JOINT ESTIMATION OF LATE REVERBERANT AND SPEECH POWER SPECTRAL DENSITIES IN NOISY ENVIRONMENTS USING FROBENIUS NORM

Ofer Schwartz, Sharon Gannot
Bar-Ilan University
Faculty of Engineering
Ramat-Gan, 52900, Israel

Emanuël A.P. Habets
International Audio Laboratories Erlangen
A joint institution of the Friedrich-Alexander University
Erlangen-Nürnberg (FAU) and Fraunhofer IIS
Am Wolfsmantel 33, 91058 Erlangen, Germany

***Abstract*—Various dereverberation and noise reduction algorithms require power spectral density estimates of the anechoic speech, reverberation, and noise. In this work, we derive a novel multichannel estimator for the power spectral densities (PSDs) of the reverberation and the speech suitable also for noisy environments. The speech and reverberation PSDs are estimated from all the entries of the received signals power spectral density (PSD) matrix. The Frobenius norm of a general error matrix is minimized to find the best fitting PSDs. Experimental results show that the proposed estimator provides accurate estimates of the PSDs, and is outperforming competing estimators. Moreover, when used in a multi-microphone noise reduction and dereverberation algorithm, the estimated reverberation and speech PSDs are shown to provide improved performance measures as compared with the competing estimators.**

I. 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 the overlap masking effects [1].

Both single- and multi-microphone techniques have been proposed to reduce reverberation and ambient noise (see [2] and the references therein). Many of these techniques require an estimate of the PSDs of the reverberation and the speech, in particular beamforming-based methods. In our previous work [3], a multi-microphone minimum mean square error (MMSE) estimator of the early speech component was implemented as a minimum variance distortionless response (MVDR) beamformer followed by a postfilter. The reverberation and the ambient noise were treated by both the MVDR stage and the postfiltering stage. The most difficult task was to provide an accurate estimation of the speech PSD required for the postfiltering stage and the reverberation PSD required for the MVDR stage and for the postfiltering stage. The reverberation PSD was estimated by averaging the marginal reverberation levels at the microphones, obtained using the single-channel estimator proposed in [4]. The speech PSD was estimated by using the decision-directed approach [5].

Maximum likelihood estimators (MLEs) of time-varying speech and noise PSD levels were derived in the past. In [6], the authors proposed to estimate the PSD level of the ambient noise from signals at the output of a blocking matrix (BM) which blocks the speech signals. A closed-form maximum likelihood estimator (MLE) of the noise level was derived w.r.t. the probability density function (p.d.f.) of the BM outputs. In [7], a closed-form solution for the MLE of the reverberation and the anechoic speech PSDs was proposed without applying any BM. The reverberation was modelled as an additive noise. Recently, in [8], an MLE for the reverberation PSD level in noisy environment was proposed. First, the received signals were filtered by a BM. Due to the complexity of the p.d.f., a closed-form solution could not be derived. Instead, an iterative Newton method for finding the maximum likelihood (ML) estimate was derived. In [9] an optimal estimator in the ML sense for the reverberation PSD in noisy environment was proposed without using a blocking stage. Instead, the reverberation and the anechoic speech PSD levels were jointly estimated. This MLE requires matrix inversions and is thus prone to instability. Beyond that, the iterative search (applied in the joint reverberant and noisy case) might not converge. Thus, in this paper we adopt a method which estimates the PSD levels by matching the observed signal PSD matrix with its model. We also circumvent the blocking stage, since it is unclear how this stage affects the accuracy of the estimates.

Speech and noise PSD estimation procedures are common practice in the design of postfilters. In [10], the author presented a practical estimator of the speech PSD suitable for spatially white noise fields. Assuming that the noise components are uncorrelated between microphones, the speech PSD was estimated by an average of the cross-PSDs. In [11], the technique was generalized to deal with an arbitrary noise field, using prior knowledge of the spatial coherence matrix of the noise. First, marginal estimates of the speech PSD were obtained from each microphone pair (using the auto- and cross-PSDs) and then the final estimate was obtained by averaging all the marginal estimates. In [12], the speech and noise PSDs are separately estimated. The speech PSD estimator identifies with the estimator in [11].

In [13], the authors estimated the speech PSD from all entries of the received signals' PSD matrix (rather than averaging the marginal microphone-pair estimates). The noise PSD matrix was assumed to be known. The best fitting value for the speech PSD was estimated by minimizing the Frobenius norm of an error matrix. Although, reverberation was not considered in this reference, the proposed technique based on the Frobenius norm minimization will be useful in the development of the proposed method.

In [14], a reverberant and noisy environment was assumed. The reverberation was modelled as a diffuse sound field with

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

time-varying level and the noise PSD matrix was assumed to be known. The authors proposed to estimate the time-varying level of the reverberation from the signals at the output of a BM in a generalized sidelobe canceller (GSC) structure. Since the speech signal was blocked, the PSD matrix of the BM outputs contains only reverberation and noise components. The best fitting value for the reverberation PSD was estimated by minimizing the Frobenius norm of an error matrix.

In this work, a joint estimator for the speech and the reverberation PSDs 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. The PSD-matrix of the received signals is first computed, and then the speech and the reverberation PSDs are jointly estimated from all entries of the PSD matrix. The Frobenius norm of a composite error matrix is minimized in order to find the best fitting speech and reverberation PSDs. Note, that unlike [13] and [14], here the error matrix depends on two variables, namely the PSDs of the speech and the reverberation. As opposed to [8], [9], a closed-form solution for the speech and reverberation PSDs is obtained.

The paper is organized as follows. In Section II, the problem is formulated. In Section III, the joint estimator for the speech and reverberation PSDs is derived. Section IV elaborates about the dereverberation and noise reduction algorithm used for the evaluation section. Section V presents the simulation setup, evaluates the performance of the proposed estimator, and compares the proposed estimator to two other estimators. Finally, in Section VI conclusions are drawn and the work is summarized.

II. 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-path speech component and a reverberation component. The i-th microphone observation can then be expressed as

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

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

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

The RDTF $G_{\mathrm{d},i}(k)$ is a pure phase depending on the time difference of arrival between the i-th microphone and the first microphone

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

where τ_i is the time difference of arrival (TDOA) between the i-th microphone and first microphone, 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 of the received microphone signals 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. The spatial characteristic of the reverberation may, however, assumed to be 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

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

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 [15], [16]

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

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 PSD matrix of the observations is given by

$$\begin{aligned}
\mathbf{\Phi}_\mathbf{y}(m,k) = \phi_S(m,k)\mathbf{g}_\mathrm{d}(k)\mathbf{g}_\mathrm{d}^\mathrm{H}(k) \\
+ \phi_R(m,k)\,\mathbf{\Gamma}(k) + \mathbf{\Phi}_\mathbf{v}(k), \quad (6)
\end{aligned}$$

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

The goal of this work is to jointly estimate the speech level $\phi_S(m,k)$ and the reverberation level $\phi_R(m,k)$ given a short-term estimate of $\mathbf{\Phi}_\mathbf{y}(m,k)$ and the noise PSD matrix $\mathbf{\Phi}_\mathbf{v}(k)$.

III. Proposed Joint Estimator of The Speech and Reverberation Levels

In this section, the parameter vector $\boldsymbol{\phi}(m,k) \equiv [\phi_S(m,k)\ \phi_R(m,k)]^T$ is estimated given the short-term estimate of the received signals PSD matrix. Whenever possible, the frequency index k is omitted for brevity. The observations PSD matrix $\mathbf{\Phi}_\mathbf{y}(m)$ can be recursively estimated:

$$\widehat{\mathbf{\Phi}}_\mathbf{y}(m) = \alpha\widehat{\mathbf{\Phi}}_\mathbf{y}(m-1) + (1-\alpha)\,\mathbf{y}(m)\mathbf{y}^{\mathrm{H}}(m), \quad (7)$$

where $0 \leq \alpha < 1$ is a smoothing factor. Matching (6) and (7), the problem at hand may be recast as a system of N^2 equations in two variables. Since there are more equations than variables, the best fitting parameter-set that minimizes the total squared error may be found by minimizing the Frobenius norm between the $\widehat{\mathbf{\Phi}}_\mathbf{y}(m)$ in (7) and its model in (6). Accordingly, $\boldsymbol{\phi}(m)$ is the minimizer of the following cost-function:

$$\widehat{\boldsymbol{\phi}}(m) = \underset{\boldsymbol{\phi}(m)}{\operatorname{argmin}}\, ||\mathbf{\Phi}_\mathrm{e}(m)||_\mathrm{F}^2, \quad (8)$$

where $\mathbf{\Phi}_\mathrm{e}(m)$ is an error matrix defined by

$$\mathbf{\Phi}_\mathrm{e}(m) = \widehat{\mathbf{\Phi}}_\mathbf{y}(m) - \left(\phi_S(m)\mathbf{g}_\mathrm{d}\mathbf{g}_\mathrm{d}^{\mathrm{H}} + \phi_R(m)\mathbf{\Gamma} + \mathbf{\Phi}_\mathbf{v}\right), \quad (9)$$

with $||\cdot||_\mathrm{F}^2$ is the squared Frobenius norm given for any arbitrary matrix $\mathbf{Z}$ by

$$||\mathbf{Z}||_\mathrm{F}^2 = \sum_{i,j}\left(\mathbf{Z}_{i,j}\right)^2 = \mathrm{Tr}\left[\mathbf{Z}^{\mathrm{H}}\mathbf{Z}\right]. \quad (10)$$

Following some algebraic steps, the cost function in (8) can be written as

$$||\mathbf{\Phi}_\mathrm{e}(m)||_\mathrm{F}^2 = \boldsymbol{\phi}^{\mathrm{T}}(m)\mathbf{A}\boldsymbol{\phi}(m) - 2\mathbf{b}^{\mathrm{T}}(m)\boldsymbol{\phi}(m) + C(m), \quad (11)$$

where $\mathbf{A}$ is time-invariant 2×2 matrix defined by

$$\mathbf{A} \equiv \begin{pmatrix} \left(\mathbf{g}_\mathrm{d}^{\mathrm{H}}\mathbf{g}_\mathrm{d}\right)^2 & \mathbf{g}_\mathrm{d}^{\mathrm{H}}\mathbf{\Gamma}\mathbf{g}_\mathrm{d} \\ \mathbf{g}_\mathrm{d}^{\mathrm{H}}\mathbf{\Gamma}\mathbf{g}_\mathrm{d} & \mathrm{Tr}\left[\mathbf{\Gamma}^{\mathrm{H}}\mathbf{\Gamma}\right] \end{pmatrix}, \quad (12)$$

$\mathbf{b}(m)$ is time-varying vector defined as

$$\mathbf{b}(m) \equiv \begin{pmatrix} \Re\left\{\mathbf{g}_\mathrm{d}^{\mathrm{H}}\left(\widehat{\mathbf{\Phi}}_\mathbf{y}(m) - \mathbf{\Phi}_\mathbf{v}\right)\mathbf{g}_\mathrm{d}\right\} \\ \Re\left\{\mathrm{Tr}\left[\left(\widehat{\mathbf{\Phi}}_\mathbf{y}(m) - \mathbf{\Phi}_\mathbf{v}\right)\mathbf{\Gamma}^{\mathrm{H}}\right]\right\} \end{pmatrix}, \quad (13)$$

where $\Re\{\cdot\}$ is the operator extracting the real-value and $C(m)$ is defined as

$$C(m) \equiv \mathrm{Tr}\left[\left(\widehat{\mathbf{\Phi}}_\mathbf{y}(m) - \mathbf{\Phi}_\mathbf{v}\right)^{\mathrm{H}}\left(\widehat{\mathbf{\Phi}}_\mathbf{y}(m) - \mathbf{\Phi}_\mathbf{v}\right)\right]. \quad (14)$$

Since the cost function $||\mathbf{\Phi}_\mathrm{e}(m)||_\mathrm{F}^2$ has a quadratic form, setting its gradient w.r.t. $\boldsymbol{\phi}(m)$ to zero yields the following minimum-point,

$$\widehat{\boldsymbol{\phi}}(m) = \mathbf{A}^{-1}\mathbf{b}(m). \quad (15)$$

Explicitly, $\phi_R(m)$ and $\phi_S(m)$ are obtained by

$$\widehat{\phi}_S(m) = \frac{\mathbf{A}_{22}\,\mathbf{b}_1(m) - \mathbf{A}_{12}\,\mathbf{b}_2(m)}{\mathbf{A}_{11}\mathbf{A}_{22} - \mathbf{A}_{12}^2}, \quad (16)$$

and

$$\widehat{\phi}_R(m) = \frac{\mathbf{A}_{22}\,\mathbf{b}_2(m) - \mathbf{A}_{21}\,\mathbf{b}_1(m)}{\mathbf{A}_{11}\mathbf{A}_{22} - \mathbf{A}_{12}^2}. \quad (17)$$

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

Compute $\mathbf{A}$ using (12).
for *all time-frames and frequency bins* m, k **do**
  Compute $\widehat{\mathbf{\Phi}}_\mathbf{y}(m)$ using (7).
  Compute $\mathbf{b}(m)$ using (13).
  Compute $\widehat{\phi}_S(m,k)$ and $\widehat{\phi}_R(m,k)$ using (16) and (17).
  Confine $\widehat{\phi}_S(m,k)$ and $\widehat{\phi}_R(m,k)$ to the range $[\epsilon, Z(m)]$.
end

The estimated PSDs $\widehat{\phi}_S(m)$ and $\widehat{\phi}_R(m)$ must be positive, and should therefore be restricted to the $(+,+)$ quadrant (or above some small positive number ϵ).

In addition, to guarantee physical plausibility, the following upper bound is applied to the estimates $\widehat{\phi}_S(m)$ and $\widehat{\phi}_R(m)$

$$Z(m) \equiv \frac{1}{N}\mathbf{y}^{\mathrm{H}}(m)\mathbf{y}(m) \quad (18)$$

which is equal to the instantaneous level of the observations. The proposed estimator is summarized in Algorithm 1.

IV. Dereverberation and Noise Reduction algorithm

In this section, the dereverberation and noise reduction algorithm used to examine the proposed estimators is briefly described. Since $S(m)$ and $\mathbf{y}(m)$ are assumed to be zero-mean complex-Gaussian random variables, the MMSE estimator of $S(m)$ can be calculated using the multichannel Wiener filter (MCWF) decomposed into an MVDR-BF filter and a subsequent postfilter [3]:

$$\widehat{S}(m) = \underbrace{\frac{\gamma(m)}{\gamma(m)+1}}_{H_\mathrm{W}(m)}\underbrace{\frac{\mathbf{g}_\mathrm{d}^{\mathrm{H}}\mathbf{\Phi}^{-1}(m)}{\mathbf{g}_\mathrm{d}^{\mathrm{H}}\mathbf{\Phi}^{-1}(m)\mathbf{g}_\mathrm{d}}}_{\mathbf{h}_{\mathrm{MVDR}}^{\mathrm{H}}(m)}\mathbf{y}(m), \quad (19)$$

where $\mathbf{\Phi}(m) = \phi_R(m)\,\mathbf{\Gamma} + \mathbf{\Phi}_\mathbf{v}$ and $\gamma(m)$ denotes the a priori speech to reverberation and noise ratio at the output of the MVDR-BF. The vector $\mathbf{h}_{\mathrm{MVDR}}(m)$ is the MVDR-BF that reduces the noise and reverberation while maintaining the direct speech undistorted and $H_\mathrm{W}(m)$ is the single-channel Wiener filter that is applied to the output of the MVDR-BF. $\gamma(m)$ is defined as $\gamma(m) = \frac{\phi_S(m)}{\phi_{RE}(m)}$ where $\phi_{RE}(m) = [\mathbf{g}_\mathrm{d}^{\mathrm{H}}\mathbf{\Phi}^{-1}(m)\mathbf{g}_\mathrm{d}]^{-1}$ denotes the residual reverberation plus noise at the output of the MVDR-BF.

In [5] (see also [3]), $\gamma(m)$ was calculated as a weighted average of a long-term estimator, obtained in the previous time-frame, and an instantaneous estimate from the current time-frame, a method known as the decision-directed approach:

$$\widehat{\gamma}(m) = \beta\frac{|\widehat{S}(m-1)|^2}{\phi_{RE}(m-1)} + (1-\beta)\frac{\widehat{\phi}_S(m)}{\phi_{RE}(m)}. \quad (20)$$

The instantaneous speech PSD is given by:

$$\widehat{\phi}_S(m) = \max\left(\left|\mathbf{h}_{\mathrm{MVDR}}^{\mathrm{H}}(m)\mathbf{y}(m)\right|^2 - \phi_{RE}(m), 0\right). \quad (21)$$

Accordingly, two alternative implementations of the decision-directed approach in (20) can be proposed: 1) estimate $\widehat{\phi}_S(m)$ and $\widehat{\phi}_R(m)$ using (16) and (17) and then compute (20); or 2) estimate $\widehat{\phi}_R(m)$ using (17), estimate $\widehat{\phi}_S(m)$ using (21) and then compute (20). In our experiments we examine the two alternatives.

V. Performance Evaluation

The performance of the proposed estimator is evaluated by: 1) examining the log-error between the estimated values of $\phi_S(m)$ and $\phi_R(m)$ versus the true speech and reverberation levels, obtained by convolving the speech signal by the direct component and by the late component of the acoustic impulse response, respectively; and 2) utilizing the estimated PSDs $\widehat{\phi}_R(m)$ and $\widehat{\phi}_S(m)$ in a speech dereverberation and noise reduction task as explained in Sec. IV.

A. Simulation setup

The experiments consist of reverberant signals plus white (sensor) noise with various SNR levels. Anechoic speech signals were convolved with room impulse responses (RIRs), downloaded from an open-source database collected in our lab. Details about the database and RIR identification method can be found in [17]. Reverberation time was set by adjusting the room panels, and was measured to be approximately $T_{60} = 0.61$ s. The spatial PSD matrix $\mathbf{\Phi_v}$ 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). The smoothing parameter was set to $\alpha = 0.7$ and $\epsilon = 10^{-10}$. All measures were computed by averaging the results obtained using 50 sentences, 4–8 s long, evenly distributed between female and male speakers.

B. Accuracy of the proposed estimator

The performance of the proposed estimator (16)-(17) compared to two existing estimators in terms of log-error between the estimated PSDs and the oracle PSDs: 1) the reverberation PSD estimator in [14], denoted henceforth Braun2013; and 2) the speech and reverberation[1] PSD estimators in [12], denoted henceforth Lefkimmiatis2006. For each algorithm, an identical lower and upper delimitation and identical smoothing were carried out.

The mean log-errors between the estimated PSD levels and the oracle PSD levels are presented. In order to calculate the oracle PSD levels of the reverberation, the anechoic speech was filtered with the reverberation tails of the RIRs. In order to calculate the oracle PSD levels of the speech, the anechoic speech was filtered with the direct path 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[2] for several SNR levels are depicted in Fig. 1. The bars are split to distinguish between underestimation errors and overestimation errors.

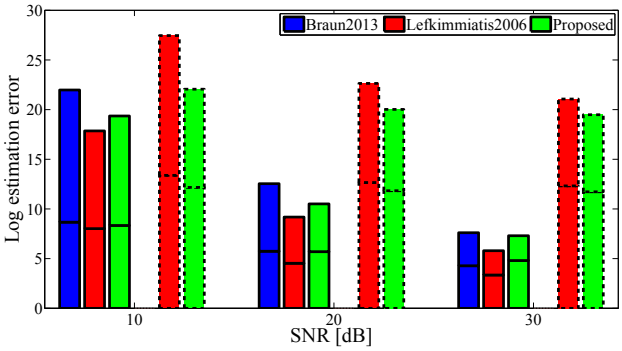

Fig. 1. Log-errors of the proposed reverberation PSD estimator compared with [14] and [12] (solid line bars) and log-errors of the proposed speech PSD estimator in comparison with [12] (dashed line bars). The upper part of each bar represents the underestimation error, while the lower part represents the overestimation error.

It is evident that the proposed speech PSD estimator outperforms Lefkimmiatis2006 [12] in terms of overall log-error for all the SNR values. For the proposed reverberation PSD estimator, the proposed estimator outperforms Braun2013 [14] for all the SNR values. However, Lefkimmiatis2006 [12] outperforms the competing estimators for all the SNR values. This result is reflected in the dereverberation performance in terms of perceptual evaluation of speech quality (PESQ) scores but not for the log-spectral distance (LSD) measure. In general, the log-error for the speech PSD estimator is higher than the log-error of the reverberation PSD.

C. Dereverberation performance

The performance of the proposed estimator is also examined by utilizing the estimated PSDs for the joint dereverberation and noise reduction task. The estimated PSDs were used to compute the MCWF presented in (19) (practical consideration can be found in [3]). All MCWF variants, listed bellow, use the decision-directed approach (20) in implementing the respective postfilter. The variants differ in the way the speech and reverberation PSDs are estimated. The weighting factor β was set to 0.9. The performance of the dereverberation algorithm was evaluated in terms of two objective measures, commonly used in the speech enhancement community, namely PESQ [18] and LSD. The clean reference for evaluation in all cases was the anechoic speech signal filtered with only the direct path of the RIR.

[1]Note that the original algorithm [12] assumes noisy environment (and no reverberation) and aims at estimating the noise PSD given the PSD matrix of the received signals, with the noise coherence matrix assumed to be known. In our implementation, the reverberation is treated in the same manner, namely its PSD should be estimated with the reverberation coherence matrix known (isotropic and diffused, in our case). The ambient noise PSD matrix (which is assumed to be known) is subtracted from the PSD matrix of the received signals and does not participate in the estimation procedure.

[2]The definition of the log-error can be found in [3].

	PESQ			LSD		
signal-to-noise ratio (SNR)	10 dB	20 dB	30 dB	10 dB	20 dB	30 dB
Unprocessed	1.57	1.86	1.94	11.14	7.86	6.04
Oracle $\phi_R(m)$ and $\phi_S(m)$	2.22	2.33	2.37	5.13	4.60	4.49
Oracle $\phi_R(m)$ and $\widehat{\phi}_S(m)$ using (21)	2.21	2.33	2.37	5.53	4.76	4.55
$\widehat{\phi}_R(m)$ using Braun2013 [14] and $\widehat{\phi}_S(m)$ using (21)	2.17	2.31	2.37	5.67	4.85	**4.63**
$\widehat{\phi}_R(m)$ and $\widehat{\phi}_S(m)$ using Lefkimmiatis2006 [12]	2.16	2.30	2.36	5.62	4.89	4.69
$\widehat{\phi}_R(m)$ using Lefkimmiatis2006 [12] and $\widehat{\phi}_S(m)$ using (21)	**2.20**	**2.34**	**2.39**	5.73	4.85	**4.63**
$\widehat{\phi}_R(m)$ and $\widehat{\phi}_S(m)$ using the proposed estimator in (17) and (16)	2.17	2.30	2.35	**5.50**	4.85	4.66
$\widehat{\phi}_R(m)$ using the proposed estimator in (17) and $\widehat{\phi}_S(m)$ using (21)	2.18	2.31	2.37	5.59	**4.83**	**4.63**

TABLE I
PESQ SCORES (LEFT) AND LSD RESULTS (RIGHT) FOR THE MCWF [3] USING VARIOUS ESTIMATORS.

The following MCWF variants were evaluated: 1) using the oracle speech and reverberation PSDs; 2) using the oracle reverberation PSD and the speech PSD estimate from (21); 3) using Braun2013's [14] reverberation PSD estimate and the speech PSD estimate from (21); 4) using Lefkimmiatis2006's reverberation and speech PSD estimates; 5) using Lefkimmiatis2006's [12] reverberation PSD estimate and the speech PSD estimate from (21); 6) using the proposed reverberation and speech PSDs estimates from (16) and (17); and 7) using the proposed reverberation PSD estimate from (17) and the speech PSD estimation from (21). In Table I 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, Lefkimmiatis2006 [12] using (21) outperforms all competing estimators.

VI. CONCLUSIONS

In this work a joint estimator for the late reverberant PSD and the anechoic speech PSD was derived based on the received signal PSD matrix. The proposed algorithm minimizes the Frobenius norm of the error between the measured PSD and its analytical model. The proposed estimation procedure, as opposed to the MLE procedures [8], [9], is closed-form and its computational load is lower. An experimental study compares the proposed PSDs estimators with other estimators when used in combination with a MCWF for joint noise reduction and dereverberation.

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," *Signals and Communication Technology*, 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] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

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

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

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

[8] 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.

[9] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.

[10] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE Int. Conf. Acoust. Speech and Sig. Proc. (ICASSP)*, NY, USA, Apr. 1988, pp. 2578–2581.

[11] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[12] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech applications," in *Proc. Interspeech-ICSLP*, 2006, pp. 2142–2145.

[13] H. Quang, D. Low, H. Dam, and S. Nordholm, "Space constrained beamforming with source psd updates," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, May 2004.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.