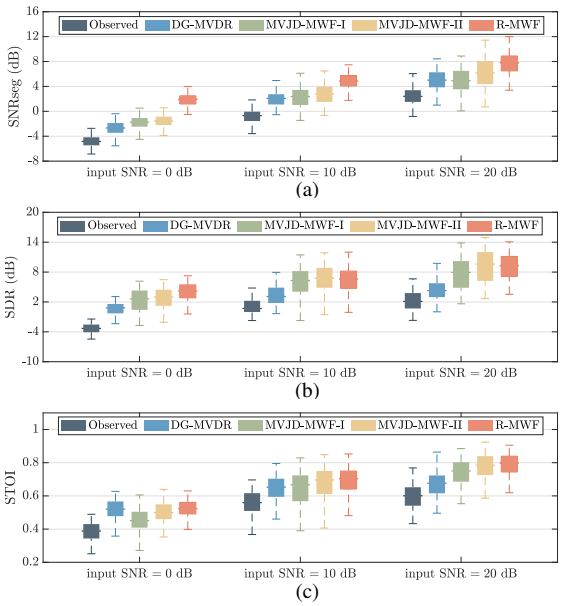

Fig. 2: Performance of the DG-MVDR, MVJD-MWF-I, MVJD-MWF-II, and R-MWF in three different input SNR levels: (a) SNRseg, (b) SDR, and (c) STOI. Conditions: $M = 4$, the input SIRs of two interfering source are randomly sampled from $[0, 10]$ dB, and the reverberation time is approximately 300 ms.

are positioned on a semicircular arc of radius 1 m centered at the array, all lying in the same horizontal plane as the array. The target source is fixed at $(5, 3, 1.5)$, corresponding to $\theta_1 = 0^\circ$ (endfire direction with respect to the array). The azimuths of the two interfering sources are randomly drawn, ensuring a minimum angular separation of 15° between any two sources.

The source signals consist of clean speech utterances from the TIMIT database, sampled at 16 kHz. Room impulse responses (RIRs) are generated using the image method [31, 32] with a reverberation time of $T_{60} \approx 300$ ms. The observed signals are obtained by convolving the source signals with the corresponding RIRs and adding white Gaussian noise at input SNRs between 0 and 20 dB. The input signal-to-interference ratios (SIRs) of the two interfering sources are uniformly sampled from $[0, 10]$ dB. Array observations are processed in the STFT domain using a frame size of 256, 75% overlap, and a Kaiser window with $\beta = 1.9\pi$.

Performance is assessed using three objective metrics: frequency-weighted segmental SNR (SNRseg), signal-to-distortion ratio (SDR), and short-time objective intelligibility (STOI) [33, 34]. The direct-path component of the target speech is taken as the reference signal for all metrics. Simulations are conducted at three SNR levels: 0 dB, 10 dB, and 20 dB, with each configuration repeated 100 times. Results are shown as box plots in Fig. 2. The results clearly show that R-MWF consistently outperforms the baseline methods in terms of SNRseg, SDR, and STOI, confirming that the proposed SCM reconstruction strategy enables robust and perceptually effective speech enhancement across different noise and interference conditions.

4.2. Experiments with Real Recordings

In this subsection, we assess the speech dereverberation performance of the proposed method using real recordings from the publicly available RealMAN dataset [35]. Recordings from three acoustic scenes

Table 1: Description of real recordings from the RealMAN dataset.

	Scene	Speaker	Distance	Azimuth
Case-1	LivingRoom6	Female	0.89 m	118.23°
Case-2	OfficeRoom1	Male	0.80 m	146.60°
Case-3	BadmintonCourt1	Female	6 m	54.82°

Table 2: Performance of the comparison methods under three different real recording scenarios.

Methods	Case-1: $T_{60} = 398$ ms			
	SNRseg (dB) ↑	SDR (dB) ↑	STOI ↑	CD ↓
Observed	1.16	6.43	0.68	4.37
DG-MVDR	2.66	7.20	0.71	3.86
MVJD-MWF-I	2.98	7.35	0.70	3.82
MVJD-MWF-II	3.07	7.20	0.70	3.93
R-MWF	**4.66**	**9.15**	**0.76**	**3.51**
Methods	Case-2: $T_{60} = 719$ ms			
	SNRseg (dB) ↑	SDR (dB) ↑	STOI ↑	CD ↓
Observed	2.11	0.02	0.75	4.75
DG-MVDR	4.15	6.03	0.80	4.00
MVJD-MWF-I	4.23	5.76	0.78	4.02
MVJD-MWF-II	4.95	6.12	0.79	**3.94**
R-MWF	**5.54**	**6.83**	**0.85**	4.11
Methods	Case-3: $T_{60} = 1577$ ms			
	SNRseg (dB) ↑	SDR (dB) ↑	STOI ↑	CD ↓
Observed	0.52	-6.00	0.41	4.73
DG-MVDR	1.49	3.40	0.45	4.50
MVJD-MWF-I	1.74	3.67	0.43	4.50
MVJD-MWF-II	1.83	3.85	0.44	**4.49**
R-MWF	**2.87**	**4.99**	**0.49**	4.66

are considered: *LivingRoom6*, *OfficeRoom1*, and *BadmintonCourt1*, corresponding to Case-1, Case-2, and Case-3, respectively. The approximate reverberation times for these scenes are 398 ms, 719 ms, and 1577 ms. The RealMAN dataset provides 32-channel microphone array recordings, from which we select channels 1, 3, 5, and 7 to form a 4-element uniform circular array (UCA) with a radius of 3 cm. Additional configuration details are provided in Table 1.

Evaluation is conducted using SNRseg, SDR, STOI, and cepstral distance (CD) [34] with the clean target speech (serving as the reference) obtained by filtering the source signal with an estimated direct-path filter [35]. Table 2 presents the dereverberation results of different methods under these real acoustic environments. Across all three cases, the proposed R-MWF consistently delivers the best or near-best results across all four metrics, demonstrating strong generalization to real-world recordings and stable performance under a wide range of reverberation conditions, rather than being limited to diffuse-field models.

5. CONCLUSIONS

This paper presented an SCM reconstruction method in reverberant, multi-source environments. By modeling the normalized array covariance as a linear combination of predefined coherence matrices, SCM reconstruction was reduced to estimating the corresponding combination weights, or variance ratios, which reflect the relative contributions of the components to the noisy observation. The variance ratios were estimated using a lightweight multiplicative update rule, enabling efficient online implementation. When incorporated into a multichannel Wiener filter, both simulation and experimental results demonstrated that the proposed method achieves competitive speech enhancement performance.

6. REFERENCES

[1] H. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation theory*. John Wiley Sons, Inc, 2002.

[2] J. Benesty, G. Huang, J. Chen, and N. Pan, *Microphone Arrays*. Berlin, Germany: Springer-Verlag, 2023, vol. 22.

[3] G. Huang, J. R. Jensen, J. Chen, J. Benesty, M. G. Christensen, A. Sugiyama, G. Elko, and T. Gaensler, "Advances in microphone array processing and multichannel speech enhancement," in *Proc. IEEE ICASSP*, 2025, pp. 1–5.

[4] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[5] W. Liu, J. Benesty, G. Huang, and J. Chen, "Beamforming in the short-time Fourier transform domain via dimensionality reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 1730–1742, Apr. 2025.

[6] G. Huang, J. Benesty, and J. Chen, "Fundamental approaches to robust differential beamforming with high directivity factors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3074–3088, Sep. 2022.

[7] A. H. Moore, S. Hafezi, R. R. Vos, P. A. Naylor, and M. Brookes, "A compact noise covariance matrix model for MVDR beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2049–2061, Jun. 2022.

[8] J. Jin, X. Luo, G. Huang, J. Chen, and J. Benesty, "Beamforming through online convex combination of differential beamformers," in *Proc. IEEE ICASSP*, 2024, pp. 8561–8565.

[9] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE Press., 2018.

[10] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1218–1234, Jul. 2006.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[12] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3461–3475, Nov. 2021.

[13] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *Proc. IEEE ICASSP*, 2019, pp. 6855–6859.

[14] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, Jan. 2022.

[15] G. Richard, P. Smaragdis, S. Gannot, P. A. Naylor, S. Makino, W. Kellermann, and A. Sugiyama, "Audio signal processing in the 21st century: The important outcomes of the past 25 years," *IEEE Signal Process. Mag.*, vol. 40, no. 5, pp. 12–26, Jul. 2023.

[16] C. Pan and J. Chen, "A framework of directional-gain beamforming and a white-noise-gain-controlled solution," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2875–2887, Aug. 2022.

[17] F. Zhang, C. Pan, J. Benesty, and J. Chen, "Directional gain based noise covariance matrix estimation for MVDR beamforming," in *Proc. IEEE ICASSP*, 2024, pp. 511–515.

[18] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, Apr. 2015.

[19] C. Li and R. C. Hendriks, "Alternating least-squares-based microphone array parameter estimation for a single-source reverberant and noisy acoustic scenario," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3922–3934, Aug. 2023.

[20] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1052–1067, Feb. 2018.

[21] O. Schwartz, S. Gannot, and E. A. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, Apr. 2016.

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

[23] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jun. 2012.

[24] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[25] M. Taseska and E. A. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2223–2236, Sep. 2017.

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

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

[28] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 755–769, Jan. 2020.

[29] C. Li and R. C. Hendriks, "Multimicrophone signal parameter estimation in a multi-source noisy reverberant scenario," *IEEE Trans. Audio, Speech, Lang. Process.*, Jan. 2025.

[30] T. E. Tuncer and B. Friedlander, *Classical and Modern Direction-of-Arrival Estimation*. Academic Press, 2009.

[31] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[32] C. Pan, L. Zhang, Y. Lu, J. Jin, L. Qiu, J. Chen, and J. Benesty, "An anchor-point based image-model for room impulse response simulation with directional source radiation and sensor directivity patterns," arXiv:2308.10543, 2023.

[33] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2007.

[34] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adva. Signal Process.*, vol. 2016, no. 1, p. 7, Jan. 2016.

[35] B. Yang, C. Quan, Y. Wang, P. Wang, Y. Yang, Y. Fang, N. Shao, H. Bu, X. Xu, and X. Li, "RealMAN: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 105997–106019, 2024.