Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise

Kuklasinski, Adam; Doclo, Simon; Jensen, Søren Holdt; Jensen, Jesper

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2016.2573591

Publication date:
2016

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Kuklasinski, A., Doclo, S., Jensen, S. H., & Jensen, J. (2016). Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise. *I E E E Transactions on Audio, Speech and Language Processing*, *24*(9), 1599-1612. Advance online publication. https://doi.org/10.1109/TASLP.2016.2573591

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise

Adam Kuklasiński, *Student Member, IEEE,* Simon Doclo, *Senior Member, IEEE,*
Søren H. Jensen, *Senior Member, IEEE,* and Jesper Jensen

***Abstract*—In this contribution we focus on the problem of power spectral density (PSD) estimation from multiple microphone signals in reverberant and noisy environments. The PSD estimation method proposed in this paper is based on the maximum likelihood (ML) methodology. In particular, we derive a novel ML PSD estimation scheme that is suitable for sound scenes which besides speech and reverberation consist of an additional noise component whose second-order statistics are known. The proposed algorithm is shown to outperform an existing similar algorithm in terms of PSD estimation accuracy. Moreover, it is shown numerically that the mean squared estimation error achieved by the proposed method is near the limit set by the corresponding Cramér-Rao lower bound. The speech dereverberation performance of a multi-channel Wiener filter (MWF) based on the proposed PSD estimators is measured using several instrumental measures and is shown to be higher than when the competing estimator is used. Moreover, we perform a speech intelligibility test where we demonstrate that both the proposed and the competing PSD estimators lead to similar intelligibility improvements.**

***Index Terms*—PSD estimation, maximum likelihood estimation, Cramér-Rao lower bound, reverberation, microphone array.**

I. Introduction

REVERBERATION and additive noise can lower the perceived quality and hinder the intelligibility of speech. This is particularly a problem in speech communication scenarios where the microphones of the receiving/recording device are at a distance from the speaker, e.g. as in hands-free telephony or in hearing aids. Clearly, noise and reverberation reduction algorithms are of practical interest.

In the literature many types of processing algorithms have been proposed for dereverberation and/or noise reduction in speech signals. Because in most scenarios both noise and reverberation are present, we focus on algorithms that can be used to jointly reduce these two types of interference (as opposed to only one of them). Moreover, we specifically focus on reduction of the late reverberation because it is believed to be particularly detrimental for speech intelligibility [3]. Following [4], speech dereverberation algorithms can be broadly divided into spectral enhancement, spatial processing, and system identification/inversion algorithms. The latter class of algorithms is generally more appropriate for dereverberation than for noise reduction (with some exceptions, e.g.: [5], [6]) and is generally used for equalization of the deterministic part of the impulse responses, rather than their stochastic (i.e. predominately late) part. On the other hand, the first two classes of algorithms (spectral enhancement and spatial processing) are well suited for noise reduction [7] and for late reverberation reduction [4]. Hence, we focus on these two types of algorithms.

Most spectral enhancement algorithms are implemented in the spectro-temporal domain and are usually based on an *a priori* statistical model of the signal components (for overviews see [7]–[9]). For example, in many noise reduction algorithms the noise power is estimated only in some spectro-temporal regions (e.g. when the signal is dominated by the noise) and is assumed to be approximately stationary between them. Speech dereverberation algorithms are mostly targeted at suppression of the late reverberation, which is often modeled as exponentially decaying and additive (e.g. [10], [11]). These and similar statistical models are used to estimate the signal-to-interference ratio in individual spectro-temporal regions, which are processed accordingly using e.g. the spectral subtraction rule or the Wiener filter [10], [11].

Spatial processing algorithms, or beamformers, work by combining the signals of an array of microphones such that it is sensitive to sounds impinging from a specific direction while suppressing sounds from other directions. Obviously, beamformers are only effective in scenarios where the interference (noise and/or reverberation) impinges on the microphone array from different directions than the target speech.

Spectral enhancement and beamforming algorithms are often combined to create a two-step algorithm where the beamformer is followed by a single channel spectral enhancement scheme (in this context referred to as the post-filter). Among the first methods of this type proposed for speech dereverberation and noise reduction were [12], [13], both composed of a delay-and-sum beamformer and a coherence-based post-filter. The beamformers and post-filters in algorithms proposed

Manuscript received MMM DD, YYYY; revised MMM DD, YYYY; accepted MMM DD, YYYY. Date of publication MMM DD, YYYY; date of current version MMM DD, YYYY.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement № ITN-GA-2012-316969. More information about the project can be found on the website: /www.dreams-itn.eu/.

A. Kuklasiński and J. Jensen are with Oticon A/S, 2765 Smørum, Denmark, and with Aalborg Univeristy, Department of Electronic Systems, Signal and Information Processing Section, 9220 Aalborg, Denmark. (e-mail: adku@oticon.com, jesj@oticon.com)

S. Doclo is with the University of Oldenburg, Department of Medical Physics and Acoustics, and Cluster of Excellence Hearing4all, Oldenburg, Germany. (e-mail: simon.doclo@uni-oldenburg.de)

S. H. Jensen is with Aalborg Univeristy, Department of Electronic Systems, Signal and Information Processing Section, 9220 Aalborg, Denmark. (e-mail: shj@aau.es.dk)

This manuscript is partially based on [1], [2]

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.xxxxxxxxxx

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

more recently are generally based on some optimality criteria, most notably the linear minimum mean square error (MMSE) resulting in the multi-channel Wiener filter (MWF) [14], [15]. The MWF depends on the inter-microphone covariance matrices of the desired (target speech) and of the interference (noise and late reverberation) components of the input signal. These matrices are usually not known but in some scenarios their structure can be modeled such that only few parameters remain to be estimated. In this paper we employ a set of assumptions that result in a signal covariance model where only the power spectral densities (PSDs) of the target speech and of the late reverberation need to be estimated.

Several methods exist for estimating the speech and the late reverberation PSDs in the considered setup. Estimators operating on a single microphone signal are generally considered inferior to PSD estimators using multiple microphones [16]. In the past, multi-microphone estimators based on the inter-microphone coherence have been proposed [12], [13]. These estimators generally are based on the assumption that the late reverberation is uncorrelated between microphones (invalid e.g. for low frequencies and finite inter-microphone distance). More recently, estimators based on optimality criteria have been proposed, e.g. by Braun and Habets [17], and by the authors of this study [18]. Both these estimators are based on the maximum likelihood (ML) methodology, and have been compared with respect to the estimation accuracy in [1]. For the special case where the signals are composed of only speech and reverberation, the estimator from [18] has been found to yield superior statistical performance compared with the estimator from [17]. In fact, in [16] it was argued that the estimator used in [18] is optimal in the minimum variance unbiased (MVU) sense.

A disadvantage of the estimator in [18] compared to the estimator in [17] is that the former does not take the additive noise into account. On the other hand, the estimator in [17] is derived using an unrealistic statistical assumption which results in its decreased estimation performance [1]. In this contribution we propose a scheme which avoids both these limitations. Specifically, we propose a novel multi-microphone PSD estimator which is ML-optimal and generalizes the method from [18] to signal models including a target signal contaminated by late reverberation and additive noise.

This paper is structured as follows. Section II presents the signal model and discusses the employed statistical assumptions. In Section III the proposed estimator is derived and several practically relevant special cases are presented for which the estimator is particularly simple. In Section IV a detailed experimental evaluation is performed and the statistical performance of the proposed estimator is compared to the estimator from [17]. It is also shown, that the mean squared error of the proposed PSD estimator is close to the lowest possible for unbiased estimators, as set by the Cramér-Rao lower bound (CRLB). In Section V the speech dereverberation performance of an MWF based on the two compared PSD estimation methods is evaluated in terms of: the frequency-weighted segmental signal-to-noise ratio (FWSegSNR) [19], the perceptual evaluation of speech quality (PESQ) [20] measure, two interference attenuation, and one speech distortion measure [21], [22]. Lastly, in Section VI, the two variants of the MWF are evaluated in a speech intelligibility (SI) test with human subjects. Section VII concludes the paper.

II. Signal model and statistical assumptions

Consider an array of M microphones in a reverberant room where a single talker is active. Speech generated by the talker reaches the microphones not only via the direct propagation path, but also via multiple reflections off the walls and other surfaces in the room. In most practical situations the microphone signals are further disrupted by the microphone self-noise and by other additive noise sources.

For a particular arrangement of a sound source and a sound receiver, acoustic properties of a room can be compactly expressed in terms of a room impulse response (RIR). We adopt an often-made assumption that RIRs are composed of three distinct parts: the direct path response, the early reflections, and the late reverberation. The direct and early components of reverberant speech are generally considered advantageous for speech intelligibility [3]; hence, we refer to their sum as the target signal. In specific scenarios it might not be desirable or practical to include all early reflections (conventionally the first 50 ms of the RIR) in the target signal model. For this reason we define the target signal as the direct path speech plus those of its early reflections whose delay relative to the direct path is less than a certain threshold t_s. The remaining early reflections are not accounted for in the signal model. All other components of the signal, i.e. the late reverberation, the microphone self-noise, and other additive noise types, are all considered an interference because of their detrimental effect on speech quality and intelligibility.

Let $y_m(t)$ denote the time-domain signal of the m-th microphone of the array ($m = 1, \ldots, M$), where t is a discrete time index. Due to the wide-band and non-stationary nature of the speech, it is often convenient to implement speech processing algorithms in the spectro-temporal domain. Thus, we express $y_m(t)$ as its short time Fourier transform (STFT) given by:

$$y_m(k,n) = \sum_{t=0}^{T-1} y_m(t+nD)w(t)e^{-2\pi i k \frac{t}{T}},$$

where k is the frequency bin index, n is the time frame index, the STFT length is denoted by T, the filterbank decimation factor is denoted by D, and $w(t)$ is the analysis window function. For notational conciseness we stack the STFT coefficients corresponding to all of the microphones in a vector $\mathbf{y}(k,n) = [\, y_1(k,n) \ldots \, y_M(k,n) \,]^T$. Furthermore, we assume that $\mathbf{y}(k,n)$ is a sum of three components:

$$\mathbf{y}(k,n) = \mathbf{s}(k,n) + \mathbf{r}(k,n) + \mathbf{x}(k,n), \tag{1}$$

where $\mathbf{s}(k,n)$ corresponds to the target signal, $\mathbf{r}(k,n)$ corresponds to the late reverberation, and $\mathbf{x}(k,n)$ is the additive noise component (i.e. sum of the microphone self-noise, ambient noise, and possibly other additive interferences).

We assume that $\mathbf{y}(k,n)$ is uncorrelated across frequency bins, which allows us to omit the frequency bin index k in

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

the subsequent presentation. All processing is performed independently in all frequency bins. Moreover, for mathematical tractability, we assume that $\mathbf{y}(n)$ is uncorrelated across time frames. In other words, we neglect the influence of any existing overlap between the time frames and any autocorrelation the microphone signals may exhibit for delays larger than the STFT length. Because reverberant speech signals are autocorrelated and the time frames do overlap, this assumption is, at best, only approximately valid. Nevertheless, it is employed in many speech processing algorithms (e.g. [10], [17], [23]) and the general success of these methods reflects that it is a useful working assumption.

Because the additive noise is generated by physical processes independent of the speech, we assume that $\mathbf{x}(n)$ is uncorrelated with $\mathbf{s}(n)$ and $\mathbf{r}(n)$. Moreover, we assume that the late reverberation $\mathbf{r}(n)$ is uncorrelated with the target signal $\mathbf{s}(n)$. This is an often used assumption (e.g. [10], [11], [17]), which can be justified by the fact that the late part of RIRs is disturbed by thermal fluctuations of the air [24] and slight movements of the source and the microphone array [25] which are unavoidable in practical scenarios. Moreover, in applications where the STFT length has to be very short (such as in hearing aids), in any time frame the reverberation can be argued to be correlated mostly with the speech component of the preceding time frames, but not of the current one.

The covariance matrix of $\mathbf{y}(n)$ is defined as:

$$\mathbf{\Phi}_{\mathbf{y}}(n) = E[\mathbf{y}(n)\mathbf{y}^H(n)], \tag{2}$$

where $E[\cdot]$ denotes the expectation operator and $(\cdot)^H$ is the Hermitian transpose. Each of the diagonal elements of $\mathbf{\Phi}_{\mathbf{y}}(n)$ is equal (up to a normalization constant) to the power spectral density (PSD) of the respective microphone signal in the particular frequency bin. Similarly, off-diagonal elements of $\mathbf{\Phi}_{\mathbf{y}}(n)$ correspond to the cross-PSDs between the respective microphones. Hence, we refer to $\mathbf{\Phi}_{\mathbf{y}}(n)$ as the cross-PSD matrix of $\mathbf{y}(n)$. Because we assume that the signal components are uncorrelated, $\mathbf{\Phi}_{\mathbf{y}}(n)$ can be decomposed into a sum of cross-PSD matrices of the individual signal components. Hence:

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \mathbf{\Phi}_{\mathbf{s}}(n) + \mathbf{\Phi}_{\mathbf{r}}(n) + \mathbf{\Phi}_{\mathbf{x}}(n), \tag{3}$$

where $\mathbf{\Phi}_{\mathbf{s}}(n)$, $\mathbf{\Phi}_{\mathbf{r}}(n)$, and $\mathbf{\Phi}_{\mathbf{x}}(n)$ denote the cross-PSD matrices of $\mathbf{s}(n)$, $\mathbf{r}(n)$, and $\mathbf{x}(n)$, respectively.

We assume that the STFT coefficients of the microphone signal and its individual components are circularly-symmetric complex Gaussian distributed, e.g: $\mathbf{y}(n) \sim \mathcal{N}_C(\mathbf{0}, \mathbf{\Phi}_{\mathbf{y}}(n))$. While it is known that the STFT coefficients, particularly of the speech component, are more accurately modeled using super-Gaussian distributions (see e.g. [26]–[28]), the resulting estimators tend to become significantly more complicated (see e.g. [22]). Thus, the Gaussian assumption appears to be a good tradeoff between accuracy and mathematical tractability.

We model the talker as a single point-source. The direct path and the early reflections can be modeled as linear filters acting on the speech emitted by the talker. In effect, the target signal received by any of the microphones is a linearly filtered version of the target signal anywhere else in the room. In order to use this property, we select a certain reference position (conventionally one of the microphones) and denote the STFT of the target signal at that position by $s(n)$ (a scalar). Next, we let $\mathbf{d}$ denote a vector of relative transfer functions (RTFs) [29] of the target signal from the chosen reference position to all of the microphones (evaluated at the center frequency of the current frequency bin). For $\mathbf{d}$ to represent the RTFs accurately, the early reflection threshold t_s must be shorter than the STFT length. Using the above definitions, we can write:

$$\mathbf{s}(n) = s(n)\mathbf{d}. \tag{4}$$

We assume that an estimate of $\mathbf{d}$ is available (e.g. because the application at hand allows its accurate off-line estimation, or, alternatively, by use of an on-line estimation scheme such as [30], [31]). Using (4) in the definition of $\mathbf{\Phi}_{\mathbf{s}}(n)$ results in:

$$\mathbf{\Phi}_{\mathbf{s}}(n) = E[\mathbf{s}(n)\mathbf{s}^H(n)] = \phi_s(n)\mathbf{d}\mathbf{d}^H. \tag{5}$$

It follows that the matrix $\mathbf{\Phi}_{\mathbf{s}}(n)$ is rank-one and constant up to a scaling factor $\phi_s(n)$, which denotes the time-varying PSD of the target speech at the reference position.

The late reverberation cross-PSD matrix may be written as:

$$\mathbf{\Phi}_{\mathbf{r}}(n) = \phi_r(n)\mathbf{\Gamma}_{\mathbf{r}}, \tag{6}$$

where $\phi_r(n)$ denotes the time-varying (scalar) PSD of the late reverberation at the reference position and $\mathbf{\Gamma}_{\mathbf{r}}$ is the cross-PSD matrix of the late reverberation normalized by $\phi_r(n)$. The proposed method is based on the assumption that $\mathbf{\Gamma}_{\mathbf{r}}$ is full-rank and known, or, equivalently, that the spatial distribution of the late reverberation is known. Drawing from statistical models employed in theoretical room acoustics (see e.g. [32]) we assume that all directions contribute equally to the late reverberant sound field, i.e. that this sound field is isotropic. In consequence, $\mathbf{\Gamma}_{\mathbf{r}}$ can be measured *a priori* as it does not depend on the position or orientation of the microphone array within the room. For free-field microphone arrays, $\mathbf{\Gamma}_{\mathbf{r}}$ can even be calculated analytically using information on the microphone array geometry [33], [34]. For other microphone arrays, $\mathbf{\Gamma}_{\mathbf{r}}$ has to be measured or modeled numerically. In many rooms, the floor and the ceiling are the most acoustically damped surfaces. In effect, the vertical component of the reverberant sound field is damped more than its horizontal components. In such rooms the reverberation is more accurately modeled as cylindrically, rather than spherically isotropic.

We assume that the third component of the signal model, $\mathbf{x}(n)$, is related to an additive noise whose statistics are varying slowly—a realistic assumption if $\mathbf{x}(n)$ is used to model the sum of the noise generated by the microphones and by other sources: ambiance, ventilation equipment, car or airplane cabin noise, etc. As a consequence, the cross-PSD matrix $\mathbf{\Phi}_{\mathbf{x}}$ can be assumed approximately constant across short spans of time (hence, we omit index n). We assume that $\mathbf{\Phi}_{\mathbf{x}}$ is known or that a reliable estimate thereof is available. In practice, an estimation scheme such as the multi-microphone speech probability estimator proposed in [35] could be used to periodically update $\mathbf{\Phi}_{\mathbf{x}}$ during time-frequency regions where speech and late reverberation levels are low compared to that of the noise (e.g. between speech utterances).

Using (5) and (6), the overall model for the microphone

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

input cross-PSD matrix can be re-written as (cf. (3)):

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \phi_s(n)\mathbf{d}\mathbf{d}^H + \phi_r(n)\mathbf{\Gamma}_{\mathbf{r}} + \mathbf{\Phi}_{\mathbf{x}}. \tag{7}$$

In this model only the scalar PSDs $\phi_s(n)$ and $\phi_r(n)$ are unknown; their estimation and application to speech dereverberation is the focus of this paper. To facilitate the derivation of the proposed estimators, we assume that $\phi_s(n)$ and $\phi_r(n)$ can be considered approximately constant across a certain number L of consecutive time frames of the STFT. For small L, such that L frames span less than 50 ms, this is analogous to the commonly made assumption of short-time speech stationarity.

The proposed PSD estimation method is intended for reverberant and noisy speech signals, and the employed assumptions are motivated by this application. However, the proposed algorithm is equally useful for other types of signals, provided that the assumptions made are satisfied, i.e. that the signals are approximately Gaussian and that their cross-PSD matrix can be modeled using (7).

III. Derivation of the proposed PSD estimators

In this section we derive the proposed maximum likelihood estimators (MLEs) of $\phi_s(n)$ and $\phi_r(n)$. We begin by formulating a probability density function (PDF) of the input signal $\mathbf{y}(n)$, which we subsequently use to define a joint likelihood function of $\phi_s(n)$ and $\phi_r(n)$.

Due to the assumptions outlined in Section II, the input signal vectors $\mathbf{y}(n)$ in any L consecutive time frames can be considered approximately independent and identically distributed. It follows, that the joint PDF of the signal in these L time frames can be calculated as the product of the PDFs of $\mathbf{y}(n)$ in individual time frames. Denoting the sample cross-PSD matrix of the input signal as:

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) = \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{y}(n-l)\mathbf{y}^H(n-l), \tag{8}$$

we can compactly express the joint complex Gaussian PDF of $\mathbf{y}(n)$ in L consecutive time frames as:

$$f = \frac{1}{\pi^{LM}|\mathbf{\Phi}_{\mathbf{y}}(n)|^L}\exp\big[-L\,\mathrm{tr}\big(\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{\Phi}_{\mathbf{y}}^{-1}(n)\big)\big]. \tag{9}$$

This joint PDF depends on ϕ_s and ϕ_r (through $\mathbf{\Phi}_{\mathbf{y}}(n)$, cf. (7)), which are regarded as deterministic but unknown.

The required joint likelihood function is obtained by interpreting the joint PDF (9) as a function of ϕ_s and ϕ_r. For mathematical convenience we will be operating on its natural logarithm $\mathcal{L} = \log(f)$. Omitting one non-essential term $(-ML\log(\pi))$, this log-likelihood $\mathcal{L}$ can be written as:

$$\mathcal{L}(\phi_s, \phi_r) = -L\log|\mathbf{\Phi}_{\mathbf{y}}(n)| - L\,\mathrm{tr}\big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{\Phi}_{\mathbf{y}}^{-1}(n)\big], \tag{10}$$

where $\mathrm{tr}[\cdot]$ denotes the matrix trace operator. The MLEs of $\phi_s(n)$ and $\phi_r(n)$ are defined as the coordinates of the global maximum of $\mathcal{L}(\phi_s, \phi_r)$ and can be found by solving a two-dimensional optimization problem:

$$\big(\hat{\phi}_{s,\mathrm{ML}}(n),\, \hat{\phi}_{r,\mathrm{ML}}(n)\big) = \arg\max_{\phi_s,\,\phi_r}\mathcal{L}(\phi_s, \phi_r), \tag{11}$$

where $\hat{\phi}_{s,\mathrm{ML}}(n)$ and $\hat{\phi}_{r,\mathrm{ML}}(n)$ denote the MLEs of $\phi_s(n)$ and $\phi_r(n)$, respectively.

A. Estimator of the target speech PSD

As shown in [36], the MLE of $\phi_s(n)$ can be analytically found by maximizing the likelihood function (10) conditioned on $\hat{\phi}_{r,\mathrm{ML}}(n)$, i.e. by solving a one-dimensional optimization problem (cf. (11)):

$$\hat{\phi}_{s,\mathrm{ML}}(n) = \arg\max_{\phi_s}\mathcal{L}(\phi_s; \hat{\phi}_{r,\mathrm{ML}}).$$

Let $\hat{\mathbf{\Phi}}_{\mathbf{v}}(n) = \hat{\phi}_{r,\mathrm{ML}}(n)\mathbf{\Gamma}_{\mathbf{r}} + \mathbf{\Phi}_{\mathbf{x}}$ denote the MLE of the cross-PSD matrix of the total interference. Then, the MLE $\hat{\phi}_{s,\mathrm{ML}}(n)$ can be written as [36, Appendix B]:

$$\hat{\phi}_{s,\mathrm{ML}}(n) = \mathbf{w}_{\mathrm{MVDR}}^H(n)\big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\mathbf{\Phi}}_{\mathbf{v}}(n)\big]\mathbf{w}_{\mathrm{MVDR}}(n), \tag{12}$$

where

$$\mathbf{w}_{\mathrm{MVDR}}(n) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}}{\mathbf{d}^H\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}} \tag{13}$$

is the weight vector of a minimum variance distortionless response (MVDR) beamformer [37]. The MLE (12) is a function of (is conditioned on) $\hat{\phi}_{r,\mathrm{ML}}(n)$ and can be interpreted as the difference between the estimates of the total PSD and the interference PSD at the output of the MVDR beamformer.

B. Estimator of the late reverberation PSD

Because $\hat{\phi}_{s,\mathrm{ML}}(n)$ and $\hat{\phi}_{r,\mathrm{ML}}(n)$ are analytically related by (12), a one-dimensional, concentrated likelihood function of ϕ_r can be defined as: $\mathcal{L}'(\phi_r) = \mathcal{L}\big(\hat{\phi}_{s,\mathrm{ML}}(\phi_r), \phi_r\big)$. The exact MLE of $\phi_r(n)$ can be found as the argument of the maximum of $\mathcal{L}'(\phi_r)$ [36]. Unfortunately, for the signal model at hand this optimization problem is not easily tractable. Instead of resorting to numerical optimization methods to find the maximum of $\mathcal{L}'(\phi_r)$, we propose a simplified MLE of $\phi_r(n)$ using a modified form of the input signal model.

The modifications consist of two steps. First, we pass the input STFT vector $\mathbf{y}(n)$ through a target-blocking matrix $\mathbf{B} \in \mathbb{C}_{M\times(M-1)}$ defined as [38]:

$$[\mathbf{B}\ \mathbf{b}] = \mathbf{I} - \mathbf{d}(\mathbf{d}^H\mathbf{d})^{-1}\mathbf{d}^H, \tag{14}$$

where $\mathbf{B}$ denotes the first $M-1$ columns and $\mathbf{b}$ denotes the last column of the matrix on the right-hand-side of (14). The columns of $\mathbf{B}$ can be interpreted as a set of $M-1$ target-canceling beamformers, i.e.: $\mathbf{B}^H\mathbf{s}(n) = \mathbf{0}$. Hence, the blocked input signal can be written as: $\mathbf{B}^H\mathbf{y}(n) = \mathbf{B}^H\mathbf{r}(n) + \mathbf{B}^H\mathbf{x}(n)$ (cf. (1)), and its cross-PSD matrix as (cf. (2)):

$$\begin{aligned} E[\mathbf{B}^H\mathbf{y}(n)\mathbf{y}^H(n)\mathbf{B}] &= \mathbf{B}^H\mathbf{\Phi}_{\mathbf{y}}(n)\mathbf{B} \\ &= \mathbf{B}^H\mathbf{\Phi}_{\mathbf{r}}(n)\mathbf{B} + \mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B}. \end{aligned}$$

The second modification of the signal model has the objective of diagonalizing $\mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B}$, i.e. the additive noise component of the blocked input cross-PSD matrix. To that end, we use a whitening matrix $\mathbf{D} \in \mathbb{C}_{(M-1)\times(M-1)}$ and define it as the Cholesky factor of the inverse of $\mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B}$:

$$\mathbf{D}\mathbf{D}^H = (\mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B})^{-1}. \tag{15}$$

It is necessary to assume that $\mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B}$ is full rank. N.b.: it is sufficient that real (and, therefore, noisy) microphones are used in the array to guarantee that $\mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B}$ is full rank, even if

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

the other noise types contributing to $\mathbf{x}(n)$ (e.g. ambient noise) do not by themselves result in a full rank cross-PSD matrix.

The blocked and whitened signal is given by $\tilde{\mathbf{y}}(n) = \mathbf{D}^H\mathbf{B}^H\mathbf{y}(n)$ and its cross-PSD matrix can be found as:

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \mathbf{D}^H\mathbf{B}^H\mathbf{\Phi}_{\mathbf{y}}(n)\mathbf{B}\mathbf{D} = \phi_r(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}} + \mathbf{I}, \tag{16}$$

where $\mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{D}^H\mathbf{B}^H\mathbf{\Gamma}_{\mathbf{r}}\mathbf{B}\mathbf{D}$.

As a result of the described modifications, the matrix $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ exhibits a useful feature: its eigenvectors are the same as that of the matrix $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$. Equivalently, the eigendecompositions of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ use the same unitary matrix $\mathbf{U}$:

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \mathbf{U}\mathbf{\Lambda}_{\mathbf{\Phi}}(n)\mathbf{U}^H, \quad \mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{U}\mathbf{\Lambda}_{\mathbf{\Gamma}}\mathbf{U}^H, \tag{17}$$

where the orthonormal columns of $\mathbf{U}$ are the eigenvectors, and where $\mathbf{\Lambda}_{\mathbf{\Phi}}(n)$ and $\mathbf{\Lambda}_{\mathbf{\Gamma}}$ are diagonal matrices of the eigenvalues of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$, respectively. Because $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ is constant, so are $\mathbf{U}$ and $\mathbf{\Lambda}_{\mathbf{\Gamma}}$. Due to (16), $\mathbf{\Lambda}_{\mathbf{\Phi}}(n)$ and $\mathbf{\Lambda}_{\mathbf{\Gamma}}$ are related as:

$$\mathbf{\Lambda}_{\mathbf{\Phi}}(n) = \phi_r(n)\mathbf{\Lambda}_{\mathbf{\Gamma}} + \mathbf{I}. \tag{18}$$

Equivalently: $\lambda_{\mathbf{\Phi},m} = \phi_r(n)\lambda_{\mathbf{\Gamma},m} + 1$, where $\lambda_{\mathbf{\Phi},m}$ and $\lambda_{\mathbf{\Gamma},m}$ denote the m-th eigenvalue of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$, respectively.

Using the blocked and whitened signal model (16) we can formulate a new and simplified log-likelihood of ϕ_r. It has a form analogous to (10) with the input cross-PSD matrix and its estimate substituted by their blocked and whitened counterparts $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)$:

$$\mathcal{L}^{''}(\phi_r) = -L\log|\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)| - L\,\mathrm{tr}\big[\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big]. \tag{19}$$

The proposed MLE of ϕ_r is defined as: $\hat{\phi}_r(n) = \arg\max_{\phi_r}\mathcal{L}^{''}(\phi_r)$. To find $\hat{\phi}_r(n)$ we must first find the derivative of $\mathcal{L}^{''}(\phi_r)$ with respect to ϕ_r. We compute it by using the fact that for any invertible matrix $\mathbf{A}(\theta)$ the following identities hold ($\mathbf{A}(\theta)$ is a function of θ) [39], [40]:

$$\frac{\mathrm{d}\log|\mathbf{A}(\theta)|}{\mathrm{d}\theta} = \mathrm{tr}\left[\mathbf{A}^{-1}(\theta)\frac{\mathrm{d}\mathbf{A}(\theta)}{\mathrm{d}\theta}\right],$$

$$\frac{\mathrm{d}\,\mathrm{tr}\left[\mathbf{A}^{-1}(\theta)\mathbf{Z}\right]}{\mathrm{d}\theta} = -\,\mathrm{tr}\left[\mathbf{A}^{-1}(\theta)\frac{\mathrm{d}\mathbf{A}(\theta)}{\mathrm{d}\theta}\mathbf{A}^{-1}(\theta)\mathbf{Z}\right].$$

We also note that the derivative of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ with respect to ϕ_r is equal to $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ (cf. (16)). The (known) result is [36, Eq. (2)]:

$$\frac{\mathrm{d}\mathcal{L}^{''}(\phi_r)}{\mathrm{d}\phi_r} = -L\,\mathrm{tr}\big[\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}} - \mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}}\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big]. \tag{20}$$

The proposed estimator is found by setting (20) to zero and solving for ϕ_r. To do so, we re-write (20) using (17):

$$\mathrm{tr}\big[\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1}(n)\mathbf{\Lambda}_{\mathbf{\Gamma}} - \mathbf{\Lambda}_{\mathbf{\Phi}}^{-1}(n)\mathbf{\Lambda}_{\mathbf{\Gamma}}\mathbf{\Lambda}_{\mathbf{\Phi}}^{-1}(n)\mathbf{U}^H\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}\big] = 0.$$

Exploiting the diagonal structure of the involved matrices and using (18), this can be written as:

$$\sum_{m=1}^{M-1}\left[\frac{\lambda_{\mathbf{\Gamma},m}}{(\phi_r\lambda_{\mathbf{\Gamma},m}+1)} - \frac{\lambda_{\mathbf{\Gamma},m}\,g_m(n)}{(\phi_r\lambda_{\mathbf{\Gamma},m}+1)^2}\right] = 0, \tag{21}$$

where $g_m(n)$ denotes the m-th diagonal element of $\mathbf{U}^H\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}$. It can be seen that (21) is a sum of $2(M-1)$ rational terms. By converting all these terms to a common denominator ($\prod_{k=1}^{M-1}(\phi_r\lambda_{\mathbf{\Gamma},m}+1)^2$), taking only the resulting numerators into account, and some additional simplifications, (21) can be expressed as a sum of $M-1$ polynomials in ϕ_r:

$$p(\phi_r) = \sum_{m=1}^{M-1} p_m(\phi_r), \quad \text{where} \tag{22}$$

$$p_m(\phi_r) = \underbrace{\left(\phi_r - \frac{g_m(n)-1}{\lambda_{\mathbf{\Gamma},m}}\right)}_{\text{order }1}\underbrace{\prod_{k=1}^{M-1,\,k\neq m}\left(\phi_r + \frac{1}{\lambda_{\mathbf{\Gamma},k}}\right)^2}_{\text{order }2(M-2)}.$$

The polynomial $p(\phi_r)$ is of odd order: $2M-3$. Hence, at least 1 and at most $2M-3$ real roots of $p(\phi_r)$ exist. When more than one real root of $p(\phi_r)$ exists, the one yielding the highest value of the likelihood (19) must be chosen as the MLE $\hat{\phi}_r(n)$.

For convenience, a pseudo-code representation of the algorithm for computing the proposed PSD estimators is provided in Figure 1. As we show in Appendix A, usually only one real root of $p(\phi_r)$ exists. Therefore, in most cases the condition in Figure 1, line 13 is satisfied, and it is not necessary to compute the numerical value of the likelihood (19).

In general, numerical methods must be applied to find the roots of $p(\phi_r)$ as no closed-form solution appears obtainable. For microphone arrays with few microphones, such as often found in hearing aids, this is computationally trivial. For large microphone arrays, solving (22) may become problematic in applications where computing power is limited.

The proposed late reverberation PSD estimator $\hat{\phi}_r(n)$ is the exact MLE of $\phi_r(n)$ *in the blocked signal domain* (16) (to within the precision of the root-finding algorithm). However, numerical simulations indicated that $\hat{\phi}_r(n)$ is not equal to

1: Define: $\mathbf{d}$, $\mathbf{\Gamma}_{\mathbf{r}}$, $\mathbf{\Phi}_{\mathbf{x}}$
2: $[\mathbf{B}\ \mathbf{b}] = \mathbf{I} - \mathbf{d}(\mathbf{d}^H\mathbf{d})^{-1}\mathbf{d}^H$ (14)
3: $\mathbf{D}\mathbf{D}^H = (\mathbf{B}^H\mathbf{\Phi}_{\mathbf{x}}\mathbf{B})^{-1}$ (15)
4: $\mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{D}^H\mathbf{B}^H\mathbf{\Gamma}_{\mathbf{r}}\mathbf{B}\mathbf{D}$ (16)
5: $\mathbf{U}\mathbf{\Lambda}_{\mathbf{\Gamma}}\mathbf{U}^H = \mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ such that: $\mathbf{U}\mathbf{U}^H = \mathbf{I}$ (17)
6: $\lambda_{\mathbf{\Gamma},m} = [\mathbf{\Lambda}_{\mathbf{\Gamma}}]_{m,m}$
7: **for all** n **do**
8: Define: $\mathbf{y}(n)$
9: Update: $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ (8)
10: $g_m(n) = [\mathbf{U}^H\mathbf{D}^H\mathbf{B}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{B}\mathbf{D}\mathbf{U}]_{m,m}$
11: Define: $p(\phi_r)$ (22)
12: $\mathcal{P}(n) = \{\phi_r : p(\phi_r) = 0\}$
13: **if** $\big|\mathcal{P}(n)\big| = 1$ **then**
14: $\hat{\phi}_r(n) = \{\mathcal{P}(n)\}$
15: **else**
16: $\hat{\phi}_r(n) = \arg\max_{\phi_r\in\mathcal{P}(n)}\mathcal{L}^{''}(\phi_r)$ (19)
17: **end if**
18: $\hat{\mathbf{\Phi}}_{\mathbf{v}}(n) = \hat{\phi}_r(n)\mathbf{\Gamma}_{\mathbf{r}} + \mathbf{\Phi}_{\mathbf{x}}$
19: $\mathbf{w}_{\mathrm{MVDR}}(n) = \hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}\big[\mathbf{d}^H\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}\big]^{-1}$ (13)
20: $\hat{\phi}_s(n) = \mathbf{w}_{\mathrm{MVDR}}^H(n)\big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\mathbf{\Phi}}_{\mathbf{v}}(n)\big]\mathbf{w}_{\mathrm{MVDR}}(n)$ (12)
21: **end for**

Fig. 1. A pseudocode representation of the proposed PSD estimation method. The presented routine is to be applied in all frequency bins (possibly in parallel). The set of roots of the polynomial $p(\phi_r)$ in the n-th time frame is denoted as $\mathcal{P}(n)$, with $|\mathcal{P}(n)|$ being its cardinality (number of elements). Relevant equation numbers are provided for cross-reference.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

the MLE $\hat{\phi}_{r,\mathrm{ML}}(n)$ defined in (11), i.e. in the unmodified signal domain (7). This is due to the loss of information about the signal induced by the blocking operation. Additionally, the target speech PSD estimator computed according to (12) but conditioned on $\hat{\phi}_r(n)$ instead of $\hat{\phi}_{r,\mathrm{ML}}(n)$ is not equal to the exact MLE $\hat{\phi}_{s,\mathrm{ML}}(n)$. Therefore, both proposed PSD estimators are only approximations of the true MLE in the unmodified signal domain. Nevertheless, experimental results reported in Section IV show that the loss of the estimation performance is very small.

C. Estimator of the late reverberation PSD for $\mathbf{x}(n) = \mathbf{0}$

A special case of the proposed late reverberation PSD estimator can be derived for signals where $\mathbf{x}(n) = \mathbf{0}$. Because $\mathbf{\Phi}_{\mathbf{x}} = \mathbf{0}$, the whitening operation is undefined and must be omitted. It follows, that (16) has to be re-written as $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \phi_r(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$. Using this in (20) a new equation for the MLE is found:

$$\phi_r^{-1}(n)\,\mathrm{tr}\big[\mathbf{I} - \mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big] = 0.$$

Unlike in the general scenario, in this special case a closed form solution for the MLE exists:

$$\hat{\phi}_{r|\mathbf{x}=\mathbf{0}}(n) = \frac{1}{M-1}\,\mathrm{tr}\big[\mathbf{\Gamma}_{\tilde{\mathbf{r}}}^{-1}\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big]. \tag{23}$$

This expression can be recognized as the multi-microphone noise PSD estimator proposed in [38]. In [16] this estimator has been shown to be minimum variance unbiased (MVU). Furthermore, (an equivalent form of) the estimator (23) was used for late reverberation PSD estimation in an earlier paper [18] by the authors of this study.

Although the assumption that $\mathbf{x}(n) = \mathbf{0}$ often does not hold in practical applications, it is approximately satisfied in scenarios where the additive noise $\mathbf{x}(n)$ is negligible compared to the late reverberation $\mathbf{r}(n)$. In some applications, the benefits of using a closed-form estimator like (23) may outweigh the benefits of modeling the signal more accurately.

D. Estimator of the late reverberation PSD for $M = 2$

Another special case may be considered for devices with only two microphones, such as some hearing aids, smartphones, and laptops. Because the blocking matrix reduces the dimensionality of the signal by one, all vectors and matrices involved in the estimation of $\phi_r(n)$ degenerate into scalars. Then, the polynomial (22) degenerates into a linear equation which is easily solved:

$$\hat{\phi}_{r|M=2}(n) = \frac{g(n)-1}{\lambda_{\mathbf{\Gamma}}} = (\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n) - \mathbf{\Phi}_{\tilde{\mathbf{x}}})\mathbf{\Gamma}_{\tilde{\mathbf{r}}}{}^{-1}. \tag{24}$$

Note that this equation is composed of scalars; we maintain the bold print for the sake of notation continuity. For $M = 2$, the proposed late reverberation PSD estimator (24) and the one proposed by Braun and Habets in [17] are equivalent (can be written as identical equations).

IV. Evaluation of the Proposed PSD Estimator in Terms of the Normalized Mean Squared Error

In this section we evaluate the proposed PSD estimator and compare it with the estimator proposed by Braun and Habets [17]. As the performance metric we use the normalized mean-squared error (MSE) of estimation defined as:

$$\mathrm{nMSE}_{\phi_s} = \frac{E\big[(\hat{\phi}_s - \phi_s)^2\big]}{\phi_s^2}, \quad \mathrm{nMSE}_{\phi_r} = \frac{E\big[(\hat{\phi}_r - \phi_r)^2\big]}{\phi_r^2}. \tag{25}$$

Because the proposed PSD estimators are not of closed form, in general it is not possible to compute their MSE analytically. Instead, we measure the MSE achieved by the considered PSD estimators in an experiment involving a test signal simulating reverberant and noisy speech. Because the proposed estimators lack closed form, we were only able to numerically verify their unbiasedness. Unbiasedness of the estimators from [17] can be shown analytically (proof omitted). As a result, the MSE of all considered PSD estimators is equal to their variance.

In the special case when the input signal contains no additive noise component ($\mathbf{x}(n) = \mathbf{0}$), the proposed PSD estimators and their MSE can be found analytically. For this restricted scenario it is also possible to analytically find the MSE of the estimators in [17]. In Appendix B we show that in the noise-free scenario the MSE of the proposed estimators is always lower than (or equal to) that of the estimators in [17].

A. Experimental setup

In the present experiment, the goal was to measure and compare the performance of the considered estimators in a synthetic scenario where all the assumptions made in Section II are precisely met. Thus, in each iteration of the experiment a test signal consisting of 25000 STFT sample vectors $\mathbf{y}(n)$, independently drawn from a circularly-symmetric, multivariate complex Gaussian distribution, was used. The covariance matrix of that distribution was modeled according to (7) (i.e. simulating a cross-PSD matrix of a reverberant and noisy speech signal) with known and constant ϕ_s and ϕ_r. Component $\mathbf{s}(n)$ was modeled using a realistic RTF vector $\mathbf{d}$, measured in an anechoic chamber using microphones of two hearing aids placed on the ears of a head and torso acoustic simulator (HATS) and a loudspeaker positioned in front of the HATS. Each of the two behind-the-ear hearing aids had two microphones spaced 1 cm apart, resulting in the total number of microphones $M = 4$. Component $\mathbf{r}(n)$ was modeled using a normalized cross-PSD matrix $\mathbf{\Gamma}_{\mathbf{r}}$ measured in a simulated cylindrically isotropic sound field using the same microphone array as before. The cross-PSD matrix of the component $\mathbf{x}(n)$ was modeled as a scaled identity matrix. Both evaluated algorithms were set to estimate the input covariance matrix (8) using the $L = 10$ most recent time frames.

The simulation experiment was repeated for two different conditions. In the first one, the MSE of the PSD estimation was evaluated as a function of frequency, and the values of ϕ_s and ϕ_r were fixed to result in a speech-to-reverberation ratio (SRR) of 0 dB (averaged over all microphones). In the second condition, the MSE was evaluated as a function of the SRR and the frequency was fixed to 1500 Hz. In both conditions, the additive noise component $\mathbf{x}(n)$ was scaled such that its power was 10 dB lower than the power of the component $\mathbf{r}(n)$ (averaged over all microphones).

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

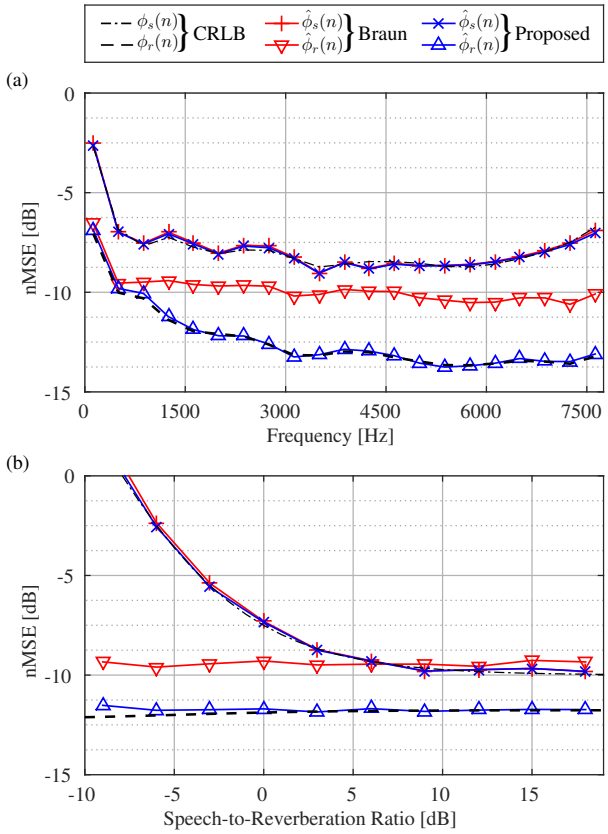


Fig. 2. Normalized MSE of PSD estimation of the proposed PSD estimators (Proposed) and the PSD estimators from [17] (Braun) as a function of: (a) frequency (SRR: 0 dB), (b) SRR (frequency: 1500 Hz). $M = 4$, $L = 10$.

B. Experimental results

Results obtained in the two described conditions are presented in Figures 2a and 2b, respectively. These results are complemented by Cramér-Rao lower bounds (CRLBs) which set a theoretical bound on the lowest possible variance any unbiased estimator of $\phi_s(n)$ and $\phi_r(n)$ can achieve in the considered signal model (7). We outline the derivation of the CRLBs in Appendix C.

From Figures 2a and 2b it may be observed that in all experimental conditions the target speech and the late reverberation PSD estimators proposed in this study (labeled as "Proposed") achieve lower MSE than the corresponding estimators from [17] ("Braun"). The difference between the MSEs yielded by the late reverberation PSD estimators was substantial. However, the difference between the two target speech PSD estimators was very small in virtually all conditions. This was expected because the two target speech estimators are conditioned on different late reverberation PSD estimators but are otherwise identical [1].

As shown in Figure 2a, the late reverberation PSD estimator by Braun and Habets achieved MSEs close to the CRLB only for frequencies below 1 kHz. For higher frequencies the MSE of estimation was up to 3.5 dB higher than the CRLB. The proposed late reverberation PSD estimator achieved MSEs close to the CRLB at all analyzed frequencies and SRRs. It is worth highlighting that this has been accomplished despite the simplifications (14)–(19) of the signal model and the likelihood function used in the derivation of the proposed estimator. It follows, that even the exact MLEs defined using the unmodified signal model (11), or any other unbiased estimator based on (11), could at best perform only slightly better that the proposed simplified method. The steep rise of the MSEs and the CRLBs for low frequencies is due to the wavelength becoming much larger than the dimensions of the array. This results in an increasing correlation of the microphone signals, which limits the attainable gain from averaging between the microphones.

The performance difference between the two compared late reverberation PSD estimators is substantial despite the fact that both estimators are derived using the maximum likelihood method and are based on similar signal models. The specific cause of this difference is the likelihood function used in [17]. This likelihood is based on the assumption that real and imaginary parts of all entries of the blocked sample cross-PSD matrix $\mathbf{B}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{B}$ are mutually independent Gaussians with equal variances. However, since sample covariance matrices are Hermitian, entries that are symmetric with respect to the main diagonal are complex conjugate pairs (and, hence, not independent). Furthermore, the distribution of diagonal elements of sample covariance matrices has a positive support (i.e. they are not Gaussian), and, generally, the elements of sample covariance matrices can have different variances. In the proposed method the likelihood function (19) is defined directly on the (modified) input signal STFT vector and a more realistic assumption on its PDF. Despite the simplifications of the signal model, this results in nearly optimal performance.

As expected, and as can be observed in Figure 2b, negative SRRs resulted in a much higher target speech PSD estimation MSE (and CRLB) than positive SRR values. Because both "Braun" and "Proposed" late reverberation PSD estimators are based on the blocked version of the input signal, their theoretical performance does not depend on the target speech component and, hence, the SRR.

V. Evaluation of an MWF based on the proposed PSD estimator: objective performance measures

The proposed PSD estimator and the estimator in [17] are both primarily intended for use with a multi-channel Wiener filter (MWF) for joint speech dereverberation and denoising. Therefore, it is of interest to evaluate the influence the PSD estimators have on the performance of the MWF. To this end, we conducted an experiment where realistically simulated reverberant and noisy speech signals were processed by the MWF based on either the proposed or the competing PSD estimator from [17]. The speech dereverberation and denoising performance of the two versions of the MWF was measured and compared in terms of the frequency-weighted segmental SNR (FWSegSNR) [19], perceptual evaluation of speech quality (PESQ) [20] measure, mean noise attenuation (NA), mean reverberation attenuation (RA), and speech-to-speech-distortion ratio (SNR-S) [21], [22].

A. Experimental setup

Both versions of the MWF were implemented as a concatenation of an MVDR beamformer and a single-channel Wiener

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

TABLE I
BASIC ACOUSTIC PARAMETERS OF THE REVERBERANT CONDITIONS

Room	T_{60} [s]	C_{50} [dB]	DRR [dB]
Bathroom	0.8	5.2	−10.1
Cellar	1.2	5.7	2.2
Staircase	2.3	11.0	4.1
Office	1.4	8.8	2.3
Auditorium	1.3	13.4	5.2
Isotropic	1.0	4.7	−0.4

post-filter. The MVDR beamformer coefficients $\mathbf{w}_{\mathrm{MVDR}}(n)$ were calculated according to (13) with the estimate of the total interference cross-PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{v}}(n)$ based on $\hat{\phi}_r(n)$. The output signal $\hat{s}(n)$ of the MWF was computed as:

$$\hat{s}(n) = \left[\frac{\hat{\phi}_{s_o}(n)}{\hat{\phi}_{s_o}(n) + \hat{\phi}_{v_o}(n)}\right] \mathbf{w}_{\mathrm{MVDR}}^{H}(n)\mathbf{y}(n),$$

where:

$$\hat{\phi}_{s_o}(n) = \hat{\phi}_s(n),$$
$$\hat{\phi}_{v_o}(n) = \mathbf{w}_{\mathrm{MVDR}}^{H}(n)\hat{\mathbf{\Phi}}_{\mathbf{v}}(n)\mathbf{w}_{\mathrm{MVDR}}(n),$$

denote the estimated PSDs of the target speech and the total interference at the output of the MVDR beamformer, respectively.

Contrary to the experiment in Section IV, in this experiment the goal was to compare the performance of the estimators in realistic conditions (violating some of the assumptions made in Section II) and for a practical application (in hearing aids). Thus, the microphone signals were generated using real speech recordings from the TIMIT database [41] and several reverberant and noisy conditions based on real room impulse responses (RIRs) and simulated microphone noise. Specifically, we used a subset of the TIMIT database containing 17 minutes of male and female speech. TIMIT sentences were convolved with RIRs measured in five real rooms using a microphone array composed of two behind-the-ear hearing aids on the HATS (same as described in Section IV). The reverberation time T_{60}, clarity index C_{50}, and the direct-to-reverberation ratio (DRR) of these five RIRs are presented in Table I. The rooms are denoted by their function as: "Bathroom", "Cellar", "Staircase", "Office", and "Auditorium", and represent a wide range of acoustic conditions a hearing aid user might encounter. A sixth, synthetic impulse response, where the reverberation was modeled as perfectly cylindrically isotropic was also used and is denoted as "Isotropic". To simulate the electrical noise that is generated by real-world microphones, spatially white and spectrally pink noise was added to the convolved speech signals. The simulations were repeated for two levels of that noise, such that at the frequency of 1 kHz the noise PSD was either 20 dB or 30 dB lower than the PSD of the target speech material.

The sampling frequency of the simulated time-domain signals was 16 kHz and the STFT length was set to 8 ms ($T = 128$ samples). This ensured a processing delay of the MWF shorter than 10 ms, which is a requirement for hearing aid systems. A square root Hann window with 50% overlap between frames was used in the analysis filterbank and in the overlap-add inverse STFT procedure used for re-synthesis of the output signal. The input cross-PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ was estimated using recursive averaging (equivalent to exponential weighting) with a time constant of 50 ms (instead of the moving average smoothing used in (8)). For processing of the signals simulated using each of the six impulse responses, the MWF algorithm and the PSD estimators were implemented using RTF vectors $\mathbf{d}$ extracted from the first 2.5 ms of the RIR in question (i.e. $t_s = 40$ samples). For the RIRs used in the experiment this resulted in $\mathbf{d}$ being based solely on the direct path response. It follows that the early reflections (particularly strong in the "Bathroom" condition) were left unaccounted for in the assumed signal model. This resulted in a realistic mismatch between the used RTF vector $\mathbf{d}$ and the actual RTF of the target speech component in the simulated signals. Moreover, because $\mathbf{d}$ depended only on the direction of the target source, the assumption that $\mathbf{d}$ is known became more realistic. The normalized cross-PSD matrix $\mathbf{\Gamma}_{\mathbf{r}}$ of the cylindrically isotropic sound field was measured *a priori* in a simulated cylindrically isotropic sound field using the same microphone array as used for measuring the RIRs. In none of the five real rooms, the late reverberation was truly isotropic which, again, resulted in a realistic mismatch between the assumed model and the actual structure of the signal. Only in the "Isotropic" condition the model of the target signal and of the reverberation component was accurate.

B. Experimental results

The results of the experiment are presented in Figure 3. Performance scores obtained by using the MWF based on the proposed PSD estimator ("Proposed") and on the estimator proposed in [17] ("Braun") are included along the scores obtained by using only the MVDR part of the two MWFs ("MVDR"). The scores calculated from the unprocessed input signal ("Input") are included for reference. The relative performance between the proposed and the competing MWFs and MVDRs was the same for the higher and the lower microphone noise level setting. Thus, we show only the results obtained for the −30 dB setting, which better corresponds to the typical microphone noise and speech levels encountered in practice.

In all simulated conditions, both versions of the MWF and the MVDR beamformer succeeded in improving FWSegSNR and PESQ. The RA was also always positive, indicating algorithms' effectiveness in reducing the reverberation. However, the NA scores were exclusively negative, indicating that *on average* all algorithms amplified the noise. This was expected because the NA measure (as well as RA and SNR-S) only accounts for those signal segments where the target speech component is active (see [22]). Because the MVDR beamformer adapts to jointly suppress the noise and the late reverberation, it was expected that during speech and, hence, reverberation activity the noise component will have negligible impact on the MVDR coefficients. Naturally, during speech and reverberation absence the MVDR beamformers adapted to primarily reduce the noise component.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

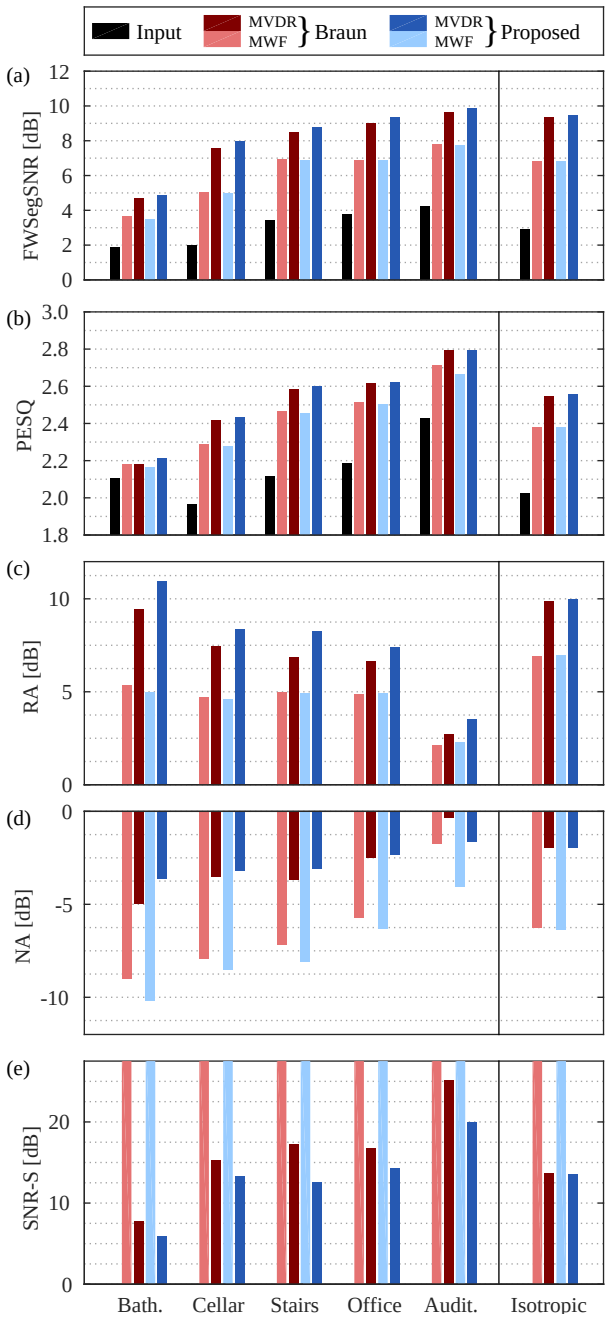


Fig. 3. (a) FWSegSNR, (b) PESQ, (c) RA, (d) NA, and (e) SNR-S scores obtained by using MWFs and MVDR beamformers based on the PSD estimators from [17] (denoted "Braun") and the proposed estimators (denoted "Proposed"). Scores obtained from the unprocessed input signal ("Input") are also included.

The total improvement of the FWSegSNR and PESQ over the unprocessed signal was greatest in the "Isotropic" and lowest in the "Bathroom" condition. This difference can be explained by the fact that in the "Isotropic" condition $\mathbf{\Gamma_r}$ and $\mathbf{d}$ accurately characterized the actual input signal whereas in the "Bathroom" condition the input signal did not match the assumed model. Prominent early reflections present in the "Bathroom" condition were unaccounted for and resulted in substantial leakage of the early speech component into the output of the blocking matrix. This lead to an overestimation of the late reverberation PSD and, ultimately, over-suppression and distortion of the target speech in the post-filter (note the very high RA and very low SNR-S in this condition).

The differences in the performance scores obtained by using the MVDR beamformers based on "Braun" and "Proposed" PSD estimators were very small. Although the performance difference between the MWFs was somewhat bigger, it was still only moderately large. For example in the "Isotropic" condition "Proposed" MWF performed only marginally better than "Braun". In the remaining conditions the difference is larger and the proposed method appears to systematically perform better than "Braun". This suggests that the proposed estimators are more robust to the mismatch between the signal model and its actual structure than the estimators from [17]. The SNR-S measure indicated stronger speech distortion when using the proposed PSD estimator. While being a clear disadvantage, low SNR-S scores are counterbalanced by higher RA and NA values.

Informal listening tests indicated similar trends as the objective performance measures. In all simulated conditions, the MWFs resulted in a decrease of the perceived reverberation and noise strength. The MVDR beamformers also reduced the amount of perceived interference, but to a smaller degree. Differences between "Braun" and "Proposed" MWFs were barely perceivable; only in specific signal scenarios a small increase in the audibility of musical noise could be noticed in the "Braun" MWF output. This was expected because the PSD estimators from [17] have higher MSE.

We close this section by noting that (when implemented in Matlab) the proposed algorithm resulted in computation times roughly 1.7 times longer than the algorithm from [17].

VI. Evaluation of an MWF based on the proposed PSD estimator: speech intelligibility improvement

In addition to the two experiments with technical/objective performance measures in Sections IV and V, we conducted a speech intelligibility (SI) test with human subjects. Dantale II [42] sentences were presented via Sennheiser HD280 pro headphones to 20 subjects, who were requested to select the words they heard from an on-screen list of options [43].

A. Experimental setup

Stimuli were constructed as follows. The Dantale II sentences were concatenated with 2 s of silence before and after the utterance and underwent the same realistic reverberation simulation as in the "Cellar" condition in Section V, corresponding to a frontal position of the target source at a distance of 2 m. Since the SI in this condition was close to 100%, speech-like interference consisting of randomly shifted and superimposed copies of the international speech test signal (ISTS) [44] was added to the reverberated Dantale II sentences. The interferer signals were convolved with 5 RIRs recorded in the same room as the target RIR but with the sound source positioned at 90°, 135°, 180°, −135°, and at −90° azimuth angle, at 2 m distance. Each of the simulated babble talkers radiated the same power as the target source. Different levels of SI were achieved by manipulating the DRR

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

of the the target source RIR. This was done by attenuating the direct part of the target speech while keeping the rest of the signal intact. In this way the DRR was offset by 0, −4, −8, and −12 dB from its original value of 2.2 dB (cf. Table I).

The RTF vector $\mathbf{d}$ and the cross-PSD matrix $\mathbf{\Gamma_r}$ were obtained in the same way, and the simulated microphone signals were processed using the same algorithms as in Section V. The additional noise cross-PSD matrix $\mathbf{\Phi_x}$ was estimated from the first 2 s of each stimuli, which was known to contain only the reverberated ISTS babble and the simulated microphone noise. In order to provide correct binaural cues of the target speech, signals presented to each of the subjects' ears were processed by separate instances of the algorithms, each using the front microphone of the corresponding hearing aid as the reference position. In the unprocessed condition ("Input") the signals of the left and right reference microphones were presented to the corresponding ears of the subject. This allowed the subjects to localize the target and the ISTS interferers at their original (simulated) positions and benefit from the binaural advantage [45]. In the processed conditions this was not possible, as all components of the enhanced signals were perceived as coming from the target direction (a known side-effect of using binaural beamformers [46]). To each of the experimental conditions five Dantale II sentences were randomly assigned (independently for each subject). The sentences were processed and then presented to subjects in a randomized order.

B. Experimental results

The word intelligibility obtained in each of the processing conditions was calculated as the percentage of words identified correctly by the subjects and is plotted in Fig. 4 as a function of the DRR offset. In order to interpret these results, we performed a two-way repeated measures ANOVA procedure [47] on the rationalized arcsine-corrected [48] subject mean word intelligibility scores. The effect of the processing type ($F_{4,76} = 232.6$), the DRR offset ($F_{3,57} = 383.8$), and the interaction term ($F_{12,228} = 5.0$) on the measured intelligibility were all found to be significant at the $p < 0.001$ level. Pairwise comparisons of the marginal means revealed that: *a)* each of the algorithms significantly improved the SI over the "Input", *b)* the MWFs outperformed their corresponding MVDR beamformers, and additionally, *c)* the "Proposed MWF" outperformed the "Braun MVDR". The familywise type I error rate was limited to 1% using Bonferroni correction.

The lack of significant differences between the SI obtained with the proposed and the competing PSD estimators was somewhat expected, given the minute instrumental performance differences of the two MWFs and MVDR beamformers obtained in Section V. On the other hand, significant improvement of SI resulting from the post-filters of the two MWFs is apparently in contrast with the general understanding that single channel spectral filters usually fail to increase SI [49]. The fact that the post-filters of the two evaluated MWFs succeeded in improving SI can be explained by the fact that they were computed using information from the multi-microphone signal (contrary to the single channel schemes discussed in [49]).

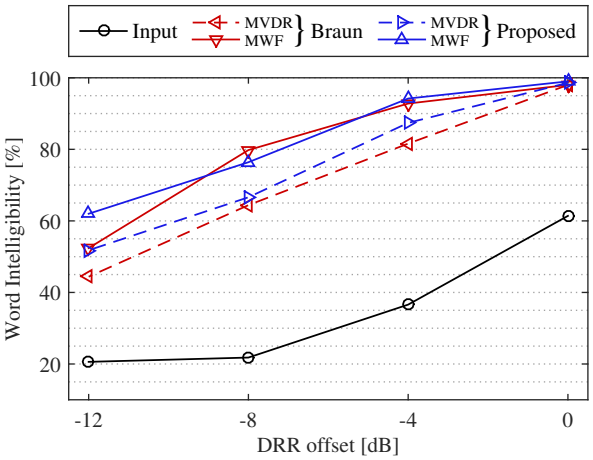


Fig. 4. Word intelligibility scores obtained in the listening test with the RIR from the "Cellar" condition and ISTS interferers, averaged across 20 subjects.

VII. Conclusion

In this paper we have proposed a pair of novel ML-based speech and late reverberation PSD estimators. The proposed method models the interference as consisting of late reverberation and additive noise; in this sense it can be seen as an extension of the method in [18] which only considers the late reverberation. We have numerically demonstrated that the proposed estimator yields lower mean squared error (MSE) of PSD estimation than the method in [17], and that this MSE is very close to the corresponding Cramér-Rao lower bound.

In an experiment with realistically simulated reverberation, we have compared speech dereverberation performance of an MWF based on the proposed estimator and on the estimator from [17]. The proposed estimator generally resulted in higher FWSegSNR, PESQ, RA, and NA scores than the estimator from [17]. However, the SNR-S indicated stronger speech distortion. In terms of speech intelligibility, the MWFs based on both PSD estimators provided statistically significant improvements over the unprocessed signal, but were not significantly different from each other. The output of both MWFs was statistically significantly more intelligible than the output of the corresponding MVDR beamformers.

Evaluation of the proposed algorithm in environments which more severely violate the assumptions made in this paper is an area for future work. In an ongoing study, we evaluate the proposed algorithm's robustness to erroneous estimates of the direction of the target speech arrival. Preliminary results for signals without the noise component have already been published in [50].

VIII. Acknowledgements

The authors would like to thank the reviewers for their thorough and insightful review of this manuscript, and Asger Heidemann Andersen, Oticon A/S, for allowing the use of his software implementation of the Dantale II matrix test.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

APPENDIX A
PROPERTIES OF THE PROPOSED LATE REVERBERATION PSD ESTIMATOR

In the following, we show that in the majority of practical cases the polynomial equation (22) (repeated below for convenience) has exactly one real-valued root and that this root is non-negative. Due to this property, the proposed MLE of $\phi_r(n)$ can be found more easily, as the likelihood (19) does not need to be calculated in order to determine which of the roots of (22) corresponds to the MLE of $\phi_r(n)$.

$$p(\phi_r) = \sum_{m=1}^{M-1} p_m(\phi_r), \quad \text{where} \tag{22}$$

$$p_m(\phi_r) = \underbrace{\left(\phi_r - \frac{g_m(n)-1}{\lambda_{\mathbf{\Gamma},m}}\right)}_{\text{order } 1} \underbrace{\prod_{k=1}^{M-1,\, k\neq m} \left(\phi_r + \frac{1}{\lambda_{\mathbf{\Gamma},k}}\right)^2}_{\text{order } 2(M-2)}.$$

We begin by noting that the order of the polynomial $p(\phi_r)$ depends linearly on the number of microphones M and it is equal to $2M-3$. Because this is always an odd number, at least one root of $p(\phi_r)$ is real. This means that for all possible input signals the proposed method will return a result.

The polynomial $p(\phi_r)$ is a sum of $M-1$ polynomials $p_m(\phi_r)$, and each $p_m(\phi_r)$ has exactly one root of algebraic multiplicity one and exactly $M-2$ roots of multiplicity two (cf. (22)). The double roots of each $p_m(\phi_r)$ are equal to $-\lambda_{\mathbf{\Gamma},k}^{-1}$. These roots are always negative because $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ is assumed positive-definite, i.e. all of its eigenvalues $\lambda_{\mathbf{\Gamma},m}$ are strictly positive. The singular root of each $p_m(\phi_r)$ is equal to $(g_m(n)-1)\lambda_{\mathbf{\Gamma},m}^{-1}$, which is non-negative if and only if $g_m(n) \geq 1$. This condition is expected to be satisfied whenever $\phi_r(n) \geq 0$, because (17), (18):

$$\begin{aligned} E[g_m(n)] &= E\big\{[\mathbf{U}^H \hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}]_{m,m}\big\} \\ &= \phi_r(n)\lambda_{\mathbf{\Gamma},m} + 1. \end{aligned} \tag{A.1}$$

The structure of the component polynomials $p_m(\phi_r)$ allows us to draft their approximate plots in Figure A.1. We note that each of the component polynomials attains a value of zero, but it does not cross it at the double roots. The $M-1$ double roots are repeated between $p_m(\phi_r)$ but each of them is absent from exactly one of the polynomials (cf. (22)). It follows that the polynomial $p(\phi_r)$ is strictly negative between $-\infty$ and the lowest of the singular roots. Analysis of the derivatives and inflection points of $p(\phi_r)$ leads to a conclusion that given $g_m(n) \geq 1$, the graph of the polynomial $p(\phi_r)$ crosses the abscissa only once at a point between the lowest and the highest of the singular roots of the component polynomials, i.e. it has exactly one real root and it is non-negative. The condition $g_m(n) \geq 1$ can be expected to be satisfied, because in reverberant scenarios $\phi_r(n)$ is almost always positive (cf. (A.1)). Our simulations confirm that; the polynomial (22) has a single positive root in over 99% of cases.

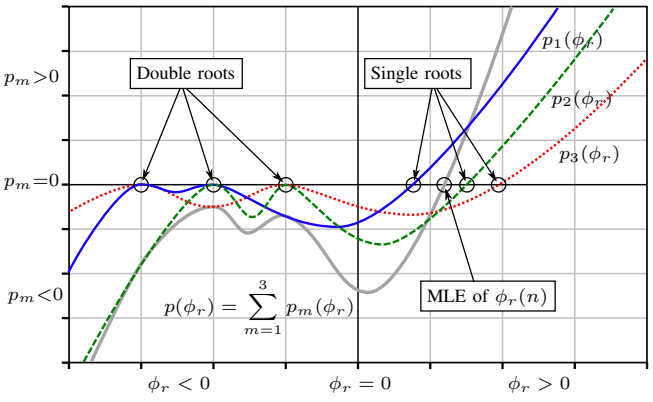


Fig. A.1. Schematic illustration of the polynomial (22) (denoted $p(\phi_r)$) and its $M-1$ components $p_m(\phi_r)$ for $M=4$.

APPENDIX B
THEORETICAL PERFORMANCE OF THE PROPOSED LATE REVERBERATION PSD ESTIMATOR IN NOISE ABSENCE

In this appendix we compare analytical expressions for the mean squared error (MSE) of the PSD estimators proposed in this study and the PSD estimators proposed by Braun and Habets in [17]. This comparison does not appear to be possible in the general case where $\mathbf{x}(n) \neq \mathbf{0}$ because of the lack of a closed-form solution for the proposed late reverberation PSD estimator. Therefore, in this appendix we are restricted to the special case where no additive noise component is present (i.e. $\mathbf{x}(n) = \mathbf{0}$). As shown in Section III-C, in such a signal scenario the proposed late reverberation PSD estimator can be written in closed-form (23).

Since the proposed PSD estimators in the special case of $\mathbf{x}(n) = \mathbf{0}$ are identical to the speech and reverberation PSD estimators proposed by us in [18], the comparison we make in this appendix is equivalent to the one presented in [1]. We outline it in the following for completeness.

The target speech PSD estimator proposed by Braun and Habets in [17] has the same form as the target speech PSD estimator (12) proposed in the present study. The difference between the estimators is that they are conditioned on different late reverberation PSD estimates. Hence, it is sufficient to compare the late reverberation PSD estimators in order to capture the difference between the two PSD estimation methods.

We start the comparison of the late reverberation PSD estimator proposed by Braun and Habets (denoted $\hat{\phi}_{r,\text{Braun}}(n)$) and the one proposed in this study (denoted $\hat{\phi}_{r,\text{Kukl.}}(n)$) by noting that they are both unbiased (proof omitted):

$$E\big[\hat{\phi}_{r,\text{Kukl.}}(n)\big] = \phi_r(n), \quad E\big[\hat{\phi}_{r,\text{Braun}}(n)\big] = \phi_r(n).$$

Therefore, the MSEs of these estimators are identical to their variances.

The variance of the proposed late reverberation PSD estimator (23) can be shown to be equal to (for proof see [16]):

$$\text{var}\big(\hat{\phi}_{r,\text{Kukl.}}(n)\big) = \phi_r^2(n)\frac{1}{L}\frac{1}{M-1}. \tag{B.1}$$

The variance of the late reverberation PSD estimator proposed by Braun and Habets [17] has been previously derived in [1]

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

and can be concisely written as:

$$\mathrm{var}\big(\hat{\phi}_{r,\mathrm{Braun}}(n)\big) = \phi_r^2(n)\frac{1}{L}\frac{1}{M-1}\left(1+\frac{\tilde{\gamma}^2}{\bar{\gamma}^2}\right), \tag{B.2}$$

where $\tilde{\gamma}$ and $\bar{\gamma}$ denote the sample variance and the mean of the squared eigenvalues of the matrix $\mathbf{\Gamma}_{\check{\mathbf{r}}} = \mathbf{B}^H\mathbf{\Gamma}_{\mathbf{r}}\mathbf{B}$, respectively.

Comparing (B.2) and (B.1) and using the fact that $\tilde{\gamma}$ and $\bar{\gamma}$ are non-negative we can conclude that the MSE of $\hat{\phi}_{r,\mathrm{Braun}}(n)$ can be either greater or equal to the MSE of $\hat{\phi}_{r,\mathrm{Kukl.}}(n)$, but can never be lower. The MSEs of these two estimators are equal only when the eigenvalues of $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ are all equal (i.e. when $\tilde{\gamma}^2 = 0$). Since $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ is Hermitian, it follows that for this special case to occur, $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ must be a scaled identity matrix [51]. In all other cases, the proposed late reverberation PSD estimator has lower MSE than the one from [17].

An important observation is that for $M = 2$ the matrix $\mathbf{\Gamma}_{\check{\mathbf{r}}}$ reduces to a scalar, such that $\tilde{\gamma}^2$ is always equal to zero. It follows that for $M = 2$ the proposed late reverberation PSD estimator (23) and the one proposed by Braun and Habets [17] achieve the same MSE. In this case they are, in fact, identical (proof omitted).

APPENDIX C
CRAMÉR RAO LOWER BOUNDS ON PSD ESTIMATION

In this appendix we outline the calculation of the Cramér-Rao lower bounds (CRLBs) included in Figures 2a and 2b. By definition, the CRLBs are equal to the elements of the inverse of the Fisher information matrix (FIM). The i,j-th element of the FIM is defined as follows [52]:

$$\mathcal{I}_{i,j} = -E\left[\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right], \tag{C.1}$$

where $\mathcal{L}$ is the log-likelihood function of the parameter vector $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_p]^T$, given the input data. For a p-parameter signal model the FIM is a $p \times p$ symmetric matrix. For L independent identically distributed circularly-symmetric complex Gaussian observations the i,j-th element of the FIM is found as [52]:

$$\mathcal{I}_{i,j} = L\,\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial\mathbf{\Phi}_{\mathbf{y}}}{\partial\theta_i}\mathbf{\Phi}_{\mathbf{y}}^{-1}\frac{\partial\mathbf{\Phi}_{\mathbf{y}}}{\partial\theta_j}\right], \tag{C.2}$$

where $\mathbf{\Phi}_{\mathbf{y}}$ is the cross-PSD matrix of the input signal. Note that because of the above equation, any invertible linear operation applied to the input signal vector (such as whitening) does not change the FIM or the CRLB.

In the signal model considered in this study (7) there are two unknown parameters ($\boldsymbol{\theta} = [\phi_s,\phi_r]^T$); hence, the FIM is a 2×2 matrix. Using the log-likelihood function (10) in (C.1), or equivalently the cross-PSD matrix (7) in (C.2) we obtain:

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{ss} & \mathcal{I}_{rs} \\ \mathcal{I}_{sr} & \mathcal{I}_{rr} \end{bmatrix} \tag{C.3}$$

$$\mathcal{I}_{ss} = L\,\mathrm{tr}\big[\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{d}\mathbf{d}^H\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{d}\mathbf{d}^H\big], \tag{C.4}$$

$$\mathcal{I}_{rr} = L\,\mathrm{tr}\big[\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{\Gamma}_{\mathbf{r}}\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{\Gamma}_{\mathbf{r}}\big], \tag{C.5}$$

$$\mathcal{I}_{rs} = \mathcal{I}_{sr} = L\,\mathrm{tr}\big[\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{\Gamma}_{\mathbf{r}}\mathbf{\Phi}_{\mathbf{y}}^{-1}\mathbf{d}\mathbf{d}^H\big]. \tag{C.6}$$

Similarly to the proposed PSD estimators, the CRLBs do not appear to be possible to be derived analytically in the general case. For the special case when $\mathbf{x}(n) = \mathbf{0}$, closed-form expressions for the CRLBs can be derived (see e.g. [16]). When $\mathbf{x}(n) \neq \mathbf{0}$ the FIM can be inverted numerically and (by definition) the CRLBs are obtained as:

$$CRLB(\phi_s) = [\mathcal{I}^{-1}]_{1,1}, \tag{C.7}$$

$$CRLB(\phi_r) = [\mathcal{I}^{-1}]_{2,2}. \tag{C.8}$$

The CRLBs included in Figures 2a and 2b were calculated using (C.2)–(C.8) and normalized by the squared parameter of interest (analogous to the normalization of MSEs in (25)).

REFERENCES

[1] A. Kuklasiński *et al.*, "Multi-channel PSD estimators for speech dereverberation – a theoretical and experimental comparison," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Brisbane, Australia, 2015, pp. 91–95.

[2] A. Kuklasiński, S. Doclo, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberant and noisy conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Shanghai, China, 2016, pp. 599–603.

[3] J. S. Bradley *et al.*, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[4] P. A. Naylor and N. D. Gaubitch, "Introduction," in *Speech dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. London, United Kingdom: Springer, 2010, ch. 1, pp. 1–15.

[5] D. Schmid *et al.*, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 8, pp. 1320–1335, Aug 2014.

[6] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multichannel equalization," in *14th Int. Workshop on Acoustic Echo and Signal Enhancement (IWAENC)*, Sept 2014, pp. 139–143.

[7] J. Benesty, S. Makino, and J. Chen, "Introduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer, 2005, ch. 1, pp. 1–8.

[8] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Taylor & Francis, 2007.

[10] K. Lebart *et al.*, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[11] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," *15th Annual Workshop on Circuits, Systems and Signal Processing*, pp. 250–254, 2004.

[12] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[13] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP)*, vol. 5, Apr 1988, pp. 2578–2581.

[14] S. Doclo *et al.*, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Wiley, 2008, pp. 269–302.

[15] ——, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7-8, pp. 636–656, 2007.

[16] J. Jensen and M. S. Pedersen, "Analysis of beamformer-directed single-channel noise reduction system for hearing aid applications," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Brisbane, Australia, 2015, pp. 5728–5732.

[17] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st Eur. Signal Process. Conf.*, Marrakech, Morocco, 2013.

[18] A. Kuklasiński *et al.*, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. 22nd Eur. Signal Process. Conf.*, Lisbon, Portugal, 2014, pp. 61–65.

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

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591

[20] "Perceptual evaluation of speech quality: an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec. P. 862*, 2001.
[21] S. Gustafsson *et al.*, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 245–256, Jul 2002.
[22] J. S. Erkelens *et al.*, "Minimum mean-square error estimation of discrete fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.
[23] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, 2004.
[24] G. W. Elko *et al.*, "Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003, pp. 67–70.
[25] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound Vibration*, vol. 102, no. 2, pp. 217–228, 1985.
[26] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, Jul. 2003.
[27] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, 2005.
[28] J. Jensen *et al.*, "A study of the distribution of time-domain speech samples and discrete fourier coefficients," in *Proc. SPS-DARTS*, vol. 1, 2005, pp. 155–158.
[29] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug 2001.
[30] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.
[31] R. Talmon *et al.*, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1420–1434, 2009.
[32] H. Kuttruff, *Room Acoustics*, 5th ed. Taylor & Francis, 2009.
[33] R. K. Cook *et al.*, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.
[34] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 61–85.
[35] M. Souden *et al.*, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 1072–1077, 2010.
[36] H. Ye and R. D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.
[37] H. Cox *et al.*, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.
[38] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf.*, Bucharest, Romania, 2012, pp. 295–299.
[39] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Nov. 2012. [Online]. Available: http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
[40] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, 2007. [Online]. Available: http://www.janmagnus.nl/misc/mdc2007-3rdedition
[41] J. S. Garofolo *et al.*, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. NIST, 1993.
[42] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, 2003.
[43] E. R. Pedersen and P. M. Juhl, "Speech in noise test based on a ten-alternative forced choice procedure," *Baltic-Nordic Acoustics Meeting*, 2012.
[44] I. Holube *et al.*, "Development and analysis of an international speech test signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, 2010.
[45] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.
[46] S. Doclo *et al.*, "Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2005, pp. 70–73.
[47] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 2011.
[48] G. A. Studebaker, "A rationalized arcsine transform," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.
[49] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, no. 3, pp. 1777–1786, 2007.
[50] A. Kuklasiński *et al.*, "Multi-channel Wiener filter for speech dereverberation in hearing aids – sensitivity to DoA errors," in *Audio Eng. Soc. 60th Int. Conf.*, Leuven, Belgium, 2016.
[51] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.
[52] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, ser. Prentice Hall Signal Processing Series. Prentice-Hall PTR, 1993.



Adam Kuklasiński (S'15) received the B.Sc. degree in audio engineering and the M.Sc. degree in acoustics from the Adam Mickiewicz University, Poznań, Poland, in 2010 and 2012, respectively. He is currently pursuing his Ph.D. degree in digital signal processing at Oticon A/S, Copenhagen, Denmark and at Aalborg University, Denmark. Mr. Kuklasiński is a Marie Skłodowska-Curie fellow in the ITN-DREAMS project and is supervised by prof. Jesper Jensen. His scientific interests include: statistical signal processing, speech dereverberation, and binaural cue preservation in hearing aids.



Simon Doclo (S'95–M'03–SM'13) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Adaptive Systems Laboratory (McMaster University, Canada). From 2007 to 2009 he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group in Leuven, Belgium. Since 2009, he has been a Full Professor at the University of Oldenburg, Germany, and Scientific Advisor for the project group Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical applications, more specifically microphone array processing, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo received the Master Thesis Award of the Royal Flemish Society of Engineers in 1997 (with Erik De Clippel), the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with Marc Moonen) and the IEEE Signal Processing Society 2008 Best Paper Award (with Jingdong Chen, Jacob Benesty, Arden Huang). He was member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2008–2013) and Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013. Prof. Doclo has served as guest editor for several special issues (IEEE Signal Processing Magazine, Elsevier Signal Processing) and is associate editor for the IEEE/ACM Transactions on Audio, Speech and Language Processing and EURASIP Journal on Advances in Signal Processing.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2573591



Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education (UNI•C), Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University. He is Full Professor and heading a research section working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, acoustics, and digital communications. Prof. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing, Elsevier Signal Processing and EURASIP Journal on Advances in Signal Processing, and is currently Associate Editor for the IEEE/ACM Transactions on Audio, Speech and Language Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Section's Signal Processing Chapter. He is member of the Danish Academy of Technical Sciences and was in January 2011 appointed as member of the Danish Council for Independent Research—Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.



Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, at Aalborg University. His main research interests are in the area of acoustic signal processing, including signal detection, estimation, and retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, including, noise reduction solutions, dereverberation algorithms, feedback cancellation algorithms, and perceptual aspects of signal processing.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.