AN INFORMED MMSE FILTER BASED ON MULTIPLE INSTANTANEOUS DIRECTION-OF-ARRIVAL ESTIMATES

Oliver Thiergart, Maja Taseska, and Emanuël A. P. Habets

International Audio Laboratories Erlangen*
Am Wolfsmantel 33, 91058 Erlangen, Germany
{oliver.thiergart, maja.taseska, emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

Sound acquisition in noisy and reverberant conditions where the acoustic scene changes rapidly remains a challenging task. In this work, we consider the problem of obtaining a desired, arbitrary spatial response for at most L sound sources being simultaneously active per time-frequency instant. We propose a minimum mean-squared error spatial filter that adapts quickly to changes in the acoustic scene by incorporating instantaneous parametric information on the sound field. In addition, an estimator for the power spectral densities of the L sources is developed that exhibits a sufficiently high temporal and spectral resolution to achieve both dereverberation and noise reduction. Simulation results demonstrate that a strong attenuation of undesired noise and interfering components can be achieved with a tolerable amount of signal distortion.

***Index Terms*—** microphone array processing, optimal beamforming, dereverberation

1. INTRODUCTION

Sound acquisition in noisy and reverberant environments with several simultaneously active sources is commonly found in modern communication systems. A large variety of spatial filtering techniques has been proposed in the last decades to accomplish this task. We can classify existing spatial filters roughly into classical linear filters [1–4] and parametric filters [5–8]. The classical linear spatial filters require estimates of the propagation vectors or second-order statistics (SOS) of the desired sources and the SOS of the interference. Some filters are derived to extract a single source signal [9–16], while others have been derived to extract the sum of two or more source signals [17, 18]. These methods require *a priori* knowledge of the directions of the desired sources or a period in which only the desired sources are active. Another drawback of these methods is the inability to adapt sufficiently quickly to new situations (e. g., source movements, competing speakers that become active when the desired source is active). Parametric spatial filters are often based on a relatively simple signal model (i. e., the received signal in the time-frequency domain consists of a single plane wave plus diffuse sound) and are computed based on instantaneous estimates of the model parameters. The advantages of parametric spatial filters are a flexible directional response, a comparatively strong suppression of noise and interferers, and the ability to quickly adapt to new situations. However, the common single plane wave signal model can easily be violated in practice which strongly degrades the performance of the parametric spatial filters [19].

To overcome these problems, we have recently proposed an informed linearly constrained minimum variance (LCMV) filter that provides an arbitrary spatial response for at most L sound sources being simultaneously active per time-frequency instant [20]. The filter adapts nearly instantaneously to changes in the acoustic scene by incorporating parametric information on the sound field, namely L direction-of-arrival (DOA) estimates and the diffuse-to-noise power ratio (DNR). The filter minimizes the diffuse and self-noise power at the filter output while providing a distortionless response for the L sources. However, the drawback of such distortionless filters is a rather poor attenuation of diffuse sound and self-noise, especially for broadside array configurations with only few microphones.

In some applications, sound acquisition with a stronger suppression of diffuse sound and self-noise is desired while a moderate amount of signal distortion can be tolerated. For this purpose, we propose to incorporate instantaneous parametric information on the acoustic scene into the design of a minimum mean-squared error (MMSE) filter, leading to an *informed* MMSE filter. The proposed filter requires estimates of the power spectral densities of the L sources, which can be obtained with sufficient accuracy as explained throughout this paper. The proposed spatial filter has similar benefits as the informed LCMV filter [20], namely an arbitrary spatial response and a very short response time, but provides a stronger attenuation of diffuse sound and self-noise at the filter output.

The paper is organized as follows: Section 2 formulates the problem. In Sec. 3, the informed LCMV filter is reviewed and the proposed informed MMSE filter is described. In Sec. 4, it is shown how the required parametric information is estimated. The performance of the proposed spatial filter is evaluated in Sec. 5. Section 6 draws the conclusions.

2. PROBLEM FORMULATION

In the following, we consider an array of M omnidirectional microphones located at $\mathbf{d}_{1\ldots M}$. The microphones capture for each time and frequency a sum of $L < M$ plane waves propagating in an isotropic and homogenous (diffuse) sound field. The microphone signals $\mathbf{x}(k,n) = [X(k,n,\mathbf{d}_1)\ldots X(k,n,\mathbf{d}_M)]^{\mathrm{T}}$ at frequency index k and time index n are written as

$$\mathbf{x}(k,n) = \mathbf{A}(k,n)\,\mathbf{x}_{\mathrm{s}}(k,n) + \mathbf{x}_{\mathrm{d}}(k,n) + \mathbf{x}_{\mathrm{n}}(k,n), \qquad (1)$$

where $\mathbf{x}_{\mathrm{s}}(k,n) = [X_1(k,n,\mathbf{d}_1)\ldots X_L(k,n,\mathbf{d}_1)]^{\mathrm{T}}$ are the microphone signals proportional to the sound pressure of the L plane waves at the first microphone, $\mathbf{x}_{\mathrm{d}}(k,n)$ denotes the measured diffuse sound field, and $\mathbf{x}_{\mathrm{n}}(k,n)$ is the uncorrelated and stationary microphone self-noise. The time and frequency dependent $M \times L$ propagation matrix $\mathbf{A}(k,n) = [\mathbf{a}(k,\varphi_1)\ldots\mathbf{a}(k,\varphi_L)]$ contains the

*A joint institution of the University Erlangen-Nuremberg and Fraunhofer IIS, Germany propagation vectors $\mathbf{a}(k,\varphi_l) = [a_1(k,\varphi_l)\ldots a_M(k,\varphi_l)]^{\mathrm{T}}$ for the L plane waves. The i-th element of $\mathbf{a}(k,\varphi_l)$,

$$a_i(k,\varphi_l) = \exp\big\{\jmath\,\kappa\, r_i \sin\varphi_l(k,n)\big\}, \tag{2}$$

is the transfer function for the l-th plane wave from the first to the i-th microphone depending on the DOA $\varphi_l(k,n)$ of the wave. Here, $\varphi_l = 0$ denotes the array broadside. Moreover, $r_i = ||\mathbf{d}_i - \mathbf{d}_1||$ is equal to the distance between the first and the i-th microphone and κ is the wavenumber. Note that the DOA $\varphi_l(k,n)$ can vary rapidly across time and frequency.

Assuming the three components in (1) are mutually uncorrelated, we can express the power spectral density (PSD) matrix of the microphone signals as

$$\begin{aligned}\boldsymbol{\Phi}_x(k,n) &= \mathrm{E}\left\{\mathbf{x}(k,n)\,\mathbf{x}^{\mathrm{H}}(k,n)\right\}\\ &= \mathbf{A}(k,n)\boldsymbol{\Phi}_{\mathrm{s}}(k,n)\mathbf{A}^{\mathrm{H}}(k,n) + \underbrace{\boldsymbol{\Phi}_{\mathrm{d}}(k,n) + \boldsymbol{\Phi}_{\mathrm{n}}(k)}_{\boldsymbol{\Phi}_{\mathrm{u}}(k,n)}. \end{aligned}\tag{3}$$

Assuming further that the L plane waves are uncorrelated, the $L \times L$ signal PSD matrix $\boldsymbol{\Phi}_{\mathrm{s}}(k,n) = \mathrm{E}\left\{\mathbf{x}_{\mathrm{s}}(k,n)\,\mathbf{x}_{\mathrm{s}}^{\mathrm{H}}(k,n)\right\}$ is diagonal and $\mathrm{diag}\{\boldsymbol{\Phi}_{\mathrm{s}}(k,n)\} = \{\phi_1(k,n),\ldots,\phi_L(k,n)\}$ are the powers $\phi_l(k,n)$ of the L plane waves at the first microphone. Moreover,

$$\boldsymbol{\Phi}_{\mathrm{n}}(k,n) = \phi_{\mathrm{n}}(k)\,\mathbf{I} \tag{4}$$

is the time-invariant PSD matrix of the stationary self-noise, where $\mathbf{I}$ is the $M \times M$ identity matrix and $\phi_{\mathrm{n}}(k)$ is the self-noise power which is assumed to be identical for all microphones. The matrix

$$\boldsymbol{\Phi}_{\mathrm{d}}(k,n) = \phi_{\mathrm{d}}(k,n)\,\boldsymbol{\Gamma}_{\mathrm{d}}(k) \tag{5}$$

is the time-variant PSD matrix of the diffuse sound. The expected power $\phi_{\mathrm{d}}(k,n)$ of the diffuse sound is strongly time and frequency dependent and is assumed to be identical for all microphones. The ij-th element of the coherence matrix $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$, denoted by $\gamma_{ij}(k)$, is the coherence between microphone i and j due to the diffuse sound. For instance for a spherically isotropic diffuse field, we have $\gamma_{ij}(k) = \mathrm{sinc}(\kappa\, r_{ij})$ [21] where $r_{ij} = ||\mathbf{d}_j - \mathbf{d}_i||$.

The aim of the paper is to filter the microphone signals $\mathbf{x}(k,n)$ such that plane waves arriving from specific spatial regions are attenuated or amplified as desired, while the diffuse sound and self-noise are suppressed. The desired signal can therefore be expressed as a weighted sum of the L plane waves at the first microphone, i. e.,

$$Y(k,n) = \mathbf{g}^{\mathrm{T}}(k,n)\,\mathbf{x}_{\mathrm{s}}(k,n). \tag{6}$$

The weights are given by $\mathbf{g}(k,n) = [G(k,\varphi_1)\ldots G(k,\varphi_L)]^{\mathrm{T}}$, where $G(k,\varphi)$ is a real-valued arbitrary directivity function which can be frequency dependent. Figure 1 shows the magnitude of an example directivity $G(k,\varphi)$ for which we attenuate a plane waves arriving outside the spatial window by 60 dB while a wave arriving inside the spatial window is not attenuated. Clearly, one can design and employ arbitrary and time-variant directivity functions, e. g., to extract moving or emerging sound sources once they have been localized.

An estimate of the desired signal $Y(k,n)$ is obtained by a linear combination of the microphone signals $\mathbf{x}(k,n)$, i. e.,

$$\widehat{Y}(k,n) = \mathbf{w}^{\mathrm{H}}(k,n)\,\mathbf{x}(k,n), \tag{7}$$

where $\mathbf{w}(k,n)$ is a complex weight vector of length M. The optimal weights are derived in the next section. In the following, the dependency of the weights $\mathbf{w}(k,n)$ on k and n is omitted for brevity.

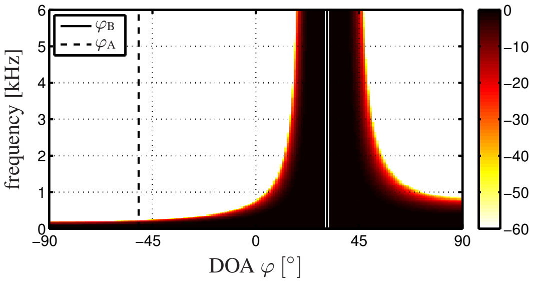


Fig. 1. Directivity function $|G(k,\varphi)|^2$ and source positions

3. OPTIMAL SPATIAL FILTERING

3.1. Informed Distortionless Spatial Filter

The informed LCMV filter in [20] provides an optimal trade-off between different state-of-the-art distortionless spatial filters. The filter is considered as reference in the following. The weights $\mathbf{w}(k,n)$ of the informed LCMV filter to estimate $Y(k,n)$ are found by minimizing the sum of the self-noise power and diffuse sound power at the filter output, i. e.,

$$\mathbf{w}_{\mathrm{iLCMV}} = \arg\min_{\mathbf{w}}\ \mathbf{w}^{\mathrm{H}}\,\boldsymbol{\Phi}_{\mathrm{u}}(k,n)\,\mathbf{w}, \tag{8}$$

subject to

$$\mathbf{w}^{\mathrm{H}}\,\mathbf{A}(k,n) = \mathbf{g}^{\mathrm{T}}(k,n). \tag{9}$$

Note that the filter weights are recomputed for each time and frequency and depend on the instantaneous DOA of the L plane waves, which define the propagation matrix $\mathbf{A}(k,n)$. Therefore, the filter adapts nearly immediately to changes in the acoustic scene. Due to the linear constraints (9), the L plane waves are captured with the correct gain according to the desired arbitrary directivity function $G(k,\varphi)$. The solution to (8) subject to (9) is [22]

$$\mathbf{w}_{\mathrm{iLCMV}} = \boldsymbol{\Phi}_{\mathrm{u}}^{-1}\mathbf{A}\left(\mathbf{A}^{\mathrm{H}}\boldsymbol{\Phi}_{\mathrm{u}}^{-1}\mathbf{A}\right)^{-1}\mathbf{g}, \tag{10}$$

where the dependencies on k and n have been omitted and $\boldsymbol{\Phi}_{\mathrm{u}}(k,n)$ is defined in (3). The estimation of $\boldsymbol{\Phi}_{\mathrm{u}}(k,n)$ is discussed in Sec. 4. In general, the performance of the distortionless filter in attenuating the diffuse sound and self-noise depends strongly on the microphone configuration and the number of microphones M. If $M \gg L+1$, the number of degrees of freedom to minimize $\boldsymbol{\Phi}_{\mathrm{u}}(k,n)$ in (8) is high. For the minimum number $M = L+1$, however, no degrees of freedom remain. In the worst case, the noise is amplified at the filter output.

3.2. Informed Minimum Mean-Squared Error Spatial Filter

In the following, we derive the optimal weights $\mathbf{w}(k,n)$ based on an MMSE criterium. The optimal weights provide the MMSE estimate of the desired signal $Y(k,n)$, i. e.,

$$\mathbf{w}_{\mathrm{iMMSE}} = \arg\min_{\mathbf{w}} \underbrace{\mathrm{E}\left\{\left|\widehat{Y}(k,n) - Y(k,n)\right|^2\right\}}_{\mathcal{J}_{\mathbf{w}}}. \tag{11}$$

Given the signal model in Sec. 2, the cost function $\mathcal{J}_{\mathbf{w}}(k,n)$ to be minimized can be written as

$$\mathcal{J}_{\mathbf{w}} = \mathbf{v}^{\mathrm{H}}(k,n)\,\boldsymbol{\Phi}_{\mathrm{s}}(k,n)\,\mathbf{v}(k,n) + \mathbf{w}^{\mathrm{H}}\,\boldsymbol{\Phi}_{\mathrm{u}}(k,n)\,\mathbf{w}, \tag{12}$$

where

$$\mathbf{v}(k,n) = \mathbf{g}(k,n) - \mathbf{A}^{\mathrm{H}}(k,n)\,\mathbf{w}. \tag{13}$$

The first term in (12) represents the speech distortion while the second term represents the power of the residual diffuse plus noise. Setting the complex derivative of $\mathcal{J}_{\mathbf{w}}$ to zero, the solution to (11) is

$$\mathbf{w}_{\text{iMMSE}} = \mathbf{W}_{\text{iMMSE}}(k,n)\,\mathbf{g}(k,n), \tag{14}$$

where $\mathbf{W}_{\text{iMMSE}}(k,n) = [\mathbf{w}_1 \ldots \mathbf{w}_L]$ is an $M \times L$ matrix given by

$$\begin{aligned}\mathbf{W}_{\text{iMMSE}} =& \big[\mathbf{A}(k,n)\,\mathbf{\Phi}_{\text{s}}(k,n)\,\mathbf{A}^{\mathrm{H}}(k,n) + \mathbf{\Phi}_{\text{u}}(k,n)\big]^{-1} \\ &\times \mathbf{A}(k,n)\,\mathbf{\Phi}_{\text{s}}(k,n).\end{aligned} \tag{15}$$

The filter weights $\mathbf{w}_{\text{iMMSE}}(k,n)$ are recomputed for each time and frequency and depend on the instantaneous DOAs $\varphi_l(k,n)$. Thus, the filter adapts quickly to changes in the acoustic scene, given the DOAs [and $\mathbf{\Phi}_{\text{s}}(k,n)$ and $\mathbf{\Phi}_{\text{u}}(k,n)$] can be estimated with a sufficiently high temporal resolution. The estimation of the PSD matrices $\mathbf{\Phi}_{\text{s}}(k,n)$ and $\mathbf{\Phi}_{\text{u}}(k,n)$ is explained in Sec. 4.

Note that each filter $\mathbf{w}_l(k,n)$ contained in $\mathbf{W}_{\text{iMMSE}}(k,n)$ provides the MMSE estimate of the corresponding source signal $X_l(k,n,\mathbf{d}_1)$ at the first microphone [23]. Since all source signals are mutually uncorrelated, i. e., $\mathbf{\Phi}_{\text{s}}(k,n)$ is diagonal, each filter $\mathbf{w}_l(k,n)$ can be represented as a minimum variance distortionless response (MVDR) filter $\mathbf{w}_{\text{MVDR},l}(k,n)$ extracting source l and a subsequent single-channel MMSE filter $H_l(k,n)$, i. e.,

$$\mathbf{w}_l = \underbrace{\frac{\mathbf{\Phi}_{\text{u},l}^{-1}\,\mathbf{a}_l}{\mathbf{a}_l^{\mathrm{H}}\,\mathbf{\Phi}_{\text{u},l}^{-1}\,\mathbf{a}_l}}_{\mathbf{w}_{\text{MVDR},l}(k,n)} \cdot \underbrace{\frac{\phi_l(k,n)}{\phi_l(k,n) + (\mathbf{a}_l^{\mathrm{H}}\,\mathbf{\Phi}_{\text{u},l}^{-1}\,\mathbf{a}_l)^{-1}}}_{H_l(k,n)}. \tag{16}$$

The PSD matrix of the noise and interference is given by

$$\mathbf{\Phi}_{\text{u},l}(k,n) = \mathbf{\Phi}_{\text{u}}(k,n) + \mathbf{A}_{\text{i},l}(k,n)\,\mathbf{\Phi}_{\text{i},l}(k,n)\,\mathbf{A}_{\text{i},l}^{\mathrm{H}}(k,n), \tag{17}$$

where the columns of $\mathbf{A}_{\text{i},l}(k,n)$ are the $L-1$ array steering vectors of the interfering plane waves and $\mathbf{\Phi}_{\text{i},l}(k,n)$ is obtained by removing the l-th row and l-th column from $\mathbf{\Phi}_{\text{s}}(k,n)$. Decomposing $\mathbf{W}_{\text{iMMSE}}(k,n)$ into the form given by (16) provides more flexibility in finding an optimum trade-off between the amount of noise reduction and speech distortion. In fact, one can apply different smoothing strategies or a lower bound to $H_l(k,n)$ to reduce speech distortion or to lower artifacts such as musical tones.

4. PARAMETER ESTIMATION

Several parameters need to be estimated for the proposed spatial filter. The DOAs $\varphi_l(k,n)$ of the L plane waves can be obtained with well-known narrowband DOA estimators such as ESPRIT [24] or root MUSIC [25], whereas the former is used throughout this work due to its lower computational complexity. The elements of the propagation matrix $\mathbf{A}(k,n)$ are computed with (2). To obtain $\mathbf{\Phi}_{\text{u}}(k,n)$ we assume that an estimate of the self-noise power $\phi_{\text{n}}(k)$ is available (e. g., estimated during silence). We then compute the DNR $\Psi(k,n) = {}^{\phi_{\text{d}}(k,n)}/_{\phi_{\text{n}}(k)}$ with the estimator in [20], which exploits the computed DOAs $\varphi_l(k,n)$. With the DNR information and with (4) and (5), an estimate of $\mathbf{\Phi}_{\text{u}}(k,n)$ can be computed as

$$\widehat{\mathbf{\Phi}}_{\text{u}}(k,n) = \phi_{\text{n}}(k)\big[\Psi(k,n)\,\mathbf{\Gamma}_{\text{d}}(k) + \mathbf{I}\big]. \tag{18}$$

To determine the signal PSDs diag$\{\mathbf{\Phi}_{\text{s}}(k,n)\}$, let us define

$$\widehat{\mathbf{\Phi}}_{\text{v}}(k,n) = \mathbf{\Phi}_x(k,n) - \widehat{\mathbf{\Phi}}_{\text{u}}(k,n), \tag{19}$$

which is an estimate of $\mathbf{A}(k,n)\,\mathbf{\Phi}_{\text{s}}(k,n)\,\mathbf{A}^{\mathrm{H}}(k,n)$ in (3), i. e.,

$$\widehat{\mathbf{\Phi}}_{\text{v}}(k,n) = \mathbf{A}(k,n)\,\mathbf{\Phi}_{\text{s}}(k,n)\,\mathbf{A}^{\mathrm{H}}(k,n) + \mathbf{\Delta}, \tag{20}$$

where $\mathbf{\Delta}$ is the estimation error. Equation (20) can be written as

$$\widehat{\mathbf{\Phi}}_{\text{v}}(k,n) = \sum_{l=1}^{L} \phi_l(k,n)\,\underbrace{\mathbf{a}(k,\varphi_l)\,\mathbf{a}^{\mathrm{H}}(k,\varphi_l)}_{\mathbf{C}_l(k,n)} + \mathbf{\Delta}. \tag{21}$$

We estimate the signal PSDs $\boldsymbol{\phi}(k,n) = [\phi_1(k,n)\ldots\phi_L(k,n)]^{\mathrm{T}}$ via the least-squares approach by minimizing the error $\mathbf{\Delta}$, i. e.,

$$\widehat{\boldsymbol{\phi}}(k,n) = \arg\min_{\boldsymbol{\phi}} \left\| \operatorname{vec}\{\widehat{\mathbf{\Phi}}_{\text{v}}(k,n)\} - \mathbf{B}(k,n)\,\boldsymbol{\phi} \right\|^2, \tag{22}$$

where vec$\{\mathbf{X}\}$ are the columns of matrix $\mathbf{X}$ stacked into one column vector and $\mathbf{B}(k,n) = \big[\operatorname{vec}\{\mathbf{C}_1(k,n)\}\ldots\operatorname{vec}\{\mathbf{C}_L(k,n)\}\big]$. The solution to the minimization problem (22) is

$$\widehat{\boldsymbol{\phi}}(k,n) = \big(\mathbf{B}^{\mathrm{H}}\,\mathbf{B}\big)^{-1}\,\mathbf{B}^{\mathrm{H}}\operatorname{vec}\{\widehat{\mathbf{\Phi}}_{\text{v}}(k,n)\}. \tag{23}$$

5. SIMULATION RESULTS

A reverberant shoebox room (6.95×5.39×2.39 m^3, $\mathrm{RT}_{60} \approx 490$ ms) and an uniform linear array with $M = 5$ omnidirectional microphones (3 cm microphone spacing) was simulated using the source-image method [26,27]. Two speech sources are located at a distance of 1.25 m at angles $\varphi_{\text{A}} = -51°$ and $\varphi_{\text{B}} = 31°$ (cf. Fig. 1). The recorded signals consist of 1 s silence, single talk (source A), double talk, and single talk (source B). White Gaussian noise was added to the microphone signals resulting in a segmental signal-to-noise ratio (SegSNR) of 28 dB. The sound was sampled at 16 kHz and transformed into the time-frequency domain using a 256-point short-time Fourier transform (STFT) with 50% overlap.

We assume $L = 2$ plane waves in the model in (1) and consider the directivity function $G(k,\varphi)$ in Fig. 1, i. e., we aim at extracting source B (desired source) without attenuation while attenuating the power of source A (interferer) by 60 dB. We compare the informed LCMV filter (Sec. 3.1) and the proposed informed MMSE filter (Sec. 3.2). The parametric information is estimated as explained in Sec. 4. The required self-noise power $\phi_{\text{n}}(k)$ is computed at the beginning of the signal when the sources are inactive. The expectation in (3) is approximated by a recursive temporal averaging filter with a time constant of $\tau = 50$ ms. With this averaging length the parameters in Sec. 4 are updated sufficiently fast to track typical changes in the acoustic scene such as moving or emerging sources.

5.1. Parameter Estimation Performance

This section studies the performance of the $\mathbf{\Phi}_{\text{s}}(k,n)$ and $\mathbf{\Phi}_{\text{u}}(k,n)$ estimation. We assume that the DOAs of the sound are given as *a priori* information, i. e., $\varphi_1(k,n) = \varphi_{\text{A}}$ and $\varphi_2(k,n) = \varphi_{\text{B}}$.

Figure 2 shows the true and estimated power $\phi_2(k,n)$ and $\widehat{\phi}_2(k,n)$ of the second source, i. e., it shows the second element of diag$\{\mathbf{\Phi}_{\text{s}}(k,n)\}$ and diag$\{\widehat{\mathbf{\Phi}}_{\text{s}}(k,n)\}$, respectively. The time domain signals at the bottom of the figure indicate which source is active when. Figure 2 shows that the source power was determined accurately for most time-frequency bins. However, at lower frequencies, power of the first source was leaking into $\widehat{\phi}_2(k,n)$ (dashed circle) or $\widehat{\phi}_2(k,n)$ was underestimated (solid circle). The leaking power (dashed circle) is the reverberation due to the first source that was not

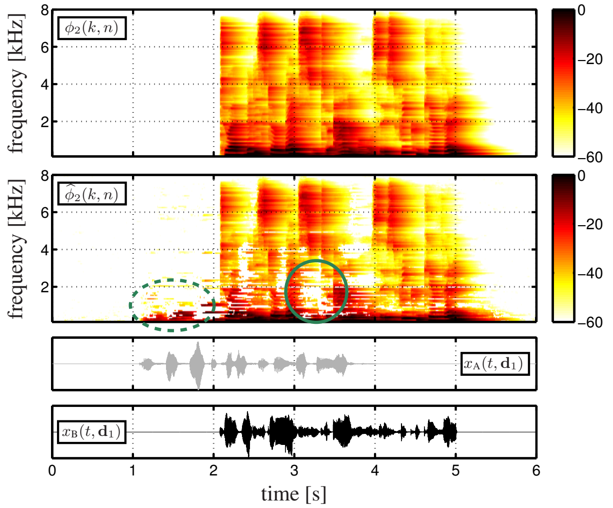


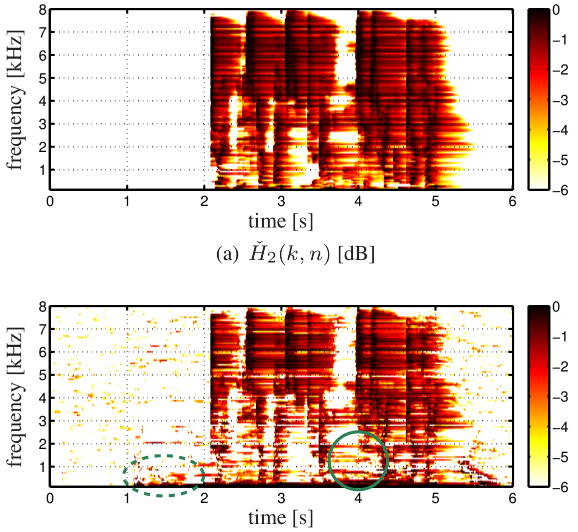
Fig. 2. Upper two plots: true and estimated power of the second source. The same temporal averaging was applied to $\phi_2(k,n)$ as used for computing $\widehat{\phi}_2(k,n)$. Lower two plots: time domain signals of the two sources.

completely subtracted in (19) due to an underestimated diffuse-plus-noise PSD matrix $\mathbf{\Phi}_\mathrm{u}(k,n)$. This underestimation resulted from an underestimated DNR $\Psi(k,n)$ in (18). Equivalently, the underestimation of $\widehat{\phi}_2(k,n)$ (solid circle) resulted from an overestimation of $\mathbf{\Phi}_\mathrm{u}(k,n)$ due to an overestimated $\Psi(k,n)$.

From the estimated parameters we can compute the optimal weights $\mathbf{w}_\mathrm{iMMSE}$, which, as described in Sec. 3.2, can be decomposed into a weighted sum of L separate filters. As shown in (16), each separate filter can be expressed as an MVDR filter and subsequent single-channel MMSE filter $H_l(k,n)$. Figure 3(a) shows the ideal filter $\check{H}_2(k,n)$ when considering the true $\mathbf{\Phi}_\mathrm{s}(k,n)$ and $\mathbf{\Phi}_\mathrm{u}(k,n)$, while Fig. 3(b) shows the filter $H_2(k,n)$ following from the estimates $\widehat{\mathbf{\Phi}}_\mathrm{s}(k,n)$ and $\widehat{\mathbf{\Phi}}_\mathrm{u}(k,n)$. Both filters attenuate strongly the output of the prior MVDR filter when mainly the noise and interferer is present. The estimated filter $H_2(k,n)$ does not differ much from the ideal filter $\check{H}_2(k,n)$, besides at the lower frequencies due to estimation errors of $\mathbf{\Phi}_\mathrm{s}(k,n)$ and $\mathbf{\Phi}_\mathrm{u}(k,n)$ mentioned before. Therefore, for some time-frequency bins, $H_2(k,n)$ does not suppress interfering power and noise as desired (dashed circle), or attenuates the desired signal (solid circle) leading to speech distortion. Nevertheless, the estimated filter is sufficiently accurate to enhance the signal, as shown in the next section.

5.2. Overall Performance

In the following, we evaluate the performance of the proposed spatial filter $\mathbf{w}_\mathrm{iMMSE}$ when the DOAs $\varphi_1(k,n)$ and $\varphi_2(k,n)$ are not given as *a priori* information, but estimated using ESPRIT [24]. The ESPRIT algorithm included a recursive temporal averaging filter with a time constant of $\tau = 50\,\mathrm{ms}$. As mentioned before, this typically yields a sufficiently high temporal resolution to track changes in the acoustic scene. Table 1 shows the performance of $\mathbf{w}_\mathrm{iMMSE}$ in terms of SegSNR, segmental signal-to-interference ratio (SegSIR), segmental signal-to-reverberation ratio (SegSRR), PESQ, and mean log spectral distortion (LSD). The values are computed over the more difficult double talk part. For comparison, we also show the results obtained with the informed LCMV filter ($\mathbf{w}_\mathrm{iLCMV}$) and the ideal informed MMSE filter ($\check{\mathbf{w}}_\mathrm{iMMSE}$), which was computed from accurate information on $\mathbf{\Phi}_\mathrm{s}(k,n)$, $\mathbf{\Phi}_\mathrm{u}(k,n)$, and the DOAs. Note that for PESQ, the direct path signal of source B as received by the first microphone was used as a reference. Moreover, the LSD given the weights $\mathbf{w}$ was computed as [28]



(a) $\check{H}_2(k,n)$ [dB]

(b) $H_2(k,n)$ [dB]

Fig. 3. True and estimated single-channel Wiener filter $H_2(k,n)$

$$\mathrm{LSD}(n)=\left[\frac{2}{K}\sum_{k=0}^{K/2-1}\left|\mathcal{L}\left\{Y_\mathrm{B}(k,n)\right\}-\mathcal{L}\left\{X_\mathrm{B}(k,n,\mathbf{d}_1)\right\}\right|^2\right]^{\frac{1}{2}}, \qquad (24)$$

where $Y_\mathrm{B}(k,n)$ is the signal of the desired source B at the filter output, i. e., $Y_\mathrm{B}(k,n) = \mathbf{w}^\mathrm{H}\mathbf{a}(k,\varphi_\mathrm{B})X_\mathrm{B}(k,n,\mathbf{d}_1)$. The log spectrum is $\mathcal{L}\{X(k,n)\} = 20\log_{10}|X(k,n)|$ which was limited to a dynamic range of 50 dB. The mean LSD is found by averaging (24) over all double talk frames.

The values in Tab. 1 show that the proposed informed MMSE filter ($\mathbf{w}_\mathrm{iMMSE}$) outperformed the informed LCMV filter ($\mathbf{w}_\mathrm{iLCMV}$) in terms of SegSIR, SegSNR, and SegSRR. The proposed MMSE filter therefore better attenuates the noise and interferer than the LCMV filter. As expected, the informed LCMV filter provides a very low LSD (i. e., nearly no distortion of the desired signal), while the distortion is higher for the MMSE-based filters. The ideal informed MMSE filter ($\check{\mathbf{w}}_\mathrm{iMMSE}$) outperforms the estimated filter ($\mathbf{w}_\mathrm{iMMSE}$) in terms of SegSIR, SegSRR, and LSD. Compared to the unprocessed signals ($*$), all filters strongly improve the signal by means of noise and interference reduction. In terms of PESQ, all spatial filters improve the signal compared to the unprocessed signal.

6. CONCLUSIONS

An informed minimum mean-squared error (MMSE) filter was proposed that provides a desired spatial response for L sources being simultaneously active for each time and frequency in a noisy and reverberant environment. The filter exploits instantaneous information on the direction-of-arrival of L plane waves and considers the power spectral density (PSD) matrices of the diffuse sound, self-noise, and source signals. Estimators for the required PSD matrices

	SegSIR	SegSNR	SegSRR	mean LSD	PESQ
$*$	13.3	27.6	-2.8	-	1.4
$\mathbf{w}_{\text{iLCMV}}$	25.8	26.2	-0.5	1.4	1.6
$\mathbf{w}_{\text{iMMSE}}$	27.2	31.7	1.0	2.9	1.6
$\check{\mathbf{w}}_{\text{iMMSE}}$	32.0	28.5	3.2	2.7	2.1

Table 1. Performance of the spatial filters [$*$ unprocessed]. Values in dB. The signals were A-weighted before computing the SegSIR, SegSRR, and SegSNR.

were proposed that are sufficiently accurate to reduce reverberation, self-noise, and interfering sounds with a tolerable amount of signal distortion. Simulations results for a highly reverberant environment demonstrate the practical applicability of the proposed filter.

7. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[2] 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. Ray Liu, Eds., chapter 9. Wiley, 2008.

[3] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., chapter 47. Springer-Verlag, 2008.

[4] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011.

[5] I. Tashev, M. Seltzer, and A. Acero, "Microphone array for headset with spatial noise suppressor," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005.

[6] M. Kallinger, G. Del Galdo, F. Kuech, D. Mahne, and R. Schultz-Amling, "Spatial filtering using directional audio coding parameters," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 217–220.

[7] M. Kallinger, G. D. Galdo, F. Kuech, and O. Thiergart, "Dereverberation in the spatial audio coding domain," in *Audio Engineering Society Convention 130*, London UK, May 2011.

[8] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, United Kingdom, May 2011.

[9] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 514–518, Nov. 1993.

[10] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[11] 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.

[12] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to real-world problems*, J. Benesty and Y. Huang, Eds., Signals and Communication Technology, chapter 6, pp. 155–194. Springer-Verlag, Berlin, Germany, 2003.

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

[14] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 206–219, Jan. 2011.

[15] E. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 945–958, May 2013.

[16] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2012.

[17] G. Reuven, S. Gannot, and I. Cohen, "Dual source transfer-function generalized sidelobe canceller," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 4, pp. 711–727, May 2008.

[18] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[19] O. Thiergart and E. A. P. Habets, "Sound field model violations in parametric spatial sound processing," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2012.

[20] O. Thiergart and E. A. P. Habets, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013.

[21] 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.

[22] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[23] H. L. van Trees, *Optimum Array Processing*, Detection, Estimation and Modulation Theory. Wiley, 2002.

[24] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 984–995, 1989.

[25] B. Rao and K. Hari, "Performance analysis of root-music*," in *Signals, Systems and Computers, 1988. Twenty-Second Asilomar Conference on*, 1988, vol. 2, pp. 578–582.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[27] E. A. P. Habets, "Room impulse response (RIR) generator," May 2008.

[28] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.