Scoring-Based ML Estimation and CRBs for Reverberation, Speech and Noise PSDs in a Spatially Homogeneous Noise-Field

Yaron Laufer, *Student Member, IEEE*, and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—Hands-free speech systems are subject to performance degradation due to reverberation and noise. Common methods for enhancing reverberant and noisy speech require the knowledge of the speech, reverberation and noise power spectral densities (PSDs). Most literature on this topic assumes that the noise PSD matrix is known. However, in many practical acoustic scenarios, the noise PSD is unknown and should be estimated along with the speech and the reverberation PSDs. In this paper, the noise is modelled as a spatially homogeneous sound field, with an unknown time-varying PSD multiplied by a known time-invariant spatial coherence matrix. We derive two maximum likelihood estimators (MLEs) for the various PSDs, including the noise: The first is a non-blocking-based estimator, that jointly estimates the PSDs of the speech, reverberation and noise components. The second MLE is a blocking-based estimator, that blocks the speech signal and estimates the reverberation and noise PSDs. Since a closed-form solution does not exist, both estimators iteratively maximize the likelihood using the Fisher scoring method. In order to compare both methods, the corresponding Cramér-Rao Bounds (CRBs) are derived. For both the reverberation and the noise PSDs, it is shown that the non-blocking-based CRB is lower than the blocking-based CRB. Performance evaluation using both simulated and real reverberant and noisy signals, shows that the proposed estimators outperform competing estimators, and greatly reduce the effect of reverberation and noise.**

***Index Terms*—Cramér-Rao Bound, Dereverberation, Maximum likelihood estimation, Noise reduction.**

I. Introduction

Real-life audio signals typically suffer from environmental noise, which may degrade the quality of speech. Apart from noise, another source of speech quality degradation is reverberation, caused by multiple reflections on the room facets and other objects in the enclosure. The presence of both reverberation and noise can significantly degrade the quality of the speech signal, and in severe case also its intelligibility [1], [2].

Speech enhancement algorithms reduce the effects of reverberation and noise, by extracting the clean speech signal from the noisy and reverberant measurements. A popular method is the multichannel Wiener filter (MCWF) beamformer, which estimates the desired speech by minimizing the mean square error (MSE). The design of the MCWF requires several model parameters, e.g. the speech, reverberation and noise PSDs. In the multichannel framework, the late reverberant signal is commonly modelled as a spatially homogeneous and spherically isotropic sound field. More specifically, the reverberation PSD matrix is factorized as a time-invariant spherical diffuse spatial coherence matrix multiplied by an unknown time-varying PSD [3]–[12]. Along with the reverberation PSD, the speech PSD is also an unknown time-varying parameter. Methods for estimating these PSDs can be divided into two classes, i.e. the non-blocking-based method [5], [8], [10] and the blocking-based method [3], [4], [6], [7], [9], [11]. While the non-blocking-based method jointly estimates the reverberation and speech PSDs, the blocking-based approach first blocks the speech signal and then estimates the reverberation PSD alone. For both classes, the estimation procedure is carried out using either the maximum likelihood (ML) approach [3], [5], [7]–[9], [11] or the least-squares (LS) criterion, by minimizing the Frobenius norm of an error PSD matrix [4], [6], [10].

In the ML framework, a closed-form solution exists for the reverberant noiseless case, i.e. when a noise-free environment is assumed [5]. A similar estimator can be applied for the non-reverberant noisy scenario, assuming that the noise coherence matrix is known [3], [13]. However, in a reverberant and noisy environment, closed-form ML solution is not available, thus requiring iterative optimization techniques [7], [8], [11]. In [7], [8], the log-likelihood was maximized iteratively using Newton's method [14], based on the first- and second-order derivatives of the log-likelihood. It is well known that the convergence of Newton's method depends on the quality of the initialization [14]. Typically, it converges very fast near the maximum value. However, if the initial estimate is not close to the optimum point, it may converge slowly or even fail to converge [15]. This lack of stability can be attributed to the fact that the Hessian matrix is not necessarily a positive definite matrix, and may be singular and non-invertible. The Fisher scoring algorithm [16]–[18] replaces the Hessian matrix with the Fisher information matrix (FIM). The FIM is always a positive definite matrix, and thus the convergence process becomes more stable. Moreover, in certain models the FIM has closed-form expression and can be computed easily.

All of the aforementioned methods do not include an estimator for the noise PSD, where either a noise-free scenario is considered [5] or an estimate of the noise PSD matrix is assumed to be available [4], [6]–[12]. As long as the noise PSD matrix is time-invariant, it can be estimated during speech-absent segments by means of a voice activity detector (VAD). However, in many practical acoustic scenarios the noise PSD is time-varying, and thus has to be included in the estimation procedure.

Yaron Laufer and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: Yaron.Laufer@biu.ac.il, Sharon.Gannot@biu.ac.il).

A useful tool to assess the quality of an estimator is the Cramér-Rao Bound (CRB), which determines a lower bound on the variance of any unbiased estimator. A fundamental property of the maximum likelihood estimator (MLE) is the asymptotic optimality, i.e. it reaches the CRB when the sample size is large [19]. Several papers derived CRB expressions for estimating the PSDs. For the noiseless case, MLEs of the reverberation and speech PSDs were proposed in [5], and the corresponding CRBs were derived in [20]. In [13], ML-based CRBs were employed in setting up a Bayesian refinement step above the ML-based speech enhancement, in a non-reverberant noisy scenario. The reverberant and noisy scenario was addressed in [21], assuming a known noise PSD matrix. Therein, CRBs on the reverberation PSD were derived for both the blocking-based MLE of [7] and the non-blocking-based MLE of [8]. Comparing these bounds, it was shown that the non-blocking-based approach yields lower CRB compared with the blocking-based method.

In contrast to previous works, in [22] the noise PSD matrix is not assumed to be known. The noise is modelled as a spatially homogeneous sound field, with the PSD matrix factorized as a time-invariant spatial coherence matrix multiplied by a time-varying PSD. The spatial coherence matrix of the noise is assumed to be known in advance, while the time-varying PSD is an unknown parameter that should be estimated. For instance, an air-conditioner with a thermostat control, which adjusts continuously its fan level. Based on the LS approach, two estimators were derived. The first is a non-blocking-based estimator that jointly estimates the reverberation, speech, and noise PSDs; and the second is a blocking-based estimator, that first blocks the speech component and then jointly estimates the noise and reverberation PSDs. However, it was stated in [11] that the ML approach produces superior PSD estimates compared to the LS method, due to the more accurate statistical model.

In our previous work [23], MLEs were derived for estimating the PSDs in the presence of rank-deficient noise field. Due to the special structure of the noise PSD matrix, a closed-form solution does exist, thus avoiding iterative techniques. The optimal strategy to estimate the reverberation level consists of generating nulls towards the speech and the directional noise sources, thus resembling the form of the the noiseless solution [5]. However, in the full-rank noise scenario similar procedure cannot be applied, and a closed-form solution is not available.

In the current contribution, we follow [22] and assume that the noise is modelled as a spatially homogeneous sound field, with an unknown time-varying PSD multiplied by a known time-invariant full-rank coherence matrix. Two novel MLEs are derived for estimating the speech, reverberation and noise PSDs, namely a non-blocking-based and a blocking-based estimators. In the absence of closed-form solutions, the scoring method is used to maximize iteratively the likelihood. In order to examine the quality of the proposed estimators, the corresponding CRBs are derived and analyzed.

To summarize, as compared to most previous studies in the field, we deal with a more general noisy and reverberant scenario, in which the noise PSD is unknown. Focusing on this scenario, the contribution of this paper is twofold. First, we derive two novel CRBs for estimating the noise, reverberation and speech PSDs, thus demonstrating the best achievable performance under the assumed data model. Second, we propose two novel ML estimators based on the scoring-based iterative algorithm, with enhanced robustness of convergence. The performance of the MLEs approaches the CRBs, thus yielding approximately optimal performance. The proposed estimators slightly outperform competing estimators, with only a moderate increase in the computational complexity.

The rest of the paper is organized as follows. Section II introduces the problem formulation, and presents the statistical model. In Section III, MLEs are derived for both the non-blocking-based and the blocking-based methods. The corresponding CRBs are derived in Section IV. Section V describes the experimental study, where the proposed estimators are evaluated using both simulated data and real audio signals. Section VI concludes the paper.

Notation

In our notation, scalars are denoted with regular lowercase letters, vectors are denoted with boldface lowercase letters and matrices are denoted with boldface uppercase letters. The superscripts $(\cdot)^\top$ and $(\cdot)^{\mathrm{H}}$ describe transposition and Hermitian transposition, respectively. The determinant of a matrix is denoted by $|\cdot|$, and the trace operator is denoted by $\mathrm{Tr}[\cdot]$.

II. Problem Formulation

A. Signal Model

A speech signal is received by N microphones, in a noisy and reverberant acoustic environment. We work in the short-time Fourier transform (STFT) domain, where $k \in [1, K]$ denote the frequency bin index, and $m \in [1, M]$ denote the time frame index. The N-channel measurement signal $\mathbf{y}(m,k) = [y_1(m,k), \cdots, y_N(m,k)]^\top$ is equal to

$$\mathbf{y}(m,k) = \mathbf{x}_{\mathrm{e}}(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k), \tag{1}$$

where $\mathbf{x}_{\mathrm{e}}(m,k)$ denotes the direct and early reverberation speech component, $\mathbf{r}(m,k) = [r_1(m,k), \cdots, r_N(m,k)]^\top$ denotes the late reverberation speech component and $\mathbf{v}(m,k) = [v_1(m,k), \cdots, v_N(m,k)]^\top$ denotes the noise. The direct and early reverberation speech component is given by $\mathbf{x}_{\mathrm{e}}(m,k) = \mathbf{g}_d(k)s(m,k)$, where $s(m,k)$ is the direct and early speech component, as received by the first microphone (assumed to be the reference microphone), and $\mathbf{g}_d(k) = [1, g_{d,2}(k), \cdots, g_{d,N}(k)]^\top$ is the time-invariant relative early transfer function (RETF) vector between the reference microphone and all microphones. In this paper, we follow previous works in the field, e.g. [4], [21], [22], [24], and neglect the early reflections. Thus, the target signal $s(m,k)$ is approximated as the direct component at the reference microphone, and $\mathbf{g}_d(k)$ reduces to the relative direct-path transfer function (RDTF) vector.

B. Probabilistic Model

The speech signal is assumed to follow a zero-mean complex Gaussian distribution with a time-varying PSD $\phi_S(m,k)$:

$$p\big(s(m,k); \phi_S(m,k)\big) = \mathcal{N}_c\big(s(m,k); 0, \phi_S(m,k)\big). \tag{2}$$

The late reverberation signal is modelled by a zero-mean complex multivariate Gaussian distribution:

$$p\big(\mathbf{r}(m,k);\boldsymbol{\Phi}_{\mathbf{r}}(m,k)\big)=\mathcal{N}_c\big(\mathbf{r}(m,k);\mathbf{0},\boldsymbol{\Phi}_{\mathbf{r}}(m,k)\big). \tag{3}$$

We assume that the reverberation PSD matrix can be modelled as a spatially homogeneous sound field with a time-varying PSD, $\boldsymbol{\Phi}_{\mathbf{r}}(m,k)=\phi_R(m,k)\boldsymbol{\Gamma}_R(k)$. The time-invariant coherence matrix $\boldsymbol{\Gamma}_R(k)$ is modelled as an ideal spherical diffuse sound field [25]:

$$\Gamma_{R,ij}(k)=\text{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{ij}}{c}\right), \tag{4}$$

where $\text{sinc}(x)=\sin(x)/x$, d_{ij} is the inter-distance between microphones i and j, f_s denotes the sampling frequency and c is the sound velocity. The noise signal is assumed to follow a zero-mean complex multivariate Gaussian distribution:

$$p\big(\mathbf{v}(m,k);\boldsymbol{\Phi}_{\mathbf{v}}(m,k)\big)=\mathcal{N}_c\big(\mathbf{v}(m,k);\mathbf{0},\boldsymbol{\Phi}_{\mathbf{v}}(m,k)\big). \tag{5}$$

The noise PSD matrix is modelled as a spatially homogeneous sound field, with a time-varying PSD multiplied by a time-invariant spatial coherence matrix, $\boldsymbol{\Phi}_{\mathbf{v}}(m,k)=\phi_V(m,k)\boldsymbol{\Gamma}_V(k)$. It is assumed that $\boldsymbol{\Gamma}_V$ is known. However, ϕ_V is an unknown parameter that should be estimated along with the speech and reverberation PSDs. Assuming that the various components in (1) are uncorrelated, the probability density function (PDF) of the measurement vector writes

$$p\big(\mathbf{y}(m,k);\boldsymbol{\Phi}_{\mathbf{y}}(m,k)\big)=\mathcal{N}_c\big(\mathbf{y}(m,k);\mathbf{0},\boldsymbol{\Phi}_{\mathbf{y}}(m,k)\big), \tag{6}$$

where

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{y}}(m,k)&=\phi_S(m,k)\mathbf{g}_d(k)\mathbf{g}_d^{\mathrm{H}}(k)+\phi_R(m,k)\,\boldsymbol{\Gamma}_R(k)\\&\quad+\phi_V(m,k)\,\boldsymbol{\Gamma}_V(k).\end{aligned} \tag{7}$$

A widely-used method for enhancing a reverberant and noisy speech is the MCWF, which produces an optimal speech estimator in the sense of minimizing the MSE [26]:

$$\hat{s}_{\text{MCWF}}(m,k)=\frac{\mathbf{g}_d^{\mathrm{H}}(k)\boldsymbol{\Phi}_i^{-1}(m,k)}{\mathbf{g}_d^{\mathrm{H}}(k)\boldsymbol{\Phi}_i^{-1}(m,k)\mathbf{g}_d(k)+\phi_S^{-1}(m,k)}\mathbf{y}(m,k), \tag{8}$$

where

$$\boldsymbol{\Phi}_i(m,k)\triangleq\phi_R(m,k)\,\boldsymbol{\Gamma}_R(k)+\phi_V(m,k)\,\boldsymbol{\Gamma}_V(k) \tag{9}$$

denotes the interference matrix. For implementing (8), it is required to estimate the time-varying PSDs of the speech, reverberation and noise components, namely ϕ_S, ϕ_R and ϕ_V. In the sequel, we derive MLEs for the various PSDs, assuming that the RDTF vector $\mathbf{g}_d$, the reverberation spatial coherence matrix $\boldsymbol{\Gamma}_R$ and the noise spatial coherence matrix $\boldsymbol{\Gamma}_V$ are known. Note that the RDTF depends only on the speaker direction of arrival (DOA) and the microphone array geometry, hence it can be constructed based on a DOA estimate. For the sake of brevity, the frame index m and the frequency bin index k are henceforth omitted whenever possible.

III. ML Estimators

In this section, two MLEs will be presented: (i) A non-blocking-based estimator, which simultaneously estimates the speech, reverberation and noise PSDs; and (ii) A blocking-based estimator, which first eliminates the speech signal and then jointly estimates the reverberation and noise PSDs.

A. Non-Blocking-Based Estimation

The set of unknown parameters is denoted by $\boldsymbol{\phi}(m)=\left[\phi_R(m),\phi_V(m),\phi_S(m)\right]^\top$. Let $\bar{\mathbf{y}}(m)$ denote a set of L successive i.i.d. snapshots of $\mathbf{y}(m)$:

$$\bar{\mathbf{y}}(m)\triangleq\left[\mathbf{y}^\top(m-L+1),\cdots,\mathbf{y}^\top(m)\right]^\top. \tag{10}$$

Using the short-time stationarity assumption [11], [20], we assume that the PSDs are approximately constant across the L segments. The PDF of $\bar{\mathbf{y}}$ writes (see e.g. [21]):

$$\begin{aligned}&p\big(\bar{\mathbf{y}}(m);\boldsymbol{\phi}(m)\big)\\&\quad=\left(\frac{1}{\pi^N|\boldsymbol{\Phi}_{\mathbf{y}}(m)|}\exp\Big(-\text{Tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{R}_{\mathbf{y}}(m)\right]\Big)\right)^L,\end{aligned} \tag{11}$$

where $\mathbf{R}_{\mathbf{y}}$ is the sample covariance matrix, given by

$$\mathbf{R}_{\mathbf{y}}(m)=\frac{1}{L}\sum_{\ell=m-L+1}^{m}\mathbf{y}(\ell)\mathbf{y}^{\mathrm{H}}(\ell). \tag{12}$$

The MLE of $\boldsymbol{\phi}(m)$ is obtained by maximizing the log likelihood

$$\boldsymbol{\phi}^{\text{ML},\bar{\mathbf{y}}}(m)=\underset{\boldsymbol{\phi}(m)}{\text{argmax}}\log p\big(\bar{\mathbf{y}}(m);\,\boldsymbol{\phi}(m)\big). \tag{13}$$

Using [19, Eqs. (15.47)–(15.48)], the derivative of the log-likelihood function w.r.t. the various PSDs writes

$$\begin{aligned}d_i^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}(m)\right)&\triangleq\frac{\partial\log p\left(\bar{\mathbf{y}}(m);\,\boldsymbol{\phi}(m)\right)}{\partial\phi_i(m)}\\&=L\,\text{Tr}\left[\left(\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{R}_{\mathbf{y}}(m)-\mathbf{I}\right)\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(m)\frac{\partial\boldsymbol{\Phi}_{\mathbf{y}}(m)}{\partial\phi_i(m)}\right],\end{aligned} \tag{14}$$

where $i\in\{R,V,S\}$, and

$$\frac{\partial\boldsymbol{\Phi}_{\mathbf{y}}(m)}{\partial\phi_R(m)}=\boldsymbol{\Gamma}_R\ ,\quad\frac{\partial\boldsymbol{\Phi}_{\mathbf{y}}(m)}{\partial\phi_V(m)}=\boldsymbol{\Gamma}_V\ ,\quad\frac{\partial\boldsymbol{\Phi}_{\mathbf{y}}(m)}{\partial\phi_S(m)}=\mathbf{g}_d\mathbf{g}_d^{\mathrm{H}}. \tag{15}$$

Setting (14) to zero and solving for $\boldsymbol{\phi}(m)$ yields the MLE. However, this is a nonlinear optimization problem, and a closed-form solution does not exist. We may therefore resort to a numerical evaluation of the MLE using an iterative maximization procedure, e.g. the Newton method [14] or the Fisher's scoring algorithm [16]. Using the iterative method of Newton, the search procedure is given by [14]

$$\boldsymbol{\phi}^{(j+1)}=\boldsymbol{\phi}^{(j)}-\left(\mathbf{H}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}^{(j)}\right)\right)^{-1}\mathbf{d}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}^{(j)}\right), \tag{16}$$

where $\mathbf{d}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}\right)\in\mathbb{R}^3$ and $\mathbf{H}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}\right)\in\mathbb{R}^{3\times3}$ denote, respectively, the gradient vector and the Hessian matrix of $\log p\left(\bar{\mathbf{y}};\boldsymbol{\phi}\right)$ w.r.t. $\boldsymbol{\phi}$. Formally, this writes

$$\mathbf{d}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}\right)=\frac{\partial\log p\big(\bar{\mathbf{y}};\boldsymbol{\phi}\big)}{\partial\boldsymbol{\phi}}\ \ ;\ \ \mathbf{H}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}\right)=\frac{\partial^2\log p\big(\bar{\mathbf{y}};\boldsymbol{\phi}\big)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^\top}. \tag{17}$$

The Hessian matrix is a function of both the parameters and the data. It is evident that when $\mathbf{H}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}\right)$ is close to be singular, the right hand term of (16) may wildly fluctuate from iteration to iteration and the search may converge slowly or even not converge. We therefore propose to use the Fisher scoring approach [16]. In this method, the Hessian matrix is replaced by its expected value, namely the negative FIM:

$$\boldsymbol{\phi}^{(j+1)} = \boldsymbol{\phi}^{(j)} + \left(\mathbf{I}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}^{(j)}\right)\right)^{-1} \mathbf{d}^{\bar{\mathbf{y}}}\left(\boldsymbol{\phi}^{(j)}\right), \tag{18}$$

where

$$\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi}) = -\mathbb{E}_{\bar{\mathbf{y}}|\boldsymbol{\phi}}\left[\frac{\partial^2 \log p\left(\bar{\mathbf{y}};\boldsymbol{\phi}\right)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^\top}\right]. \tag{19}$$

The stability of the iteration is increased, since the FIM is a positive definite matrix and hence invertible [19]. Note that the expectation is taken over the data given the parameters, and therefore the FIM is only a function of the parameters and is not sensitive to the data. The elements of $\mathbf{d}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})$ are given in (14), and an explicit expression for $\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})$ will be derived in Section IV-A. Note that to obtain $\boldsymbol{\phi}^{(j+1)}$, both $\mathbf{d}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})$ and $\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})$ in (18) are computed using $\boldsymbol{\phi}^{(j)}$, namely the value of $\boldsymbol{\phi}$ obtained in the previous iteration.

B. Blocking-Based Estimation

In the blocking-based approach, a blocking matrix (BM) is applied to the input vector in order to block the speech signal. Let $\mathbf{B} \in \mathbb{C}^{N\times(N-1)}$ denote the BM, satisfying $\mathbf{B}^{\mathrm{H}}\mathbf{g}_d = 0$. The output of the BM is given by

$$\mathbf{z}(m) \triangleq \mathbf{B}^{\mathrm{H}}\,\mathbf{y}(m) = \mathbf{B}^{\mathrm{H}}\big(\mathbf{r}(m) + \mathbf{v}(m)\big), \tag{20}$$

with the PDF

$$p\big(\mathbf{z}(m); \boldsymbol{\Phi}_{\mathbf{z}}(m)\big) = \mathcal{N}_c\big(\mathbf{z}(m); \mathbf{0}, \boldsymbol{\Phi}_{\mathbf{z}}(m)\big). \tag{21}$$

Using (7) and (9), the PSD matrix of the blocked signal writes

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{z}}(m) &= \mathbf{B}^{\mathrm{H}}\boldsymbol{\Phi}_i(m)\mathbf{B}\\ &= \phi_R(m)\,\mathbf{B}^{\mathrm{H}}\boldsymbol{\Gamma}_R\mathbf{B} + \phi_V(m)\,\mathbf{B}^{\mathrm{H}}\boldsymbol{\Gamma}_V\mathbf{B}.\end{aligned} \tag{22}$$

Under this model, the parameter set of interest is $\tilde{\boldsymbol{\phi}}(m) = [\phi_R(m), \phi_V(m)]^\top$. Let $\bar{\mathbf{z}}$ be defined similarly to $\bar{\mathbf{y}}$ in (10), as a concatenation of L i.i.d. consecutive frames. Similarly to $\boldsymbol{\Phi}_{\mathbf{y}}(m)$, we assume that $\boldsymbol{\Phi}_{\mathbf{z}}(m)$ is fixed over the entire segment. The PDF of $\bar{\mathbf{z}}$ therefore writes

$$\begin{aligned}&p\big(\bar{\mathbf{z}}(m); \tilde{\boldsymbol{\phi}}(m)\big)\\ &\quad= \left(\frac{1}{\pi^{N-1}|\boldsymbol{\Phi}_{\mathbf{z}}(m)|}\exp\left(-\operatorname{Tr}\left[\boldsymbol{\Phi}_{\mathbf{z}}^{-1}(m)\mathbf{R}_{\mathbf{z}}(m)\right]\right)\right)^L\end{aligned} \tag{23}$$

where $\mathbf{R}_{\mathbf{z}}(m)$ is defined similarly to (12). The MLE of $\tilde{\boldsymbol{\phi}}(m)$ is obtained by solving:

$$\tilde{\boldsymbol{\phi}}^{\mathrm{ML},\bar{\mathbf{z}}}(m) = \underset{\tilde{\boldsymbol{\phi}}(m)}{\operatorname{argmax}} \log p\big(\bar{\mathbf{z}}(m);\, \tilde{\boldsymbol{\phi}}(m)\big). \tag{24}$$

To the best of our knowledge, this problem does not have a closed-form solution. Again, we use the scoring method for iterative maximization of the log-likelihood:

$$\tilde{\boldsymbol{\phi}}^{(j+1)} = \tilde{\boldsymbol{\phi}}^{(j)} + \left(\mathbf{I}^{\bar{\mathbf{z}}}\left(\tilde{\boldsymbol{\phi}}^{(j)}\right)\right)^{-1}\mathbf{d}^{\bar{\mathbf{z}}}\left(\tilde{\boldsymbol{\phi}}^{(j)}\right), \tag{25}$$

where $\mathbf{d}^{\bar{\mathbf{z}}}\left(\tilde{\boldsymbol{\phi}}\right) \in \mathbb{R}^2$ and $\mathbf{I}^{\bar{\mathbf{z}}}\left(\tilde{\boldsymbol{\phi}}\right) \in \mathbb{R}^{2\times 2}$ denote, respectively, the gradient vector and the FIM of $\log p\left(\bar{\mathbf{z}};\, \tilde{\boldsymbol{\phi}}\right)$ w.r.t. $\tilde{\boldsymbol{\phi}}$. The elements of $\mathbf{d}^{\bar{\mathbf{z}}}(\boldsymbol{\phi})$ are given by

$$d_i^{\bar{\mathbf{z}}}\left(\tilde{\boldsymbol{\phi}}\right) = L\operatorname{Tr}\left[\left(\boldsymbol{\Phi}_{\mathbf{z}}^{-1}(m)\mathbf{R}_{\mathbf{z}}(m) - \mathbf{I}\right)\boldsymbol{\Phi}_{\mathbf{z}}^{-1}(m)\frac{\partial\boldsymbol{\Phi}_{\mathbf{z}}(m)}{\partial\tilde{\phi}_i(m)}\right], \tag{26}$$

where $i \in \{R, V\}$, and

$$\frac{\partial\boldsymbol{\Phi}_{\mathbf{z}}(m)}{\partial\phi_R(m)} = \mathbf{B}^{\mathrm{H}}\boldsymbol{\Gamma}_R\mathbf{B}\ ,\ \frac{\partial\boldsymbol{\Phi}_{\mathbf{z}}(m)}{\partial\phi_V(m)} = \mathbf{B}^{\mathrm{H}}\boldsymbol{\Gamma}_V\mathbf{B}. \tag{27}$$

The FIM $\mathbf{I}^{\bar{\mathbf{z}}}\left(\tilde{\boldsymbol{\phi}}\right)$ will be derived explicitly in Section IV-B.

In Section V-B, the MCWF will be used for enhancing a reverberant and noisy speech. The implementation of the MCWF requires (among others) an estimate of the speech PSD (see (8)), which is missing in the blocking-based framework. By substituting the obtained blocking-based reverberation and noise estimates, namely $\phi_R^{\mathrm{ML},\bar{\mathbf{z}}}$ and $\phi_V^{\mathrm{ML},\bar{\mathbf{z}}}$, into the general likelihood function in (11), the maximization becomes a one-dimensional optimization problem, and a closed-form solution is available [11], [27]:

$$\begin{aligned}\phi_S^{\mathrm{ML},\bar{\mathbf{z}}}(m) = \mathbf{w}_{\mathrm{MVDR}}^{\mathrm{H}}(m)\Big(&\mathbf{R}_{\mathbf{y}}(m) - \phi_R^{\mathrm{ML},\bar{\mathbf{z}}}(m)\boldsymbol{\Gamma}_R\\ &- \phi_V^{\mathrm{ML},\bar{\mathbf{z}}}(m)\boldsymbol{\Gamma}_V\Big)\mathbf{w}_{\mathrm{MVDR}}(m),\end{aligned} \tag{28}$$

where

$$\mathbf{w}_{\mathrm{MVDR}}(m) = \frac{\hat{\boldsymbol{\Phi}}_i^{-1}(m)\mathbf{g}_d}{\mathbf{g}_d^{\mathrm{H}}\hat{\boldsymbol{\Phi}}_i^{-1}(m)\mathbf{g}_d}. \tag{29}$$

IV. CRB Derivation

In this section, CRBs are derived for both the non-blocking-based and the blocking-based estimators of Section III.

A. CRB for the Non-Blocking-Based Estimation

Under the non-blocking-based method, the Fisher information matrix (FIM) of $\boldsymbol{\phi}$ writes

$$\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi}) = \begin{pmatrix} \mathrm{I}_{RR}^{\bar{\mathbf{y}}} & \mathrm{I}_{RV}^{\bar{\mathbf{y}}} & \mathrm{I}_{RS}^{\bar{\mathbf{y}}}\\ \mathrm{I}_{VR}^{\bar{\mathbf{y}}} & \mathrm{I}_{VV}^{\bar{\mathbf{y}}} & \mathrm{I}_{VS}^{\bar{\mathbf{y}}}\\ \mathrm{I}_{SR}^{\bar{\mathbf{y}}} & \mathrm{I}_{SV}^{\bar{\mathbf{y}}} & \mathrm{I}_{SS}^{\bar{\mathbf{y}}}\end{pmatrix}, \tag{30}$$

where $\mathrm{I}_{ij}^{\bar{\mathbf{y}}} \triangleq [\mathrm{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})]_{ij} = -\mathbb{E}\left[\frac{\partial^2\log p\left(\bar{\mathbf{y}};\boldsymbol{\phi}\right)}{\partial\phi_i\partial\phi_j}\right]$ and $i, j \in \{R, V, S\}$. Since $\mathbf{y}(m) \sim \mathcal{N}_c\big(\mathbf{0}, \boldsymbol{\Phi}_{\mathbf{y}}(m)\big)$, the FIM elements are given by [28], [29]:

$$\left[\mathrm{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})\right]_{ij} = L\operatorname{Tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\frac{\partial\boldsymbol{\Phi}_{\mathbf{y}}}{\partial\phi_i}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\frac{\partial\boldsymbol{\Phi}_{\mathbf{y}}}{\partial\phi_j}\right]. \tag{31}$$

Upon inverting the FIM in (30) we have that

$$\mathrm{CRB}_{\phi_R}^{\bar{\mathbf{y}}} = \left[\left(\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})\right)^{-1}\right]_{11} = \frac{\mathrm{I}_{VV}^{\bar{\mathbf{y}}}\mathrm{I}_{SS}^{\bar{\mathbf{y}}} - \mathrm{I}_{VS}^{\bar{\mathbf{y}}}\mathrm{I}_{SV}^{\bar{\mathbf{y}}}}{\mathrm{I}_\Delta}, \tag{32a}$$

$$\mathrm{CRB}_{\phi_V}^{\bar{\mathbf{y}}} = \left[\left(\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})\right)^{-1}\right]_{22} = \frac{\mathrm{I}_{RR}^{\bar{\mathbf{y}}}\mathrm{I}_{SS}^{\bar{\mathbf{y}}} - \mathrm{I}_{RS}^{\bar{\mathbf{y}}}\mathrm{I}_{SR}^{\bar{\mathbf{y}}}}{\mathrm{I}_\Delta}, \tag{32b}$$

$$\mathrm{CRB}_{\phi_S}^{\bar{\mathbf{y}}} = \left[\left(\mathbf{I}^{\bar{\mathbf{y}}}(\boldsymbol{\phi})\right)^{-1}\right]_{33} = \frac{\mathrm{I}_{RR}^{\bar{\mathbf{y}}}\mathrm{I}_{VV}^{\bar{\mathbf{y}}} - \mathrm{I}_{RV}^{\bar{\mathbf{y}}}\mathrm{I}_{VR}^{\bar{\mathbf{y}}}}{\mathrm{I}_\Delta}, \tag{32c}$$

where

$$\begin{aligned}\mathrm{I}_\Delta = \mathrm{I}_{RR}^{\bar{\mathbf{y}}}\left(\mathrm{I}_{VV}^{\bar{\mathbf{y}}}\mathrm{I}_{SS}^{\bar{\mathbf{y}}} - \mathrm{I}_{VS}^{\bar{\mathbf{y}}}\mathrm{I}_{SV}^{\bar{\mathbf{y}}}\right) - \mathrm{I}_{RV}^{\bar{\mathbf{y}}}\left(\mathrm{I}_{VR}^{\bar{\mathbf{y}}}\mathrm{I}_{SS}^{\bar{\mathbf{y}}} - \mathrm{I}_{VS}^{\bar{\mathbf{y}}}\mathrm{I}_{SR}^{\bar{\mathbf{y}}}\right)\\ + \mathrm{I}_{RS}^{\bar{\mathbf{y}}}\left(\mathrm{I}_{VR}^{\bar{\mathbf{y}}}\mathrm{I}_{SV}^{\bar{\mathbf{y}}} - \mathrm{I}_{VV}^{\bar{\mathbf{y}}}\mathrm{I}_{SR}^{\bar{\mathbf{y}}}\right).\end{aligned} \tag{33}$$

Using (31) and (15), we have

$$
\begin{aligned}
\mathrm{I}^{\bar{\mathbf{y}}}_{RR} &= L\ \mathrm{Tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_R\right], && \text{(34a)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{RV} &= I^{\bar{\mathbf{y}}}_{VR} = L\ \mathrm{Tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_V\right], && \text{(34b)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{RS} &= I^{\bar{\mathbf{y}}}_{SR} = L\ \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\mathbf{g}_d, && \text{(34c)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{VV} &= L\ \mathrm{Tr}\left[\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_V\right], && \text{(34d)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{VS} &= I^{\bar{\mathbf{y}}}_{SV} = L\ \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\mathbf{g}_d, && \text{(34e)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{SS} &= L\left(\mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_{\mathbf{y}}^{-1}\mathbf{g}_d\right)^2. && \text{(34f)}
\end{aligned}
$$

To simplify (34), $\boldsymbol{\Phi}_{\mathbf{y}}$ can be recast as $\boldsymbol{\Phi}_{\mathbf{y}} = \phi_S\mathbf{g}_d\mathbf{g}_d^{\mathrm{H}} + \boldsymbol{\Phi}_i$ (see (7) and (9)), and then $\boldsymbol{\Phi}_{\mathbf{y}}^{-1}$ is obtained by the Woodbury matrix inversion identity [30]. Substituting into (34), yields

$$
\begin{aligned}
\mathrm{I}^{\bar{\mathbf{y}}}_{RR} &= L\left(\gamma_0 - \frac{2\gamma_3}{\gamma_1+\phi_S^{-1}} + \frac{\gamma_2^2}{\left(\gamma_1+\phi_S^{-1}\right)^2}\right), && \text{(35a)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{RV} &= L\left(\tilde{\gamma}_0 - \frac{2\tilde{\gamma}_3}{\gamma_1+\phi_S^{-1}} + \frac{\gamma_2\hat{\gamma}_2}{\left(\gamma_1+\phi_S^{-1}\right)^2}\right), && \text{(35b)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{RS} &= \frac{L\gamma_2}{\left(1+\gamma_1\phi_S\right)^2}, && \text{(35c)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{VV} &= L\left(\hat{\gamma}_0 - \frac{2\hat{\gamma}_3}{\gamma_1+\phi_S^{-1}} + \frac{\hat{\gamma}_2^2}{\left(\gamma_1+\phi_S^{-1}\right)^2}\right), && \text{(35d)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{VS} &= \frac{L\hat{\gamma}_2}{\left(1+\gamma_1\phi_S\right)^2}, && \text{(35e)}\\
\mathrm{I}^{\bar{\mathbf{y}}}_{SS} &= \frac{L\gamma_1^2}{\left(1+\gamma_1\phi_S\right)^2}, && \text{(35f)}
\end{aligned}
$$

where, in a similiar manner to [21], the following auxiliary variables are introduced:

$$
\begin{aligned}
\gamma_0 &= \mathrm{Tr}\left[\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\right], && \text{(36a)}\\
\gamma_1 &= \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_i^{-1}\mathbf{g}_d, && \text{(36b)}\\
\gamma_2 &= \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_i^{-1}\mathbf{g}_d, && \text{(36c)}\\
\gamma_3 &= \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_i^{-1}\mathbf{g}_d, && \text{(36d)}\\
\hat{\gamma}_0 &= \mathrm{Tr}\left[\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\right], && \text{(36e)}\\
\hat{\gamma}_2 &= \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\mathbf{g}_d, && \text{(36f)}\\
\hat{\gamma}_3 &= \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\mathbf{g}_d, && \text{(36g)}\\
\tilde{\gamma}_0 &= \mathrm{Tr}\left[\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\right], && \text{(36h)}\\
\tilde{\gamma}_3 &= \mathbf{g}_d^{\mathrm{H}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R\boldsymbol{\Phi}_i^{-1}\mathbf{g}_d. && \text{(36i)}
\end{aligned}
$$

In the derivation of (35b), we used the fact that $\boldsymbol{\Gamma}_V\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_R = \boldsymbol{\Gamma}_R\boldsymbol{\Phi}_i^{-1}\boldsymbol{\Gamma}_V$, which can be verified using (9). Note that all quantities defined in (36) are independent of ϕ_S. Substituting (35) into (32) yields the CRBs, given in (36)–(38) at the bottom of the page, where

$$
\begin{aligned}
\delta &= \gamma_1\gamma_3 - \gamma_2^2, && \text{(40a)}\\
\hat{\delta} &= \gamma_1\hat{\gamma}_3 - \hat{\gamma}_2^2, && \text{(40b)}\\
\tilde{\delta} &= \gamma_1\tilde{\gamma}_3 - \gamma_2\hat{\gamma}_2. && \text{(40c)}
\end{aligned}
$$

It should be noted that the reverberation CRB in (36) resembles the one derived in [21, Eq. (39)] for the case of known noise PSD, except an additional term in the denominator.

B. CRB for the Blocking-Based Estimation

Under the blocking-based method, the FIM of $\tilde{\phi}$ writes

$$
\mathbf{I}^{\bar{\mathbf{z}}}\left(\tilde{\phi}\right) = \begin{pmatrix} \mathrm{I}^{\bar{\mathbf{z}}}_{RR} & \mathrm{I}^{\bar{\mathbf{z}}}_{RV} \\ \mathrm{I}^{\bar{\mathbf{z}}}_{VR} & \mathrm{I}^{\bar{\mathbf{z}}}_{VV} \end{pmatrix}. \tag{41}
$$

We have upon inversion that

$$
\begin{aligned}
\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R} &= \left[\left(\mathbf{I}^{\bar{\mathbf{z}}}\left(\tilde{\phi}\right)\right)^{-1}\right]_{11} = \frac{\mathrm{I}^{\bar{\mathbf{z}}}_{VV}}{\mathrm{I}^{\bar{\mathbf{z}}}_{RR}\mathrm{I}^{\bar{\mathbf{z}}}_{VV} - \mathrm{I}^{\bar{\mathbf{z}}}_{RV}\mathrm{I}^{\bar{\mathbf{z}}}_{VR}}, && \text{(42a)}\\
\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_V} &= \left[\left(\mathbf{I}^{\bar{\mathbf{z}}}\left(\tilde{\phi}\right)\right)^{-1}\right]_{22} = \frac{\mathrm{I}^{\bar{\mathbf{z}}}_{RR}}{\mathrm{I}^{\bar{\mathbf{z}}}_{RR}\mathrm{I}^{\bar{\mathbf{z}}}_{VV} - \mathrm{I}^{\bar{\mathbf{z}}}_{RV}\mathrm{I}^{\bar{\mathbf{z}}}_{VR}}. && \text{(42b)}
\end{aligned}
$$

$$
\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}\left(\phi_S\right) = \frac{1}{L}\,\frac{1}{\gamma_0 - 2\frac{\gamma_3}{\gamma_1} + \frac{\gamma_2^2}{\gamma_1^2} + \left(\frac{2\delta}{\gamma_1^2(1+\gamma_1\phi_S)} - \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2} + \frac{2\tilde{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} + \frac{2\hat{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}}\right)} \tag{36}
$$

$$
\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_V}\left(\phi_S\right) = \frac{1}{L}\,\frac{1}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} + \left(\frac{2\hat{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)} - \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2} + \frac{2\tilde{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}\right)^2}{\gamma_0 - 2\frac{\gamma_3}{\gamma_1} + \frac{\gamma_2^2}{\gamma_1^2} + \frac{2\delta}{\gamma_1^2(1+\gamma_1\phi_S)}}\right)} \tag{37}
$$

$$
\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_S} = \frac{\left(1+\gamma_1\phi_S\right)^2}{L\gamma_1^2}\left(1 + \frac{1}{\gamma_1^2\frac{\left(\gamma_0\hat{\gamma}_0 - \tilde{\gamma}_0^2\right)(1+\gamma_1\phi_S)^3 - (2\gamma_3\hat{\gamma}_0 + 2\gamma_0\hat{\gamma}_3 - 4\tilde{\gamma}_0\tilde{\gamma}_3)(1+\gamma_1\phi_S)^2\phi_S + \left(4\gamma_3\hat{\gamma}_3 - 4\tilde{\gamma}_3^2\right)(1+\gamma_1\phi_S)\phi_S^2}{\left(\gamma_0\hat{\gamma}_2^2 + \gamma_2^2\hat{\gamma}_0 - 2\gamma_2\hat{\gamma}_2\tilde{\gamma}_0\right)(1+\gamma_1\phi_S) - \left(2\gamma_3\hat{\gamma}_2^2 + 2\gamma_2^2\hat{\gamma}_3 - 4\gamma_2\hat{\gamma}_2\tilde{\gamma}_3\right)\phi_S} - \left(1-\gamma_1^2\phi_S^2\right)}\right) \tag{38}
$$

Since $\mathbf{z}(m) \sim \mathcal{N}_c\big(\mathbf{0}, \mathbf{\Phi}_{\mathbf{z}}(m)\big)$, the FIM writes [28], [29]:

$$\left[\mathrm{I}^{\bar{\mathbf{z}}}\left(\tilde{\phi}\right)\right]_{ij} = L \;\mathrm{Tr}\left[\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\frac{\partial \mathbf{\Phi}_{\bar{\mathbf{z}}}}{\partial \tilde{\phi}_i}\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\frac{\partial \mathbf{\Phi}_{\bar{\mathbf{z}}}}{\partial \tilde{\phi}_j}\right], \quad i,j \in \{R,V\}\,. \tag{43}$$

Using (43) and (27), we have

$$\mathrm{I}^{\bar{\mathbf{z}}}_{RR} = L \;\mathrm{Tr}\left[\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_R\mathbf{B}\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_R\mathbf{B}\right], \tag{44a}$$

$$\mathrm{I}^{\bar{\mathbf{z}}}_{RV} = I^{\bar{\mathbf{z}}}_{VR} = L \;\mathrm{Tr}\left[\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_R\mathbf{B}\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_V\mathbf{B}\right], \tag{44b}$$

$$\mathrm{I}^{\bar{\mathbf{z}}}_{VV} = L \;\mathrm{Tr}\left[\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_V\mathbf{B}\mathbf{\Phi}_{\bar{\mathbf{z}}}^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{\Gamma}_V\mathbf{B}\right]. \tag{44c}$$

Using (22) and following the identity from [21, Eq. (45)]:

$$\mathbf{B}\left(\mathbf{B}^{\mathrm{H}}\mathbf{\Phi}_i\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{H}} = \mathbf{\Phi}_i^{-1} - \frac{\mathbf{\Phi}_i^{-1}\mathbf{g}_d\mathbf{g}_d^{\mathrm{H}}\mathbf{\Phi}_i^{-1}}{\mathbf{g}_d^{\mathrm{H}}\mathbf{\Phi}_i^{-1}\mathbf{g}_d}, \tag{45}$$

the FIM elements of (44) can be written as

$$\mathrm{I}^{\bar{\mathbf{z}}}_{RR} = L\left(\gamma_0 - 2\frac{\gamma_3}{\gamma_1} + \frac{\gamma_2^2}{\gamma_1^2}\right), \tag{46a}$$

$$\mathrm{I}^{\bar{\mathbf{z}}}_{RV} = L\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2}\right), \tag{46b}$$

$$\mathrm{I}^{\bar{\mathbf{z}}}_{VV} = L\left(\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2}\right). \tag{46c}$$

Substituting (46) into (42) yields the CRBs:

$$\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R} = \frac{1}{L}\frac{1}{\gamma_0 - 2\frac{\gamma_3}{\gamma_1} + \frac{\gamma_2^2}{\gamma_1^2} - \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2}}}, \tag{47}$$

$$\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_V} = \frac{1}{L}\frac{1}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} - \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2}\right)^2}{\gamma_0 - 2\frac{\gamma_3}{\gamma_1} + \frac{\gamma_2^2}{\gamma_1^2}}}. \tag{48}$$

Again, the reverberation CRB in (47) is similar to the one derived in [21, Eq. (48)] for the case of known noise PSD, except an extra term in the denominator.

C. Comparing the CRBs

In this section, we compare the non-blocking-based and blocking-based CRBs derived in the previous sections.

1) Comparing the reverberation CRBs: First, the reverberation CRBs in (36) and (47) are compared. Both expressions are identical except the last terms in the denominator. In the sequel, these terms will be carefully examined.

For the sake of convenience, let us denote the last term in the denominator of $\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S)$ by

$$\alpha_{\bar{\mathbf{y}}}(\phi_S) \triangleq \frac{2\delta}{\gamma_1^2(1+\gamma_1\phi_S)} - \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2} + \frac{2\tilde{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} + \frac{2\hat{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}}. \tag{49}$$

The last term in the denominator of $\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R}$ is denoted by

$$\alpha_{\bar{\mathbf{z}}} \triangleq -\frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2}}. \tag{50}$$

Obviously, $\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R}$ is independent of ϕ_S. However, for the sake of comparison, the behaviour of $\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S)$ can be analyzed as a function of ϕ_S [21]. Specifically, we consider the following three possible regions of ϕ_S:

1) When $\phi_S = 0$, $\alpha_{\bar{\mathbf{y}}}(\phi_S)$ is reduced to

$$\alpha_{\bar{\mathbf{y}}}(\phi_S = 0) = \frac{2\delta}{\gamma_1^2} - \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2} + \frac{2\tilde{\delta}}{\gamma_1^2}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} + \frac{2\hat{\delta}}{\gamma_1^2}}. \tag{51}$$

In Appendix A, it shown that $\hat{\delta} \geq 0$ and $\tilde{\delta} \leq 0$. It follows that

$$\frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2} + \frac{2\tilde{\delta}}{\gamma_1^2}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} + \frac{2\hat{\delta}}{\gamma_1^2}} \leq \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2}}. \tag{52}$$

Using (51), (52) and the fact that $\delta \geq 0$ (see Appendix A), it follows that $\alpha_{\bar{\mathbf{y}}}(\phi_S = 0) \geq \alpha_{\bar{\mathbf{z}}}$ and thus

$$\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S = 0) \leq \mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R}, \tag{53}$$

with equality if and only if $\delta = \hat{\delta} = \tilde{\delta} = 0$. As shown in Appendix A, this condition is satisfied only when $\phi_R = \phi_V = 0$ or $\mathbf{\Gamma}_R = \mathbf{\Gamma}_V$. It should be noted that in the latter case, i.e. when the spatial fields of the reverberation and noise are identical, the problem essentially reduces to the noiseless scenario, since only two PSDs have to be estimated: the speech PSD and the combined reverberation-plus-noise PSD. In this case, the CRB becomes simpler and identical for both methods [20], [21].

2) The behaviour in the range $0 < \phi_S < \infty$: Let the two terms of $\alpha_{\bar{\mathbf{y}}}(\phi_S)$ be denoted by $\eta_1 \triangleq \frac{2\delta}{\gamma_1^2(1+\gamma_1\phi_S)}$ and $\eta_2 \triangleq \frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2} + \frac{2\tilde{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2} + \frac{2\hat{\delta}}{\gamma_1^2(1+\gamma_1\phi_S)}}$, s.t. $\alpha_{\bar{\mathbf{y}}} = \eta_1 - \eta_2$. To analyze the trend of η_1 versus ϕ_S, note that $\delta \geq 0$ (see Appendix A) and γ_1 is positive, since $\mathbf{\Phi}_i$ is a positive definite matrix (see (36b)). Accordingly, η_1 is non-negative and therefore a monotonically decreasing function of ϕ_S. Using the fact that $\hat{\delta} \geq 0$ and $\tilde{\delta} \leq 0$, η_2 is monotonically increasing with ϕ_S. It follows that $\alpha_{\bar{\mathbf{y}}}(\phi_S)$ is monotonically decreasing in ϕ_S, and thus $\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S)$ is a monotonic increasing function of ϕ_S.

3) When ϕ_S goes to infinity, $\alpha_{\bar{\mathbf{y}}}(\phi_S)$ approaches $\alpha_{\bar{\mathbf{z}}}$:

$$\lim_{\phi_S\to\infty}\alpha_{\bar{\mathbf{y}}}(\phi_S) = -\frac{\left(\tilde{\gamma}_0 - 2\frac{\tilde{\gamma}_3}{\gamma_1} + \frac{\gamma_2\hat{\gamma}_2}{\gamma_1^2}\right)^2}{\hat{\gamma}_0 - 2\frac{\hat{\gamma}_3}{\gamma_1} + \frac{\hat{\gamma}_2^2}{\gamma_1^2}} = \alpha_{\bar{\mathbf{z}}}, \tag{54}$$

and thus

$$\lim_{\phi_S\to\infty}\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S) = \mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R}. \tag{55}$$

Interestingly, as $\phi_S \to \infty$ the non-blocking-based reverberation CRB becomes independent of ϕ_S, and approaches the blocking-based CRB.

We have shown that $\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S)$ is smaller than $\mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R}$ at $\phi_S = 0$, and that both CRBs coincide as $\phi_S \to \infty$. By the monotonicity of $\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S)$ in ϕ_S, we conclude that

$$\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_R}(\phi_S) \leq \mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_R}. \tag{56}$$

2) Comparing the noise CRBs: Using a similar set of arguments for comparing the noise CRBs in (37) and (48), it can be deduced that

$$\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_V}(\phi_S) \leq \mathrm{CRB}^{\bar{\mathbf{z}}}_{\phi_V}. \tag{57}$$

D. Analysis of the Speech CRB

In this section, we present the two limiting cases of the speech CRB in (38). When $\phi_S = 0$, the speech CRB reduces to

$$\mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_S}\Big|_{\phi_S=0} = \frac{1}{L\gamma_1^2}\left(1 + \frac{1}{\frac{\gamma_1^2\left(\gamma_0\hat{\gamma}_0 - \tilde{\gamma}_0^2\right)}{\gamma_0\hat{\gamma}_2^2 + \gamma_2^2\hat{\gamma}_0 - 2\gamma_2\hat{\gamma}_2\tilde{\gamma}_0} - 1}\right). \tag{58}$$

When $\phi_S \to \infty$, the speech CRB is simplified to

$$\lim_{\phi_S \to \infty} \mathrm{CRB}^{\bar{\mathbf{y}}}_{\phi_S} = \frac{\phi_S^2}{L}. \tag{59}$$

V. Experimental Study

In this section, we assess the performance of the proposed MLEs using both simulated data and real-life audio signals. Section V-A deals with simulated data, where the signals are generated synthetically according to the assumed statistical model. A large number of Monte-Carlo trials are carried out in order to examine the effect of the different model parameters, while comparing the MSEs to the corresponding CRBs. In Section V-B, real-life audio signals are generated using room impulse responses (RIRs) and noise signals that are recorded in a reverberant environment. Based on the proposed PSD estimators, an MCWF is constructed, aiming to enhance reverberant and noisy speech. The performance is evaluated by means of objective speech quality measures.

A. Monte-Carlo Simulation

1) Simulation Setup: Synthetic data is generated based on the assumed signal model (1), by simulating L i.i.d. snapshots of a single-tone signal at $f = 2000$ Hz. The signals are captured by a uniform linear array (ULA) of N microphones with inter-distance spacing of $d = 6$ cm. The desired signal s is generated according to (2), then multiplied by the RDTF $\mathbf{g}_d = \exp\left(-j2\pi f \boldsymbol{\tau}\right)$, where $\boldsymbol{\tau} = \frac{d\sin(\theta)}{c} \times [0, \cdots, N-1]^\top$ is the time difference of arrival (TDOA) w.r.t. the first microphone, and θ is the DOA, measured w.r.t. the broadside angle. The reverberation signal $\mathbf{r}$ was drawn according to (3), where $\boldsymbol{\Gamma}_R$ is modelled as an ideal spherical diffuse sound field, given by $\Gamma_{R,ij} = \mathrm{sinc}\left(2\pi f \frac{d|i-j|}{c}\right)$, $i, j \in 1, \ldots, N$. The noise signal $\mathbf{v}$ was drawn according to (5), where $\boldsymbol{\Gamma}_V$ is modelled as a white spatial field.

The various PSDs were estimated using both the non-blocking-based method of (18) and the blocking-based method of (25), where the parameters were initialized with $\boldsymbol{\phi}^{(0)} = 10^{-10}$. For comparison, we also evaluated the LS non-blocking-based and blocking-based estimators proposed in [22]. The accuracy of the estimators was assessed using the normalized mean square error (nMSE) criterion,

$$\mathrm{nMSE}_{\phi_i} = \frac{\mathbb{E}\left[\left(\phi_i - \hat{\phi}_i\right)^2\right]}{\phi_i^2}, \quad i \in \{R, V, S\}, \tag{60}$$

by averaging over 2000 Monte-Carlo trials. As a benchmark, normalized versions of the CRBs in (36)-(38) and (47)-(48) were calculated, against which the nMSEs will be compared.

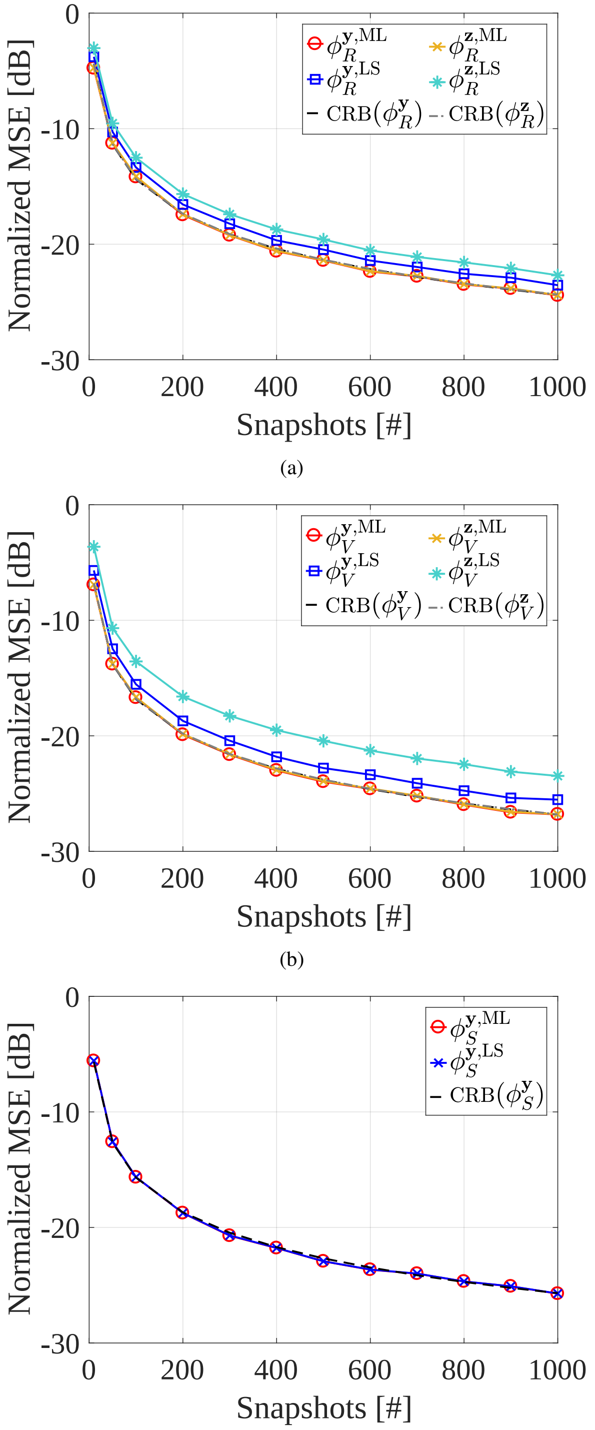
TABLE I: Nominal Parameters

Definition	Symbol	Nominal value
Frequency	f	2 kHz
Direction of arrival	θ	0°
Number of Snapshots	L	100
Speech PSD	ϕ_S	0.5
Reverberation PSD	ϕ_R	0.5
Noise PSD	ϕ_V	0.5
Number of Microphones	N	4
Inter-sensor spacing	d	6 cm
Number of iterations	J	2

The examined estimators and the CRBs are inspected as a function of the following model parameters: i) number of snapshots L; ii) reverberation PSD ϕ_R; iii) noise PSD ϕ_V; iv) speech PSD ϕ_S; and v) number of microphones N. In each experiment, the value of one parameter is changed, while keeping the rest fixed to the nominal values, shown in Table I. In the sequel, the following notation is used: The signal-to-reverberation ratio (SRR) is defined as $\mathrm{SRR} \triangleq 10\log\left(\frac{\phi_S}{\phi_R}\right)$, the signal-to-noise ratio (SNR) as $\mathrm{SNR} \triangleq 10\log\left(\frac{\phi_S}{\phi_V}\right)$ and the signal-to-reverberation-plus-noise ratio (SRNR) as $\mathrm{SRNR} \triangleq 10\log\left(\frac{\phi_S}{\phi_R+\phi_V}\right)$.

2) Simulation Results: The first experiment examines the effect of increasing the number of snapshots, L. The nMSEs and CRBs of the reverberation, noise and speech PSDs are shown in Fig. 1(a), 1(b) and 1(c), respectively. Clearly, the nMSEs of all the estimators decrease as the number of snapshots increases. In comparison with the LS method presented in [22], it is evident that the proposed reverberation and noise MLEs outperform the corresponding LS estimators. For the speech PSD, the results of both estimators are quite similar. Finally, the MSEs of the proposed MLEs coincide with the corresponding CRBs. Under the parameter choice of this experiment, i.e. $\phi_S = \phi_R = \phi_V$, the non-blocking-based and the blocking-based methods seem to be indistinguishable. A deviation between the CRBs will be demonstrated in Fig. 4, where the effect of changing ϕ_S is examined.

In the next experiment, the effect of changing the reverberation level is investigated. We change ϕ_R and hold the other parameters fixed s.t. the resulting SRR ranges between -20 dB and 20 dB. In Fig. 2, the various nMSEs are presented versus

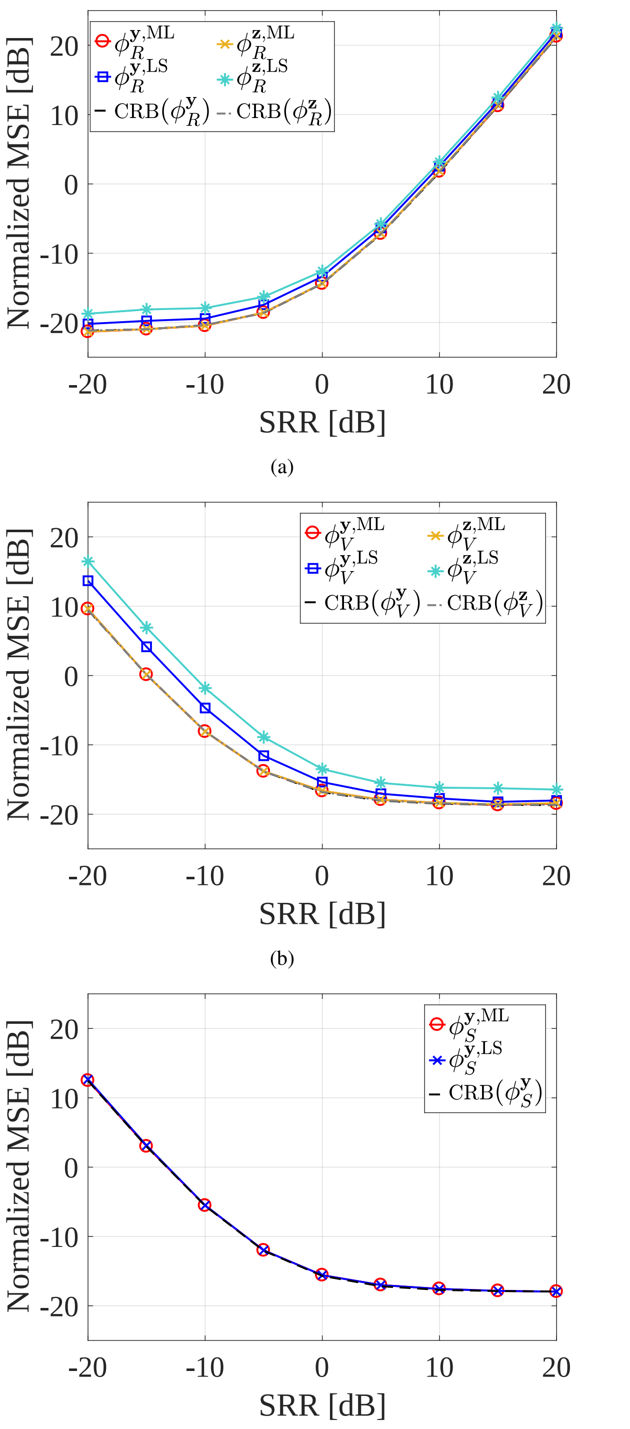


(a)

(b)

(c)

Fig. 1: Normalized MSEs and CRBs versus Number of snapshots, for estimating: (a) ϕ_R, (b) ϕ_V, and (c) ϕ_S.



(a)

(b)

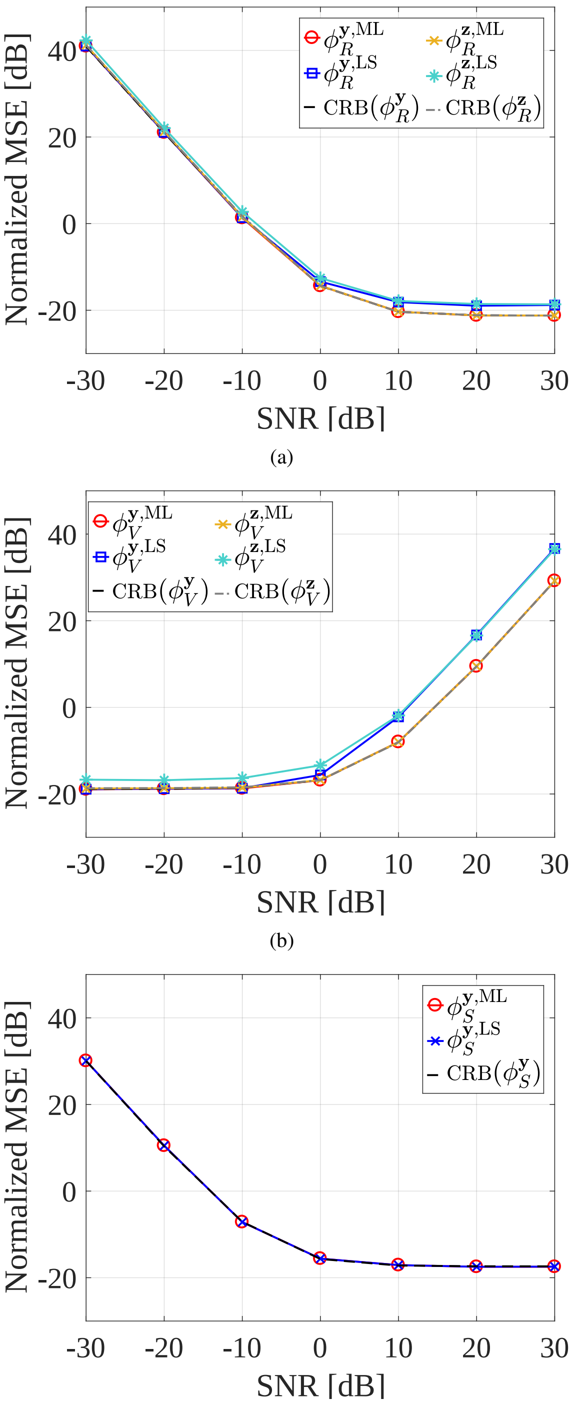
(c)

Fig. 2: Normalized MSEs and CRBs versus SRR, for estimating: (a) ϕ_R, (b) ϕ_V, and (c) ϕ_S.

SRR. It is evident that the performance of the reverberation PSD estimators improves as the reverberation level increases (i.e. when the SRR decreases). The trend for the speech and noise PSDs is reversed; the nMSEs decrease as the reverberation level decreases (i.e. SRR increases).

Next, we study the effect of varying the noise level. In this experiment, ϕ_V is changed s.t. the SNR varies between -30 dB and 30 dB. Fig. 3 shows the nMSEs against SNR. The performance of the noise PSD estimators improves as the noise level increases (SNR decreases), while for the speech and reverberation PSD estimators, the nMSEs decrease as ϕ_V decreases (SNR increases).

The effect of varying the speech PSD level is now inspected, by setting ϕ_S to several values s.t. the SRNR ranges between



(a)

(b)

(c)

Fig. 3: Normalized MSEs and CRBs versus SNR, for estimating: (a) ϕ_R, (b) ϕ_V, and (c) ϕ_S.

-20 dB and 20 dB. In Fig. 4, the nMSEs are presented versus SRNR. As expected from the analytical study, for small ϕ_S the non-blocking-based method yields a lower MSE compared to the blocking-based method. The gap is, however, very small (on the order of 0.1 dB).

Obviously, the blocking-based estimators are not affected by the value of ϕ_S. For the non-blocking-based method, it is shown that the speech PSD estimators are improved as the speech level increases, until reaching the limiting value of $\frac{1}{L}$, which is inline with (59). For the non-blocking-based reverberation and noise PSD estimators, there is a fundamental difference between the ML and the LS methods: The LS estimators degrade significantly as ϕ_S is increased. Similar trends were observed in Figs. 2 and 3, where increase in the level of one PSD value deteriorates the estimation accuracy of the other PSDs. In contrast, the reverberation and noise MLEs (and the corresponding CRBs) start with a very small gap below the blocking-based estimators (as described before), and as ϕ_S increases they approach the blocking-based performance (i.e. become independent on ϕ_S), as manifested by (55). This behaviour may be attributed to the fact that the LS method finds the parameter values to best fit the assumed model to the observed data. As the level of one parameter increases, the fit becomes less sensitive to errors in other parameters, and vice versa.

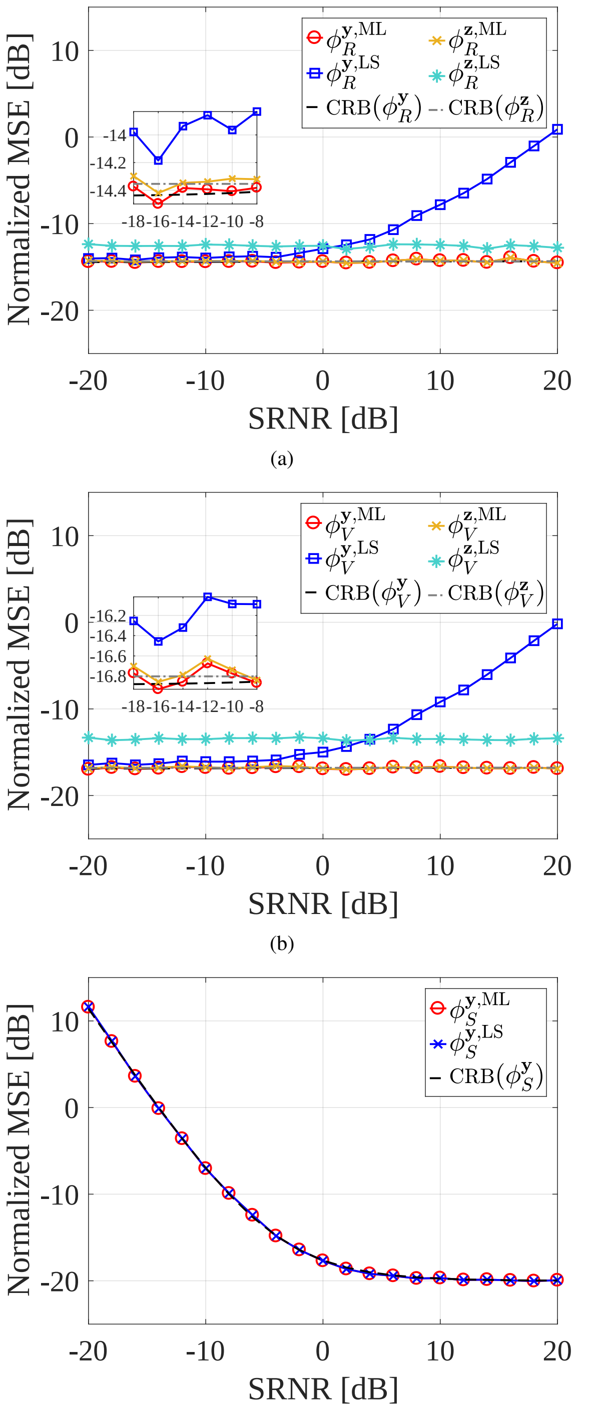
In contrast, the ML method optimally takes into account the different spatial characteristics of the various components, namely the rank-1 nature of the speech component as opposed to the full-rankness of the reverberation and noise components. By (55), the non-blocking-based CRB coincides with the blocking-based CRB for large ϕ_S, which implies that the optimal strategy to estimate ϕ_R includes a preliminary step of generating a null towards the rank-1 speech signal. This phenomenon resembles the noiseless case, where the reverberation non-blocking-based MLE is identical to the blocking-based MLE [20]. The MLE in that case can be interpreted as consisting of two steps: i) applying a projection matrix onto the subspace orthogonal to the speech subspace, which blocks the direction of the speech signal; and ii) averaging the resulting normalized variance. We therefore presume that in our noisy scenario, as ϕ_S becomes large the non-blocking-based MLE of the reverberation (or noise) imitates the blocking-based operation and generates a null towards the speech direction. Therefore, it is not affected by any further change in ϕ_S.

Figure 5 depicts the nMSEs versus the number of microphones, N. Obviously, the nMSEs of the proposed MLEs decrease as the number of microphones increases. While the non-blocking-based LS estimators show a similar trend, the blocking-based LS estimators demonstrate much weaker dependency on N. The reverberation estimator produces only a moderate improvement with N, while the noise estimator approaches a constant nMSE for $N > 6$.

We conclude that the proposed MLEs for the reverberation and noise PSDs outperform the competing LS estimators, for both the non-blocking-based and the blocking-based methods. The performance of the speech PSD estimators is quite similar. It is further demonstrated that the proposed MLEs achieve the corresponding CRBs derived in Section IV.

3) Implementation Issues: In this section we briefly discuss three practical issues, namely the effect of bad initial guess on the required number of iterations, a comparison of techniques to estimate the speech PSD for the blocking-based method, and the computational complexity of both the proposed and the competing methods.

a) Convergence: Both scoring and Newton methods employ an iterative process to maximize the likelihood. It

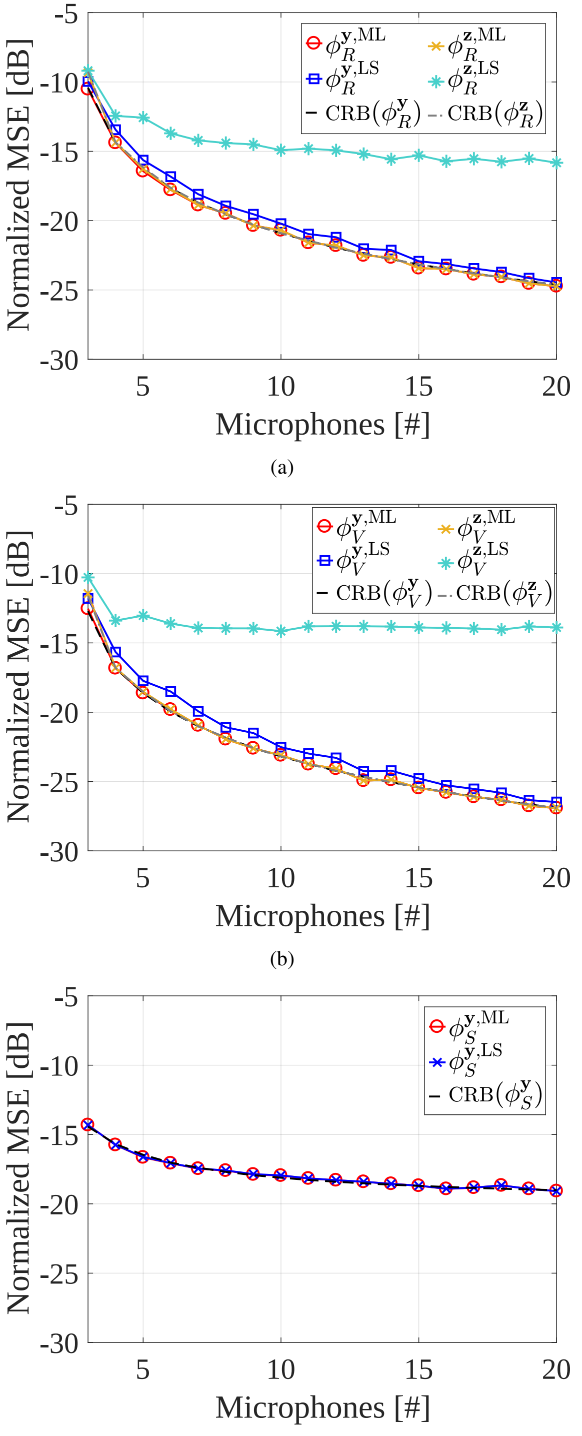


(a)

(b)

(c)

Fig. 4: Normalized MSEs and CRBs versus SRNR, for estimating: (a) ϕ_R, (b) ϕ_V, and (c) ϕ_S.



(a)

(b)

(c)

Fig. 5: Normalized MSEs and CRBs versus N, for estimating: (a) ϕ_R, (b) ϕ_V, and (c) ϕ_S.

is well known that the Newton method converges quickly near the maximum value [15]. However, if started with bad initial values it may converge very slowly. In the following experiment, we examine the convergence of the Newton method compared to the scoring method, as a function of the number of iterations. For the sake of brevity, let the nMSE of the non-blocking-based noise PSD estimator obtained by Newton method be denoted as $\phi_V^{\mathbf{y},\mathrm{N}}$, and the corresponding scoring-based estimator as $\phi_V^{\mathbf{y},\mathrm{S}}$. In Fig. 6, $\phi_V^{\mathbf{y},\mathrm{N}}$ is presented versus the number of iterations for various initial points $\boldsymbol{\phi}^{(0)}$, while $\phi_V^{\mathbf{y},\mathrm{S}}$ is depicted only for the worse initial point. As a benchmark, the CRB in (37) is also depicted. It is shown that for the scoring method, convergence to a very high accuracy is achieved in one iteration, even for a remote initial point. In contrast, the Newton method requires a few dozens of

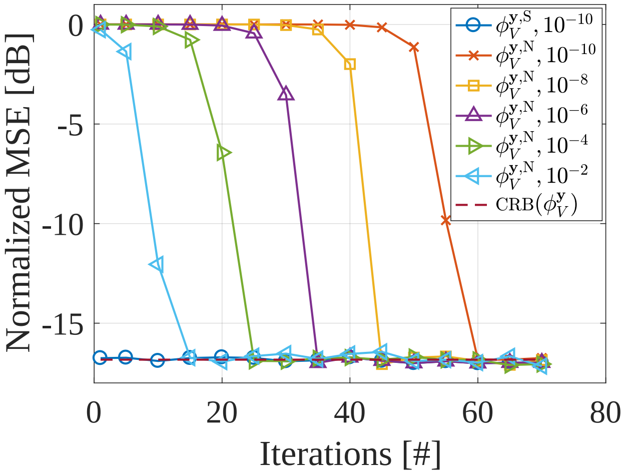


Fig. 6: Convergence behaviour of the non-blocking-based noise MLE implemented by both Newton and scoring methods, for various initial points.

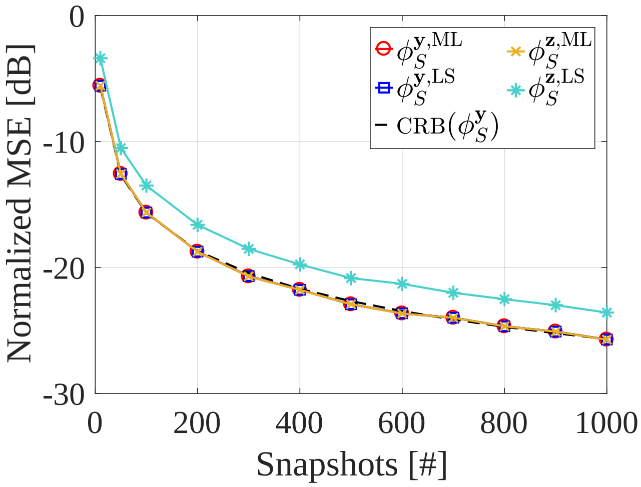


Fig. 7: Normalized MSEs and CRB for estimating ϕ_S, as a function of Number of snapshots.

iterations for convergence, unless a very good initialization is provided.

b) Speech PSD estimate for the blocking-based method: In (28), a speech PSD estimate was proposed for the blocking-based ML estimator. For the LS estimator, the blocking-based speech PSD can be estimated as [22]

$$\phi_S^{\mathrm{LS},\bar{\mathbf{z}}}(m) = \frac{1}{\mathbf{g}_d^{\mathrm{H}}\mathbf{g}_d}\mathrm{Tr}\Big[\mathbf{R}_{\mathbf{y}}(m) - \phi_R^{\mathrm{LS},\bar{\mathbf{z}}}(m)\mathbf{\Gamma}_R - \phi_V^{\mathrm{LS},\bar{\mathbf{z}}}(m)\mathbf{\Gamma}_V\Big], \tag{61}$$

where $\phi_R^{\mathrm{LS},\bar{\mathbf{z}}}$ and $\phi_V^{\mathrm{LS},\bar{\mathbf{z}}}$ are defined therein. In Fig. 7, the nMSEs of both estimators are depicted versus the number of snapshots. For comparison, the non-blocking-based estimators and CRB are depicted as well. It is shown that the proposed blocking-based MLE in (28) outperforms the LS estimator of (61), and approaches the performance of the non-blocking-based method.

c) Computational Complexity: In this section, we compare the computational complexity and the running time of the proposed ML estimators, compared to the LS estimators [22]. The complexity of the various estimators is summarized in Table II, where K denotes the number of frequency bins, M is the number of STFT frames, N denotes the number of microphones and J is the number of iterations for the scoring method. The computational complexity of the proposed estimators is therefore about $2J$ times higher than that of the competing LS estimators. For demonstration, we present also the running time for each method, when required to process a 3 sec recording at sampling rate of 16 kHz, with $K = 257, M = 372, N = 8$ and $J = 1$. The experiments were run in MATLAB R2016b on a HP Compaq Elite 8300 PC, with an Intel 4-core (8 threads) i7-3770 CPU at 3.4 GHz and 32 GB RAM. From these results, it can be deduced that both methods can be implemented in real-time applications.

B. Experiments with Measured Room Impulse Responses

Real noisy and reverberant audio signals are used for assessing the performance of the proposed PSD estimators, when utilized for speech dereverberation and noise reduction. Two different acoustic scenarios are considered. In the sequel, we first describe the experimental setup and the implementation details. Then, we present the performance measures used for assessing the quality of speech, and show the obtained results.

1) Experimental Setup: We consider two different real-life acoustic scenarios, recorded at the acoustic lab of the Engineering Faculty at Bar-Ilan University (BIU). Both scenarios consist of measured RIRs and recorded noises. A first series of experiments was carried out using an eight microphones non-uniform array, with inter-distances of $[8, 6, 4, 3, 3, 4, 6]$ cm. The room panels were adjusted to create reverberation level of $T_{60} = 400$ msec. A 30 sec periodic chirp signal was played from a Head and Torso Simulator (HATS) mannequin with built-in mouth, which was positioned a 1 m from the array, at an angle of $90°$ (see Fig. 8). More details on the setup can be found in [31]. The recorded signals were utilized for identifying the RIR, using the technique described in [32]. For the additive noise, an air-conditioner noise was recorded under the same conditions. The noisy and reverberant signals were constructed by convolving clean speech utterances from the TIMIT database [33] with the RIR, and then adding the noise with several reverberant signal-to-noise ratio (RSNR) levels.

In the second series of experiments, RIRs were downloaded from the RIR database [34]. The reverberation time was set to $T_{60} \in \{360, 610\}$ msec. Measurements were carried out by a ULA of 8 microphones with inter-distance of 8 cm between adjacent microphones. A loudspeaker was located at 1 m distance in front of the array center (angle of $0°$). For the additive noise, we used a babble noise signal from the NOISEX-92 database [35]. Different segments from the babble signal were simultaneously played from 4 loudspeakers located in the room corners (facing the wall), as illustrated in Fig. 8. The microphone signals were synthesized using the same procedure described in the first experiment.

The proposed method assumes the knowledge of the noise spatial coherence matrix. To this end, a 1 sec noise segment

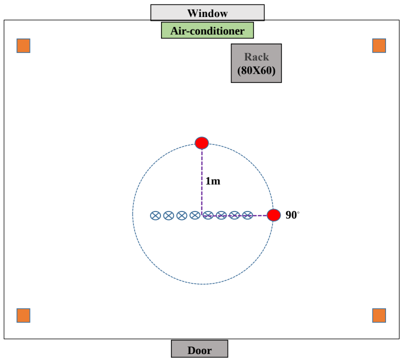


Fig. 8: The room sketch: microphone locations are depicted with blue ‘x’ marks, positions of the speaker are marked by black circle, and position of noise loudspeakers are depicted by orange squares.

TABLE II: Computational Complexity

Method	Complexity	Running Time
Non-blocking LS [22]	$\mathcal{O}\left(5KMN^3\right)$	10.2 sec
Blocking LS [22]	$\mathcal{O}\left(5KM(N-1)^3+4KN^3\right)$	8.3 sec
Non-blocking ML	$\mathcal{O}\left(10KMN^3J\right)$	18.6 sec
Blocking ML	$\mathcal{O}\left(10KM(N-1)^3J+4KN^3\right)$	17.1 sec

is preceded to the reverberant and noisy speech signal. This noise-only segment was utilized for estimating the noise spatial coherence matrix $\boldsymbol{\Gamma}_V$. Note that $\boldsymbol{\Gamma}_V$ is a full-rank matrix, due to reverberation and the short frame size.

The various parameters were set as follows. The sampling rate was set to 16 KHz, and the STFT frame length was 32 msec with 75% overlap. Due to non-stationarity of speech, the sample covariance matrices $\mathbf{R}_{\mathbf{y}}(m)$ and $\mathbf{R}_{\mathbf{z}}(m)$ were estimated using recursive averaging [11] with a smoothing parameter $\alpha = 0.7$, instead of the moving-window averaging of (12). The scoring method was initialized with $\boldsymbol{\phi}^{(0)} = 10^{-10}$, and the number of iterations was fixed to 2.

2) Dereverberation and Noise Reduction Algorithm: A widely-used method for enhancing a reverberant and noisy speech is the MCWF, which produces an optimal speech estimator in the sense of minimizing the MSE [26]. The MCWF, given in (8), can be decomposed into a multichannel minimum variance distortionless response (MVDR) beamformer followed by a single-channel Wiener postfilter [36], [37]:

$$\hat{s}_{\text{MCWF}}(m) = \underbrace{\frac{\hat{\gamma}(m)}{\hat{\gamma}(m)+1}}_{H_W(m)} \underbrace{\frac{\mathbf{g}_d^{\text{H}}\hat{\boldsymbol{\Phi}}_i^{-1}(m)}{\mathbf{g}_d^{\text{H}}\hat{\boldsymbol{\Phi}}_i^{-1}(m)\mathbf{g}_d}}_{\mathbf{w}_{\text{MVDR}}^{\text{H}}(m)} \mathbf{y}(m), \qquad (62)$$

where

$$\hat{\gamma}(m) = \frac{\hat{\phi}_S(m)}{\hat{\phi}_{RE}(m)} \qquad (63)$$

denotes the SRNR at the output of the MVDR, and $\hat{\phi}_{RE}(m) \triangleq \left(\mathbf{g}_d^{\text{H}}\hat{\boldsymbol{\Phi}}_i^{-1}(m)\mathbf{g}_d\right)^{-1}$ is the residual interference at the MVDR output. As an alternative to (63), the decision-directed approach [38] suggests to smooth the estimate of $\hat{\gamma}$ by

$$\hat{\gamma}_{\text{DD}}(m) = \beta\frac{|\hat{s}(m-1)|^2}{\hat{\phi}_{RE}(m-1)} + (1-\beta)\frac{\hat{\phi}_{S_i}(m)}{\hat{\phi}_{RE}(m)}, \qquad (64)$$

where β is a weighting factor, and $\hat{\phi}_{S_i}$ is an instantaneous estimate based on the MVDR output [10]:

$$\hat{\phi}_{S_i}(m) = \max\left(|\mathbf{w}_{\text{MVDR}}^{\text{H}}(m)\mathbf{y}(m)|^2 - \hat{\phi}_{RE}(m), 0\right). \qquad (65)$$

In order to calculate the MCWF in (62), the various PSDs were estimated using both the non-blocking-based method of (18) and the blocking-based method of (25). The proposed estimators are compared to the LS non-blocking-based and blocking-based estimators derived in [22]. Therein, it was shown that the decision-directed approach in (64)–(65) yields improved performance compared to (63). We therefore examine both versions of computing $\hat{\gamma}$: i) The direct implementation in (63), denoted henceforth as Dir; and ii) the decision-directed implementation in (64)–(65), which will be referred to as DD. The smoothing factor for the decision-directed was set to $\beta = 0.8$, and the gain of the single channel postfilter was lower bounded to -15 dB.

3) Performance Measures: Three commonly used objective measures were used for evaluating the speech quality: perceptual evaluation of speech quality (PESQ) [39], frequency-weighted segmental SNR (fwSNRseg) [40] and log-spectral distance (LSD). The various measures were calculated by comparing $\hat{s}_{\text{MCWF}}(m)$ to the clean reference $s(m)$, i.e. the direct speech signal as measured by the reference microphone. The reference signal was obtained by convolving the anechoic speech with the direct path component of the RIR. The measures were averaged over five male and five female TIMIT speakers.

To demonstrate the efficiency of the proposed estimators, the scores obtained by an *oracle* MCWF with true parameters were also computed. The oracle speech PSD was calculated from the clean reference $s(m)$, the oracle reverberation PSD was computed by convolving the anechoic speech signal with the reverberation component of the RIR (assumed to start 2 msec after the direct path), and the oracle noise PSD was calculated from the noise signal.

4) Experimental Results: PESQ, fwSNRseg and LSD scores for the various methods are presented in Tables III, IV and V for both acoustic scenarios. The best results are highlighted in boldface. Note that low LSD indicates a high speech quality. For both acoustic scenarios, it is evident that the proposed estimators provide significant improvement with respect to the noisy and reverberant signal. For the first acoustic scenario, Table III shows that the proposed blocking-based ML DD method obtains the best PESQ and LSD scores, while the proposed non-blocking-based ML DD method yields the best fwSNRseg results. For the second acoustic scenario, Tables IV and V demonstrate that the proposed blocking-based ML DD method yields the best scores in terms of all performance measures, for almost all cases. Note that the decision-directed implementation is superior to the direct implementation for all the considered scenarios. It should be further emphasized that in most cases, each of the proposed

TABLE III: Speech Quality Measures for Scenario 1: Air-Conditioner Noise, $T_{60} = 400$ msec

	PESQ					fwSNRseg					LSD				
Alg.\RSNR	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB
Unprocessed	1.38	1.52	1.67	1.80	1.90	-13.33	-10.02	-7.26	-5.34	-4.05	12.81	10.65	8.98	7.77	6.66
Blocking LS [22] Dir	1.97	2.15	2.26	2.32	2.35	-3.46	-1.75	-0.61	0.17	0.88	6.36	5.17	4.37	4.01	3.87
Blocking LS [22] DD	2.31	2.52	2.63	2.68	2.69	0.39	1.12	1.75	2.23	2.55	**5.08**	4.52	4.13	3.93	3.81
Blocking ML Dir	2.14	2.25	2.31	2.34	2.36	-2.07	-1.04	-0.23	0.49	1.09	5.13	4.49	4.12	3.92	3.82
Blocking ML DD	**2.35**	**2.55**	**2.66**	**2.70**	**2.71**	0.55	1.32	1.93	2.37	**2.70**	**5.08**	**4.47**	**4.07**	**3.86**	**3.74**
Non-blocking LS [22] Dir	2.10	2.25	2.33	2.37	2.38	-1.92	-0.68	0.13	0.74	1.40	5.51	4.65	4.15	3.90	3.78
Non-blocking LS [22] DD	2.30	2.50	2.62	2.68	2.69	0.56	1.32	1.91	2.37	2.67	5.14	4.59	4.20	3.97	3.83
Non-blocking ML Dir	2.17	2.32	2.39	2.42	2.43	-0.72	-0.03	0.51	1.14	1.60	5.10	4.51	4.11	3.92	3.81
Non-blocking ML DD	2.32	2.53	2.64	2.69	**2.71**	**0.60**	**1.36**	**1.96**	**2.39**	**2.70**	5.15	4.53	4.13	3.91	3.78
Oracle MCWF	3.04	3.13	3.19	3.30	3.33	2.20	2.83	3.30	3.82	4.12	3.93	3.58	3.39	3.20	3.14

TABLE IV: Speech Quality Measures for Scenario 2: Babble Noise, $T_{60} = 360$ msec

	PESQ					fwSNRseg					LSD				
Alg.\RSNR	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB
Unprocessed	1.41	1.59	1.78	1.96	2.10	-10.42	-7.52	-5.00	-3.34	-2.27	10.53	8.80	7.52	6.39	5.26
Blocking LS [22] Dir	1.93	2.23	2.49	2.70	2.85	-1.66	-0.01	1.10	1.88	2.44	5.66	4.78	4.14	3.66	**3.43**
Blocking LS [22] DD	2.22	2.58	2.86	3.11	3.24	0.70	1.68	2.35	2.82	3.15	4.82	4.23	3.82	**3.62**	3.51
Blocking ML Dir	1.89	2.17	2.42	2.63	2.80	-2.07	-0.48	0.57	1.34	1.94	5.81	5.08	4.46	4.12	3.93
Blocking ML DD	**2.28**	**2.63**	**2.89**	**3.16**	**3.26**	**0.88**	**1.77**	**2.37**	**2.90**	**3.20**	**4.78**	**4.21**	**3.80**	**3.62**	3.53
Non-blocking LS [22] Dir	2.03	2.31	2.54	2.74	2.85	-1.00	0.22	1.16	1.85	2.35	5.17	4.49	4.01	3.72	3.58
Non-blocking LS [22] DD	2.23	2.56	2.81	3.10	3.21	0.63	1.53	2.12	2.63	2.93	4.96	4.44	4.04	3.94	3.76
Non-blocking ML Dir	2.07	2.36	2.58	2.77	2.88	-0.83	0.36	1.27	1.98	2.42	5.05	4.38	3.91	**3.62**	3.51
Non-blocking ML DD	**2.28**	2.62	2.87	3.15	3.25	0.81	1.64	2.23	2.72	3.02	4.85	4.32	3.93	3.81	3.65
Oracle MCWF	3.07	3.23	3.33	3.59	3.68	2.83	3.32	3.68	4.12	4.35	3.46	3.26	3.14	2.88	2.81

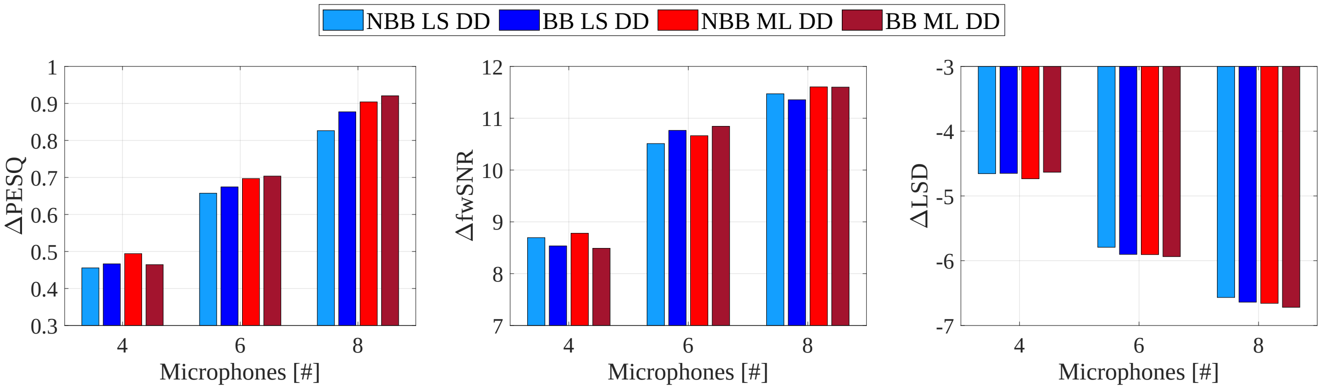
TABLE V: Speech Quality Measures for Scenario 2: Babble Noise, $T_{60} = 610$ msec

	PESQ					fwSNRseg					LSD				
Alg.\RSNR	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB
Unprocessed	1.35	1.47	1.59	1.68	1.74	-14.53	-11.78	-9.53	-7.99	-6.92	13.06	11.31	9.51	7.71	6.28
Blocking LS [22] Dir	1.80	2.01	2.17	2.29	2.36	-4.79	-3.08	-1.86	-1.08	-0.41	6.63	5.54	4.72	4.24	4.07
Blocking LS [22] DD	2.07	2.34	2.53	2.68	2.74	-1.41	-0.43	0.24	0.79	1.06	5.31	4.67	4.30	**4.13**	4.03
Blocking ML Dir	1.73	1.93	2.09	2.21	2.29	-4.66	-3.34	-2.20	-1.40	-0.86	6.48	5.56	4.88	4.70	4.55
Blocking ML DD	**2.11**	**2.39**	**2.58**	**2.74**	**2.80**	**-1.11**	**-0.18**	**0.42**	**0.90**	**1.20**	**5.16**	**4.59**	**4.25**	**4.13**	**4.02**
Non-blocking LS [22] Dir	1.89	2.08	2.22	2.32	2.37	-3.79	-2.50	-1.62	-0.93	-0.39	5.79	4.97	4.43	4.23	4.14
Non-blocking LS [22] DD	2.04	2.29	2.49	2.67	2.75	-1.23	-0.31	0.26	0.78	1.09	5.28	4.74	4.41	4.30	4.17
Non-blocking ML Dir	1.94	2.14	2.27	2.37	2.42	-3.39	-2.17	-1.42	-0.69	-0.19	5.56	4.79	4.32	4.14	4.06
Non-blocking ML DD	2.10	2.37	2.56	**2.74**	**2.80**	-1.12	**-0.18**	0.36	0.85	1.13	5.18	4.65	4.30	4.21	4.08
Oracle MCWF	2.87	2.99	3.07	3.28	3.35	1.33	1.66	1.88	2.33	2.51	3.69	3.50	3.38	3.19	3.15

ML implementations outperforms the competing LS implementation, except for the blocking ML Dir in scenario 2.

We now examine the influence of the number of microphones on the performance. Fig. 9 depicts the measures obtained with different number of microphones, i.e. $N \in \{4, 6, 8\}$, for acoustic scenario 2 with $T_{60} = 610$msec and RSNR $= 5$ dB. For both the ML and the LS approaches, the best 2 implementations (i.e. with DD) are depicted, where NBB denotes the non-blocking-based method, and BB refers to the blocking-based method. To emphasize the score differences, we present the improvement in the performance measures with respect to the input signal. As expected, the performance is improved as the number of microphones is increased. It is evident that the proposed methods outperform the baseline methods for all values of N. Note that for $N = 4$, the non-blocking-based method outperforms the blocking-based method. This may be attributed to the fact that the blocking stage reduces one dimension, which becomes more meaningful when the dimensions number of the data is small.

The capability of the proposed method to jointly reduce reverberation and noise while maintaining low speech distortion is further demonstrated in Fig. 10, where we depict some sonogram examples of the various signals in acoustic scenario 2 with $T_{60} = 610$msec. Fig. 10(a) depicts s, i.e. the clean direct speech signal, as received by the first microphone. Fig. 10(b) shows y_1, the noisy and reverberant signal at the first microphone. Figs. 10(c) and 10(d) present the enhanced signal at the output of the MCWF, computed using either the blocking ML Dir or the blocking ML DD, respectively. Sonograms for the LS estimators are quite similar. However, the proposed method yields a slightly better performance. It can be concluded that the proposed MLEs, when used to



(a) (b) (c)

Fig. 9: Results versus N, for Scenario 2 with $T_{60} = 610$msec and RSNR $= 5$dB: (a) ΔPESQ, (b) ΔfwSNR, and (c) ΔLSD.

construct an MCWF, provide significant interference reduction while keeping low speech distortion. Audio examples can be found in our website.[1]

VI. Conclusions

In this paper, we addressed the problem of speech dereverberation and noise reduction in a spatially homogeneous noise-field. Based on the Fisher scoring algorithm, we derived a non-blocking-based and a blocking-based ML estimators of the various PSDs required for implementing the MCWF. As opposed to state-of-the-art methods which assume the knowledge of the noise PSD, our procedure includes an estimator for the noise PSD, along with the reverberation and speech PSDs. Furthermore, CRBs on the various PSDs were derived for the two proposed MLEs. For both the reverberation and the noise PSD estimators, it is shown that the non-blocking-based CRB is smaller than the blocking-based CRB. The discussion is supported by an experimental study, based on both simulated data and real-life audio signals, demonstrating the advantage of the proposed estimators over competing estimators.

Appendix A

In [21, Appendix B], it was shown that $\delta \geq 0$, with equality if $\phi_V = 0$ or $\mathbf{\Gamma}_V = \mathbf{\Gamma}_R$. Following the same lines of proof, we now show that $\hat{\delta} \geq 0$ and $\tilde{\delta} \leq 0$.

We start by showing that $\hat{\delta} \geq 0$. To this end, we derive an explicit expression for $\hat{\delta}$. Using the Cholesky decomposition, the reverberation spatial coherence matrix is decomposed as $\mathbf{\Gamma}_R = \mathbf{R}\mathbf{R}^{\mathrm{H}}$, where $\mathbf{R}$ is an $N \times N$ lower triangular matrix. Using (9), the prewhitened interference PSD matrix writes

$$\mathbf{\Phi}_{i\mathrm{W}} \triangleq \mathbf{R}^{-1}\mathbf{\Phi}_i\mathbf{R}^{-\mathrm{H}} = \phi_R\mathbf{I} + \phi_V\mathbf{R}^{-1}\mathbf{\Gamma}_V\mathbf{R}^{-\mathrm{H}}. \tag{66}$$

Applying the eigenvalue decomposition (EVD) to $\mathbf{R}^{-1}\mathbf{\Gamma}_V\mathbf{R}^{-\mathrm{H}}$ yields

$$\mathbf{R}^{-1}\mathbf{\Gamma}_V\mathbf{R}^{-\mathrm{H}} = \mathbf{U}^{\mathrm{H}}\mathbf{\Lambda}\mathbf{U}, \tag{67}$$

where $\mathbf{U}$ is the $N \times N$ eigenvectors matrix and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, denoted by $\lambda_i \triangleq \Lambda_{ii}, i = 1, \ldots, N$. Substituting (67) into (66) and using the orthonormality of $\mathbf{U}$, $\mathbf{\Phi}_i^{-1}$ writes

$$\mathbf{\Phi}_i^{-1} = \mathbf{R}^{-\mathrm{H}}\mathbf{\Phi}_{i\mathrm{W}}^{-1}\mathbf{R}^{-1} = \mathbf{R}^{-\mathrm{H}}\mathbf{U}^{\mathrm{H}}\mathbf{\Upsilon}^{-1}\mathbf{U}\mathbf{R}^{-1}, \tag{68}$$

where in the last step we defined the diagonal matrix $\mathbf{\Upsilon} \triangleq \phi_R\mathbf{I}+\phi_V\mathbf{\Lambda}$. Substituting (68) into (36b), (36f) and (36g) yields

$$\gamma_1 = \mathbf{d}^{\mathrm{H}}\mathbf{\Upsilon}^{-1}\mathbf{d}, \tag{69a}$$
$$\hat{\gamma}_2 = \mathbf{d}^{\mathrm{H}}\mathbf{\Upsilon}^{-1}\mathbf{\Lambda}\mathbf{\Upsilon}^{-1}\mathbf{d}, \tag{69b}$$
$$\hat{\gamma}_3 = \mathbf{d}^{\mathrm{H}}\mathbf{\Upsilon}^{-1}\mathbf{\Lambda}\mathbf{\Upsilon}^{-1}\mathbf{\Lambda}\mathbf{\Upsilon}^{-1}\mathbf{d}, \tag{69c}$$

where $\mathbf{d} \triangleq \mathbf{U}\mathbf{R}^{-1}\mathbf{g}_d$. By construction, $\mathbf{\Lambda}$ and $\mathbf{\Upsilon}$ are diagonal matrices, and thus (69) can be recast as

$$\gamma_1 = \sum_{i=1}^{N} \frac{|d_i|^2}{\phi_R + \phi_V\lambda_i}, \tag{70a}$$
$$\hat{\gamma}_2 = \sum_{i=1}^{N} \frac{|d_i|^2\lambda_i}{(\phi_R + \phi_V\lambda_i)^2}, \tag{70b}$$
$$\hat{\gamma}_3 = \sum_{i=1}^{N} \frac{|d_i|^2\lambda_i^2}{(\phi_R + \phi_V\lambda_i)^3}. \tag{70c}$$

Substituting (70) into $\hat{\delta}$ in (40b) and using the double-sum identity in [21, Eq. B.68], yields

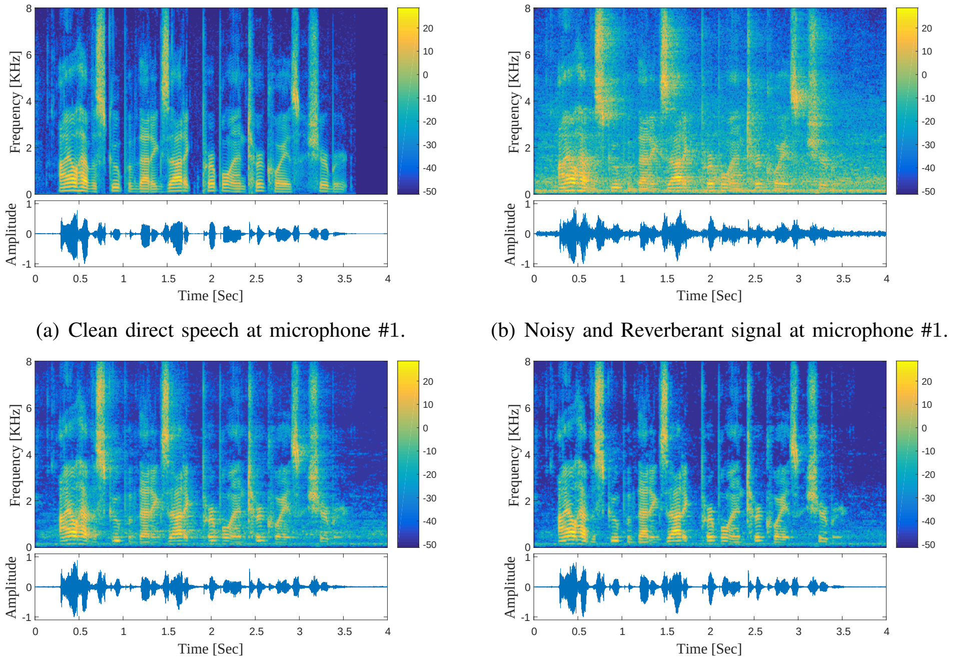
$$\hat{\delta} = \frac{\phi_R^2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{|d_i|^2|d_j|^2\left(\lambda_i - \lambda_j\right)^2}{(\phi_R + \phi_V\lambda_i)^3\,(\phi_R + \phi_V\lambda_j)^3}. \tag{71}$$

Note that $\lambda_i > 0$ since they are eigenvalues of a positive definite matrix. It follows that $\hat{\delta} \geq 0$, with equality if: i) $\phi_R = 0$; or ii) $\lambda_i = \lambda_j$, $\forall 1 \leq i, j \leq N$. It can be shown that the latter case occurs only if $\mathbf{\Gamma}_V = \mathbf{\Gamma}_R$.

The proof that $\tilde{\delta} \leq 0$ is similar. Substituting (68) into (36c) and (36i) yields

$$\gamma_2 = \mathbf{d}^{\mathrm{H}}\mathbf{\Upsilon}^{-2}\mathbf{d} = \sum_{i=1}^{N} \frac{|d_i|^2}{(\phi_R + \phi_V\lambda_i)^2}, \tag{72a}$$
$$\tilde{\gamma}_3 = \mathbf{d}^{\mathrm{H}}\mathbf{\Upsilon}^{-1}\mathbf{\Lambda}\mathbf{\Upsilon}^{-2}\mathbf{d} = \sum_{i=1}^{N} \frac{|d_i|^2\lambda_i}{(\phi_R + \phi_V\lambda_i)^3}. \tag{72b}$$

[1] http://www.eng.biu.ac.il/gannot/speech-enhancement/



(a) Clean direct speech at microphone #1.

(b) Noisy and Reverberant signal at microphone #1.

(c) Output of the MCWF using the proposed blocking-based ML Dir.

(d) Output of the MCWF using the proposed blocking-based ML DD.

Fig. 10: Sonogram examples for Scenario 2, with $T_{60} = 610$ msec and RSNR $= 10$ dB.

Substituting (70a), (70b) and (72) into $\tilde{\delta}$ in (40c), it can be shown that

$$\tilde{\delta} = -\frac{\phi_R \phi_V}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{|d_i|^2 |d_j|^2 \left(\lambda_i - \lambda_j\right)^2}{\left(\phi_R + \phi_V \lambda_i\right)^3 \left(\phi_R + \phi_V \lambda_j\right)^3}, \quad (73)$$

and thus $\hat{\delta} \leq 0$, with equality if: i) $\phi_R = 0$; or ii) $\phi_V = 0$; or iii) $\mathbf{\Gamma}_V = \mathbf{\Gamma}_R$.

References

[1] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1581–1592, 1994.

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

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

[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)*, 2013, pp. 1–5.

[5] 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.

[6] S. Braun and E. A. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 34, 2015.

[7] O. Schwartz, S. Braun, S. Gannot, and E. A. 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)*, 2015, pp. 1–5.

[8] O. Schwartz, S. Gannot, and E. A. 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)*, 2016, pp. 151–155.

[9] A. Kuklasinski, S. Doclo, and J. Jensen, "Maximum likelihood psd estimation for speech enhancement in reverberant and noisy conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 599–603.

[10] O. Schwartz, S. Gannot, and E. A. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm," in *24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1123–1127.

[11] 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. 1599–1612, 2016.

[12] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 6, pp. 1052–1067, 2018.

[13] P. Thüne and G. Enzner, "Maximum-likelihood approach with Bayesian refinement for multichannel-Wiener postfiltering," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3399–3413, 2017.

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

[15] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 5, pp. 1210–1217, 1983.

[16] C. R. Rao, *Linear statistical inference and its applications*. Wiley, New York, 1965.

[17] J. C. Ogilvie and C. D. Creelman, "Maximum-likelihood estimation of receiver operating characteristic curve parameters," *Journal of mathematical psychology*, vol. 5, no. 3, pp. 377–391, 1968.

[18] D. D. Dorfman and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory—a direct solution," *Psychometrika*, vol. 33, no. 1, pp. 117–124, 1968.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[20] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr.*, 2015.

[21] O. Schwartz, S. Gannot, and E. A. Habets, "Cramér–rao bound analysis of reverberation level estimators for dereverberation and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1680–1693, 2017.

[22] I. Kodrasi and S. Doclo, "Joint late reverberation and noise power spectal density estimation in a spatially homogenous noise field," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[23] Y. Laufer, B. Laufer-Goldshtein, and S. Gannot, "ML estimation and CRBs for reverberation, speech and noise PSDs in rank-deficient noise-field," *preprint*, 2019, arXiv:1907.09250.

[24] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, 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)*, 2015, pp. 91–95.

[25] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1732–1736, 1962.

[26] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation and modulation theory*. Wiley, 2002.

[27] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 938–949, 1995.

[28] W. J. Bangs, "Array processing with generalized beamformers," Ph.D. dissertation, Yale University, New Haven, CT, 1972.

[29] H. Hung and M. Kaveh, "On the statistical sufficiency of the coherently averaged covariance matrix for the estimation of the parameters of wideband sources," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 12, 1987, pp. 33–36.

[30] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," *Technical University of Denmark*, 2008.

[31] Y. Laufer and S. Gannot, "A Bayesian hierarchical model for speech enhancement with time-varying audio channel," *IEEE/ACM Tran. on Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 225–239, 2019.

[32] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," Disc, 1993.

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

[35] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[36] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[37] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Process. Workshop*, 2002, pp. 209–213.

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

[39] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[40] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.