JOINT ESTIMATION OF RETF VECTOR AND POWER SPECTRAL DENSITIES FOR SPEECH ENHANCEMENT BASED ON ALTERNATING LEAST SQUARES

Marvin Tammen, Simon Doclo

University of Oldenburg, Dept. of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All, Oldenburg, Germany
{marvin.tammen,simon.doclo}@uol.de

Ina Kodrasi

Speech and Audio Processing Group
Idiap Research Institute, Martigny, Switzerland
ina.kodrasi@idiap.ch

ABSTRACT

The multi-channel Wiener filter (MWF) is a well-known multi-microphone speech enhancement technique, aiming at improving the quality of the recorded speech signals in noisy and reverberant environments. Assuming that reverberation and ambient noise can be modeled as a diffuse sound field and the spatial coherence of the residual noise is known, the MWF requires estimates of the relative early transfer function (RETF) vector of the target speaker as well as the power spectral densities (PSDs) of the target, diffuse and residual noise component. RETF vector and PSD estimation is often decoupled, where one quantity is estimated independently of the other quantity. In this paper, we propose to jointly estimate the RETF vector and all PSDs by minimizing the Frobenius norm of a model-based error matrix using an alternating least squares method. Experimental results using different dynamic acoustic scenarios with a moving speaker show that the proposed method leads to a larger MWF performance than a state-of-the-art method based on covariance whitening.

***Index Terms*—** MWF, dereverberation, noise reduction, PSD estimation, RETF estimation

1. INTRODUCTION

In many hands-free speech communication applications, such as teleconferencing and hearing aids, the recorded speech signals are corrupted by reverberation and ambient noise. Since this may lead to a decreased speech quality and a performance degradation of automatic speech recognition systems [1], [2], speech enhancement techniques that are capable of suppressing late reverberation as well as ambient noise are required. Multi-microphone techniques are generally preferred over single-microphone techniques due to their capability of exploiting spatial information. A frequently used technique is the multi-channel Wiener filter (MWF) [3]–[5], which minimizes the mean square error between a target signal and the output signal. Modeling late reverberation and ambient noise as a diffuse sound field and residual noise (e.g., sensor noise) as a spatially homogeneous sound field with time-varying power spectral densities (PSDs), the MWF requires estimates of the (possibly time-varying) relative early transfer function (RETF) vector of the target speaker as well as the time-varying target, diffuse and residual noise PSDs.

Various RETF vector and diffuse PSD estimators have been proposed in the literature. The RETF vector has been estimated, e.g., based on the least squares method [6], [7], the covariance subtraction method [8]–[10], or the covariance whitening (CW) method [9]–[12]. The PSD of diffuse sound fields has been estimated, e.g., using maximum likelihood-based estimators [13]–[15], Frobenius norm-based estimators [15], [16], or an eigenvalue decomposition (EVD)-based estimator [17]. While most RETF vector and diffuse PSD estimators are decoupled, the combination of different estimators as well as joint estimators have been proposed. In [18] it has been shown that estimating the RETF vector using the CW method and the diffuse PSD using the EVD-based method results in a high dereverberation and noise reduction performance, both in stationary as well as moving speaker scenarios. Based on the minimization of the Frobenius norm of a model-based error matrix, in [19] an alternating least squares (ALS) method has been proposed to jointly estimate the (possibly time-varying) RETF vector as well as the target and the diffuse PSDs. Simulation results for a spatially stationary speaker in a perfectly diffuse noise field show that this method outperforms the CW-based method in [18].

All previously described estimators assume the residual noise PSD matrix or the residual noise PSD to be known. The residual noise PSD matrix is typically estimated during speech pauses detected by a voice activity detector, generally requiring the residual noise PSD to be rather stationary. In [20] a Frobenius norm-based estimator for joint diffuse and residual noise PSD estimation has been proposed, where the residual noise is modeled by a spatially homogeneous sound field with a time-varying PSD. Motivated by [20], in this paper we propose an extension of the method in [19], jointly estimating not only the RETF vector and the target and diffuse PSDs, but also the residual noise PSD in an ALS fashion. For different dynamic acoustic scenarios with a moving speaker in a reverberant environment the performance of an MWF using the RETF vector and PSD estimates of either the proposed method, the method in [19], or the CW method is compared, showing that the proposed method leads to the highest performance in the presence of residual noise.

2. SIGNAL MODEL

We consider an acoustic scenario with one (possibly moving) target speaker as well as diffuse and residual noise in a reverberant environment. In the short-time Fourier transform (STFT) domain, the stacked vector of noisy and reverberant microphone signals

$$\mathbf{y}(k,l) = [Y_1(k,l),\, Y_2(k,l),\, ...,\, Y_M(k,l)]^T, \tag{1}$$

with M the number of microphones, k the frequency index, and l the frame index, can be written as

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{d}(k,l) + \mathbf{v}(k,l), \tag{2}$$

where $\mathbf{x}(k,l)$ denotes the speech component, $\mathbf{d}(k,l)$ denotes the diffuse component (representing late reverberation as well as diffuse noise), and $\mathbf{v}(k,l)$ denotes the residual noise component (e.g., sensor noise).

The target component can be modeled as

$$\mathbf{x}(k,l) = \mathbf{a}(k,l) X_{1,e}(k,l), \tag{3}$$

where $X_{1,e}(k,l)$ denotes the early reverberant speech component in the first microphone signal (i.e., the target signal) and

This work was supported by the Cluster of Excellence 1077 Hearing4all, funded by the German Research Foundation (DFG), and by the joint Lower Saxony-Israeli Project ATHENA.

$\mathbf{a}(k,l) = [1,\ A_2(k,l),\ \ldots,\ A_M(k,l)]^T$ denotes the RETF vector of the target speaker, with the first microphone selected as the reference microphone without loss of generality. Assuming that all components in (2) are uncorrelated, the $M \times M$-dimensional microphone PSD matrix $\mathbf{\Phi_y}(k,l) = \mathcal{E}\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\}$, with $\mathcal{E}\{\cdot\}$ the expectation operator, can be written as

$$\boxed{\mathbf{\Phi_y}(k,l) = \underbrace{\phi_s(k,l)\mathbf{a}(k,l)\mathbf{a}^H(k,l)}_{\mathbf{\Phi_x}(k,l)} + \underbrace{\phi_d(k,l)\mathbf{\Gamma}(k)}_{\mathbf{\Phi_d}(k,l)} + \underbrace{\phi_v(k,l)\mathbf{\Psi}(k)}_{\mathbf{\Phi_v}(k,l)},} \quad (4)$$

where $\phi_s(k,l) = \mathcal{E}\left\{|X_{1,e}(k,l)|^2\right\}$ denotes the target PSD, $\phi_d(k,l)$ denotes the diffuse PSD, $\phi_v(k,l)$ denotes the residual noise PSD, and the matrices $\mathbf{\Gamma}(k)$ and $\mathbf{\Psi}(k)$ denote the spatial coherence matrices of the diffuse and residual noise components, which are assumed to be time-invariant and known. As defined in (4), $\mathbf{\Phi_x}(k,l)$, $\mathbf{\Phi_d}(k,l)$ and $\mathbf{\Phi_v}(k,l)$ denote the PSD matrices of the target, diffuse and residual noise component.

The MWF $\mathbf{w}_{\mathrm{MWF}}(k,l)$ produces the minimum mean square error estimate of the target signal $X_{1,e}(k,l)$, solving the optimization problem

$$\mathbf{w}_{\mathrm{MWF}}(k,l) = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{E}\left\{|\mathbf{w}^H(k,l)\mathbf{y}(k,l) - X_{1,e}(k,l)|^2\right\}. \quad (5)$$

Using the signal model in (3) and (4), the MWF is given as [3], [5]

$$\mathbf{w}_{\mathrm{MWF}}(k,l) = \underbrace{\frac{[\phi_d(k,l)\mathbf{\Gamma}(k) + \phi_v(k,l)\mathbf{\Psi}(k)]^{-1}\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)[\phi_d(k,l)\mathbf{\Gamma}(k) + \phi_v(k,l)\mathbf{\Psi}(k)]^{-1}\mathbf{a}(k,l)}}_{\mathbf{w}_{\mathrm{MVDR}}(k,l)} \frac{\xi(k,l)}{1+\xi(k,l)}, \quad (6)$$

where $\mathbf{w}_{\mathrm{MVDR}}(k,\, l)$ denotes the minimum variance distortionless response (MVDR) beamformer, and $\xi(k,l)$ denotes the a-priori signal-to-noise ratio (SNR) at the output of the MVDR beamformer, i.e.,

$$\xi(k,l) = \frac{\phi_s(k,l)}{\mathbf{w}_{\mathrm{MVDR}}^H(k,l)[\phi_d(k,l)\mathbf{\Gamma}(k) + \phi_v(k,l)\mathbf{\Psi}(k)]\mathbf{w}_{\mathrm{MVDR}}(k,l)}. \quad (7)$$

As can be observed from (6) and (7), the MWF requires estimates of the RETF vector $\mathbf{a}(k,l)$ as well as the PSDs $\{\phi_s(k,l), \phi_d(k,l), \phi_v(k,l)\}$. In the following sections, two methods are discussed for joint RETF vector and PSD estimation. In Section 3 we briefly review the baseline CW method [17], [18], which estimates the RETF vector $\mathbf{a}(k,l)$ and the target and diffuse PSDs $\phi_s(k,l)$ and $\phi_d(k,l)$, but requires an estimate of the residual noise PSD matrix $\mathbf{\Phi_v}(k,l)$ to be known. In Section 4 we present a novel method to jointly estimate the RETF vector and all PSDs $\{\phi_s(k,l), \phi_d(k,l), \phi_v(k,l)\}$ based on an ALS approach to minimize the Frobenius norm of a model-based error matrix. This method is an extension of [19], which only yields estimates of the RETF vector $\mathbf{a}(k,l)$ as well as the target and diffuse PSDs $\phi_s(k,l)$ and $\phi_d(k,l)$. Please note that in practice, instead of directly using the estimated target PSD, the decision-directed approach (DDA) as described in [21], [22] will be used to estimate the a-priori SNR, as has also been proposed in [7], [18]–[20].

In addition, in practice the microphone PSD matrix $\mathbf{\Phi_y}(k,l)$ will be estimated from the microphone signals using recursive averaging, i.e.,

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(l) = \alpha\hat{\mathbf{\Phi}}_{\mathbf{y}}(l-1) + (1-\alpha)\mathbf{y}(l)\mathbf{y}^H(l). \quad (8)$$

For conciseness, the frequency and frame indices k and l will be omitted in the remainder of this paper whenever possible.

3. COVARIANCE WHITENING (CW) METHOD

Assuming that an estimate of the residual noise PSD matrix $\mathbf{\Phi_v}$ is available, the microphone PSD matrix $\mathbf{\Phi_y}$ in (4) can be modified as

$$\tilde{\mathbf{\Phi}}_{\mathbf{y}} = \mathbf{\Phi_y} - \mathbf{\Phi_v} = \phi_s\mathbf{a}\mathbf{a}^H + \phi_d\mathbf{\Gamma}. \quad (9)$$

Using the Cholesky decomposition of the spatial coherence matrix $\mathbf{\Gamma}$, i.e., $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H$, with $\mathbf{L}$ an $M \times M$-dimensional lower triangular matrix, the prewhitened (modified) microphone PSD matrix is equal to

$$\tilde{\mathbf{\Phi}}_{\mathbf{y}}^{w} = \mathbf{L}^{-1}\tilde{\mathbf{\Phi}}_{\mathbf{y}}\mathbf{L}^{-H} = \phi_s(\mathbf{L}^{-1}\mathbf{a})(\mathbf{L}^{-1}\mathbf{a})^H + \phi_d\mathbf{I}_M, \quad (10)$$

where $\mathbf{I}_M$ denotes the $M \times M$-dimensional identity matrix. The EVD of $\tilde{\mathbf{\Phi}}_{\mathbf{y}}^{w}$ is equal to

$$\tilde{\mathbf{\Phi}}_{\mathbf{y}}^{w} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \quad (11)$$

where $\mathbf{U}$ and $\mathbf{\Lambda}$ are $M \times M$-dimensional matrices containing the eigenvectors and the eigenvalues of $\tilde{\mathbf{\Phi}}_{\mathbf{y}}^{w}$, respectively. As shown in [17], [18], the RETF vector $\mathbf{a}$ is equal to a scaled version of the transformed principal eigenvector $\mathbf{L}\mathbf{u}_1$, and all eigenvalues except the principal eigenvalue λ_1 are equal to the diffuse PSD. Hence, an estimate of the RETF vector and the diffuse PSD can be obtained as

$$\begin{cases} \hat{\mathbf{a}}_{\mathrm{CW}} = \mathbf{L}\mathbf{u}_1 / \left(\mathbf{e}^T\mathbf{L}\mathbf{u}_1\right) & (12) \\ \hat{\phi}_{d,\mathrm{CW}} = \left(\mathrm{trace}\{\tilde{\mathbf{\Phi}}_{\mathbf{y}}^{w}\} - \lambda_1\right)/(M-1), & (13) \end{cases}$$

with the M-dimensional selection vector $\mathbf{e} = [1,0,\ldots,0]^T$.

A drawback of the CW method is the fact that the noise component which can be addressed is completely determined by the spatial coherence matrix $\mathbf{\Gamma}$ used for prewhitening. Hence, to apply this method in the presence of residual noise, an estimate of the residual noise PSD matrix $\mathbf{\Phi_v}$ needs to be available and subtracted from the estimated microphone PSD matrix, i.e., $\hat{\mathbf{\Phi}}_{\mathbf{y}} - \hat{\mathbf{\Phi}}_{\mathbf{v}}$ (cf. (9)).

4. ALTERNATING LEAST SQUARES METHOD

Based on the model in (4), in this section we propose a method to jointly estimate the RETF vector $\mathbf{a}$ and the PSDs $\boldsymbol{\phi} = [\phi_s,\ \phi_d,\ \phi_v]^T$ by minimizing the Frobenius norm of a model-based error matrix, i.e.,

$$\left(\hat{\mathbf{a}}_{\mathrm{ALS}},\, \hat{\boldsymbol{\phi}}_{\mathrm{ALS}}\right) = \underset{\mathbf{a},\ \boldsymbol{\phi}}{\operatorname{argmin}} \| \underbrace{\hat{\mathbf{\Phi}}_{\mathbf{y}} - (\phi_d\mathbf{\Gamma} + \phi_v\mathbf{\Psi})}_{=:\hat{\mathbf{\Phi}}_{\mathbf{x}}} - \phi_s\mathbf{a}\mathbf{a}^H \|_F^2. \quad (14)$$

Since, to the best of our knowledge, the optimization problem in (14) has no closed-form solution, we propose to use a two-step ALS method to obtain the estimates $\hat{\mathbf{a}}_{\mathrm{ALS}}$ and $\hat{\boldsymbol{\phi}}_{\mathrm{ALS}}$, similarly to the method proposed in [19].

First, the RETF vector $\mathbf{a}$ is assumed to be fixed to the estimate from the previous iteration $\hat{\mathbf{a}}^{(i-1)}$, with i the iteration index, and the minimization is performed with respect to the PSDs, i.e., [15], [20]

$$\boxed{\hat{\boldsymbol{\phi}}_{\mathrm{ALS}}^{(i)} = \left(\mathbf{A}^{(i-1)}\right)^{-1}\mathbf{b}^{(i)},} \quad (15)$$

with the matrix $\mathbf{A}^{(i)}$ defined as

$$\mathbf{A}^{(i)} = \begin{bmatrix} \left(\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i),H}\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)}\right)^2 & \hat{\mathbf{a}}_{\mathrm{ALS}}^{(i),H}\mathbf{\Gamma}\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} & \hat{\mathbf{a}}_{\mathrm{ALS}}^{(i),H}\mathbf{\Psi}\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} \\ \hat{\mathbf{a}}_{\mathrm{ALS}}^{(i),H}\mathbf{\Gamma}\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} & \mathrm{trace}\{\mathbf{\Gamma}^H\mathbf{\Gamma}\} & \mathrm{trace}\{\mathbf{\Gamma}^H\mathbf{\Psi}\} \\ \hat{\mathbf{a}}_{\mathrm{ALS}}^{(i),H}\mathbf{\Psi}\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} & \mathrm{trace}\{\mathbf{\Gamma}^H\mathbf{\Psi}\} & \mathrm{trace}\{\mathbf{\Psi}^H\mathbf{\Psi}\} \end{bmatrix} \quad (16)$$

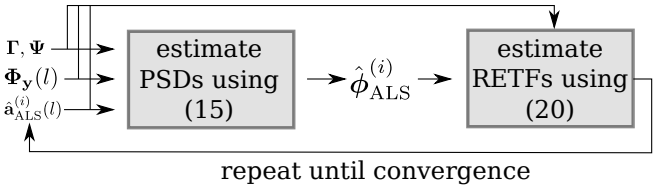

Fig. 1: Block diagram of ALS-based RETF vector and PSD estimation.

and the vector $\mathbf{b}^{(i)}$ defined as

$$\mathbf{b}^{(i)} = \begin{bmatrix} \hat{\mathbf{a}}_{\mathrm{ALS}}^{(i),H} \hat{\mathbf{\Phi}}_{\mathbf{y}} \hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} \\ \mathrm{trace}\{\hat{\mathbf{\Phi}}_{\mathbf{y}} \mathbf{\Gamma}^H\} \\ \mathrm{trace}\{\hat{\mathbf{\Phi}}_{\mathbf{y}} \mathbf{\Psi}^H\} \end{bmatrix}. \tag{17}$$

Second, the PSDs $\boldsymbol{\phi}$ are assumed to be fixed to the estimate $\hat{\boldsymbol{\phi}}_{\mathrm{ALS}}^{(i)}$ at iteration i, and the minimization is performed with respect to the RETF vector, i.e.,

$$\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} = \underset{\mathbf{a}}{\mathrm{argmin}} \left\| \hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)} - \hat{\phi}_{s,\mathrm{ALS}}^{(i)} \mathbf{a}\mathbf{a}^H \right\|_F^2, \tag{18}$$

where $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$ is the estimated target PSD matrix in iteration i defined in (14), i.e.,

$$\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)} = \hat{\mathbf{\Phi}}_{\mathbf{y}} - \left(\phi_{\mathrm{d,ALS}}^{(i)} \mathbf{\Gamma} + \phi_{\mathrm{v,ALS}}^{(i)} \mathbf{\Psi} \right). \tag{19}$$

Interpreting (18) as the best rank-1 approximation of the estimated target PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$, the solution is given by the scaled principal eigenvector [23], i.e.,

$$\boxed{\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)} = \sqrt{\frac{\hat{\lambda}_1^{(i)}}{\hat{\phi}_{s,\mathrm{ALS}}^{(i)}}} \hat{\boldsymbol{\nu}}_1^{(i)},} \tag{20}$$

where $\hat{\lambda}_1^{(i)}$ and $\hat{\boldsymbol{\nu}}_1^{(i)}$ are the principal eigenvalue and eigenvector of $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$, respectively.

A block diagram of the proposed ALS method is depicted in Figure 1. Additionally, the structure of the implementation is provided in Algorithm 1. The RETF vector $\hat{\mathbf{a}}_{\mathrm{ALS}}^{(0)}$ can be initialized either randomly or, e.g., based on an estimate of the direction of arrival of the target speaker. The ALS iterations are repeated until a convergence criterion (e.g., a fixed number of iterations) is reached.

A special case of the proposed ALS method was presented in [19], which only estimates the RETF vector and the target and diffuse PSDs, disregarding the residual noise, i.e., $\mathbf{\Phi}_{\mathbf{v}} \equiv \mathbf{0}$. This estimator (denoted ALS2) is also based on the minimization of the Frobenius norm of a model-based error matrix, i.e.,

$$\left(\hat{\mathbf{a}}_{\mathrm{ALS2}}, \hat{\boldsymbol{\phi}}_{\mathrm{ALS2}} \right) = \underset{\mathbf{a},\, \boldsymbol{\phi}}{\mathrm{argmin}} \| \hat{\mathbf{\Phi}}_{\mathbf{y}} - \phi_d \mathbf{\Gamma} - \phi_s \mathbf{a}\mathbf{a}^H \|_F^2, \tag{21}$$

where $\boldsymbol{\phi} = [\phi_s,\, \phi_d]^T$, and leads to similar update equations as in (15) and (20).

5. EXPERIMENTAL VALIDATION

To evaluate the performance of the considered methods for realistic acoustic scenarios, recordings were performed in the variable acoustics laboratory at the University of Oldenburg using a uniform linear microphone array with $M = 6$ omni-directional microphones and

Algorithm 1: ALS method to jointly estimate the RETF vector and PSDs.

Input: $\mathbf{\Gamma}(k)$, $\mathbf{\Psi}(k)$, $\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)$, iterations N, init. $\hat{\mathbf{a}}^{(0)}(k,1)$
Output: $\hat{\mathbf{a}}_{\mathrm{ALS}}(k,l)$, $\hat{\boldsymbol{\phi}}_{\mathrm{ALS}}(k,l)$
for all (k,l) **do**
  for $i = 1:N$ **do**
    compute $\mathbf{A}^{(i-1)}(k,l)$ using (16) and $\mathbf{b}^{(i-1)}(k,l)$ using (17)
    $\hat{\boldsymbol{\phi}}_{\mathrm{ALS}}^{(i)}(k,l) = \left(\mathbf{A}^{(i-1)}(k,l)\right)^{-1} \mathbf{b}^{(i-1)}(k,l)$ (15)
    constrain $\hat{\boldsymbol{\phi}}_{\mathrm{ALS}}^{(i)}(k,l)$ using (25)
    $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}(k,l) =$
      $\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l) - (\hat{\phi}_{d,\mathrm{ALS}}^{(i)}(k,l)\mathbf{\Gamma}(k) + \hat{\phi}_{v,\mathrm{ALS}}^{(i)}(k,l)\mathbf{\Psi}(k))$
    $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}(k,l) = \hat{\mathbf{N}}^{(i)}(k,l)\hat{\mathbf{\Lambda}}^{(i)}(k,l)\hat{\mathbf{N}}^{(i),H}(k,l)$ (EVD)
    $\hat{\mathbf{a}}_{\mathrm{ALS}}^{(i)}(k,l) = \sqrt{\hat{\lambda}_1^{(i)}(k,l)/\hat{\phi}_{s,\mathrm{ALS}}^{(i)}(k,l)}\,\hat{\boldsymbol{\nu}}_1^{(i)}(k,l)$ (20)
  end
  $\hat{\mathbf{a}}_{\mathrm{ALS}}^{(1)}(k,l+1) = \hat{\mathbf{a}}_{\mathrm{ALS}}^{(N)}(k,l)/(\mathbf{e}^T \hat{\mathbf{a}}_{\mathrm{ALS}}^{(N)}(k,l))$ (for next frame)
end

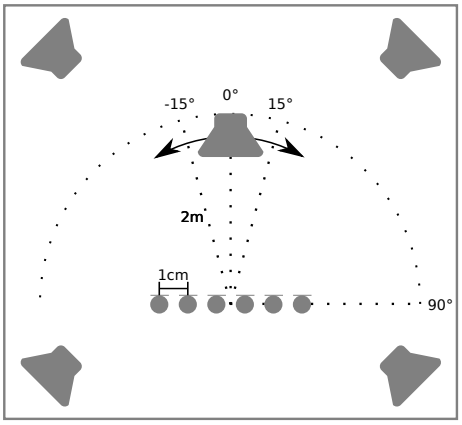

Fig. 2: Recording setup.

$d = 1\mathrm{cm}$ spacing. See Figure 2 for an overview of the recording setup. Using absorber panels on the walls, the reverberation time was set to $T_{60} \approx 350\mathrm{ms}$. A male English speech signal (length of 16 s) played back by a loudspeaker located about 2 m away from the microphone array served as the target signal. Three different dynamic scenarios are considered where the loudspeaker was moved by hand, i.e.,

(i) slowly moving approximately from 0 ° to 90 °,

(ii) normally moving approximately from 0 ° to 90 °, then standing still, and

(iii) moving between −15 ° and 15 °, simulating the motion of a person presenting in front of an audience.

In all considered scenarios, the direction of arrival at the starting position was 0 °, i.e., orthogonal to the microphone array axis (broadside), and the movement was performed keeping approximately the same distance to the microphone array. Pseudo-diffuse babble noise was generated using four loudspeakers facing the corners of the laboratory and playing back different multi-talker recordings.

The microphone signals were then obtained by mixing the target signal component, the pseudo-diffuse babble noise component, and an artificially generated uncorrelated noise component, such that the desired signal-to-diffuse ratio (SDR) and diffuse-to-noise ratio (DNR) values were

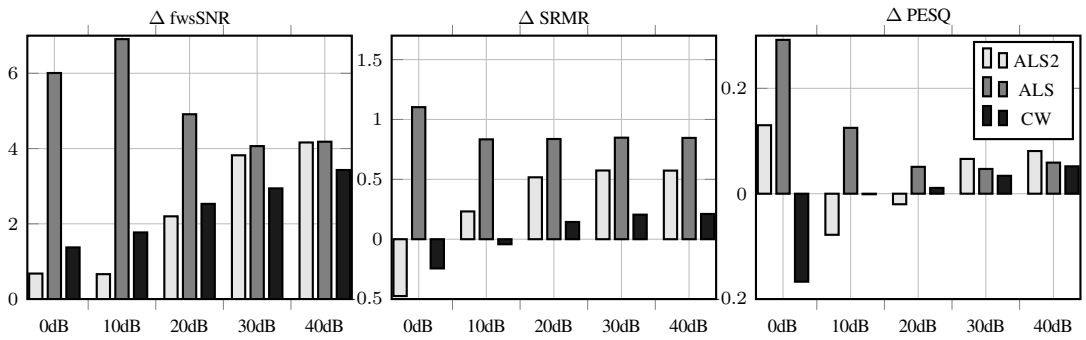

Fig. 3: Average MWF performance over three dynamic scenarios for different DNRs in terms of fwsSNR, SRMR, and PESQ (left to right column).

obtained in the reference microphone, i.e.,

$$\begin{cases} \mathrm{SDR/dB} = 10\log_{10} \dfrac{\sum_k \sum_l |X_1(k,l)|^2}{\sum_k \sum_l |D_1(k,l)|^2} & (22) \\ \mathrm{DNR/dB} = 10\log_{10} \dfrac{\sum_k \sum_l |D_1(k,l)|^2}{\sum_k \sum_l |V_1(k,l)|^2}. & (23) \end{cases}$$

The signals are processed in the STFT domain at a sampling frequency of 16 kHz using weighted overlap-add (WOLA) processing with a frame length of 512 samples (corresponding to 32 ms), an overlap of 75 %, and a Hamming analysis and synthesis window. The microphone PSD matrix is estimated via recursive averaging with $\alpha = 0.67$ (corresponding to about $20\mathrm{ms}$). The diffuse coherence matrix $\mathbf{\Gamma}$ is constructed based on the microphone geometry and assuming spherically diffuse noise [24]. To avoid numerical issues (e.g., for the Cholesky decomposition), it is regularized as $\mathbf{\Gamma} \leftarrow \mathbf{\Gamma} + \mu \mathbf{I}_M$, with $\mu = 10^{-3}$ the regularization constant.

The CW method (cf. Section 3), the ALS2 method and the proposed ALS method (cf. Section 4) are evaluated by using the estimated quantities in the MWF formulation in (6). The DDA with smoothing constant 0.98 is used to estimate the a-priori SNR as described in [22]. Furthermore, a minimum gain of -10 dB is used for the postfilter. For the proposed ALS method, the spatial coherence matrix of the residual noise is chosen as the identity matrix ($\mathbf{\Psi} = \mathbf{I}_M$), modeling uncorrelated noise with equal power at each microphone such as microphone self-noise. Depending on the acoustic scenario that is considered, different choices may be more suitable.

The number of ALS iterations is equal to 10, which has been observed to ensure convergence in preliminary simulations. For the ALS2 and the CW method, which do not inherently model any residual noise, the residual noise PSD matrix is estimated from the first second of the residual noise component as

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(k) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{v}(k,l) \mathbf{v}^H(k,l) \qquad (24)$$

and subtracted from the estimated microphone PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}$. Note that in case of the ALS method, this subtraction is obviously not performed.

Since PSDs can only assume positive values, the PSD estimates of all considered methods are lower-bounded by the machine precision `eps`. Furthermore, since none of the PSDs can be larger than the microphone signal PSD, also an upper bound given by the average microphone PSD is applied, i.e.,

$$\mathtt{eps} \leq \{\hat{\phi}_s, \hat{\phi}_d, \hat{\phi}_v\} \leq \frac{1}{M} \mathbf{y}^H \mathbf{y}. \qquad (25)$$

The MWF output signal is evaluated using the following performance measures: the frequency-weighted segmental signal-to-noise ratio (fwsSNR) [25], the speech-to-reverberation modulation energy ratio (SRMR) [25], and the perceptual evaluation of speech quality (PESQ) [26] measure. The reference signal used for these intrusive performance measures is the anechoic target signal. The MWF performance is assessed as the performance improvement between the output signal and the signal at the reference microphone.

Figure 3 displays the obtained results for different DNRs, averaged over the three considered dynamic acoustic scenarios. In terms of all considered performance measures, the proposed ALS method performs either similarly or significantly better than the ALS2 and the CW method, with the difference being greater at low DNRs.

Summarizing, these simulation results demonstrate the advantages of the proposed method in realistic dynamic acoustic scenarios, significantly outperforming existing methods in scenarios where the DNR is low, while leading to a similar performance for high DNRs.

6. CONCLUSION AND OUTLOOK

In this paper a multi-channel approach to jointly estimate the RETF vector as well as the diffuse and residual noise PSDs has been proposed. The estimates are obtained by minimizing the Frobenius norm of a model-based error matrix using an alternating least squares method. The proposed method yields a high performance when used in a multi-channel Wiener filter, especially outperforming state-of-the-art estimators in scenarios where the non-diffuse noise power is high.

References

[1] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms," in *Proc. International Workshop on Acoustic Signal Enhancement*, Juan-les-Pins, France, Sep. 2014, pp. 333–337.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[4] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[5] 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, Apr. 2017.

[6] 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, Aug. 2001.

[7] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-Microphone 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, Feb. 2015.

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

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

[10] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, San Francisco, USA, Mar. 2017, pp. 11–15.

[11] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

[12] S. Markovich-Golan, 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, Aug. 2009.

[13] A. Kuklasinski, 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. 1595–1608, Sep. 2016.

[14] O. Schwartz, S. Gannot, and E. A. P. Habets, "An Expectation-Maximization Algorithm for Multimicrophone Speech Dereverberation and Noise Reduction With Coherence Matrix Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, Jun. 2016.

[15] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 6, pp. 1052–1067, Jun. 2018.

[16] S. Braun and E. A. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Applied Signal Processing*, vol. 2015, no. 1, Dec. 2015.

[17] 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, Jun. 2018.

[18] ——, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, San Francisco, USA, Mar. 2017, pp. 116–120.

[19] M. Tammen, I. Kodrasi, and S. Doclo, "Iterative Alternating Least-Squares Approach to Jointly Estimate the RETFs and the Diffuse PSD," in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 221–225.

[20] I. Kodrasi and S. Doclo, "Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Canada, Apr. 2018, pp. 441–445.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[22] R. C. Hendriks, R. Heusdens, and J. Jensen, "Forward–backward decision directed approach for speech enhancement," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005, pp. 109–112.

[23] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[24] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[25] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. New Jersey, USA, 1988.

[26] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs P.862*. Feb. 2001.