MAXIMUM LIKELIHOOD PSD ESTIMATION FOR SPEECH ENHANCEMENT IN REVERBERANT AND NOISY CONDITIONS

Adam Kuklasiński[*,†], *Simon Doclo*[‡], *Jesper Jensen*[*,†]

[*]Oticon A/S, 2765 Smørum, Denmark
[†]Aalborg University, Department of Electronic Systems, 9220 Aalborg, Denmark
[‡]University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

We propose a novel Power Spectral Density (PSD) estimator for multi-microphone systems operating in reverberant and noisy conditions. The estimator is derived using the maximum likelihood approach and is based on a blocked and pre-whitened additive signal model. The intended application of the estimator is in speech enhancement algorithms, such as the Multi-channel Wiener Filter (MWF) and the Minimum Variance Distortionless Response (MVDR) beamformer. We evaluate these two algorithms in a speech dereverberation task and compare the performance obtained using the proposed and a competing PSD estimator. Instrumental performance measures indicate an advantage of the proposed estimator over the competing one. In a speech intelligibility test all algorithms significantly improved the word intelligibility score. While the results suggest a minor advantage of using the proposed PSD estimator, the difference between algorithms was found to be statistically significant only in some of the experimental conditions.

Index Terms— multi-microphone, PSD estimation, maximum likelihood estimator, isotropic sound field, speech dereverberation.

1. INTRODUCTION

In many speech communication scenarios reverberation and noise degrade the intelligibility and the perceived quality of speech signals. In order to reduce the negative impact of these interferences, various speech enhancement algorithms have been proposed for devices such as hearing aids and conferencing systems. Many of these algorithms require the Power Spectral Densities (PSDs) of the target and the interference signal components. Estimation of these (usually unknown) PSDs is the topic of this paper.

One general class of PSD estimators used for speech enhancement is based on the assumption that the interference is more stationary than the target. This enables the use of PSD estimators based on e.g. Voice Activity Detection (VAD) [1] or on Minimal Statistics (MS) [2]. In many scenarios (notably for reverberation) the interference is highly non-stationary, which requires the use of PSD estimators other than those solely based on VAD or MS.

The class of PSD estimators that is of interest in this paper differentiates between the target and the interference based on their *spatial* characteristics. These methods can be implemented only in multi-microphone systems, but allow for simultaneous on-line estimation of the target and the interference PSDs. Many modern speech communication devices have more than one microphone, which makes spatial processing schemes of particular interest.

In this study we focus on PSD estimators that are suitable for reverberant and noisy conditions; specifically, on the maximum likelihood estimators proposed in [3,4]. In both these estimators the signal components are assumed to be Gaussian and uncorrelated, and reverberation is modeled as isotropic. The estimator in [3] also accounts for an additive noise component which is assumed to be stationary. The estimator in [4] does not include the additional noise component, but it follows the Maximum Likelihood Estimation (MLE) methodology more closely than [3]. As shown in [5], in the absence of the additive noise, [4] performs better then [3]. However, when the additive noise is present, [4] generally performs worse than [3].

In this paper we first derive an extended version of the estimator from [4], which explicitly includes the reverberation and the additive stationary noise in the signal model, while still attempting to closely follow the MLE methodology. We then evaluate four speech dereverberation algorithms based on the proposed PSD estimator and on the estimator from [3]. The evaluation is based on instrumental performance measures and on a speech intelligibility test.

2. SIGNAL MODEL AND STATISTICAL ASSUMPTIONS

Consider an array of M microphones in a reverberant room where a single talker is active. Due to the wide-band and non-stationary nature of speech signals, we implement the proposed method in the Short Time Fourier Transform (STFT) domain. For notational conciseness the complex-valued STFT coefficients of all microphone signals are stacked in an $M \times 1$ vector $\mathbf{y}(k,n)$, where k is the frequency bin index and n is the time frame index. We assume that $\mathbf{y}(k,n)$ is uncorrelated across n and k. This allows us to omit the frequency bin index k in the following without loss of generality.

We model the vector $\mathbf{y}(n)$ as a sum of three components: the target signal $\mathbf{s}(n)$ (i.e. the direct speech and early reverberation), late reverberation $\mathbf{r}(n)$, and additive noise $\mathbf{x}(n)$,

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

We define the cross-PSD matrix of the input $\mathbf{y}(n)$ as $\mathbf{\Phi}_{\mathbf{y}}(n) = E[\mathbf{y}(n)\mathbf{y}^H(n)]$, where $\cdot^H$ denotes the Hermitian transpose. Assuming that $\mathbf{s}(n)$, $\mathbf{r}(n)$, and $\mathbf{x}(n)$ are uncorrelated with each other, $\mathbf{\Phi}_{\mathbf{y}}(n)$ can be modeled as a sum of the cross-PSD matrices of these individual signal components (defined analogously to $\mathbf{\Phi}_{\mathbf{y}}(n)$). We

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

The authors would like to thank Asger Heidemann Andersen for allowing the use of his software implementation of the Dantale II matrix test.

assume that the microphone signal and its components are Gaussian distributed; hence, their STFT coefficients are circularly-symmetric complex Gaussian distributed, e.g: $\mathbf{y}(n) \sim \mathcal{N}_C(\mathbf{0}, \mathbf{\Phi_y}(n))$.

We model the talker as a single point-source and the early reflections as linear filters acting on the speech emitted by the talker. Hence, $\mathbf{s}(n)$ can be written as $\mathbf{d}s(n)$, where $s(n)$ is the scalar STFT coefficient of the direct-path speech at a chosen reference position, and $\mathbf{d}$ is a vector of Relative Transfer Functions (RTFs) [6] of the target signal from the reference position to all microphones. In terms of the cross-PSD matrix of the target component, this leads to: $\mathbf{\Phi_s}(n) = \phi_s(n)\mathbf{d}\mathbf{d}^H$, where $\phi_s(n)$ denotes the scalar PSD of the target speech component at the reference position. We assume that an estimate of $\mathbf{d}$ is available (e.g., because the application at hand allows accurate off-line estimation of $\mathbf{d}$, or alternatively, by use of an on-line estimation scheme such as [7, 8]).

We also assume that the late reverberation is cylindrically isotropic, i.e. that the late reverberant energy impinges on the microphone array from all horizontal directions with equal intensity. It allows us to write the cross-PSD matrix of $\mathbf{r}(n)$ as: $\mathbf{\Phi_r}(n) = \phi_r(n)\mathbf{\Gamma_r}$, where $\phi_r(n)$ denotes the PSD of the reverberation at the reference position and $\mathbf{\Gamma_r}$ denotes the cross-PSD matrix of a cylindrically isotropic sound field normalized by $\phi_r(n)$. The matrix $\mathbf{\Gamma_r}$ is constant and may be measured *a priori* (see e.g. [9, 10]).

We finally assume that the third signal component, $\mathbf{x}(n)$, is related to an additive noise whose statistics are slowly varying (e.g. microphone noise, car cabin noise). It follows that the cross-PSD matrix $\mathbf{\Phi_x}$ can be assumed approximately constant (hence, we omit the index n). We assume that $\mathbf{\Phi_x}$ is known or that a reliable estimate thereof is available (e.g., periodically updated during speech- and reverberation-free fragments of the signal).

Based on the aforementioned assumptions the overall model of the microphone input cross-PSD matrix may be written as:

$$\mathbf{\Phi_y}(n) = \phi_s(n)\mathbf{d}\mathbf{d}^H + \phi_r(n)\mathbf{\Gamma_r} + \mathbf{\Phi_x}. \tag{2}$$

In this model only the scalar PSDs $\phi_s(n)$ and $\phi_r(n)$ are unknown and time-varying; their estimation is the focus of this paper.

3. DERIVATION OF THE PROPOSED PSD ESTIMATOR

In this section we derive the proposed Maximum Likelihood Estimator (MLE) of $\phi_s(n)$ and $\phi_r(n)$. To facilitate the derivation we assume that $\phi_s(n)$ and $\phi_r(n)$ are approximately constant across a certain number L of consecutive time frames. We define the sample cross-PSD matrix based on the L most recent input vectors $\mathbf{y}(n)$ as:

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) = \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{y}(n-l)\mathbf{y}^H(n-l). \tag{3}$$

It has been shown [11, 12] that the joint log-likelihood function of $\phi_s(n)$ and $\phi_r(n)$ is given by:

$$\mathcal{L}(\phi_s, \phi_r) = -L\log|\mathbf{\Phi_y}(n)| - L\,\mathrm{tr}\big(\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\mathbf{\Phi_y}^{-1}(n)\big), \tag{4}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace operator. The MLEs of $\phi_s(n)$ and $\phi_r(n)$ are defined as the values maximizing $\mathcal{L}(\phi_s, \phi_r)$.

3.1. Estimator of the target speech PSD

The class of MLEs considered in this paper has been studied in [11], from where it follows that the MLE of $\phi_s(n)$ can be found as: $\arg\max_{\phi_s}\mathcal{L}(\phi_s, \hat{\phi}_r(n))$, where $\hat{\phi}_r(n)$ is the MLE of $\phi_r(n)$.

Denoting the MLE of the total cross-PSD matrix of the interference by: $\hat{\mathbf{\Phi}}_{\mathbf{v}}(n) = \hat{\phi}_r(n)\mathbf{\Gamma_r} + \mathbf{\Phi_x}$, and applying the results of [11] to the signal model in (2), the MLE of $\phi_s(n)$ can be found as:

$$\hat{\phi}_s(n) = \mathbf{w}_{\mathrm{mvdr}}^H(n)\big[\hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\mathbf{\Phi}}_{\mathbf{v}}(n)\big]\mathbf{w}_{\mathrm{mvdr}}(n), \tag{5}$$

where the vector:

$$\mathbf{w}_{\mathrm{mvdr}}(n) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}}{\mathbf{d}^H\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(n)\mathbf{d}} \tag{6}$$

contains the Minimum Variance Distortionless Response (MVDR) beamformer weights. Obviously, in order to compute $\hat{\phi}_s(n)$, the MLE $\hat{\phi}_r(n)$ must first be found (i.e. $\hat{\phi}_s(n)$ is a function of $\hat{\phi}_r(n)$).

3.2. Modification of the signal model by blocking and whitening

Since $\hat{\phi}_s(n)$ and $\hat{\phi}_r(n)$ are analytically related by (5), the joint likelihood in (4) can now be written as a function of only one variable, i.e.: $\mathcal{L}'(\phi_r) = \mathcal{L}\big(\hat{\phi}_s(\phi_r), \phi_r\big)$. The MLE of $\phi_r(n)$ is determined by maximizing $\mathcal{L}'(\phi_r)$. Unfortunately, for the signal model at hand this optimization problem is not easily tractable. Instead of resorting to numerical optimization methods, we propose an alternative method and approximate the MLE of $\phi_r(n)$ by using a likelihood function based on a simplified form of the signal model.

The simplification is achieved in 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 in [12]:

$$[\mathbf{B}\ \mathbf{b}] = \mathbf{I} - \mathbf{d}(\mathbf{d}^H\mathbf{d})^{-1}\mathbf{d}^H. \tag{7}$$

The columns of $\mathbf{B}$ can be interpreted as a set of $M-1$ target-canceling beamformers, i.e.: $\mathbf{B}^H\mathbf{s}(n) = \mathbf{0}_{(M-1)\times 1}$. The second modification of the signal model has the objective of diagonalizing $\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B}$, i.e. the cross-PSD matrix of the blocked additive noise component $\mathbf{B}^H\mathbf{x}(n)$. This is achieved by using a whitening matrix $\mathbf{D} \in \mathbb{C}_{(M-1)\times(M-1)}$ defined as the Cholesky factor of the inverse of $\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B}$ (we assume that $\mathbf{\Phi_x}$ is full rank):

$$\mathbf{D}\mathbf{D}^H = (\mathbf{B}^H\mathbf{\Phi_x}\mathbf{B})^{-1}. \tag{8}$$

We indicate the blocked and whitened signals with a tilde, e.g.: $\tilde{\mathbf{y}}(n) = \mathbf{D}^H\mathbf{B}^H\mathbf{y}(n)$. The cross-PSD matrix of $\tilde{\mathbf{y}}(n)$ is given by:

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \phi_r(n)\mathbf{\Gamma}_{\tilde{\mathbf{r}}} + \mathbf{I}, \tag{9}$$

where $\mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{D}^H\mathbf{B}^H\mathbf{\Gamma_r}\mathbf{B}\mathbf{D}$. Thanks to the blocking operation the number of unknowns in the signal model is now reduced to one (compare (2) and (9)). However, the blocking operation is not invertible and some information is lost in the process. It follows, that the proposed estimator is an *approximation* of the exact MLE.

Due to the whitening operation the matrices $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$ can be diagonalized by the same unitary matrix $\mathbf{U}$:

$$\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n) = \mathbf{U}\mathbf{\Lambda}_{\mathbf{\Phi}}(n)\mathbf{U}^H,\ \ \mathbf{\Gamma}_{\tilde{\mathbf{r}}} = \mathbf{U}\mathbf{\Lambda}_{\mathbf{\Gamma}}\mathbf{U}^H, \tag{10}$$

where the orthonormal columns of $\mathbf{U}$ are the eigenvectors, and $\mathbf{\Lambda}_{\mathbf{\Phi}}(n)$ and $\mathbf{\Lambda}_{\mathbf{\Gamma}}$ are diagonal matrices containing the eigenvalues of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$, respectively. Let $\lambda_{\mathbf{\Phi},m}$ and $\lambda_{\mathbf{\Gamma},m}$ denote the m-th eigenvalue of $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\mathbf{\Gamma}_{\tilde{\mathbf{r}}}$, respectively. Then, due to (9) and (10) the following holds:

$$\lambda_{\mathbf{\Phi},m} = \phi_r(n)\lambda_{\mathbf{\Gamma},m} + 1\,. \tag{11}$$

3.3. Estimator of the late reverberation PSD

Using the blocked and whitened signal model (9) we can formulate a new and simplified log-likelihood function of ϕ_r. By substituting the input cross-PSD matrix and its estimate in (4) by their modified counterparts $\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)$ and $\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)$ we obtain:

$$\mathcal{L}^{''}(\phi_r) = -L\log|\mathbf{\Phi}_{\tilde{\mathbf{y}}}(n)| - L\,\mathrm{tr}\big(\mathbf{\Phi}_{\tilde{\mathbf{y}}}^{-1}(n)\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\big). \qquad (12)$$

Thanks to the modifications introduced in Section 3.2, solving for the maximum of $\mathcal{L}^{''}(\phi_r)$ is more easily tractable than for $\mathcal{L}^{'}(\phi_r)$. It can be shown (proof omitted) that the extrema of the log-likelihood function in (12) are the roots of the following polynomial equation:

$$p(\phi_r) = \sum_{m=1}^{M-1} p_m(\phi_r), \quad \text{where:} \qquad (13)$$

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

where $g_m(n)$ denotes the m-th diagonal element of $\mathbf{U}^H\hat{\mathbf{\Phi}}_{\tilde{\mathbf{y}}}(n)\mathbf{U}$.

Unfortunately, no general closed-form solution for the roots of $p(\phi_r)$ exists. Therefore, numerical methods for root finding need to be used. Since the polynomial $p(\phi_r)$ is of odd order $2M-3$, at least 1 and at most $2M-3$ real-valued roots of $p(\phi_r)$ exist. When more than one real-valued root of $p(\phi_r)$ exists, the one yielding the highest value of the likelihood (12) must be chosen as the MLE $\hat{\phi}_r(n)$. It can be shown that typically in practice $g_m(n) \geq 1$ for all m. When this condition is met, only one real-valued root of $p(\phi_r)$ exists and it is non-negative (proof omitted). Then, computation and comparison of numerical values of the likelihood (12) is not necessary. For $M = 2$, the polynomial (13) has exactly one solution, which is equal to the late reverberation PSD estimator in [3].

4. EVALUATION OF THE PROPOSED PSD ESTIMATOR

The intended application of the proposed estimator is in speech dereverberation algorithms based on the Multi-channel Wiener Filter (MWF) [13,14] and on the MVDR beamformer. We evaluated the speech dereverberation performance of an MVDR- and an MWF-based algorithm using the proposed PSD estimator and compared it to the performance of these algorithms when used with the estimator from [3].

The MVDR beamformer coefficients $\mathbf{w}_{\mathrm{mvdr}}(n)$ were calculated according to (6). The MWF was implemented as a concatenation of the MVDR beamformer and a single-channel Wiener post-filter. Hence, the MWF output $\hat{s}(n)$ is given by:

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

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.

To investigate the performance of the MVDR- and MWF-based dereverberation algorithms, two separate experiments have been conducted: one using instrumental performance measures (Section 4.1) and another using a speech intelligibility test (Section 4.2).

4.1. Experiment A — instrumental performance measures

The first experiment comprised of a realistic simulation of microphone signals in a reverberant and noisy single-talker scenario. The microphone signals were simulated by concatenating TIMIT sentences [15] and convolving them with Room Impulse Responses (RIRs) measured using a microphone array composed of two behind-the-ear hearing aids. Each of the hearing aids had 2 microphones 1 cm apart ($M = 4$ microphones in total) and was placed on one of the ears of a Head and Torso Acoustic Simulator (HATS). The RIRs were measured in five reverberant rooms denoted: "Bathroom", "Cellar", "Stairs", "Office", and "Auditorium". A sixth, synthetic impulse response ("Isotropic") where the reverberation was modeled as perfectly cylindrically isotropic was also used. Microphone noise was simulated by a spatially white, spectrally pink noise whose PSD at 1kHz was 30 dB lower than the PSD of the target speech averaged over TIMIT sentences.

The sampling frequency of the simulated time-domain signals was 16 kHz and the STFT frame length was set to 8 ms ($T = 128$ samples). This ensured that the processing delay was below 10 ms, which is a requirement for hearing aid systems. A square-root Hanning window with 50% overlap was used in the STFT filterbank and for signal re-synthesis. The input cross-PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ was estimated recursively with a time constant of 50 ms.

In each of the six simulated reverberant conditions, the MWF algorithm, the MVDR beamformer, and the PSD estimators were implemented using RTF vectors $\mathbf{d}$ extracted from the first 2.5 ms of the RIR in question. This was done to ensure that $\mathbf{d}$ is based only on the direct path response and does not contain information on the early reflections present in the RIR. This resulted in a realistic mismatch between the used RTF vector $\mathbf{d}$ and the actual RTF of the target speech component. The normalized cross-PSD matrix $\mathbf{\Gamma}_{\mathbf{r}}$ was measured *a priori* in a simulated cylindrically isotropic sound field. 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 signal model was accurate. The matrix $\mathbf{\Phi}_{\mathbf{x}}$ was set to the identity matrix scaled by the PSD of the simulated microphone noise.

We evaluated the algorithms using the Frequency-Weighted Segmental SNR (FWSegSNR) [16] and the Perceptual Evaluation of Speech Quality (PESQ) [17] instrumental measures with the direct path speech $s(n)$ as the reference signal. Fig. 1 shows the performance scores obtained by the MWF and the MVDR beamformer based on the two PSD estimators (denoted "Braun" and "Proposed"). The scores calculated from the unprocessed input signal ("Input") are included for reference.

In most conditions all algorithms succeed in improving PESQ and FWSegSNR. Only in the "Bathroom" condition, the post-filtering stage of the two MWFs decreased the PESQ score compared to the corresponding MVDR beamformers. This was caused by strong early reflections present in this condition, which caused leakage of the early speech signal to the output of the blocking matrix. This resulted in overestimation of the late reverberation PSD and distortion of the target speech by the post-filter.

The performance difference between the MWF based on "Braun" and "Proposed" PSD estimators is small but consistent. This suggests that the proposed estimator is more robust to mismatches between the signal model and its actual structure than the estimator from [3]. The performance scores of the MVDR beamformers are very similar. The difference between the "Proposed" and "Braun" MWFs arises mostly in the post-filtering stage.

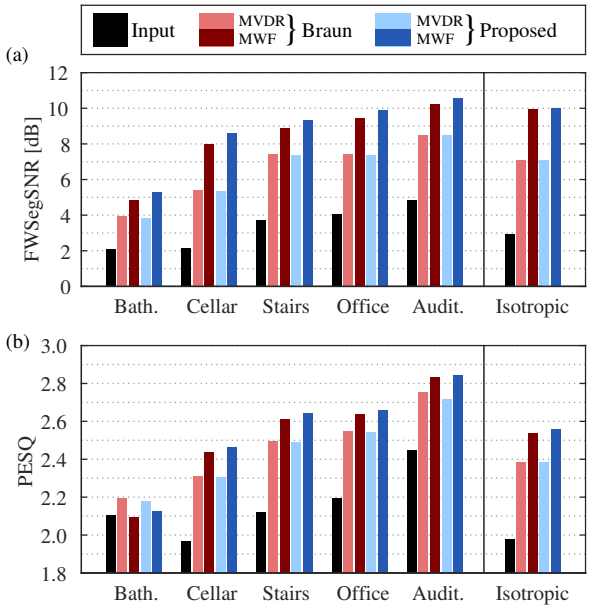

Fig. 1. (a) FWSegSNR and (b) PESQ scores obtained by using MWF and MVDR beamformer based on the PSD estimators from [3] ("Braun") and the proposed estimators ("Proposed").

4.2. Experiment B — speech intelligibility

The second experiment was a Dantale II [18] Speech Intelligibility (SI) test with 10 subjects. Sentences were presented via Sennheiser HD280 pro headphones and subjects were requested to select the words they heard from an on-screen list of options, as in [19].

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 simulation as in the "Cellar" condition in Section 4.1, corresponding to a frontal position of the target source at a distance of 2 m. Since the SI of this condition is (close to) 100%, speech-like interference consisting of randomly shifted and superimposed copies of the International Speech Test Signal (ISTS) [20] was added to the reverberated Dantale 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 Direct to Reverberant Ratio (DRR) 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.

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 4.1. 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 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

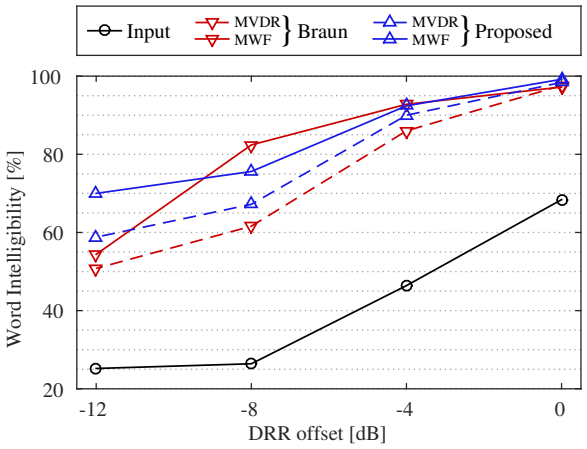

Fig. 2. Word intelligibility scores obtained in a Dantale II test with the RIR from the "Cellar" condition, averaged across 10 subjects.

the target and the ISTS interferers at their original (simulated) positions and benefit from the binaural advantage [21]. In the processed conditions this was not possible, as all components of the enhanced signals were perceived as coming from the target position (a known side-effect of using binaural beamformers [22]).

To each of the experimental conditions five Dantale sentences were randomly assigned (independently for each subject). The sentences were processed and then presented to subjects in a randomized order. The percentage of words identified correctly was averaged across subjects and is presented in Fig. 2.

Subject mean SI scores were corrected using the rationalized arcsine method [23] and analyzed using the repeated measures ANOVA procedure [24]. The influence of the processing type ($F_{4,36} = 110.2$, $p \ll 0.001$), the DRR offset ($F_{3,27} = 134.3$, $p \ll 0.001$), and the interaction term ($F_{12,108} = 2.2$, $p < 0.05$) were all found to be significant. Bonferroni-corrected pairwise comparisons of the marginal means revealed that: *a)* each of the algorithms significantly improved the SI over the "Input", *b)* the "Braun MWF" outperformed the "Braun MVDR", and *c)* the "Proposed MWF" outperformed the "Braun MVDR". The familywise type I error rate was limited to 5%.

5. CONCLUSION

In this paper we have proposed a new method for PSD estimation in reverberant and noisy conditions. The estimator is based on the maximum likelihood methodology and is computed by finding roots of a low order polynomial. We have evaluated the proposed estimator and compared it with a competing estimator [3] using two experiments. The first experiment revealed that MWF- and MVDR-based speech dereverberation algorithms using the proposed PSD estimator result in higher instrumental performance scores than when the PSD estimator from [3] is used. In the second experiment a speech intelligibility test with 10 subjects was conducted. All evaluated dereverberation algorithms significantly improved the speech intelligibility. However, only few significant differences between the algorithms were found. Notably, the MWF using the proposed PSD estimator outperformed the MVDR based on the estimator from [3].

6. REFERENCES

[1] J. Sohn, N. Soo Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[3] 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.

[4] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. 22nd Eur. Signal Process. Conf.*, Lisbon, Portugal, 2014, pp. 61–65.

[5] A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen, and J. Jensen, "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.

[6] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[7] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[8] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1420–1434, 2009.

[9] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 61–85. Springer, 2001.

[10] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[11] 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.

[12] 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.

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

[14] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. Ray Liu, Eds., pp. 269–302. Wiley, 2008.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, NIST, 1993.

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

[17] "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.

[18] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. of Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[19] E. R. Pedersen and P. M. Juhl, "Speech in noise test based on a ten-alternative forced choice procedure," *Baltic-Nordic Acoustics Meeting*, 2012.

[20] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ISTS)," *Int. J. of Audiology*, vol. 49, no. 12, pp. 891–903, 2010.

[21] B. C. J. Moore, *An introduction to the psychology of hearing*, Brill, 2012.

[22] S. Doclo, R. Dong, T. J. Klasen, J. Wouters, S. Haykin, and M. Moonen, "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.

[23] G. A. Studebaker, "A rationalized arcsine transform," *J. of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.

[24] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, CRC Press, 2011.