OPTIMAL BINAURAL LCMV BEAMFORMING IN COMPLEX ACOUSTIC SCENARIOS: THEORETICAL AND PRACTICAL INSIGHTS

Nico Gößling[1], *Daniel Marquardt*[1,2], *Ivo Merks*[2], *Tao Zhang*[2], *Simon Doclo*[1]

[1]University of Oldenburg, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, Oldenburg, Germany
[2]Starkey Hearing Technologies, Eden Prairie, MN, 55344, USA

ABSTRACT

Binaural beamforming algorithms for head-mounted assistive listening devices are crucial to improve speech quality and speech intelligibility in noisy environments, while maintaining the spatial impression of the acoustic scene. While the well-known BMVDR beamformer is able to preserve the binaural cues of one desired source, the BLCMV beamformer uses additional constraints to also preserve the binaural cues of interfering sources. In this paper, we provide theoretical and practical insights on how to optimally set the interference scaling parameters in the BLCMV beamformer for an arbitrary number of interfering sources. In addition, since in practice only a limited temporal observation interval is available to estimate all required beamformer quantities, we provide an experimental evaluation in a complex acoustic scenario using measured impulse responses from hearing aids in a cafeteria for different observation intervals. The results show that even rather short observation intervals are sufficient to achieve a decent noise reduction performance and that a proposed threshold on the optimal interference scaling parameters leads to smaller binaural cue errors in practice.

Index Terms— Hearing aids, binaural cues, noise reduction, beamforming, BLCMV, RTF

1. INTRODUCTION

For head-mounted assistive listening devices (e.g., hearing aids, cochlear implants), algorithms that use the microphone signals from both the left and the right hearing device are effective techniques to improve speech intelligibility, as the spatial information captured by all microphones can be exploited [1, 2]. Besides reducing undesired sources and limiting speech distortion, another important objective of binaural speech enhancement algorithms is the preservation of the listener's perception of the acoustical scene, in order to exploit the binaural hearing advantage [3] and to reduce confusions due to a mismatch between acoustical and visual information. To achieve binaural noise reduction with binaural cue preservation, two main concepts have been developed. In the first concept, a common real-valued spectro-temporal gain is applied to the reference microphone signals in the left and the right hearing device [4–10], ensuring perfect preservation of the instantaneous binaural cues but inevitably introducing speech distortion. The second concept, which is considered in this paper, is to apply a complex-valued filter to all available microphone signals on the left and the right hearing device using binaural extensions of spatial filtering techniques [11–19].

While the well-known binaural minimum variance distortionless response (BMVDR) beamformer [1] preserves the binaural cues (i.e., the interaural level difference (ILD) and interaural time difference (ITD)) of one desired source, the binaural linearly constrained minimum variance (BLCMV) beamformer [15] is also able to preserve the binaural cues of interfering sources. This is achievable by imposing interference scaling constraints for these sources. It should be noted that the BMVDR and BLCMV beamformers require an estimate of the correlation matrix that should be minimized and an estimate of the *relative transfer functions* (RTFs) of the desired and interfering sources. The performance of these beamformers may significantly deteriorate in case of estimation errors. Such estimation errors occur if only short temporal observation intervals for estimation can be used, e.g., due to dynamic spatial scenarios such as moving sources or head movement.

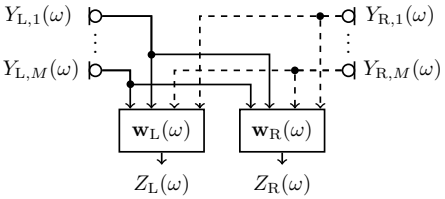

Fig. 1: Binaural hearing device configuration.

In this paper, we first derive optimal values for the interference scaling parameters in the BLCMV beamformer based on the BMVDR beamformer with RTF preservation (BMVDR-RTF) [13, 14] for an arbitrary number of interfering sources. Secondly, since these values are optimal in the sense of noise reduction but not robust against RTF estimation errors in practice, we propose to apply an upper and lower threshold on them. We evaluate the performance of the BMVDR beamformer and the BLCMV beamformer using the two different interference scaling parameters and measured impulse responses from hearing aids in a cafeteria [20] for several temporal observation intervals. The results show that even rather short temporal temporal observation intervals lead to sufficient noise reduction performance and that the imposed threshold on the optimal interference scaling parameters can significantly reduce binaural cue errors.

2. CONFIGURATION AND NOTATION

Consider the binaural hearing device configuration in Fig. 1, consisting of a microphone array with M microphones on the left and the right hearing device. For an acoustic scenario with one desired source, P interfering sources and incoherent background noise, the m-th microphone signal of the left hearing device $Y_{\mathrm{L},m}(\omega)$ can be written in the frequency-domain

This work was supported by the Collaborative Research Centre 1330 Hearing Acoustics and the Cluster of Excellence 1077 Hearing4all, funded by the German Research Foundation (DFG), by the joint Lower Saxony-Israeli Project ATHENA, and a research gift from Starkey Hearing Technologies.

978-1-5386-8151-0/18/$31.00

as

$$Y_{\mathrm{L},m}(\omega)=X_{\mathrm{L},m}(\omega)+\sum_{p=1}^{P}U_{\mathrm{L},p,m}(\omega)+N_{\mathrm{L},m}(\omega), \tag{1}$$

with $X_{\mathrm{L},m}(\omega)$ the desired speech component, $U_{\mathrm{L},p,m}(\omega)$ the p-th interference component and $N_{\mathrm{L},m}(\omega)$ the background noise component (e.g., diffuse noise) in the m-th microphone signal. The m-th microphone signal of the right hearing device $Y_{\mathrm{R},m}(\omega)$ is defined similarly. For conciseness we will omit the frequency variable ω in the remainder of the paper. We define the $2M$-dimensional stacked signal vector $\mathbf{y}$ as

$$\mathbf{y}=[Y_{\mathrm{L},1}...Y_{\mathrm{L},M}Y_{\mathrm{R},1}...Y_{\mathrm{R},M}]^T, \tag{2}$$

where $(\cdot)^T$ denotes the transpose and which can be written as

$$\mathbf{y}=\mathbf{x}+\mathbf{v}, \quad \mathbf{v}=\sum_{p=1}^{P}\mathbf{u}_p+\mathbf{n}, \tag{3}$$

where $\mathbf{x}$, $\mathbf{u}_p$ and $\mathbf{n}$ are defined similarly as in (2) and $\mathbf{v}$ denotes the overall undesired component, i.e., interference plus background noise components. For the coherent desired source S_{x} and the coherent interfering sources $S_{\mathrm{u},p}$, with $p\in\{1,...,P\}$, the vectors $\mathbf{x}$ and $\mathbf{u}_p$ can be written as

$$\mathbf{x}=S_{\mathrm{x}}\mathbf{a}, \qquad \mathbf{u}_p=S_{\mathrm{u},p}\mathbf{b}_p, \tag{4}$$

with $\mathbf{a}$ and $\mathbf{b}_p$ the *acoustic transfer functions* (ATFs) between all microphones and the desired and the p-th interfering source, respectively. Without loss of generality, we choose the first microphones on the left and the right hearing device as reference microphones, i.e.,

$$Y_{\mathrm{L}}=\mathbf{e}_{\mathrm{L}}^T\mathbf{y}, \quad Y_{\mathrm{R}}=\mathbf{e}_{\mathrm{R}}^T\mathbf{y}, \tag{5}$$

where $\mathbf{e}_{\mathrm{L}}$ and $\mathbf{e}_{\mathrm{R}}$ are $2M$-dimensional vectors with one element equal to 1 and the other elements equal to 0, i.e., $\mathbf{e}_{\mathrm{L}}[1]=1$ and $\mathbf{e}_{\mathrm{R}}[M+1]=1$. The correlation matrices of the background noise component, the desired speech component, the p-th interference component and all interference components are defined as

$$\mathbf{R}_{\mathrm{n}}=\mathcal{E}\left\{\mathbf{n}\mathbf{n}^H\right\}, \quad \mathbf{R}_{\mathrm{x}}=\mathcal{E}\left\{\mathbf{x}\mathbf{x}^H\right\}=\Phi_{\mathrm{x}}\mathbf{a}\mathbf{a}^H, \tag{6}$$

$$\mathbf{R}_{\mathrm{u},p}=\mathcal{E}\left\{\mathbf{u}_p\mathbf{u}_p^H\right\}=\Phi_{\mathrm{u},p}\mathbf{b}_p\mathbf{b}_p^H, \quad \mathbf{R}_{\mathrm{u}}=\sum_{p=1}^{P}\mathbf{R}_{\mathrm{u},p}, \tag{7}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, $(\cdot)^H$ denotes the conjugate transpose and Φ_{x} and $\Phi_{\mathrm{u},p}$ denote the *power spectral density* (PSD) of the desired source and the p-th interfering source, respectively. Assuming statistical independence between the components in (1), the correlation matrix of the microphone signals $\mathbf{R}_{\mathrm{y}}$ can be written as

$$\mathbf{R}_{\mathrm{y}}=\mathbf{R}_{\mathrm{x}}+\mathbf{R}_{\mathrm{u}}+\mathbf{R}_{\mathrm{n}}=\mathbf{R}_{\mathrm{x}}+\mathbf{R}_{\mathrm{v}}, \tag{8}$$

with $\mathbf{R}_{\mathrm{v}}$ the correlation matrix of the overall undesired component.
The output signal at the left hearing device Z_{L} is obtained by filtering the microphone signals with the $2M$-dimensional filter $\mathbf{w}_{\mathrm{L}}$, i.e.,

$$Z_{\mathrm{L}}=\mathbf{w}_{\mathrm{L}}^H\mathbf{y}=\mathbf{w}_{\mathrm{L}}^H\mathbf{x}+\sum_{p=1}^{P}\mathbf{w}_{\mathrm{L}}^H\mathbf{u}_p+\mathbf{w}_{\mathrm{L}}^H\mathbf{n}, \tag{9}$$

The output signal at the right hearing aid Z_{R} is similarly defined. Furthermore, we define the $4M$-dimensional filter vector $\mathbf{w}$ as

$$\mathbf{w}=\begin{bmatrix}\mathbf{w}_{\mathrm{L}}\\ \mathbf{w}_{\mathrm{R}}\end{bmatrix}. \tag{10}$$

The RTF vectors of the desired and the interfering sources are defined by relating the ATF vectors to the ATF of the reference microphone on the left and the right hearing device, i.e.,

$$\mathbf{a}_{\mathrm{L}}=\frac{\mathbf{a}}{A_{\mathrm{L}}}, \quad \mathbf{a}_{\mathrm{R}}=\frac{\mathbf{a}}{A_{\mathrm{R}}}, \quad \mathbf{b}_{\mathrm{L},p}=\frac{\mathbf{b}_p}{B_{\mathrm{L},p}}, \quad \mathbf{b}_{\mathrm{R},p}=\frac{\mathbf{b}_p}{B_{\mathrm{R},p}}. \tag{11}$$

The $2M\times P$-dimensional matrices $\mathbf{B}_{\mathrm{L}}$ and $\mathbf{B}_{\mathrm{R}}$ containing the RTF vectors of all interfering sources are defined as

$$\mathbf{B}_{\mathrm{L}}=\left[\mathbf{b}_{\mathrm{L},1},...,\mathbf{b}_{\mathrm{L},P}\right], \quad \mathbf{B}_{\mathrm{R}}=\left[\mathbf{b}_{\mathrm{R},1},...,\mathbf{b}_{\mathrm{R},P}\right]. \tag{12}$$

The binaural input and output *signal-to-noise ratio* (SNR) is defined as the ratio of the average input and output PSDs of the desired speech component and the background noise component, i.e.,

$$\mathrm{SNR}^{\mathrm{i}}=\frac{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{x}}\mathbf{e}_{\mathrm{L}}+\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{x}}\mathbf{e}_{\mathrm{R}}}{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{n}}\mathbf{e}_{\mathrm{L}}+\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{n}}\mathbf{e}_{\mathrm{R}}}, \quad \mathrm{SNR}^{\mathrm{o}}=\frac{\mathbf{w}_{\mathrm{L}}^H\mathbf{R}_{\mathrm{x}}\mathbf{w}_{\mathrm{L}}+\mathbf{w}_{\mathrm{R}}^H\mathbf{R}_{\mathrm{x}}\mathbf{w}_{\mathrm{R}}}{\mathbf{w}_{\mathrm{L}}^H\mathbf{R}_{\mathrm{n}}\mathbf{w}_{\mathrm{L}}+\mathbf{w}_{\mathrm{R}}^H\mathbf{R}_{\mathrm{n}}\mathbf{w}_{\mathrm{R}}}. \tag{13}$$

The binaural input and output *signal-to-interference ratio* (SIR) is defined as the ratio of the average input and output PSDs of the desired speech component and the interference components, i.e.,

$$\mathrm{SIR}^{\mathrm{i}}=\frac{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{x}}\mathbf{e}_{\mathrm{L}}+\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{x}}\mathbf{e}_{\mathrm{R}}}{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{u}}\mathbf{e}_{\mathrm{L}}+\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{u}}\mathbf{e}_{\mathrm{R}}}, \quad \mathrm{SIR}^{\mathrm{o}}=\frac{\mathbf{w}_{\mathrm{L}}^H\mathbf{R}_{\mathrm{x}}\mathbf{w}_{\mathrm{L}}+\mathbf{w}_{\mathrm{R}}^H\mathbf{R}_{\mathrm{x}}\mathbf{w}_{\mathrm{R}}}{\mathbf{w}_{\mathrm{L}}^H\mathbf{R}_{\mathrm{u}}\mathbf{w}_{\mathrm{L}}+\mathbf{w}_{\mathrm{R}}^H\mathbf{R}_{\mathrm{u}}\mathbf{w}_{\mathrm{R}}}. \tag{14}$$

The binaural input and output *signal-to-interference-plus-noise ratio* (SINR) is defined as the ratio of the average input and output PSDs of the desired speech component and the overall undesired component, i.e.,

$$\mathrm{SINR}^{\mathrm{i}}=\frac{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{x}}\mathbf{e}_{\mathrm{L}}+\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{x}}\mathbf{e}_{\mathrm{R}}}{\mathbf{e}_{\mathrm{L}}^T\mathbf{R}_{\mathrm{v}}\mathbf{e}_{\mathrm{L}}+\mathbf{e}_{\mathrm{R}}^T\mathbf{R}_{\mathrm{v}}\mathbf{e}_{\mathrm{R}}}, \quad \mathrm{SINR}^{\mathrm{o}}=\frac{\mathbf{w}_{\mathrm{L}}^H\mathbf{R}_{\mathrm{x}}\mathbf{w}_{\mathrm{L}}+\mathbf{w}_{\mathrm{R}}^H\mathbf{R}_{\mathrm{x}}\mathbf{w}_{\mathrm{R}}}{\mathbf{w}_{\mathrm{L}}^H\mathbf{R}_{\mathrm{v}}\mathbf{w}_{\mathrm{L}}+\mathbf{w}_{\mathrm{R}}^H\mathbf{R}_{\mathrm{v}}\mathbf{w}_{\mathrm{R}}}. \tag{15}$$

3. BINAURAL NOISE REDUCTION ALGORITHMS

In Section 3.1 and 3.2 we briefly review the BMVDR beamformer [1, 2, 12] and the BLCMV beamformer [15]. Based on the optimality of the BMVDR-RTF beamformer [14] in optimizing the SINR (or SNR) while preserving the binaural cues of all sources, in Section 3.3 we derive optimal values for the interference scaling parameters in the BLCMV beamformer in the case of an arbitrary number of interfering sources. Furthermore, in order to achieve a robust binaural cue preservation performance in case of estimation errors of the correlation matrices and the RTF vectors (Section 3.4), we propose to threshold these interference scaling parameters.

3.1. BMVDR beamformer

The BMVDR beamformer aims at minimizing the output PSD in both hearing devices, while preserving the desired speech component in the reference microphone signals. The corresponding constrained optimization problem is given by

$$\min_{\mathbf{w}}\ \mathbf{w}^H\widetilde{\mathbf{R}}\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H\mathbf{C}=\mathbf{g}, \tag{16}$$

with

$$\widetilde{\mathbf{R}}=\begin{bmatrix}\mathbf{R} & \mathbf{0}_{2M\times 2M}\\ \mathbf{0}_{2M\times 2M} & \mathbf{R}\end{bmatrix}, \tag{17}$$

with $\mathbf{R}$ either equal to the correlation matrix $\mathbf{R}_{\mathrm{y}}$ of the microphone signals, the correlation matrix $\mathbf{R}_{\mathrm{v}}$ of the overall undesired component or the correlation matrix $\mathbf{R}_{\mathrm{n}}$ of the background noise component. The constraint set in (16) is given by

$$\mathbf{C}=\begin{bmatrix}\mathbf{a}_{\mathrm{L}} & \mathbf{0}_{2M\times 1}\\ \mathbf{0}_{2M\times 1} & \mathbf{a}_{\mathrm{R}}\end{bmatrix}, \qquad \mathbf{g}=\begin{bmatrix}1 & 1\end{bmatrix}, \tag{18}$$

requiring the RTF vectors of the desired source. The solution to the optimization problem in (16) using the constraint set in (18) is equal to [1, 12, 21]

$$\mathbf{w}_{\mathrm{MVDR,L}} = \frac{\mathbf{R}^{-1}\mathbf{a}_{\mathrm{L}}}{\mathbf{a}_{\mathrm{L}}^{H}\mathbf{R}^{-1}\mathbf{a}_{\mathrm{L}}}, \quad \mathbf{w}_{\mathrm{MVDR,R}} = \frac{\mathbf{R}^{-1}\mathbf{a}_{\mathrm{R}}}{\mathbf{a}_{\mathrm{R}}^{H}\mathbf{R}^{-1}\mathbf{a}_{\mathrm{R}}}. \tag{19}$$

From a theoretical point of view, in the case of perfectly estimated quantities (i.e., correlation matrices and RTF vector), using $\mathbf{R}=\mathbf{R}_{\mathrm{y}}$ or $\mathbf{R}=\mathbf{R}_{\mathrm{v}}$ in (19) is optimal in the SINR sense, whereas using $\mathbf{R}=\mathbf{R}_{\mathrm{n}}$ in (19) is optimal in the SNR sense. While the BMVDR beamformer preserves the binaural cues of the desired source, its major drawback is the distortion of the binaural cues of the interfering sources (and background noise), such that all sources are perceived as coming from the direction of the desired source. In practice, it should also be realized that using $\mathbf{R}=\mathbf{R}_{\mathrm{y}}$ may lead to target cancellation in the case of RTF estimation errors of the desired source [21] and that $\mathbf{R}_{\mathrm{v}}$ is not straightforward to estimate.

3.2. BLCMV beamformer

In order to also take binaural cue preservation of the interfering sources into account as well as control the amount of interference suppression, it has been proposed in [15] to add interference scaling constraints to the BMVDR beamformer, leading to the BLCMV beamformer. This corresponds to the constrained optimization problem in (18) with the constraint set

$$\mathbf{C}_1 = \begin{bmatrix} \mathbf{a}_{\mathrm{L}} & \mathbf{B}_{\mathrm{L}} & \mathbf{0}_{2M\times 1} & \mathbf{0}_{2M\times P} \\ \mathbf{0}_{2M\times 1} & \mathbf{0}_{2M\times P} & \mathbf{a}_{\mathrm{R}} & \mathbf{B}_{\mathrm{R}} \end{bmatrix}, \quad \mathbf{g}_1 = \begin{bmatrix} 1 & \boldsymbol{\delta}_{\mathrm{L}} & 1 & \boldsymbol{\delta}_{\mathrm{R}} \end{bmatrix}, \tag{20}$$

requiring the RTF vectors of the desired source and all interfering sources. The P-dimensional vectors $\boldsymbol{\delta}_{\mathrm{L}} = \begin{bmatrix}\delta_{\mathrm{L},1}...\delta_{\mathrm{L},P}\end{bmatrix}$ and $\boldsymbol{\delta}_{\mathrm{R}} = \begin{bmatrix}\delta_{\mathrm{R},1}...\delta_{\mathrm{R},P}\end{bmatrix}$ contain the *interference scaling parameters*, which control the suppression and the binaural cue preservation of the P interfering sources. The BLCMV beamformer is given by

$$\mathbf{w}_{\mathrm{LCMV}} = \widetilde{\mathbf{R}}^{-1}\mathbf{C}_1\left(\mathbf{C}_1^{H}\widetilde{\mathbf{R}}^{-1}\mathbf{C}_1\right)^{-1}\mathbf{g}_1^{H}. \tag{21}$$

Setting $\delta_{\mathrm{L},p}=\delta_{\mathrm{R},p}$ ensures binaural cue preservation of the p-th interfering source, while the absolute values of $\delta_{\mathrm{L},p}$ and $\delta_{\mathrm{R},p}$ directly determine the SIR improvement for the p-th interfering source. From a theoretical point of view, in the case of perfectly estimated quantities (i.e., correlation matrices and RTF vectors), setting $\delta_{\mathrm{L},p} = \delta_{\mathrm{R},p} = 0$ in the BLCMV beamformer is optimal in the SIR sense, but not necessarily in the SINR or SNR sense. Moreover, in contrast to the BMVDR beamformer, the choice of the correlation matrix $\mathbf{R}$ has no impact on the SINR, SNR and SIR improvement and the binaural cue preservation as these are completely determined by the interference scaling parameters. In practice, in the case of estimation errors the choice of the correlation matrix $\mathbf{R}$ will obviously have an influence on the performance of the BLCMV beamformer (cf. Section 4).

3.3. Interference scaling parameters

As an extension of the method presented in [22] for an arbitrary number of interfering sources, in this section we propose a method to determine the interference scaling parameters that maximize the SINR or the SNR while preserving the binaural cues of the interfering sources. To this end, we will use the BMVDR beamformer with RTF preservation [14], denoted as BMVDR-RTF beamformer, which is a special case of the BLCMV beamformer. In the BMVDR-RTF beamformer the constraints related to the interfering sources only control the binaural cue preservation while the amount of desired interference suppression is not specified, i.e.,

$$\frac{\mathbf{w}_{\mathrm{L}}^{H}\mathbf{b}_p}{\mathbf{w}_{\mathrm{R}}^{H}\mathbf{b}_p} = \frac{B_{\mathrm{L},p}}{B_{\mathrm{R},p}} \Rightarrow \frac{\mathbf{w}_{\mathrm{L}}^{H}\mathbf{b}_{\mathrm{L},p}}{\mathbf{w}_{\mathrm{R}}^{H}\mathbf{b}_{\mathrm{R},p}} = 1, \tag{22}$$

leading to the constraint set

$$\mathbf{C}_2 = \begin{bmatrix} \mathbf{a}_{\mathrm{L}} & \mathbf{B}_{\mathrm{L}} & \mathbf{0}_{2M\times 1} \\ \mathbf{0}_{2M\times 1} & -\mathbf{B}_{\mathrm{R}} & \mathbf{a}_{\mathrm{R}} \end{bmatrix}, \quad \mathbf{g}_2 = \begin{bmatrix} 1 & \mathbf{0}_{1\times P} & 1 \end{bmatrix}. \tag{23}$$

The BMVDR-RTF beamformer is given by [14]

$$\mathbf{w}_{\mathrm{RTF}} = \widetilde{\mathbf{R}}^{-1}\mathbf{C}_2\left(\mathbf{C}_2^{H}\widetilde{\mathbf{R}}^{-1}\mathbf{C}_2\right)^{-1}\mathbf{g}_2^{H}, \tag{24}$$

and either maximizes the SINR ($\mathbf{R}=\mathbf{R}_{\mathrm{y}}$ or $\mathbf{R}=\mathbf{R}_{\mathrm{v}}$) or the SNR ($\mathbf{R}=\mathbf{R}_{\mathrm{n}}$), while preserving the binaural cues of all sources.
Hence, the optimal interference scaling parameters for the BLCMV beamformer (in the SINR or SNR sense) can be determined as

$$\boxed{\delta_p^{\mathrm{opt}} = \delta_{\mathrm{L},p} = \delta_{\mathrm{R},p} = \mathbf{w}_{\mathrm{RTF,L}}^{H}\mathbf{b}_{\mathrm{L},p} = \mathbf{w}_{\mathrm{RTF,R}}^{H}\mathbf{b}_{\mathrm{R},p}} \tag{25}$$

However, using the optimal interference scaling parameters may lead to problems in practice due to estimation errors of the correlation matrices and RTF vectors. More in particular, in the case of SINR maximization, the corresponding interference scaling parameters may be rather small, leading to a decreased binaural cue preservation performance (cf. simulations in Section 4). On the other hand, in the case of SNR maximization, the corresponding interference scaling parameters may be rather large, depending on the position of the interfering source, leading to an unsatisfying SINR improvement. Hence, we propose to enforce an upper and lower threshold on the optimal interference scaling parameters, i.e.,

$$\delta_p^{\mathrm{thr}} = \begin{cases} |\delta_p^{\mathrm{opt}}|, & \text{if} \quad \delta^{\min} < |\delta_p^{\mathrm{opt}}| < \delta^{\max}, \\ \delta^{\min}, & \text{if} \quad |\delta_p^{\mathrm{opt}}| \le \delta^{\min}, \\ \delta^{\max}, & \text{if} \quad |\delta_p^{\mathrm{opt}}| \ge \delta^{\max}. \end{cases} \tag{26}$$

The thresholds have been experimentally obtained as $\delta^{\min} = 0.2$ and $\delta^{\max} = 0.4$, limiting the theoretically possible SIR improvement for each interfering source between $8\,\mathrm{dB}$ and $14\,\mathrm{dB}$.

3.4. Estimation of correlation matrices and RTFs

All considered binaural beamformers require an estimate of the RTF vectors $\mathbf{a}_{\mathrm{L}}$ and $\mathbf{a}_{\mathrm{R}}$ of the desired source (cf. (11)). In addition, the BLCMV and BMVDR-RTF beamformers require an estimate of the RTF vectors $\mathbf{b}_{\mathrm{L},p}$ and $\mathbf{b}_{\mathrm{R},p}$ of each interfering source. In this paper, we will estimate these RTFs using the covariance whitening approach [23, 24], which is based on the generalized eigenvalue decomposition (GEVD) of the speech + noise correlation matrix $\mathbf{R}_{\mathrm{xn}} = \mathbf{R}_{\mathrm{x}} + \mathbf{R}_{\mathrm{n}}$ and the background noise correlation matrix $\mathbf{R}_{\mathrm{n}}$ or the GEVD of the interference + noise correlation matrix $\mathbf{R}_{\mathrm{v},p} = \mathbf{R}_{\mathrm{u},p} + \mathbf{R}_{\mathrm{n}}$ and $\mathbf{R}_{\mathrm{n}}$. While $\mathbf{R}_{\mathrm{n}}$ can be estimated exploiting the assumed stationarity of the background noise, estimating $\mathbf{R}_{\mathrm{xn}}$ and $\mathbf{R}_{\mathrm{v},p}$ from the available mixture is not straightforward. Due to limited source activity and possible spatial changes of the acoustic scenarios, the *temporal observation interval* that is available in practice for estimating these correlation matrices is typically limited. We assume that the correlation matrix $\mathbf{R}_{\mathrm{xn}}$ can be estimated from an observation interval consisting of T_L frames (corresponding to L seconds) where only the desired source and the background noise are active, i.e.,

$$\hat{\mathbf{R}}_{\mathrm{xn}} = \frac{1}{T_L}\sum_{t=1}^{T_L}\Big(\mathbf{x}(t)+\mathbf{n}(t)\Big)\Big(\mathbf{x}(t)+\mathbf{n}(t)\Big)^{H}, \tag{27}$$

where t is the frame index. Similarly, we assume that the correlation matrix $\mathbf{R}_{\mathrm{v},p}$ can be estimated from an observation interval of T_L frames where only the p-th interfering source and the background noise are active.

4. EXPERIMENTAL RESULTS

In this section, we experimentally investigate the effect of the temporal observation interval on the performance of the BMVDR beamformer ($\mathbf{w}_{\mathrm{MVDR}}$) and the BLCMV beamformer using either the optimal interference scaling parameters ($\mathbf{w}_{\mathrm{LCMV}}(\delta^{\mathrm{opt}})$) or the proposed thresholded interference scaling parameters ($\mathbf{w}_{\mathrm{LCMV}}(\delta^{\mathrm{thr}})$) (cf. Section 3.3).

We consider three different acoustic scenarios comprising of one desired source, one or two interfering sources and diffuse background noise (cf. Table 1 for source positions). The desired source was a male German speaker, the first interfering source was a male Dutch speaker and the second interfering source was a male English speaker. The desired speech and interference components were generated by convolving the desired and interfering source signals with measured impulse responses of binaural behind-the-ear hearing aids mounted on a dummy head in a cafeteria ($T_{60} \approx 1250\mathrm{ms}$) [20], with $M = 2$ microphones per hearing aid. For background noise we used real ambient noise recorded in the same cafeteria with the same setup. The sampling frequency was 16 kHz. All signals start with 2 s of noise-only, followed by about 20 s of all sources being active. The broadband input SNR was set to 5 dB and the SIRs were set to 0 dB.

The noise correlation matrix $\mathbf{R}_{\mathrm{n}}$ was estimated using the 2 s noise-only segment. To estimate the correlation matrices $\mathbf{R}_{\mathrm{y}}$, $\mathbf{R}_{\mathrm{v}}$, $\mathbf{R}_{\mathrm{xn}}$ and $\mathbf{R}_{\mathrm{v},p}$, we considered different temporal observation intervals (starting at 2 s), whose length L ranged between 0.1 s and 3 s. To estimate the correlation matrices $\mathbf{R}_{\mathrm{v}}$, $\mathbf{R}_{\mathrm{xn}}$ and $\mathbf{R}_{\mathrm{v},p}$ the algorithm had access to the respective mixtures. The RTF vectors of the desired source and the interfering source(s) were then calculated based on these estimated correlation matrices (cf. Section 3.4). Please note that shorter temporal observation intervals correspond to larger estimation errors.

The microphone signals were processed using a weighted overlap-add framework with a block length of 256 with 50% overlap and a square-root Hann window. The BMVDR and BLCMV beamformers were calculated using three different correlation matrices, i.e., $\mathbf{R} = \mathbf{R}_{\mathrm{y}}$ (maximizing SINR with possible target cancellation), $\mathbf{R} = \mathbf{R}_{\mathrm{v}}$ (maximizing SINR) and $\mathbf{R} = \mathbf{R}_{\mathrm{n}}$ (maximizing SNR). The filters were used as fixed filters over the whole signal.

As performance measures we used the binaural SINR improvement and the binaural cues errors, i.e., ILD and ITD errors, that we calculated using a model of binaural auditory processing [25]. All performance measures were averaged over all frequencies and all acoustic scenarios.

Figure 2 depicts the SINR improvement for different lengths of the temporal observation interval and for different correlation matrices, while Figure 3 depicts the binaural cue errors of the first interfering source for the same temporal observation intervals and $\mathbf{R} = \mathbf{R}_{\mathrm{v}}$. First, it can be observed that when using $\mathbf{R}_{\mathrm{y}}$ or $\mathbf{R}_{\mathrm{v}}$ the SINR improvement is generally larger than when using $\mathbf{R}_{\mathrm{n}}$. This is expected because using the noise correlation matrix $\mathbf{R}_{\mathrm{n}}$ is maximizing the SNR and not the SINR. Second, when using $\mathbf{R}_{\mathrm{y}}$ or $\mathbf{R}_{\mathrm{v}}$, an apparent difference can be seen for small observation intervals below 200 ms. The small observation intervals lead to larger estimation errors for the correlation matrices and hence also for the RTF vectors, such that the drop in SINR improvement observed when using $\mathbf{R}_{\mathrm{y}}$ is probably attributed to target cancellation. For longer observation intervals and hence smaller estimation errors, the difference between using $\mathbf{R}_{\mathrm{y}}$ and $\mathbf{R}_{\mathrm{v}}$ is smaller. As expected, the SINR improvement of the BLCMV beamformer using the thresholded interference scaling parameters δ^{thr} is smaller than for the BLCMV beamformer using the optimal interference scaling parameters δ^{opt}. Although, looking at the binaural cue errors, using δ^{thr} in the BLCMV beamformer leads to much better binaural cue preservation, while using δ^{opt} leads to similar binaural cue errors as for the BMVDR beamformer. This difference is especially visible for the ITD error at small observation intervals and is also confirmed by informal listening tests. Third, when using $\mathbf{R}_{\mathrm{n}}$, the BLCMV beamformer outperforms the BMVDR beamformer for longer observation intervals above 300ms because of the additional constraints. Additionally, using δ^{thr} in the BLCMV beamformer apparently leads to marginally better SINR improvement in this case. Because $\mathbf{R}_{\mathrm{v}}$ is in practice very hard to accurately estimate, it should be recommended to use $\mathbf{R}_{\mathrm{n}}$ when short observation intervals are required (e.g., in dynamic acoustic scenarios) and to use δ^{thr} in the BLCMV beamformer to prevent binaural cue errors.

Scenario	1	2	3
Desired	-35°	0°	0°
Interfering	150°	-35°	-35°, 150°

Table 1: Spatial scenarios (0°: frontal direction. -90°: left hand side. 90°: right hand side).

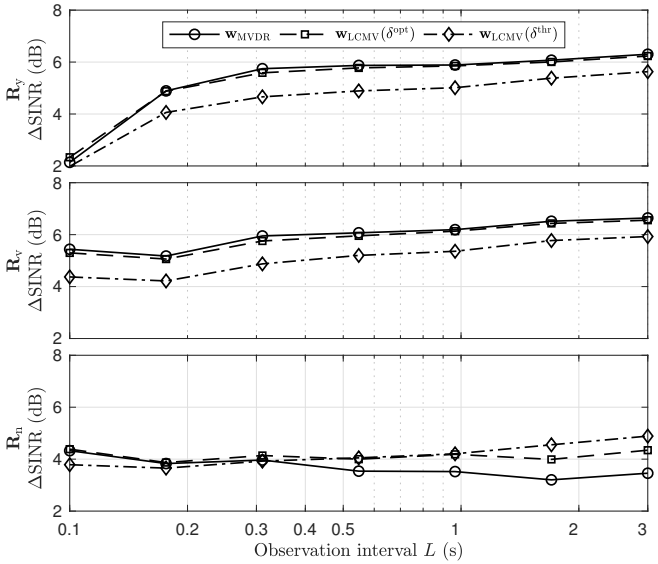

Fig. 2: SINR improvement for different temporal observation intervals for $\mathbf{R} = \mathbf{R}_{\mathrm{y}}$ (top), $\mathbf{R} = \mathbf{R}_{\mathrm{v}}$ (mid) and $\mathbf{R} = \mathbf{R}_{\mathrm{n}}$ (bottom).

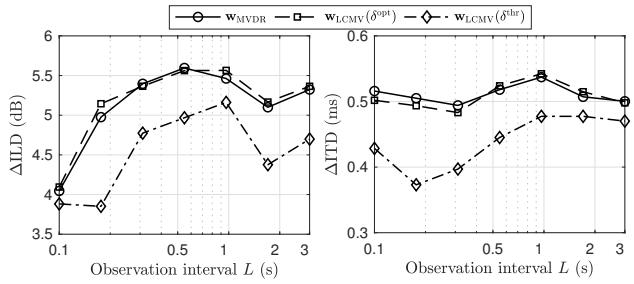

Fig. 3: Binaural cue errors of the first interfering source ($\mathbf{R} = \mathbf{R}_{\mathrm{v}}$).

5. CONCLUSIONS

In this paper, we proposed optimal values for the interference scaling parameters in the BLCMV beamformer for an arbitrary number of interfering sources based on the BMVDR-RTF beamformer. We showed how to set these parameters in practice such that a robust performance in the case of estimation errors can be achieved. Evaluation results in a complex acoustic scenario showed that even short temporal observation intervals for estimating the required correlation matrices and RTF vectors are sufficient to achieve a decent noise reduction performance and binaural cue preservation.

6. REFERENCES

[1] S. Doclo, W. Kellermann, S. Makino, and S.E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[2] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural Speech Processing with Application to Hearing Devices," in *Audio Source Separation and Speech Enhancement*, chapter 18. Wiley, 2018.

[3] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

[4] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–14, 2006.

[5] G. Grimm, V. Hohmann, and B. Kollmeier, "Increase and Subjective Evaluation of Feedback Stability in Hearing Aids by a Binaural Coherence-based Noise Reduction Scheme," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1408–1419, Sep. 2009.

[6] A. H. Kamkar-Parsi and M. Bouchard, "Improved Noise Power Spectrum Density Estimation for Binaural Hearing Aids Operating in a Diffuse Noise Field Environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 521–533, May 2009.

[7] A. H. Kamkar-Parsi and M. Bouchard, "Instantaneous Binaural Target PSD Estimation for Hearing Aid Noise Reduction in Complex Acoustic Environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 4, pp. 1141–1154, Apr. 2011.

[8] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, 2013.

[9] R. Baumgärtel, M Krawczyk-Becker, D. Marquardt, C. Völker, H. Hu, T. Herzke, G. Coleman, K. Adiloğlu, S. M. A. Ernst, T. Gerkmann, S. Doclo, B. Kollmeier, V. Hohmann, and M. Dietz, "Comparing binaural signal processing strategies I: Instrumental evaluation.," *Trends in Hearing*, vol. 19, pp. 1–16, 2015.

[10] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct. 2017, pp. 234–238.

[11] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann, "Multichannel source separation preserving spatial information," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu HI, USA, Apr. 2007, pp. 5–8.

[12] B. Cornelis, S. Doclo, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural multi-microphone noise reduction techniques," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[13] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical Analysis of Linearly Constrained Multi-channel Wiener Filtering Algorithms for Combined Noise Reduction and Binaural Cue Preservation in Binaural Hearing Aids," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2384–2397, Dec. 2015.

[14] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical Analysis of Binaural Transfer Function MVDR Beamformers with Interference Cue Preservation Constraints," *IEEE/ACM Trans. Audio, Speech and Language Proc.*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[15] E. Hadad, S. Doclo, and S. Gannot, "The Binaural LCMV Beamformer and its Performance Analysis," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 24, no. 3, pp. 543–558, 2016.

[16] E. Hadad, D. Marquardt, W. Pu, S. Gannot, S. Doclo, Z.-Q. Luo, I. Merks, and T. Zhang, "Comparison of two binaural beamforming approaches for hearing aids," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 236–240.

[17] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 25, no. 1, pp. 137–152, Jan. 2017.

[18] W. Pu, J. Xiao, T. Zhang, and Z.-Q. Luo, "A penalized inequality-constrained minimum variance beamformer with applications in hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct. 2017, pp. 175–179.

[19] D. Marquardt and S. Doclo, "Interaural Coherence Preservation for Binaural Noise Reduction Using Partial Noise Estimation and Spectral Postfiltering," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 26, no. 7, pp. 1257–1270, 2018.

[20] H. Kayser, S. Ewert, J. Annemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel In-Ear and Behind-The-Ear Head-Related and Binaural Room Impulse Responses," *Eurasip Journal on Advances in Signal Processing*, vol. 2009, pp. 10 pages, 2009.

[21] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[22] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Optimal binaural LCMV beamformers for combined noise reduction and binaural cue preservation," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, Sep. 2014, pp. 288–292.

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

[24] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 544–548.

[25] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, pp. 592–605, 2011.