JOINT MAXIMUM LIKELIHOOD ESTIMATION OF LATE REVERBERANT AND SPEECH POWER SPECTRAL DENSITY IN NOISY ENVIRONMENTS

Ofer Schwartz[⋆]*, Sharon Gannot*[⋆]*, and Emanuël A.P. Habets*[†]

[⋆]Bar-Ilan University, Faculty of Engineering, Ramat-Gan, 52900, Israel
[†]International Audio Laboratories Erlangen[∗], Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT

An estimate of the power spectral density (PSD) of the late reverberation is often required by dereverberation algorithms. In this work, we derive a novel multichannel maximum likelihood (ML) estimator for the PSD of the reverberation that can be applied in noisy environments. Since the anechoic speech PSD is usually unknown in advance, it is estimated as well. As a closed-form solution for the maximum likelihood estimator is unavailable, a Newton method for maximizing the ML criterion is derived. Experimental results show that the proposed estimator provides an accurate estimate of the PSD, and outperforms competing estimators. Moreover, when used in a multi-microphone dereverberation and noise reduction algorithm, the best performance in terms of the log-spectral distance is achieved when employing the proposed PSD estimator.

1. INTRODUCTION

Reverberation and ambient noise may degrade the ability of mobile devices, smart TVs and audio conferencing systems to process speech signals. While intelligibility does not degrade in presence of early speech reflections, it can be significantly deteriorated in reverberant environments due to overlap masking effects [1].

Both single- and multi-microphone techniques have been proposed to reduce reverberation (see [2] and the references therein). Many of these techniques require an estimate of the PSD of the reverberation (e.g. [3]). It should be noted that the estimation of the reverberation PSD is a much more challenging task than the estimation of the ambient noise PSD, since it is highly non-stationary and since speech-absence periods cannot be utilized.

In [4], reverberation was modelled as a diffuse sound field with time-varying level. Similarly to [5], the authors proposed to estimate the time-varying level of the reverberation from the signals at the output of a blocking matrix (BM) in a generalized sidelobe canceller (GSC) structure. The so-called error matrix of the reverberant PSD at the output of the BM was assumed to be normally distributed with zero-mean. The time-varying reverberation level was estimated by maximizing the log-likelihood. In [6], the authors considered two microphones and used a blind source separation algorithm to separate the early speech component and the late reverberation component. The estimated late reverberant signal was then used to compute one single-channel Wiener filter that is applied to both microphone signals. In [7] dereverberation for hearing aids applications is addressed, assuming a noise-free environment. A closed-form solution for the ML estimate of the time-varying reverberation PSD and of the anechoic speech PSD is then derived without using any BM. In [8], the authors proved that the maximum likelihood estimator (MLE) derived in [7], which circumvents the BM, has a lower mean squared error than the MLE derived in [4], which uses the BM. Recently, in [9], an optimal estimator for the reverberation PSD in noisy environment was proposed. First, the received signals are filtered by a BM to block the anechoic speech. Then, the likelihood of the reverberation PSD given the signals at the output of the BM, is maximized. However, since the BM processes the data and reduces the number of available signals, applying the MLE at the output of the BM might be sub-optimal.

In this work, an optimal estimator in the ML sense for the reverberation PSD in noisy environment is derived. The reverberation PSD is modelled as a diffuse sound field with time-varying level, while the noise PSD is assumed to be known. Since the anechoic speech PSD is changing rapidly across time and is unknown in advance, the anechoic speech should be either blocked or estimated. Since the blocking operation reduces the amount of information useful for the estimation, we prefer to circumvent the blocking. Therefore, the reverberation PSD and the anechoic speech PSD will be jointly estimated. Due to the complexity of the probability density function (p.d.f.), a closed-form solution cannot be derived. Instead, an iterative Newton method for maximizing the ML is derived. For the application of Newton's iterations, the first- and second-order derivatives of the log-likelihood are calculated in closed-form. An experimental study using recorded noisy and reverberant speech signals demonstrates the benefits of the proposed algorithm in terms of the late reverberation estimation accuracy. Moreover, it is shown that when used in the multichannel Wiener filter for joint noise reduction and dereverberation, the proposed estimator outperforms competing estimators.

2. PROBLEM FORMULATION

Consider N microphone observations consisting of reverberant speech and additive noise. The reverberant speech can be decomposed into two components, i.e., a direct speech component and a reverberation component. The i-th microphone observation can then be expressed as

$$Y_i(m,k) = X_{\mathrm{d},i}(m,k) + X_{\mathrm{r},i}(m,k) + V_i(m,k), \qquad (1)$$

where $Y_i(m,k)$ denotes the i-th microphone observation with time-index m and frequency index k, $X_{\mathrm{d},i}(m,k)$ denotes the direct speech component, $X_{\mathrm{r},i}(m,k)$ denotes the reverberation, and $V_i(m,k)$ denotes the ambient noise. Here $X_{\mathrm{d},i}(m,k)$ is modeled as a multiplication of the anechoic speech $S(m,k)$ (as received

This research was partially supported by a Grant from the GIF, the German-Israeli Foundation for Scientific Research and Development.

[∗]A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

by the first microphone that was arbitrary chosen as the reference microphone) and the relative direct-path transfer function (RDTF) of the i-th microphone $G_{\mathrm{d},i}(k)$, i.e.,

$$X_{\mathrm{d},i}(m,k) = G_{\mathrm{d},i}(k)S(m,k). \tag{2}$$

For a plane wave the RDTF $G_{\mathrm{d},i}(k)$ depends only on the time difference of arrival (TDOA) between the i-th microphone and the first microphone that is denoted by τ_i, i.e.,

$$G_{\mathrm{d},i}(k) = \exp\left(-j\frac{2\pi k}{K}\frac{\tau_i}{T_{\mathrm{s}}}\right), \tag{3}$$

where $j = \sqrt{-1}$, T_{s} is the sampling time, and K is the number of frequency bins. The estimation of τ_i is beyond the scope of this paper. The N microphone signals can be concatenated in a vector

$$\mathbf{y}(m,k) = \mathbf{x}_{\mathrm{d}}(m,k) + \mathbf{x}_{\mathrm{r}}(m,k) + \mathbf{v}(m,k)$$

where

$$\begin{aligned}
\mathbf{y}(m,k) &= \left[\begin{array}{cccc} Y_1(m,k) & \dots & Y_N(m,k) \end{array}\right]^{\mathrm{T}} \\
\mathbf{x}_{\mathrm{d}}(m,k) &= \left[\begin{array}{cccc} X_{\mathrm{d},1}(m,k) & \dots & X_{\mathrm{d},N}(m,k) \end{array}\right]^{\mathrm{T}} \\
&= \mathbf{g}_{\mathrm{d}}(k)S(m,k), \\
\mathbf{g}_{\mathrm{d}}(k) &= \left[\begin{array}{cccc} G_{\mathrm{d},1}(k) & \dots & G_{\mathrm{d},N}(k) \end{array}\right]^{\mathrm{T}} \\
\mathbf{x}_{\mathrm{r}}(m,k) &= \left[\begin{array}{cccc} X_{\mathrm{r},1}(m,k) & \dots & X_{\mathrm{r},N}(m,k) \end{array}\right]^{\mathrm{T}} \\
\mathbf{v}(m,k) &= \left[\begin{array}{cccc} V_1(m,k) & \dots & V_N(m,k) \end{array}\right]^{\mathrm{T}}.
\end{aligned}$$

The speech signal is modeled as a complex Gaussian process with $S(m,k) \sim \mathcal{N}_C(0, \phi_S(m,k))$. The reverberation and the noise components are assumed to be uncorrelated and may be modelled by zero-mean multivariate Gaussian probability density functions. The PSD matrix of the noise is assumed to be time-invariant and known in advance (or can be accurately estimated during speech-absent periods). The PSD matrix of the reverberation is naturally time-variant, since the reverberation originates from the speech source. On the other hand, the spatial characteristic of the reverberation may be assumed constant, as long as the speaker and microphones positions do not change. Therefore, it is reasonable to model the PSD matrix of the reverberation as a time-invariant normalized matrix with time-varying level. Finally, the reverberation is modelled as

$$\mathbf{x}_{\mathrm{r}}(m,k) \sim \mathcal{N}_C\left(0,\, \phi_R(m,k)\,\mathbf{\Gamma}(k)\right), \tag{4}$$

where $\mathbf{\Gamma}(k)$ is the time-invariant spatial coherence matrix of the reverberation and $\phi_R(m,k)$ is the temporal level of the reverberation. In the current contribution we assume that the reverberation can be modelled using a spatially homogenous and spherically isotropic sound field and determine $\mathbf{\Gamma}(k)$ accordingly [10, 11]

$$\mathbf{\Gamma}_{ij}(k) = \mathrm{sinc}\left(\frac{2\pi k}{K}\frac{d_{i,j}}{T_{\mathrm{s}}c}\right), \tag{5}$$

where $\mathrm{sinc}(x) = \sin(x)/x$, $d_{i,j}$ is the inter-distance between microphones i and j and c is the sound velocity. Collecting all definitions, the microphone signal vector is modelled as

$$\begin{aligned}
\mathbf{y}(m,k) \sim \mathcal{N}_C\big(0, &\phi_S(m,k)\mathbf{g}_{\mathrm{d}}(k)\mathbf{g}_{\mathrm{d}}^{\mathrm{H}}(k) \\
&+ \phi_R(m,k)\,\mathbf{\Gamma}(k) + \mathbf{\Phi}_{\mathbf{v}}(k)\big),
\end{aligned} \tag{6}$$

where $\mathbf{\Phi}_{\mathbf{v}}(k)$ is the PSD matrix of the noise.

The goal in this work is to estimate the late reverberation PSD $\phi_R(m,k)$. Since the speech PSD $\phi_S(m,k)$ is unknown, it should be estimated as well. Therefore, $\phi_R(m,k)$ and $\phi_S(m,k)$ will be jointly estimated as a parameter set:

$$\boldsymbol{\phi}(m,k) = \left[\begin{array}{cc} \phi_R(m,k) & \phi_S(m,k) \end{array}\right]^{\mathrm{T}}. \tag{7}$$

3. PROPOSED MAXIMUM LIKELIHOOD ESTIMATOR

In this section we derive the maximum likelihood estimator of the parameter set $\boldsymbol{\phi}(m,k)$. Whenever possible, the frequency index k is omitted for brevity. The ML estimator of $\boldsymbol{\phi}(m)$ is given by

$$\boldsymbol{\phi}^{\mathrm{ML}}(m) = \underset{\boldsymbol{\phi}(m)}{\mathrm{argmax}} \log f\left(\mathbf{y}(m);\, \boldsymbol{\phi}(m)\right), \tag{8}$$

where

$$\begin{aligned}
&f\left(\mathbf{y}(m);\, \boldsymbol{\phi}(m)\right) = \\
&\qquad \frac{1}{\pi^N |\mathbf{\Phi}_{\mathbf{y}}(m)|} \exp\left(-\mathbf{y}^{\mathrm{H}}(m)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{y}(m)\right),
\end{aligned} \tag{9}$$

and $\mathbf{\Phi}_{\mathbf{y}}(m) = \phi_S(m)\mathbf{g}_{\mathrm{d}}\mathbf{g}_{\mathrm{d}}^{\mathrm{H}} + \phi_R(m)\,\mathbf{\Gamma} + \mathbf{\Phi}_{\mathbf{v}}$. Since there is no closed-form solution for the ML of $\boldsymbol{\phi}(m)$, we propose to iteratively determine the solution.

3.1. ML estimation using Newton's method

In this work, the ML solution is obtained using the Newton's method (c.f. [12]):

$$\boldsymbol{\phi}^{(\ell+1)}(m) = \boldsymbol{\phi}^{(\ell)}(m) - \mathbf{H}^{-1}\left(\boldsymbol{\phi}^{(\ell)}(m)\right)\mathbf{d}\left(\boldsymbol{\phi}^{(\ell)}(m)\right), \tag{10}$$

where $\mathbf{d}\left(\boldsymbol{\phi}(m)\right)$ is the first-order derivative of the log-likelihood with respect to $\boldsymbol{\phi}(m)$ and $\mathbf{H}\left(\boldsymbol{\phi}(m)\right)$ is the corresponding Hessian matrix, i.e.,

$$\begin{aligned}
\mathbf{d}\left(\boldsymbol{\phi}(m)\right) &\equiv \frac{\partial \log f\left(\mathbf{y}(m);\, \boldsymbol{\phi}(m)\right)}{\partial \boldsymbol{\phi}(m)}, \\
\mathbf{H}\left(\boldsymbol{\phi}(m)\right) &\equiv \frac{\partial^2 \log f\left(\mathbf{y}(m);\, \boldsymbol{\phi}(m)\right)}{\partial \boldsymbol{\phi}(m)\partial \boldsymbol{\phi}^{\mathrm{T}}(m)}.
\end{aligned} \tag{11}$$

The first-order derivative $\mathbf{d}\left(\boldsymbol{\phi}(m)\right)$ is a 2-dimensional vector

$$\mathbf{d}\left(\boldsymbol{\phi}(m)\right) \equiv \left[\begin{array}{cc} D_R\left(\boldsymbol{\phi}(m)\right) & D_S\left(\boldsymbol{\phi}(m)\right) \end{array}\right]^{\mathrm{T}}, \tag{12}$$

with elements

$$D_i\left(\boldsymbol{\phi}(m)\right) = \mathrm{Tr}\left[\left(\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{R}(m) - \mathbf{I}\right)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_i(m)}\right], \tag{13}$$

for $i \in \{R, S\}$ where

$$\mathbf{R}(m) \equiv \mathbf{y}(m)\mathbf{y}^{\mathrm{H}}(m), \tag{14}$$

$\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_R(m)} = \mathbf{\Gamma}$ and $\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_S(m)} = \mathbf{g}_{\mathrm{d}}\mathbf{g}_{\mathrm{d}}^{\mathrm{H}}$. The Hessian is a 2×2 matrix:

$$\mathbf{H}\left(\phi(m)\right) \equiv \left[\begin{array}{cc} H_{RR}\left(\boldsymbol{\phi}(m)\right) & H_{SR}\left(\boldsymbol{\phi}(m)\right) \\ H_{RS}\left(\boldsymbol{\phi}(m)\right) & H_{SS}\left(\boldsymbol{\phi}(m)\right) \end{array}\right]. \tag{15}$$

Applying second derivative on $D_R\left(\boldsymbol{\phi}(m)\right)$ and $D_S\left(\boldsymbol{\phi}(m)\right)$ yields the elements of $\mathbf{H}\left(\boldsymbol{\phi}(m)\right)$:

$$\begin{aligned}
&H_{ij}\left(\boldsymbol{\phi}(m)\right) = \\
&-\mathrm{Tr}\Big[\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_j(m)}\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{R}(m)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_i(m)} + \\
&\left(\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{R}(m) - \mathbf{I}\right)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_j(m)}\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\frac{\partial \mathbf{\Phi}_{\mathbf{y}}(m)}{\partial \phi_i(m)}\Big],
\end{aligned} \tag{16}$$

where $i,j \in \{R,S\}$. Collecting all terms, the Newton's method for estimating $\phi_R(m)$ and $\phi_S(m)$ can be reformulated as

$$\boldsymbol{\phi}_R^{(\ell+1)}(m) = \boldsymbol{\phi}_R^{(\ell)}(m) - \Big[H_{SS}\left(\boldsymbol{\phi}^{(\ell)}(m)\right) D_R\left(\boldsymbol{\phi}^{(\ell)}(m)\right) - H_{SR}\left(\boldsymbol{\phi}^{(\ell)}(m)\right) D_S\left(\boldsymbol{\phi}^{(\ell)}(m)\right)\Big] \frac{1}{\left|\mathbf{H}\left(\boldsymbol{\phi}^{(\ell)}(m)\right)\right|} \quad (17)$$

and

$$\boldsymbol{\phi}_S^{(\ell+1)}(m) = \boldsymbol{\phi}_S^{(\ell)}(m) - \Big[H_{RR}\left(\boldsymbol{\phi}^{(\ell)}(m)\right) D_S\left(\boldsymbol{\phi}^{(\ell)}(m)\right) - H_{RS}\left(\boldsymbol{\phi}^{(\ell)}(m)\right) D_R\left(\boldsymbol{\phi}^{(\ell)}(m)\right)\Big] \frac{1}{\left|\mathbf{H}\left(\boldsymbol{\phi}^{(\ell)}(m)\right)\right|}, \quad (18)$$

where

$$\left|\mathbf{H}\left(\boldsymbol{\phi}^{(\ell)}(m)\right)\right| = H_{RR}\left(\boldsymbol{\phi}^{(\ell)}(m)\right) H_{SS}\left(\boldsymbol{\phi}^{(\ell)}(m)\right) - H_{RS}^2\left(\boldsymbol{\phi}^{(\ell)}(m)\right) \quad (19)$$

is the determinant of $\mathbf{H}\left(\boldsymbol{\phi}^{(\ell)}(m)\right)$. According to our experiments, the Newton's method converges after about 10-20 iterations.

3.2. Practical considerations

When implementing the iterative solution proposed above, some practical issues should be considered:

Lower and Upper Bounds: The estimated PSDs $\phi_R^{(\ell)}(m)$ and $\phi_S^{(\ell)}(m)$ must be positive, and should therefore be restricted to the $(+,+)$ quadrant.

There is a minimum-point of the likelihood function when $\phi_R{}^{(\ell)}(m)$ or $\phi_S^{(\ell)}(m)$ approaches infinity. Also, the first-derivative in (13) approaches zero when $\phi_R{}^{(\ell)}(m)$ or $\phi_S^{(\ell)}(m)$ approaches infinity. Thus, for large values of $\phi_R^{(\ell)}(m)$ or $\phi_S^{(\ell)}(m)$ there is a hazard that Newton's method will step to infinity. To prevent this from occurring, we apply the following upper bound to $\phi_R^{(\ell)}(m)$ and $\phi_S^{(\ell)}(m)$:

$$Z(m) \equiv \frac{1}{N}\mathbf{y}^H(m)\mathbf{y}(m) - \frac{1}{N}\mathrm{Tr}\left[\boldsymbol{\Phi}_{\mathbf{v}}\right], \quad (20)$$

which is equal to the instantaneous level of the observations minus the average noise level.

Initialization: According to our experience, the log-likelihood function exhibits only a single two-dimensional peak for positive values of $\phi_R(m)$ and $\phi_S(m)$. However, for extreme cases the peak may be located at negative values of $\phi_R(m)$ or $\phi_S(m)$. If the peak is located in the $(-,-)$ quadrant, $\phi_R^{\mathrm{ML}}(m)$ and $\phi_S^{\mathrm{ML}}(m)$ should be set to zero and Newton's method should becomes inactive. However, even when the peak is located in the $(-,+)$ quadrant or in the $(+,-)$ quadrant, Newton's method should becomes active to find the point with the highest likelihood in the $(+,+)$ quadrant. We have therefore applied a simple initialization step. If $D_R(\mathbf{0})$ and $D_S(\mathbf{0})$ are negative, we postulate that the peak is located in the $(-,-)$ quadrant. Then, $\phi_R^{\mathrm{ML}}(m)$ and $\phi_S^{\mathrm{ML}}(m)$ are set to zero (or a small pre-defined value ϵ) and Newton's procedure is skipped. Otherwise, the Newton procedure becomes active and its initial value is set to $\boldsymbol{\phi}^{(0)}(m) = \mathbf{0}$. As mentioned above, the search is confined to the $(+,+)$ quadrant.

Algorithm 1: Multi-microphone reverberation and speech PSD estimation in noisy environment.

```
for all time frames and frequency bins m, k do
    Compute R̄(m) using (14) and (21).
    Initialize by φ^(0)(m) = 0.
    if (D_R(0) < 0) & (D_S(0) < 0) then
        φ_R^ML(m) = φ_S^ML(m) = ε
    else
        for ℓ = 0 to L − 1 do
            Calculate φ_R^(ℓ+1)(m) and φ_S^(ℓ+1)(m) using (17) and (18).
            Confine φ_R^(ℓ+1)(m) and φ_S^(ℓ+1)(m) to the range [ε, Z(m)].
        end
    end
end
```

Smoothing: The late reverberation PSD is expected to be smooth over time. Smoothing stage may be carried out by time-averaging of the instantaneous PSD matrix, i.e.,

$$\bar{\mathbf{R}}(m) = \alpha_R\,\bar{\mathbf{R}}(m) + (1-\alpha_R)\,\mathbf{R}(m). \quad (21)$$

where α_R ($0 \le \alpha_R < 1$) is a smoothing factor. $\bar{\mathbf{R}}(m)$ is used to calculate the first- and second-order derivatives in (13) and (16) instead of $\mathbf{R}(m)$. The proposed ML estimator is summarized in Algorithm 1.

4. PERFORMANCE EVALUATION

The performance of the proposed estimator is evaluated by: 1) examining the log-error between the estimated value of $\phi_R^{\mathrm{ML}}(m)$ and the true reverberation level, obtained by convolving the speech signal by the late component of the acoustic impulse response; and 2) utilizing the estimated PSD $\phi_R^{\mathrm{ML}}(m)$ in a speech dereverberation task.

4.1. Simulation setup

The experiments consist of reverberant signals plus directional noise with various signal-to-noise ratio (SNR) levels. Spatially white (sensor) noise was also added, with power 20 dB lower than the directional noise power. Anechoic speech signals were convolved by room impulse responses (RIRs), downloaded from an open-source database of our lab. Details about the database and RIRs identification method can be found in [13]. Reverberation time was set by adjusting the room panels, and was measured to be approximately $T_{60} = 0.61$ s. The reverberant speech signals were mixed with directional noise signals with several SNR levels. The spatial PSD matrix $\boldsymbol{\Phi}_{\mathbf{v}}$ was estimated using periods in which the desired speech source was inactive. The loudspeaker was positioned in front of a four microphone linear array such that the steering vector was set to $\mathbf{g}_{\mathrm{d}} = \left[\begin{array}{cccc} 1 & 1 & 1 & 1 \end{array}\right]^{\mathrm{T}}$. The inter-distances between the microphones were [3, 8, 3] cm. The sampling frequency was 16 kHz, the frame length of the short-time Fourier transform (STFT) was 32 ms with 8 ms between successive time-frames (i.e., 75% overlap). In Algorithm 1, the smoothing parameter was set to $\alpha_R = 0.95$ and $\epsilon = 10^{-10}$. The number of iterations was $L = 10$. All measures were computed by averaging the results obtained using 50 sentences, 4–8 s long, evenly distributed between female and male speakers. In Fig. 1, a contour plot of the observed signals p.d.f. (9) is depicted

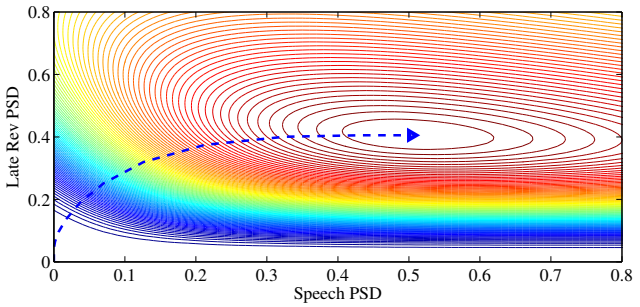

Fig. 1. Example of the observed signals p.d.f. (9) (for single T-F bin) w.r.t. the late reverberation PSD and the speech PSD.

(for a single T-F bin). It is evident that only one maximum as a function of the late reverberation PSD and the speech PSD exists. The dashed line depicts the convergence of the Newton iteration to this maximum.

4.2. Accuracy of the ML estimator

The performance of the proposed estimator was compared to three existing estimators in terms of log-error between the estimated PSD and the oracle PSD: 1) the estimator in [4], denoted henceforth Braun2013, 2) the estimator in [14][1], denoted henceforth Lefkimmiatis2006 (assuming the signals are time-aligned), and 3) the ML estimator in [9] which initially blocks the direct path. For each algorithm, an identical lower and upper delimitation and identical smoothing were carried out as explained in Section 3.2. It should be noted that in [14] it is explicitly assumed that the direct-paths are time aligned prior to the PSD estimation.

The mean log-errors between the estimated PSD levels and the oracle PSD levels are presented. In order to calculate the oracle PSD levels, the anechoic speech was filtered with the reverberation tails of the RIRs. The reverberation tails were set to start 2 ms after the arrival time of the direct-path. To reduce the variance of the oracle PSD, the mean value of the oracle PSDs over all microphones was computed. The log-error results for several SNR levels are depicted in Fig. 2. The results bars are split to distinguish between underestimation errors and overestimation errors.

It is evident that the proposed estimator outperforms the ML estimator in [9] in terms of overall log-error for all evaluated SNRs. The proposed estimator also outperforms Braun2013 [4] for an SNR of 5 and 10 dB. Lefkimmiatis2006 [14] outperforms all competing estimators for all evaluated SNRs. For a yet unknown reason, this result is not reflected to the dereverberation performance as shown in the next section.

4.3. Dereverberation performance

The performance of the proposed estimator is also examined by utilizing the estimated PSDs for joint dereverberation and noise reduction. The estimated PSDs were used to compute the multichannel Wiener filter presented in [3]. The multichannel Wiener filter (MCWF) was designed to jointly suppress the power of the total interference (e.g. the reverberation and the noise) by estimating

[1]Note that the algorithm in [14] aims to estimate the noise variance given the received signals PSD matrix while the noise coherence matrix is assumed to be known. In our implementation, we treat the reverberation as an additive noise, and we estimate its level while we subtract the ambient noise PSD matrix from the received signals PSD matrix.

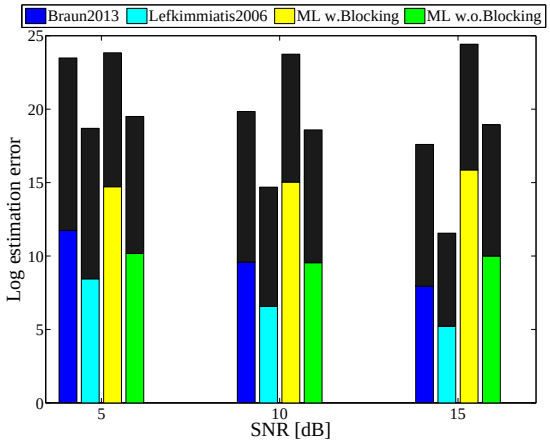

Fig. 2. Log-errors of the proposed late reverberation PSD estimator in comparison with [4], [14] and [9]. The upper part of each bar represents the underestimation error, while the lower part represents the overestimation error.

SNR	5 dB	10 dB	15 dB
Unprocessed	1.33 (9.42)	1.53 (8.08)	1.73 (6.91)
Oracle $\phi_R(m)$	1.93 (5.79)	2.04 (5.36)	2.10 (5.09)
Braun2013 [4]	1.81 (6.32)	1.94 (5.71)	2.05 (5.36)
Lefkimmiatis2006 [14]	1.84 (6.15)	2.01 (5.67)	2.10 (5.35)
ML with blocking [9]	**1.92** (6.21)	**2.04** (5.66)	**2.12** (5.30)
ML without blocking	1.89 (**6.18**)	1.99 (**5.61**)	2.08 (**5.27**)

Table 1. PESQ (and LSD in brackets) scores for the MCWF [3] using various estimators.

the minimum mean square error (MMSE) estimation of the direct-path. The MCWF was implemented by a two stage approach: a minimum variance distortionless response beamformer followed by a corresponding post-filter. The performance of the dereverberation algorithm was evaluated in terms of two objective measures, commonly used in the speech enhancement community, namely perceptual evaluation of speech quality (PESQ) [15] and log-spectral distance (LSD). The clean reference for evaluation in all cases was the anechoic speech signal filtered only with the direct path of the RIR. In Table 1 the performance measures for several input SNR levels are depicted. The proposed estimator outperforms all competing estimators with respect to the LSD measures. As for the PESQ scores, the ML estimator in [9] that explicitly blocks the direct-path outperforms all competing estimators.

5. CONCLUSIONS

In this work a joint ML estimator for the late reverberant PSD and the anechoic speech PSD was derived that can be used in noisy environments. The proposed algorithm maximizes the log-likelihood of the received signals using Newton iterations. In contrast to previous work [9] the speech PSD was estimated rather than blocked. The fact that the proposed PSD estimator uses all N microphone signals, rather than $N-1$ signals due to the blocking matrix, might explain why the proposed estimator performs better. An experimental study demonstrated the advantage of the proposed PSD estimator when used in combination with a multichannel Wiener filter for joint noise reduction and dereverberation.

6. REFERENCES

[1] A. Kjellberg, "Effects of reverberation time on the cognitive load in speech communication : Theoretical considerations," *Noise and Health*, vol. 7, no. 25, pp. 11–21, 2004.

[2] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[3] 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*, pp. 240–251, Feb. 2015.

[4] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO), Marrakech, Morocco, Aug.*, 2013, pp. 1–5.

[5] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 295–299.

[6] A. Schwarz, K. Reindl, and W. Kellermann, "A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 113–116.

[7] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 61–65.

[8] S. Doclo A. Kuklasinski, S. H. Jensen T. Gerkmann, and J. Jensen, "Multi-channel psd estimators for speech dereverberation-a theoretical and experimental comparison," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr.* IEEE, 2015, pp. 91–95.

[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 (WASPAA), New-Paltz, NY, USA, Oct.*, 2015.

[10] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, no. 1, pp. 43–56, 1988.

[11] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, pp. 3464–3470, Dec. 2007.

[12] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[13] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC), Aachen, Germany, Sep.*, 2014, pp. 313–317.

[14] S. Lefkimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech applications.," in *Proc. Interspeech Conf.*, 2006.

[15] 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," Feb. 2001.