EVD-BASED MULTI-CHANNEL DEREVERBERATION OF A MOVING SPEAKER USING DIFFERENT RETF ESTIMATION METHODS

Ina Kodrasi, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics,
and Cluster of Excellence Hearing4All, Oldenburg, Germany
{ina.kodrasi,simon.doclo}@uni-oldenburg.de

ABSTRACT

The multi-channel Wiener filter (MWF) for dereverberation relies on estimating the late reverberant power spectral density (PSD) and the relative early transfer functions (RETFs) of the target signal from a reference microphone to all microphones. State-of-the-art multi-channel late reverberant PSD estimators also require an estimate of the RETFs, which may be difficult to estimate accurately, particularly in highly reverberant and noisy scenarios. Recently we proposed a more advantageous late reverberant PSD estimator based on an eigenvalue decomposition (EVD) which does not require knowledge of the RETFs, thereby avoiding the propagation of any RETF estimation errors into the PSD estimate. However, the performance of the proposed EVD-based estimator was analyzed by using it in an MWF with simulated RETF estimation errors for a fixed speaker position in noiseless scenarios. In this paper the EVD-based estimator is combined with several practical RETF estimation methods, i.e., the covariance whitening, covariance subtraction, and least-squares methods. The performance of the MWF using the EVD-based estimator and the considered RETF estimation methods is then investigated for a fixed and a moving speaker in different noiseless and noisy scenarios. Experimental results show that while combining the EVD-based estimator with any of the considered RETF estimation methods yields a high performance, in noiseless scenarios the covariance whitening and subtraction methods result in the best performance, whereas in noisy scenarios the least-squares method results in the best performance.

Index Terms— dereverberation, EVD, RETF estimation, MWF

1. INTRODUCTION

In many speech communication applications such as teleconferencing applications, voice-controlled systems, and hearing aids, the microphone signals are corrupted by reverberation, typically leading to decreased speech quality and intelligibility [1, 2] and performance deterioration in speech recognition systems [3]. Since late reverberation is the major cause of speech quality and intelligibility degradation, effective enhancement techniques that reduce the late reverberation are required. A commonly used speech enhancement technique is multi-channel Wiener filtering (MWF), which yields a minimum mean-square error estimate of the target signal [4–6]. Implementing the MWF for speech dereverberation requires estimates of i) the late reverberant power spectral density (PSD), and ii) the relative early transfer functions (RETFs) of the target signal from the reference microphone to all microphones. Many multi-channel late reverberant PSD estimators [5–10] also require an estimate of the RETFs, which may be difficult to estimate accurately, particularly in highly reverberant and noisy scenarios. As experimentally validated in [11, 12], RETF estimation errors degrade the PSD estimation accuracy, resulting in an additional degradation in the dereverberation performance of the speech enhancement system. Recently we have proposed a multi-channel late reverberant PSD estimator based on an eigenvalue decomposition (EVD) which does not require knowledge of the RETFs [13]. By decoupling the PSD estimation from the RETF estimation it is avoided that any RETF estimation errors propagate into the PSD estimate. The advantage of decoupling the PSD estimation from the RETF estimation has been illustrated in [13], where it is shown that using the RETF-independent EVD-based PSD estimator in a MWF outperforms using the RETF-dependent maximum-likelihood PSD estimator from [6], both when the true RETFs are known as well as in the presence of RETF estimation errors.

However, in [13] the performance of the EVD-based PSD estimator was analyzed by using it in an MWF with simulated RETF estimation errors for a fixed speaker position. Moreover, the effect of additive noise was neglected. In this paper we investigate the dereverberation and noise reduction performance of the MWF for a fixed and a moving speaker using the EVD-based PSD estimator and RETFs estimated with several practical methods, i.e., the covariance whitening method [14, 15], the covariance subtraction method [15, 16], and the least-squares method [17]. It is shown that combining the EVD-based PSD estimator with the covariance whitening and subtraction methods is straightforward and does not introduce any significant additional computations. Furthermore, it is shown that while any of the considered RETF estimation methods results in a high performance, the covariance whitening and subtraction methods yield the best performance in noiseless scenarios, whereas the least-squares method yields the best performance in noisy scenarios.

2. PROBLEM FORMULATION

Consider a reverberant and noisy system with a single source and $M \geq 2$ microphones. In the short-time Fourier transform domain, the M-dimensional vector of the microphone signals $\mathbf{y}(k,l) = [Y_1(k,l) \; \ldots \; Y_M(k,l)]^T$, with k the frequency index and l the frame index, is given by

$$\mathbf{y}(k,l) = \underbrace{\mathbf{x}_{\rm d}(k,l) + \mathbf{x}_{\rm r}(k,l)}_{\mathbf{x}(k,l)} + \mathbf{v}(k,l), \tag{1}$$

with $\mathbf{x}(k,l)$ the speech component, $\mathbf{x}_{\rm d}(k,l)$ the direct and early reverberant speech component, $\mathbf{x}_{\rm r}(k,l)$ the late reverberant speech component, and $\mathbf{v}(k,l)$ the additive noise component. Assuming a moving speaker, the direct speech component $\mathbf{x}_{\rm d}(k,l)$ can be described by

$$\mathbf{x}_{\rm d}(k,l) = \mathbf{d}(k,l)S(k,l), \tag{2}$$

This work was supported in part by the Cluster of Excellence 1077 “Hearing4All”, funded by the German Research Foundation (DFG), the Marie Curie Initial Training Network DREAMS (Grant no. 316969), and the joint Lower Saxony-Israeli Project ATHENA, funded by the State of Lower Saxony.

with $S(k,l)$ the target signal (direct and early reverberant speech component) received by a reference microphone and $\mathbf{d}(k,l)$ the vector of time-varying RETFs of the target signal from the reference microphone to all microphones. Without loss of generality, we assume that the first microphone is the reference microphone such that the RETF vector is given by $\mathbf{d}(k,l) = [1 \; D_2(k,l) \; \ldots \; D_M(k,l)]^T$. Since the processing is done independently in each frequency, in the following the frequency index k is omitted.

Assuming that the speech and noise components are uncorrelated, the PSD matrix of the microphone signals is equal to

$$\mathbf{R}_{\mathbf{y}}(l) = \mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} = \underbrace{\mathcal{E}\{\mathbf{x}(l)\mathbf{x}^H(l)\}}_{\mathbf{R}_{\mathbf{x}}(l)} + \underbrace{\mathcal{E}\{\mathbf{v}(l)\mathbf{v}^H(l)\}}_{\mathbf{R}_{\mathbf{v}}(l)}, \quad (3)$$

with $\mathcal{E}$ the expected value operator. Furthermore, assuming that the direct and early reverberant speech component is uncorrelated to the late reverberant speech component, the PSD matrix $\mathbf{R}_{\mathbf{x}}(l)$ can be written as

$$\mathbf{R}_{\mathbf{x}}(l) = \underbrace{\mathcal{E}\{\mathbf{x}_{\mathrm{d}}(l)\mathbf{x}_{\mathrm{d}}^H(l)\}}_{\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}(l)} + \underbrace{\mathcal{E}\{\mathbf{x}_{\mathrm{r}}(l)\mathbf{x}_{\mathrm{r}}^H(l)\}}_{\mathbf{R}_{\mathbf{x}_{\mathrm{r}}}(l)} \quad (4)$$

$$= \mathbf{d}(l)\mathbf{d}^H(l)\Phi_{\mathrm{s}}(l) + \mathbf{R}_{\mathbf{x}_{\mathrm{r}}}(l), \quad (5)$$

with $\Phi_s(l) = \mathcal{E}\{|S(l)|^2\}$ the target signal PSD. Modeling the late reverberation as a diffuse sound field, the PSD matrix of the late reverberant speech component $\mathbf{R}_{\mathbf{x}_{\mathrm{r}}}(l)$ may be written as [5, 7, 9, 10]

$$\mathbf{R}_{\mathbf{x}_{\mathrm{r}}}(l) = \Phi_{\mathrm{r}}(l)\mathbf{\Gamma}, \quad (6)$$

with $\Phi_{\mathrm{r}}(l)$ the late reverberant PSD and $\mathbf{\Gamma}$ the diffuse spatial coherence matrix which can be analytically computed given the geometry of the microphone array.

The MWF $\mathbf{w}(l) = [W_1(l) \; \ldots \; W_M(l)]^T$ is designed such that the mean-square error between the output signal $Z(l) = \mathbf{w}^H(l)\mathbf{y}(l)$ and the target signal $S(l)$ is minimized. It is well known that the MWF can be decomposed into a Minimum Variance Distortionless Response (MVDR) Beamformer $\mathbf{w}_{\mathrm{MVDR}}(l)$ and a single-channel Wiener postfilter $G(l)$ applied to the MVDR output [18], i.e.,

$$\mathbf{w}(l) = \underbrace{\frac{\mathbf{R}_{\mathbf{u}}^{-1}(l)\mathbf{d}(l)}{\mathbf{d}^H(l)\mathbf{R}_{\mathbf{u}}^{-1}(l)\mathbf{d}(l)}}_{\mathbf{w}_{\mathrm{MVDR}}(l)} \underbrace{\frac{\Phi_{\mathrm{s}}(l)}{\Phi_{\mathrm{s}}(l) + (\mathbf{d}^H(l)\mathbf{R}_{\mathbf{u}}^{-1}(l)\mathbf{d}(l))^{-1}}}_{G(l)}, \quad (7)$$

with $\mathbf{R}_{\mathbf{u}}(l)$ the PSD matrix of the undesired signal components (i.e., late reverberation and noise) given by

$$\mathbf{R}_{\mathbf{u}}(l) = \Phi_{\mathrm{r}}(l)\mathbf{\Gamma} + \mathbf{R}_{\mathbf{v}}(l). \quad (8)$$

As illustrated in (7) and (8), the implementation of the MWF requires knowledge of the late reverberant PSD $\Phi_{\mathrm{r}}(l)$, coherence matrix $\mathbf{\Gamma}$, noise PSD matrix $\mathbf{R}_{\mathbf{v}}(l)$, RETF vector $\mathbf{d}(l)$, and target signal PSD $\Phi_{\mathrm{s}}(l)$. In previous work it has been assumed that $\mathbf{\Gamma}$, $\mathbf{R}_{\mathbf{v}}(l)$, and $\mathbf{d}(l)$ are known such that only the PSDs $\Phi_{\mathrm{s}}(l)$ and $\Phi_{\mathrm{r}}(l)$ need to be estimated [5–10]. In this work we only assume that $\mathbf{\Gamma}$ and $\mathbf{R}_{\mathbf{v}}(l)$ are known, such that the PSDs $\Phi_{\mathrm{s}}(l)$ and $\Phi_{\mathrm{r}}(l)$ and the RETF vector $\mathbf{d}(l)$ need to be estimated. While $\mathbf{\Gamma}$ can be analytically computed based on the array geometry [5, 7, 9, 10], the PSD matrix $\mathbf{R}_{\mathbf{v}}(l)$ can be periodically updated in time-frequency regions where the speech level is low in comparison to the noise level, e.g., using the multi-channel speech presence probability estimator in [19]. Note that since the noise PSD matrix $\mathbf{R}_{\mathbf{v}}(l)$ is assumed to be available, also the speech PSD matrix $\mathbf{R}_{\mathbf{x}}(l)$ can be estimated, e.g., as $\mathbf{R}_{\mathbf{x}}(l) = \mathbf{R}_{\mathbf{y}}(l) - \mathbf{R}_{\mathbf{v}}(l)$, with $\mathbf{R}_{\mathbf{y}}(l)$ estimated from the microphone signals.

3. EVD-BASED LATE REVERBERANT PSD ESTIMATION

In the following we briefly review the recently proposed EVD-based estimator from [13] which estimates the late reverberant PSD using the eigenvalues of the whitened PSD matrix $\mathbf{R}_{\mathbf{x}}(l)$. In order to whiten $\mathbf{R}_{\mathbf{x}}(l)$, the coherence matrix $\mathbf{\Gamma}$ and its inverse $\mathbf{\Gamma}^{-1}$ are decomposed using the Cholesky decomposition as

$$\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H, \qquad \mathbf{\Gamma}^{-1} = \mathbf{L}^{-H}\mathbf{L}^{-1}, \quad (9)$$

with $\mathbf{L}$ an $M \times M$-dimensional lower triangular matrix. Using (9), the whitened PSD matrix $\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l)$ is computed as

$$\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l) = \mathbf{L}^{-1}\mathbf{R}_{\mathbf{x}}(l)\mathbf{L}^{-H}. \quad (10)$$

Using (5), (6), and (9), the matrix $\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l)$ can be written as

$$\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l) = \Phi_{\mathrm{s}}(l)\mathbf{L}^{-1}\mathbf{d}(l)\mathbf{d}^H(l)\mathbf{L}^{-H} + \Phi_{\mathrm{r}}(l)\mathbf{L}^{-1}\mathbf{\Gamma}\mathbf{L}^{-H} \quad (11)$$

$$= \Phi_{\mathrm{s}}(l)\mathbf{b}(l)\mathbf{b}^H(l) + \Phi_{\mathrm{r}}(l)\mathbf{I}, \quad (12)$$

with $\mathbf{I}$ the $M \times M$-dimensional identity matrix and $\mathbf{b}(l)$ the whitened RETF vector, i.e.,

$$\mathbf{b}(l) = \mathbf{L}^{-1}\mathbf{d}(l). \quad (13)$$

Computing the EVD of $\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l)$ yields

$$\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l) = \mathbf{U}(l)\mathbf{S}(l)\mathbf{U}^H(l), \quad (14)$$

with $\mathbf{U}(l)$ an $M \times M$-dimensional matrix of eigenvectors and $\mathbf{S}(l)$ the $M \times M$-dimensional diagonal matrix of eigenvalues. Since $\mathbf{R}_{\mathbf{x}}^{\mathrm{w}}(l)$ is the sum of a rank-1 matrix and a scaled identity matrix, cf. (12), $\mathbf{S}(l)$ is given by

$$\mathbf{S}(l) = \mathrm{diag}\{[\sigma(l) + \Phi_{\mathrm{r}}(l) \; \Phi_{\mathrm{r}}(l) \; \ldots \; \Phi_{\mathrm{r}}(l)]^T\}, \quad (15)$$

with $\sigma(l)$ the only non-zero eigenvalue of the rank-1 matrix $\Phi_{\mathrm{s}}(l)\mathbf{b}(l)\mathbf{b}^H(l)$. Based on (15) and using the fact that the trace of a matrix is equal to the sum of its eigenvalues, in [13] we have proposed to estimate the late reverberant PSD as

$$\hat{\Phi}_{\mathrm{r}}(l) = \frac{1}{M-1}\left(\mathrm{tr}\{\mathbf{R}_x^{\mathrm{w}}(l)\} - \lambda_1\{\mathbf{R}_x^{\mathrm{w}}(l)\}\right), \quad (16)$$

where $\mathrm{tr}\{\cdot\}$ denotes the matrix trace operator and $\lambda_1\{\cdot\}$ denotes the first (maximum) eigenvalue. Using $\hat{\Phi}_{\mathrm{r}}(l)$, an estimate of the clean speech PSD $\hat{\Phi}_{\mathrm{s}}(l)$ is obtained based on the decision directed approach [20].

Unlike other state-of-the-art multi-channel late reverberant PSD estimators [5–10], the EVD-based estimator in (16) does not require knowledge of the RETF vector $\mathbf{d}(l)$. An RETF-independent PSD estimator is advantageous in order to avoid the propagation of any RETF estimation errors into the PSD estimate. As is experimentally validated in [13], using the EVD-based PSD estimator in the MWF in (7) yields a better dereverberation performance than the maximum likelihood estimator in [6], both for perfectly estimated RETFs as well as in the presence of RETF estimation errors. However, the RETF estimation errors considered in [13] were simulated using knowledge of the true RETFs.

4. RETF ESTIMATION

In the following several practical RETF estimation methods which can be used together with the EVD-based PSD estimator are described, i.e., the covariance whitening method [14, 15], the covariance subtraction method [15, 16], and the least-squares method [17]. Note that the covariance whitening and subtraction methods typically have been used to estimate the complete relative transfer functions instead of the RETFs [15], hence, in the following these methods are formulated for RETF estimation.

4.1. Covariance whitening method

In order to estimate the RETFs using the covariance whitening method, the EVD in (14) can readily be used. Let us denote by $\mathbf{u}(l)$ the eigenvector corresponding to $\lambda_1\{\mathbf{R}^{\rm w}_{\mathbf{x}}(l)\}$ (i.e., the first column of $\mathbf{U}(l)$). Since $\mathbf{R}^{\rm w}_{\mathbf{x}}(l)$ is given by (12) with $\mathbf{b}(l)$ in (13), the RETF vector $\mathbf{d}(l)$ is a scaled and rotated version of $\mathbf{u}(l)$. The estimated RETF vector using the covariance whitening method is obtained by transforming $\mathbf{u}(l)$ back from the whitened domain and normalizing by its first entry (since the first microphone is the reference microphone), i.e.,

$$\hat{\mathbf{d}}_{\rm CW}(l) = \frac{\mathbf{L}\mathbf{u}(l)}{\mathbf{e}^T\mathbf{L}\mathbf{u}(l)}, \tag{17}$$

with $\mathbf{e} = [1\ 0\ \ldots\ 0]^T$.

4.2. Covariance subtraction method

In order to estimate the RETFs using the covariance subtraction method, first an estimate of the late reverberant PSD matrix $\hat{\mathbf{R}}_{\mathbf{x}_{\rm r}}(l)$ is constructed using $\hat{\Phi}_{\rm r}(l)$ in (16) and the coherence matrix $\mathbf{\Gamma}$, i.e., $\hat{\mathbf{R}}_{\mathbf{x}_{\rm r}}(l) = \hat{\Phi}_{\rm r}(l)\mathbf{\Gamma}$. By subtracting $\hat{\mathbf{R}}_{\mathbf{x}_{\rm r}}(l)$ from $\mathbf{R}_{\mathbf{x}}(l)$ in (5), an estimate of the direct and early reverberant speech component PSD matrix $\hat{\mathbf{R}}_{\mathbf{x}_{\rm d}}(l)$ is obtained, i.e., $\hat{\mathbf{R}}_{\mathbf{x}_{\rm d}}(l) = \mathbf{R}_{\mathbf{x}}(l) - \hat{\mathbf{R}}_{\mathbf{x}_{\rm r}}(l)$. The estimated RETF vector using the covariance subtraction method is then computed by normalizing the first column of $\hat{\mathbf{R}}_{\mathbf{x}_{\rm d}}(l)$ by its first entry, i.e.,

$$\hat{\mathbf{d}}_{\rm CS}(l) = \frac{\hat{\mathbf{R}}_{\mathbf{x}_{\rm d}}(l)\mathbf{e}}{\mathbf{e}^T\hat{\mathbf{R}}_{\mathbf{x}_{\rm d}}(l)\mathbf{e}}. \tag{18}$$

Note that for perfect knowledge of the speech component PSD matrix, late reverberant PSD, and late reverberation coherence matrix, both the covariance whitening and subtraction method yield the true RETF vector according to the signal model in (5). However, due to unavoidable errors in the estimation of $\mathbf{R}_{\mathbf{x}}(l)$ and since the late reverberation is not truly isotropic (i.e., the assumed coherence matrix is erroneous) the covariance whitening and subtraction methods can yield different RETFs.

4.3. Least-squares method

In order to estimate the RETFs using the least-squares method, consider that the direct and early reverberant speech components $X_{{\rm d},m}(l)$ are related by the RETFs $D_m(l)$ according to (cf. (2))

$$X_{{\rm d},m}(l) = D_m(l)X_{{\rm d},1}(l), \quad m = 2,\ \ldots,\ M. \tag{19}$$

Multiplying both sides of (19) by $X^*_{{\rm d},1}(l)$ and taking the expectation yields

$$\Phi_{{\rm d}_{m,1}}(l) = D_m(l)\Phi_{{\rm d}_1}(l), \tag{20}$$

with $\Phi_{{\rm d}_{m,1}}(l)$ the cross-PSD of $X_{{\rm d},m}(l)$ and $X_{{\rm d},1}(l)$, i.e., $\Phi_{{\rm d}_{m,1}}(l) = \mathcal{E}\{X_{{\rm d},m}(l)X^*_{{\rm d},1}(l)\}$, and $\Phi_{{\rm d}_1}(l)$ the PSD of $X_{{\rm d},1}(l)$, i.e., $\Phi_{{\rm d}_1}(l) = \mathcal{E}\{|X_{{\rm d},1}(l)|^2\}$. Using (20), a least-squares criterion can be formulated to estimate the RETFs $D_m(l)$. Assuming that the RETFs are time-invariant during the latest T frames, a least-squares estimate of $D_m(l)$ is given by

$$\hat{D}_{m,{\rm LS}}(l) = \frac{\sum_{l'=l-T+1}^{l}\Phi_{{\rm d}_{m,1}}(l')\Phi_{{\rm d}_1}(l')}{\sum_{l'=l-T+1}^{l}\Phi^2_{{\rm d}_1}(l')}. \tag{21}$$

In order to estimate $X_{{\rm d},m}(l)$, single-channel dereverberation and denoising filters are applied to each microphone signal. For details on the derivation of these filters, the reader is referred to [17]. The PSDs required in (21) are then computed from the estimated $X_{{\rm d},m}(l)$ by recursive averaging.

Table 1: Characteristics of the considered acoustic systems.

System	T_{60} [ms]	$d_{\rm im}$ [cm]
AS$_1$	610	8
AS$_2$	800	5

5. EXPERIMENTAL RESULTS

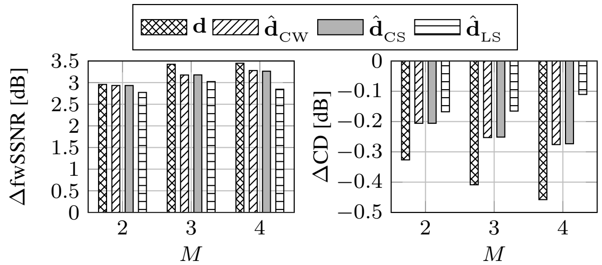
In this section the performance of the MWF in (7) using the EVD-based PSD estimator described in Section 3 and the different RETF estimation methods described in Section 4 is evaluated by means of objective performance measures. In Section 5.2 the performance in noiseless scenarios is investigated, whereas in Section 5.3 the performance in noisy scenarios is investigated.

5.1. Setup

We consider two multi-channel acoustic systems consisting of a linear microphone array with $M \in \{2,\ 3,\ 4\}$ microphones. Table 1 presents the reverberation time T_{60} and the inter-microphone distance $d_{\rm im}$ for each acoustic system. For system AS$_1$ the speaker was located at fixed positions of 0°, 15°, 30°, 45°, and 60°, with the speech components generated by convolving an 18 s long anechoic signal with measured room impulse responses (RIRs) [21]. For system AS$_2$ the speaker was moving from 0° to 60°, with the speech components simulated with the signal generator from [22] using a 3 s long anechoic signal. For both acoustic systems, the anechoic signals were taken from the TIMIT database [23]. It should be noted that although the performance of the considered techniques has been analyzed for a wide range of acoustic systems (different anechoic signals and RIRs), due to space constraint only two exemplary acoustic systems are presented in this paper. The noise components consist of a stationary directional interference at -30° and spatially uncorrelated noise. The considered signal-to-interference ratios (SIRs) are 10, 20, and 30 dB, and the signal-to-noise ratio is 20 dB. To be able to estimate the noise PSD matrix $\mathbf{R}_{\mathbf{v}}$ during speech absence, a 500 ms long noise-only signal precedes the speech signal.

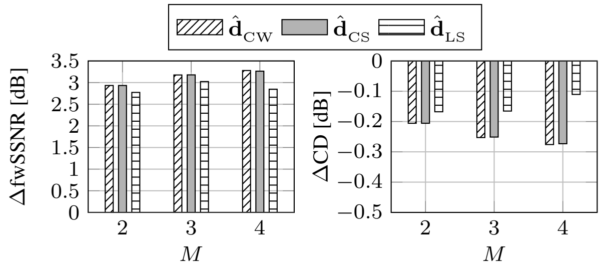
The signals are processed using a weighted overlap-add framework with a frame size of 1024 samples and an overlap of 75% at a sampling frequency of 16 kHz. The PSD matrices are computed using recursive averaging with a time constant of 50 ms. For the noiseless scenarios in Section 5.2, the PSD matrix $\mathbf{R}_{\mathbf{x}}(l)$ is directly estimated from the microphone signals. For the noisy scenarios in Section 5.3, the PSD matrix $\mathbf{R}_{\mathbf{x}}(l)$ is estimated as $\mathbf{R}_{\mathbf{y}}(l) - \mathbf{R}_{\mathbf{v}}$, with $\mathbf{R}_{\mathbf{y}}(l)$ estimated from the microphone signals during the speech-plus-noise period and $\mathbf{R}_{\mathbf{v}}$ estimated during the noise-only period. Due to PSD matrix estimation errors, computing $\mathbf{R}_{\mathbf{x}}(l)$ as $\mathbf{R}_{\mathbf{y}}(l) - \mathbf{R}_{\mathbf{v}}$ may not yield a positive definite matrix, particularly at low input SIRs. The estimated $\mathbf{R}_{\mathbf{x}}(l)$ is forced to be positive semi-definite by computing its eigenvalue decomposition and setting the negative eigenvalues to 0. The number of frames used for the least-squares average in (21) is $T = 3$. A minimum gain of -20 dB is used for the single-channel Wiener postfilter.

The performance is evaluated in terms of the improvement in frequency-weighted segmental signal-to-noise-ratio (ΔfwSSNR) [24] and cepstral distance (ΔCD) [25] between the output signal and the first microphone signal. The fwSSNR and CD measures are intrusive measures, comparing the output signal to a reference signal. The reference signal used in this paper is the anechoic signal. Note that a positive ΔfwSSNR and a negative ΔCD indicate a performance improvement.



(a) (b)

Fig. 1: Dereverberation performance for system AS$_1$ using the true and the estimated RETFs: (a) ΔfwSSNR and (b) ΔCD.



(a) (b)

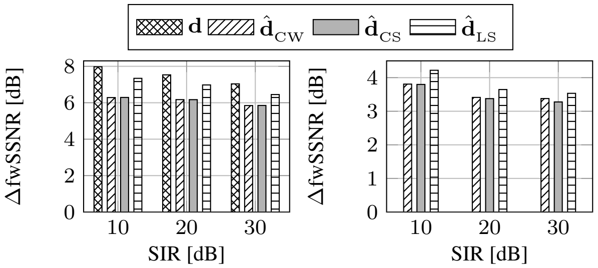
Fig. 2: Dereverberation performance for system AS$_2$ using the estimated RETFs: (a) ΔfwSSNR and (b) ΔCD.

5.2. Performance in noiseless scenarios

In this section the dereverberation performance of the MWF using the EVD-based PSD estimator and the considered RETF estimation methods is investigated for both acoustic systems and all considered array configurations. Since for system AS$_1$ the measured RIRs are available, the true RETFs can be constructed using the frequency response of the truncated direct path and early reflections of the measured RIRs (up to 10 ms). Hence, for system AS$_1$ also the performance when using the true RETFs is investigated, representing the optimal achievable performance.

Fig. 1 depicts the dereverberation performance for system AS$_1$ when using the true RETFs and the RETFs estimated with the considered methods. It can be observed that using the RETFs estimated with the covariance whitening and subtraction methods yields a very similar performance. In addition, it can be observed that using these methods results in a slightly better performance than using the least-squares method, both in terms of ΔfwSSNR and ΔCD. Finally, it can be observed that using the true RETFs yields only a slightly better performance than using the RETFs estimated with the covariance whitening and subtraction methods, with an insignificant performance difference in the order of 0.20 dB for ΔfwSSNR and 0.15 dB for ΔCD. Fig. 2 depicts the dereverberation performance for system AS$_2$ when using the RETFs estimated with the considered methods. Similarly as for system AS$_1$, it is illustrated that the covariance whitening and subtraction methods yield a very similar performance, slightly outperforming the least-squares method. Comparing the results presented in Figs. 1 and 2 it can be observed that although system AS$_2$ represents a more challenging system where the speaker position changes fast, the performance improvement obtained for both systems and all considered array configurations is similar.

In summary these results demonstrate the suitability of combining any of the considered RETF estimation methods with the EVD-based PSD estimator in a MWF to successfully dereverberate a fixed as well as a moving speaker. Since the covariance whitening and subtraction methods yield the best performance and since they do not introduce any additional significant computations, it can be said that they are the preferred methods to be used in reverberant scenarios.



(a) (a)

Fig. 3: Dereverberation and noise reduction performance using the true and the estimated RETFs for (a) system AS$_1$ and (b) system AS$_2$ (ΔfwSSNR, $M = 4$).

5.3. Performance in noisy scenarios

In this section the dereverberation and noise reduction performance of the MWF using the EVD-based PSD estimator and the considered RETF estimation methods is investigated for both acoustic systems and $M = 4$ microphones. As in Section 5.2, for system AS$_1$ also the performance when using the true RETFs is depicted. Since similar conclusions are derived by analyzing the ΔfwSSNR and the ΔCD values, Fig. 3 depicts only the ΔfwSSNR values for both acoustic systems. Fig. 3a shows that for system AS$_1$ using the RETFs estimated with the covariance whitening and subtraction methods yields a very similar performance. In addition, it is illustrated that in the presence of additive noise these methods result in a worse performance than the least-squares method, particularly for lower input SIRs. This can be explained by the fact that the covariance whitening and subtraction methods rely on the speech PSD matrix $\mathbf{R}_{\mathbf{x}}(l)$ to estimate the RETFs, which is unavoidably erroneous when computed as $\mathbf{R}_{\mathbf{y}}(l) - \mathbf{R}_{\mathbf{v}}$ (particularly for lower input SIRs). Finally, it can be observed that using the true RETFs yields only a slightly better performance than using the RETFs estimated with the least-squares method. Similar conclusions can be derived by analyzing the performance for system AS$_2$ depicted in Fig. 3b, i.e., the covariance whitening and subtraction methods yield a very similar performance. Furthermore, the least-squares method outperforms the covariance whitening and subtraction methods. However, the difference in performance between the different RETF estimation methods is smaller than for system AS$_1$, since system AS$_2$ represents a rather challenging system for any of the considered RETF estimation methods.

In summary these results demonstrate the suitability of combining any of the considered RETF estimation methods with the EVD-based PSD estimator to successfully dereverberate and denoise a fixed as well as a moving speaker. Since the least-squares method yields the best performance, it is the preferred method to be used in reverberant and noisy scenarios.

6. CONCLUSION

In this paper the performance of the MWF using the EVD-based late reverberant PSD estimator and several RETF estimation methods has been investigated for noiseless and noisy scenarios as well as for a fixed and a moving a speaker. It has been shown that using the EVD-based PSD estimator and any of the considered RETF estimation methods, i.e., the covariance whitening, covariance subtraction, and least-squares methods, yields a high dereverberation and noise reduction performance. While the covariance whitening and subtraction methods yield the best performance in reverberant scenarios, the least-squares method yields the best performance in reverberant and noisy scenarios.

7. REFERENCES

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, July 2006.

[2] 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 algorithm," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sept. 2014, pp. 333–337.

[3] 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.

[4] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, July 2007.

[5] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sept. 2013.

[6] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. European Signal Processing Conference*, Lisbon, Portugal, Sept. 2014.

[7] 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.

[8] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum Likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1595–1608, Sept. 2016.

[9] 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 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2015.

[10] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 151–155.

[11] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, June 2015.

[12] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Multichannel Wiener filter for speech dereverberation in hearing aids - sensitivity to DoA errors," in *Proc. AES 60th Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, Feb. 2016.

[13] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, Mar. 2017, accepted for publication.

[14] 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.

[15] 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.

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

[17] 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.

[18] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, Berlin, Germany, 2001.

[19] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, July 2010.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[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, Sept. 2014, pp. 313–317.

[22] E. A. P. Habets, "Signal generator for a moving sound source," `https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator`.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic phonetic continuous speech corpus," Philadelphia: Linguistic Data Consortium, 1993.

[24] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

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