Joint Estimation of Relative Early Transfer Function Vector and Diffuse Power Spectral Density Estimation Using an Alternating Least-Squares Approach

Marvin Tammen[⋆] *Ina Kodrasi*[†] *Simon Doclo*[⋆]

[⋆] Department of Medical Physics and Acoustics & Cluster of Excellence Hearing4All, University of Oldenburg, Germany
[†] Idiap Research Institute, Speech and Audio Processing Group, Martigny, Switzerland
Email: marvin.tammen@uol.de, ina.kodrasi@idiap.ch, simon.doclo@uol.de

Abstract

The multi-channel Wiener filter (MWF) is a commonly used speech enhancement technique for improving speech quality and intelligibility in reverberant and noisy environments. The MWF is typically implemented as a minimum variance distortionless response (MVDR) beamformer followed by a single-channel Wiener postfilter. Assuming that reverberation and ambient noise can be modeled as diffuse sound fields, estimates of the relative early transfer function (RETF) vector of the target speaker and of the diffuse power spectral density (PSD) are required to implement the MWF. RETF vector and diffuse PSD estimation methods are often decoupled, i.e., one of the quantities is estimated assuming that the other quantity is known. In this paper, we aim at jointly estimating the RETF vector and the diffuse PSD by minimizing the Frobenius norm of an error matrix based on the presumed signal model. To solve this minimization problem, we propose to use an alternating least-squares approach. Simulation results using artificial and real data show that the proposed method leads to a better performance than a state-of-the-art method based on covariance whitening.

1 Introduction

In many hands-free speech communication applications such as hearing aids and teleconferencing, the recorded microphone signals are corrupted by reverberation and ambient noise. This may result in a decreased speech quality and intelligibility and in a performance deterioration of automatic speech recognition systems. Hence, enhancement techniques are required that are able to suppress both reverberation and ambient noise. To this end, many single- and multi-channel techniques have been proposed, with multi-channel techniques generally being preferred since they are capable of exploiting spatial information. The multi-channel Wiener filter (MWF) is a commonly used speech enhancement technique, which minimizes the mean square error between a target signal and the output signal [1–3]. The MWF can be implemented as a minimum variance distortionless response (MVDR) beamformer followed by a single-channel Wiener postfilter [4]. Assuming that reverberation and ambient noise can be modeled as diffuse sound fields, the implementation of the MVDR beamformer and the Wiener postfilter requires estimates of the relative early transfer function (RETF) vector of the target speaker and of the diffuse power spectral density (PSD).

On the one hand, several RETF vector estimation procedures have been proposed, e.g., based on the least-squares method [5, 6], the covariance subtraction method [7–9], or the covariance whitening (*CW*) method [8, 10, 11]. On the other hand, several multi-channel diffuse PSD estimators have been proposed, e.g., maximum likelihood-based estimators [12–14], Frobenius norm-based estimators [14–16], or an eigenvalue decomposition (EVD)-based estimator [17]. Many RETF vector and PSD estimation methods are decoupled, i.e.: a) the RETF vector is estimated either assuming that the diffuse PSD is known [6] or without requiring knowledge of the diffuse PSD [5, 7–11]; b) the diffuse PSD is estimated either assuming that the RETF vector is known [12, 13, 15, 16] or without requiring knowledge of the RETF vector [17]. In [18] it has been shown that jointly estimating both the RETF vector and the diffuse PSD based on CW results in a high dereverberation and noise reduction performance. In addition, in [19] a batch expectation-maximization method has been proposed to jointly estimate the (time-invariant) RETF vector, the diffuse PSD, and the spatial coherence matrix in a maximum likelihood framework.

As an extension of the Frobenius norm-based PSD estimator in [16], in this paper we present a method to jointly estimate the (time-varying) RETF vector and diffuse PSD by minimizing the Frobenius norm of an error matrix constructed from the presumed signal model. Since no closed-form solution exists for the RETF vector *and* the diffuse PSD, we propose to perform the minimization in an iterative fashion using an alternating least-squares approach. By coupling the RETF vector and PSD estimation procedures we expect to obtain estimates which fit the signal model more accurately. Experimental results based on simulated data confirm a high RETF vector and diffuse PSD estimation accuracy and demonstrate the robustness of the proposed approach to deviations from the assumed signal model. In addition, experimental results based on real data for two different acoustic scenarios show that using the proposed RETF vector and diffuse PSD estimates in an MWF yields a better performance than using the estimates based on covariance whitening.

2 Signal Model and Notation

We consider a noisy and reverberant acoustic scenario with one speech source and $M \geq 2$ microphones. In the short-time Fourier transform (STFT) domain, the M-dimensional vector of the microphone signals $\mathbf{y}(k,l) = [Y_1(k,l),...,Y_M(k,l)]^T$, with k the frequency bin index and l the frame index, is given by

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{d}(k,l), \tag{1}$$

where $\mathbf{x}(k,\ l)$ denotes the direct and early reverberation speech component and $\mathbf{d}(k,l)$ denotes the diffuse component, representing both late reverberation as well as ambient noise. The vectors $\mathbf{x}(k,\ l)$ and $\mathbf{d}(k,\ l)$ are defined similarly to $\mathbf{y}(k,\ l)$. Although no non-diffuse noise (e.g., uncorrelated sensor noise) is present in (1), in the simulations we will also consider non-diffuse noise. Since processing is performed independently for each frequency bin, in the remainder of this paper the index k is omitted wherever possible.

The direct and early reverberation speech component $\mathbf{x}(l)$ can be expressed as

$$\mathbf{x}(l) = S(l)\mathbf{a}(l) = S(l)[1, A_2(l), ..., A_M(l)]^T, \tag{2}$$

where $S(l)$ is the target signal, i.e., the direct and early reverberation speech component in a reference microphone (chosen to be the first microphone), and $\mathbf{a}(l)$ is a vector of (possibly time-varying) RETFs between all microphones and the reference microphone. Assuming that the direct and early reverberation component $\mathbf{x}(l)$ is uncorrelated with the diffuse component $\mathbf{d}(l)$, which is a common assumption, we can write the $M \times M$-dimensional microphone PSD matrix $\boldsymbol{\Phi}_\mathbf{y}(l)$ as

$$\boldsymbol{\Phi}_\mathbf{y}(l) = \mathbb{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} = \boldsymbol{\Phi}_\mathbf{x}(l) + \boldsymbol{\Phi}_\mathbf{d}(l), \tag{3}$$

This work was supported in part by the Cluster of Excellence Hearing4All, funded by the German Research Foundation (DFG), and the joint Lower Saxony-Israeli Project ATHENA, funded by the State of Lower Saxony.

with $\mathbb{E}\{\cdot\}$ the expected value operator, $\{\cdot\}^H$ the Hermitian operator, and $\Phi_{\mathbf{x}}(l)$ and $\Phi_{\mathbf{d}}(l)$ the PSD matrices of $\mathbf{x}(l)$ and $\mathbf{d}(l)$, respectively. Using (2), the PSD matrix $\Phi_{\mathbf{x}}(l)$ can be expressed as

$$\Phi_{\mathbf{x}}(l) = \phi_s(l)\mathbf{a}(l)\mathbf{a}^H(l), \tag{4}$$

with $\phi_s(l) = \mathbb{E}\{|S(l)|^2\}$ the target signal PSD. For a diffuse sound field, the PSD matrix $\Phi_{\mathbf{d}}(l)$ can be expressed as

$$\Phi_{\mathbf{d}}(l) = \phi_d(l)\Gamma, \tag{5}$$

with $\phi_d(l)$ the time-varying diffuse PSD and Γ the (assumed to be) time-invariant spatial coherence matrix of a diffuse sound field, which can be analytically computed based on the microphone array geometry. Hence, using (4) and (5), the PSD matrix $\Phi_{\mathbf{y}}(l)$ can be expressed using the following signal model:

$$\boxed{\Phi_{\mathbf{y}}(l) = \phi_s(l)\mathbf{a}(l)\mathbf{a}^H(l) + \phi_d(l)\Gamma.} \tag{6}$$

In order to achieve dereverberation and noise reduction, the MWF can be applied to the microphone signals, i.e., $\hat{S}(l) = \mathbf{w}_{\text{MWF}}^H(l)\mathbf{y}(l)$, with $\mathbf{w}_{\text{MWF}}(l)$ the M-dimensional filter vector. The MWF minimizes the mean square error between the output signal $\hat{S}(l)$ and the target signal $S(l)$ and can be decomposed as an MVDR beamformer $\mathbf{w}_{\text{MVDR}}$ and a single-channel Wiener postfilter $G(l)$ [1, 4], i.e.,

$$\mathbf{w}_{\text{MWF}}(l) = \underbrace{\frac{\Gamma^{-1}\mathbf{a}(l)}{\mathbf{a}^H(l)\Gamma^{-1}\mathbf{a}(l)}}_{\mathbf{w}_{\text{MVDR}}(l)} \underbrace{\frac{\phi_s(l)}{\phi_s(l) + \phi_d(l)/\left(\mathbf{a}^H(l)\Gamma^{-1}\mathbf{a}(l)\right)}}_{G(l)}. \tag{7}$$

As can be observed from (7), estimates of the RETF vector $\mathbf{a}(l)$ and the PSDs $\phi_s(l)$ and $\phi_d(l)$ are required to implement the MWF.

In practice, the presumed signal model in (6) does not perfectly hold. First, the assumption of a diffuse sound field for the reverberation and the ambient noise to compute Γ is not perfectly true. Second, the microphone PSD matrix in (3) is estimated using recursive averaging of a single realization of the microphone signals, i.e.,

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

with α a smoothing factor, such that the estimated PSD matrix $\hat{\Phi}_{\mathbf{y}}(l)$ will differ from the true PSD matrix. Third, in addition to diffuse noise also uncorrelated noise, e.g., sensor noise, is typically present. In Sections 3 and 4, we will present an existing and a novel approach to jointly estimate the RETF vector and the PSDs from the estimated PSD matrix. In Section 5, we will compare the performance of both approaches and their sensitivity to model deviations. For conciseness, in the remainder of this paper also the frame index l is omitted wherever possible.

3 Estimation Based on Covariance Whitening (CW)

In this section, we briefly review the baseline CW-based method to jointly estimate the RETF vector and the diffuse PSD.

Using the Cholesky decomposition of the spatial coherence matrix $\Gamma = \mathbf{L}\mathbf{L}^H$ and (6), the prewhitened microphone PSD matrix is given by

$$\Phi_{\mathbf{y}}^w = \mathbf{L}^{-1}\Phi_{\mathbf{y}}\mathbf{L}^{-H} = \phi_s(\mathbf{L}^{-1}\mathbf{a})(\mathbf{L}^{-1}\mathbf{a})^H + \phi_d\mathbf{I}. \tag{9}$$

The EVD of $\Phi_{\mathbf{y}}^w$ is equal to

$$\Phi_{\mathbf{y}}^w = \mathbf{U}\Lambda\mathbf{U}^H, \tag{10}$$

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

$$\boxed{\begin{cases} \hat{\mathbf{a}}_{\text{CW}} = \mathbf{L}\hat{\mathbf{u}}_1 / \left(\mathbf{e}^T\mathbf{L}\hat{\mathbf{u}}_1\right) & (11) \\ \hat{\phi}_{d,\text{CW}} = \left(\text{trace}\{\hat{\Phi}_{\mathbf{y}}^w\} - \hat{\lambda}_1\right)/(M-1), & (12) \end{cases}}$$

with $\mathbf{u}_1$ and λ_1 denoting the principal eigenvector and eigenvalue of the prewhitened estimated PSD matrix $\hat{\Phi}_{\mathbf{y}}^w$, and $\mathbf{e}$ a selection vector containing zeros and one element equal to 1, i.e., $\mathbf{e}(1) = 1$.

4 Frobenius Norm-Based Estimation

In this section, we propose an alternative approach to jointly estimate the time-varying RETF vector and the diffuse PSD by minimizing the Frobenius norm of the error matrix, constructed by subtracting the signal model in (6) from the estimated PSD matrix in (8), i.e.,

$$\mathbf{E} = \hat{\Phi}_{\mathbf{y}} - (\phi_s\mathbf{a}\mathbf{a}^H + \phi_d\Gamma). \tag{13}$$

We now define the PSD vector $\boldsymbol{\phi} := [\phi_s, \phi_d]^T$, containing the target signal and the diffuse PSD. Assuming the RETF vector to be equal to $\bar{\mathbf{a}}$, in [16] it has been proposed to estimate the PSD vector $\boldsymbol{\phi}$ by minimizing the Frobenius norm $\|\cdot\|_F$ of the error matrix in (13), i.e.,

$$\boxed{\hat{\boldsymbol{\phi}}_{\text{LS}} = \underset{\boldsymbol{\phi}}{\text{argmin}} \left\| \hat{\Phi}_{\mathbf{y}} - \left(\phi_s\bar{\mathbf{a}}\bar{\mathbf{a}}^H + \phi_d\Gamma\right)\right\|_F^2.} \tag{14}$$

It has been shown in [16] that $\hat{\boldsymbol{\phi}}_{\text{LS}}$ can be computed as $\hat{\boldsymbol{\phi}} = \mathbf{A}^{-1}\mathbf{b}$ with

$$\mathbf{A} = \begin{bmatrix} \left(\bar{\mathbf{a}}^H\bar{\mathbf{a}}\right)^2 & \bar{\mathbf{a}}^H\Gamma\bar{\mathbf{a}} \\ \bar{\mathbf{a}}^H\Gamma\bar{\mathbf{a}} & \text{trace}\{\Gamma^H\Gamma\} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} \text{Re}\{\bar{\mathbf{a}}^H\hat{\Phi}_{\mathbf{y}}\bar{\mathbf{a}}\} \\ \text{Re}\{\text{trace}\{\hat{\Phi}_{\mathbf{y}}\Gamma^H\}\} \end{bmatrix}. \tag{15}$$

As an extension of this method, we propose to jointly estimate the RETF vector $\bar{\mathbf{a}}$ and the PSD vector $\boldsymbol{\phi}$, i.e.,

$$\left(\hat{\mathbf{a}}_{\text{LS}}, \hat{\boldsymbol{\phi}}_{\text{LS}}\right) = \underset{\mathbf{a},\,\boldsymbol{\phi}}{\text{argmin}} \left\| \hat{\Phi}_{\mathbf{y}} - \left(\phi_s\mathbf{a}\mathbf{a}^H + \phi_d\Gamma\right)\right\|_F^2. \tag{16}$$

To the best of our knowledge, no closed-form solution for the optimization problem in (16) exists. Hence, we propose to perform this minimization in an iterative fashion using an alternating least-squares approach.

In the first step, the minimization is performed w.r.t. the PSD vector $\boldsymbol{\phi}$, assuming that the RETF vector is fixed to the estimate from the i-th iteration $\hat{\mathbf{a}}^{(i)}$. Since this is similar to the assumption made in [16], the PSD vector can be estimated as

$$\boxed{\hat{\boldsymbol{\phi}}_{\text{LS}}^{(i)} = \left(\mathbf{A}^{(i)}\right)^{-1}\mathbf{b}^{(i)},} \tag{17}$$

where $\mathbf{A}^{(i)}$ and $\mathbf{b}^{(i)}$ are constructed similarly to (15) by replacing $\bar{\mathbf{a}}$ with $\hat{\mathbf{a}}^{(i)}$.

In the second step, the minimization is performed w.r.t. the RETF vector $\mathbf{a}$, assuming that the PSD vector is fixed to $\hat{\boldsymbol{\phi}}_{\text{LS}}^{(i)} = [\hat{\phi}_{s,\text{LS}}^{(i)}, \hat{\phi}_{d,\text{LS}}^{(i)}]^T$. By defining $\hat{\Phi}_{\mathbf{x}}^{(i)} = \hat{\Phi}_{\mathbf{y}} - \hat{\phi}_{d,F}^{(i)}\Gamma$, the RETF vector can be estimated as

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

which can be interpreted as the best rank-1 approximation of the matrix $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$. Based on the Eckart-Young-Mirsky theorem [20] and assuming that $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$ is positive definite, the best rank-1 approximation of $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$ is equal to $\hat{\lambda}_1^{(i)}\hat{\mathbf{u}}_1^{(i)}\hat{\mathbf{u}}_1^{(i),H}$, with $\hat{\lambda}_1^{(i)}$ and $\hat{\mathbf{u}}_1^{(i)}$ the principal eigenvalue and eigenvector of $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}$. Hence, the solution to (18) is equal to a scaled version of the principal eigenvector, i.e.,

$$\boxed{\hat{\mathbf{a}}_{\text{LS}}^{(i)} = \sqrt{\frac{\hat{\lambda}_1^{(i)}}{\hat{\phi}_{s,\text{LS}}^{(i)}}}\,\hat{\mathbf{u}}_1^{(i)}.} \tag{19}$$

Algorithm 1 describes the complete implementation of the proposed alternating least-squares approach. In the first frame, the RETF vector estimate is initialized, e.g., with complex-valued components and the first element set to 1. For each time-frequency bin, an estimate of the RETF vector $\hat{\mathbf{a}}(k,\,l)$ and the PSD vector $\hat{\boldsymbol{\phi}}(k,\,l)$ is obtained by performing N iterations. The resulting RETF vector $\hat{\mathbf{a}}^{(N)}(k,l)$ is then normalized w.r.t. the first element and used as the initial estimate for the next frame. Since PSDs can only assume positive values, the PSD estimates are lower-bounded by the machine precision `eps`. Furthermore, since neither the diffuse nor the target PSD can be larger than the microphone signal PSD, also an upper bound is applied, i.e.,

$$\texttt{eps} \leq \{\hat{\phi}_s, \hat{\phi}_d\} \leq \frac{1}{M}\mathbf{y}^H\mathbf{y}. \tag{20}$$

Algorithm 1: Alternating least-squares approach to jointly estimate the RETF vector and PSDs.

```
Input: $\Gamma(k)$, $\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)$, num. iterations $N$, init. $\hat{\mathbf{a}}^{(1)}(k,1)$
Output: $\hat{\mathbf{a}}(k,l)$, $\hat{\boldsymbol{\phi}}_{\text{LS}} = [\hat{\phi}_{s,\text{LS}}, \hat{\phi}_{d,\text{LS}}]^T$
for all $k$ do
    for all $l$ do
        for $i = 1:N$ do
            compute $\mathbf{A}^{(i)}(k,l)$ and $\mathbf{b}^{(i)}(k,l)$ using (15)
            $\hat{\boldsymbol{\phi}}^{(i)}(k,l) = \left(\mathbf{A}^{(i)}\right)^{-1}(k,l)\mathbf{b}^{(i)}(k,l)$ (17)
            constrain $\hat{\boldsymbol{\phi}}^{(i)}(k,l)$ using (20)
            $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}(k,l) = \hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l) - \hat{\phi}_d^{(i)}(k,l)\Gamma(k)$
            $\hat{\mathbf{\Phi}}_{\mathbf{x}}^{(i)}(k,l) = \mathbf{U}^{(i)}(k,l)\Lambda^{(i)}(k,l)\mathbf{U}^{(i),H}(k,l)$
             (EVD)
            $\hat{\mathbf{a}}^{(i)}(k,l) = \sqrt{\lambda_1^{(i)}(k,l)/\hat{\phi}_s^{(i)}(k,l)}\mathbf{u}_1^{(i)}(k,l)$ (19)
        end
        $\hat{\mathbf{a}}^{(1)}(k,l+1) = \hat{\mathbf{a}}^{(N)}(k,l)/(\mathbf{e}^T\hat{\mathbf{a}}^{(N)}(k,l))$
         (for next frame)
    end
end
```

5 Experimental Validation

In this section, the performance of the proposed method is validated, both using simulated artificial data (Section 5.1) and using real recordings (Section 5.2).

5.1 Artificial Data

To evaluate the estimation accuracy and the convergence speed of the proposed method, artificial data is generated according to the assumed signal model in (6), such that oracle information is available. In total, $M = 4$ microphones are simulated for $K = 513$ frequency bins and $L = 100$ time frames. Specifically, the target and diffuse PSDs at each time-frequency bin are drawn from a scaled and squared normal distribution, i.e., $\phi_s(k,l),\, \phi_d(k,l) \sim 10^{-7}(\mathcal{N}(\mu=0,\, \sigma=1))^2$. The RETF vectors $\mathbf{a}(k)$ are assumed to be time-invariant and are generated using normally distributed complex-valued components, and the first element is set equal to 1. For the spatial coherence matrix $\Gamma(k)$, a random positive-definite matrix is added to a random diagonal matrix with positive values, and the resulting matrix is scaled such that the diagonal elements are equal to 1.

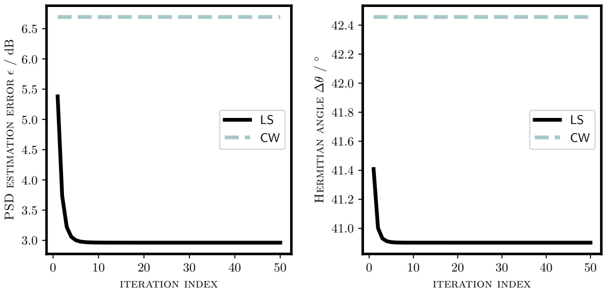

Figure 1: Diffuse PSD and RETF vector estimation errors vs. the iteration index ($M = 4$, DNR = 0 dB).

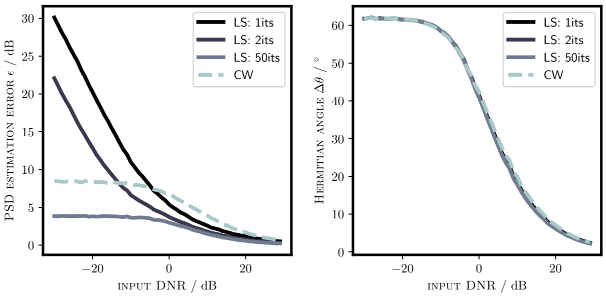

Figure 2: Diffuse PSD and RETF vector estimation errors vs. input DNR for different numubers of iterations ($M = 4$).

Since in typical acoustical scenarios the late reverberation and the ambient noise are not perfectly diffuse and since the early reverberation and diffuse components are not perfectly uncorrelated, the signal model in (6) is typically violated. In order to evaluate the robustness of the proposed method to model mismatches, an $M \times M$-dimensional scaled error matrix Ξ is added, i.e.,

$$\mathbf{\Phi}_{\mathbf{y}}(k,l) = \phi_s(k,l)\mathbf{a}(k)\mathbf{a}^H(k) + \phi_d(k,l)\Gamma(k) + \delta\Xi(k,l), \tag{21}$$

where $\delta = \frac{10^{-7}}{10^{\text{input DNR}/10}}$ determines the model mismatch, with *DNR* representing the diffuse-to-noise ratio. The error matrix $\Xi(k,l)$ is generated as $\Xi(k,\,l) = \mathbf{n}(k,\,l)\mathbf{n}^H(k,\,l)$, with $\mathbf{n}(k,l)$ an M-dimensional vector with normally distributed complex-valued components. The considered DNR values range from -50 dB to 50 dB.

The PSD estimation accuracy is evaluated using the average PSD estimation error over all time-frequency bins [23], i.e.,

$$\epsilon = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} 10\log_{10}\frac{\phi(k,l)}{\hat{\phi}(k,l)}, \tag{22}$$

The RETF vector estimation accuracy is evaluated using the average Hermitian angle between the oracle vector $\mathbf{a}(k,l)$ and the RETF vector estimate $\hat{\mathbf{a}}(k,l)$ as [9]

$$\Delta\theta = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\arccos\left(\frac{\left|\hat{\mathbf{a}}^H(k,l)\mathbf{a}(k)\right|}{\|\hat{\mathbf{a}}(k,l)\|_2\|\mathbf{a}(k)\|_2}\right)\frac{360^\circ}{2\pi}, \tag{23}$$

which is a measure disregarding the length difference of both vectors. Figure 1 depicts the diffuse PSD estimation error and the Hermitian angle versus the number of iterations for a DNR of 0 dB. In addition, the performance of the CW method is depicted. It can

	ΔPESQ				ΔfwsSNR			
uncorrelated noise SNR / dB	0	10	20	30	0	10	20	30
CW	0.0547	0.1632	-0.0561	-0.1037	1.7112	2.2500	3.6697	3.4651
LS	**0.3823**	**0.5264**	**0.1818**	**0.2331**	**4.5457**	**4.9639**	**4.7897**	**3.7432**

Table 1: Performance on realistic data, averaged over both considered acoustical systems in terms of ΔPESQ and ΔfwsSNR.

be observed that while there is no significant change in the Hermitian angle for an increasing number of iterations, the diffuse PSD estimation error is decreased by about 2 dB after convergence, which is reached after approximately 5 iterations. Furthermore, it can be observed that using the proposed method, lower PSD and RETF vector estimation errors are obtained than using CW.

For different DNRs, Figure 2 compares the diffuse PSD and RETF vector estimation accuracy of the CW method and the proposed LS method for different numbers of iterations ($N = 1,\ 2,\ 50$). In terms of the diffuse PSD estimation error it can be observed that for only a few iterations (i.e., 1 or 2), the proposed method performs significantly worse than the CW method. After convergence, however, the LS method clearly outperforms the CW method for low input DNRs, while resulting in a similar estimation accuracy at high input DNRs. In terms of the Hermitian angle, there is no significant difference between the methods for all considered DNRs. Hence, the alternating least-squares approach improves the PSD estimation accuracy with increasing number of iterations, while one iteration seems to suffice in terms of RETF vector estimation accuracy.

In summary, the proposed method exhibits a large robustness to model noise in artificial data, outperforming the CW method in terms of diffuse PSD estimation accuracy, while leading to a comparable performance in terms of RETF vector estimation accuracy.

5.2 Recorded Data

	array geometry	d/cm	θ/°	T_{60}/s
AS_1 [21]	linear	8	45	0.61
AS_2 [22]	linear	6	-15	1.25

Table 2: Configuration of considered acoustic scenarios;
d: inter-microphone distance, θ: speaker direction of arrival.

To evaluate the performance in realistic acoustic scenarios, we consider a spatially stationary speech source in the presence of reverberation and diffuse babble noise. The reverberant multi-channel speech signals are obtained by convolving an 8.8 s long anechoic speech signal with measured room impulses (RIRs) with different reverberation times (described in Table 2). As for the simulated data, $M = 4$ microphones are used. Diffuse babble noise is generated as described in [24] and added at 0 dB input SNR w.r.t. the reference microphone.

As already mentioned, it should be noted that in realistic acoustic scenarios, deviations from the signal model in (6) are to be expected, since the perfectly diffuse sound field model is generally violated and since the individual components are not perfectly uncorrelated. To investigate the impact of additional model mismatch, spatially uncorrelated noise is added to the microphone signals at different input SNRs ranging from 0 dB to 30 dB.

The signals are processed in the STFT domain at a sampling frequency of 16 kHz, using a frame length of 1024 samples (corresponding to 64 ms), an overlap of 75 %, and using a Hamming window. The estimated microphone PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(l)$ is obtained using (8) with a smoothing constant $\alpha = 0.67$, corresponding to approximately 40 ms. Note that the uncorrelated noise PSD matrix is not estimated and subtracted from the microphone PSD matrix, as, e.g., done in [15–17], such that the sensitivity to uncorrelated noise can be evaluated.

For the realistic scenario, we compare the performance of the LS method and the CW method when using the obtained RETF vector and diffuse PSD estimates in an MWF. The performance of the MWF output signal is evaluated using the perceptual evaluation of speech quality [25] (*PESQ*) and the frequency-weighted segmental SNR [26] (*fwsSNR*), using the anechoic speech signal as reference signal.

It was shown in [16, 27] that using the decision-directed approach [28] to obtain an estimate of the a-priori SNR results in a better MWF performance than directly utilizing the Frobenius norm-based target signal PSD estimate. For this reason, the target signal PSD estimation accuracy has not been evaluated in Section 5.1, and the decision-directed approach is used to implement the MWF in this section. The a-priori SNR $\xi(l)$ is estimated as

$$\hat{\xi}(l) = \rho \frac{|\hat{X}(l-1)|^2}{\hat{\phi}_d(l-1)} + (1-\rho)\max\left\{ \frac{|\hat{X}(l)|^2}{\hat{\phi}_d(l)} - 1, 0 \right\}, \qquad (24)$$

with the smoothing constant $\rho = 0.98$. The MVDR beamformer coefficients $\mathbf{w}_{\mathrm{MVDR}}(l)$ are computed as in (7) and the postfilter $G(l)$ is computed using the a-priori SNR estimate as

$$G(l) = \frac{\hat{\xi}(l)}{1+\hat{\xi}(l)}. \qquad (25)$$

A minimum gain of -20 dB is used for the postfilter.

The MWF performance is presented in Table 1 for several SNRs, where the average performance over both considered acoustical systems is presented. It can be observed that the LS method outperforms the CW method for all considered SNRs in terms of both performance measures. In terms of PESQ, a better performance of up to 0.36 is obtained using the LS method, whereas in terms of fwsSNR, a better performance of up to 2.83 dB is obtained. In terms of fwsSNR, the performance difference between both methods becomes smaller for larger SNRs, which is in line with the results from Figure 2. In terms of PESQ, the CW method is not able to achieve a large performance improvement, while the LS method leads to a significant improvement.

In summary, the proposed LS method yields a significantly better performance than the CW method when used in an MWF, confirming the advantages of coupling the RETF vector and diffuse PSD estimation in an alternating least-squares approach.

6 Conclusions & Outlook

In this paper an alternating least-squares approach has been proposed, in which both the RETF vector and the diffuse PSD are iteratively estimated by minimizing the Frobenius norm of an error matrix constructed based on the assumed signal model. It is shown that the proposed method yields a higher PSD estimation accuracy and a similar RETF vector estimation accuracy than the CW method. In addition, it is shown that using the proposed estimates in an MWF significantly outperforms using the estimates obtained with the CW method. For future research, it will be investigated how to extend the proposed approach to also jointly estimate non-diffuse noise PSDs.

References

[1] 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, pp. 18–30, Mar. 2015.

[2] E. A. P. Habets and J. Benesty, “A Two-Stage Beamforming Approach for Noise Reduction and Dereverberation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 945–958, May 2013.

[3] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, 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.

[4] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), Berlin, Germany: Springer, 2001.

[5] S. Gannot, D. Burshtein, and E. E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Transactions on Signal Processing*, vol. 49, pp. 1614–1626, Aug. 2001.

[6] 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, pp. 240–251, Feb. 2015.

[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 *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 544–548, IEEE, Apr. 2015.

[9] 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), pp. 11–15, 2017.

[10] E. Warsitz and R. Haeb-Umbach, “Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1529–1539, July 2007.

[11] 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, pp. 1071–1086, Aug. 2009.

[12] O. Schwartz, S. Braun, S. Gannot, and E. A. P. Habets, “Maximum likelihood estimation of the late reverberant power spectral density in noisy environments,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New York, USA), Oct. 2015.

[13] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood PSD estimation for speech enhancement in reverberation and noise,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1595–1608, Sept. 2016.

[14] 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, pp. 1052–1067, June 2018.

[15] 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, Dec. 2015.

[16] O. Schwartz, S. Gannot, and E. A. P. Habets, “Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm,” in *Proc. European Signal Processing Conference*, (Budapest, Hungary), pp. 1123–1127, Sept. 2016.

[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, pp. 1106–1118, June 2018.

[18] I. Kodrasi and S. Doclo, “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), pp. 116–120, IEEE, Mar. 2017.

[19] 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. 24, pp. 1495–1510, Sept. 2016.

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

[21] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” in *Proc. International Workshop on Acoustic Echo and Noise Control*, (Antibes, France), pp. 313–317, Sept. 2014.

[22] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, “The ACE challenge - Corpus description and performance evaluation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New York, USA), pp. 1–5, Oct. 2015.

[23] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[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, pp. 2911–2917, Nov. 2008.

[25] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*. International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[26] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. New Jersey, USA: Prentice-Hall, 1988.

[27] 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), pp. 441–445, Apr. 2018.

[28] 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, pp. 1109–1121, Dec. 1984.