COMBINING LINEAR SPATIAL FILTERING AND NON-LINEAR PARAMETRIC PROCESSING FOR HIGH-QUALITY SPATIAL SOUND CAPTURING

Oliver Thiergart, Guendalina Milano, and Emanuël A. P. Habets

International Audio Laboratories Erlangen*, Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT

Flexible spatial sound capturing and reproduction can be achieved with multiple microphones by using linear spatial filtering or non-linear parametric processing. The non-linear approaches usually provide a superior spatial resolution compared to the linear approaches but can result in artifacts due to violations of the sound field model. In this paper, we combine both approaches to achieve a high robustness against model violations and a high spatial resolution. We assume linear spatial filters that approximate the spatial responses of the desired output format and compensate remaining deviations with an optimal post filter. The post filter is computed such that the proposed approach behaves like a linear system when the spatial filters achieve the desired spatial response, and scales towards a non-linear system otherwise. Experimental results show that the proposed approach can significantly reduce distortions of existing parametric processing schemes especially when a sufficiently high number of microphones is available.

***Index Terms*—** Microphone arrays, spatial sound acquisition, optimum filtering

1. INTRODUCTION

Spatial sound capturing and reproduction is of increasing relevance in recent audio applications. A popular approach to parametric spatial sound capturing and reproduction, which can provide a higher spatial resolution compared to classical linear spatial sound reproduction methods [1], is represented by directional audio coding (DirAC) [2]. The non-linear DirAC processing assumes that for each time and frequency, the sound scene can be decomposed into a direct sound component and a diffuse sound component. Together with parametric side-information, such as the direction-of-arrival (DOA) of the direct sound, it is possible to synthesize the loudspeaker signals that recreate the original spatial impression of the sound scene.

Assuming a single direct sound component per time and frequency requires that the source signals composing the sound scene are sufficiently sparse such that at most one source is dominant for each time-frequency point. While this model assumption is typically fulfilled when recording speech sources, it often is violated for more complex natural sound scenes which degrades the spatial rendering [3]. To improve the robustness against such model violations, some parametric approaches assume a signal model with multiple direct sounds per time and frequency [4, 5]. The drawback of these approaches is a high computational complexity due to the estimation of multiple DOAs and the recomputation of the signal-adaptive multi-channel filters. To enable robust spatial sound capturing with lower computational complexity, we have used a fixed multi-channel filtering approach together with an optimal post filter in [6]. The post filter is derived assuming a single-wave sound field model, however, its effect on the processing vanishes if the multi-channel filters closely approximate the desired target responses of the spatial sound reproduction system. The approach was able to reduce the effect of model violations and DOA estimation errors. However, as only the direct sound capturing and reproduction was considered, it cannot be used for high-quality spatial sound capturing.

In this paper, we extend the work in [6] and derive an approach for the robust capturing and reproduction of both direct sounds and diffuse sound (e. g., ambience). Similarly as in [6], we combine the non-linear DirAC processing with classical linear spatial filtering to achieve a high robustness against model violations and a high spatial resolution. We assume linear spatial filters that approximate the spatial responses of the desired output format and compensate remaining deviations using optimal post filters. The post filters are computed such that the proposed approach behaves like a linear system (which is robust against model violations and parameter estimation errors) when the spatial filters achieve the desired spatial response, and scales towards a non-linear system otherwise.

The paper is structured as follows: Section 2 introduces the sound field model and reviews the fundamentals of perceptually motivated parametric spatial sound processing. Section 3 briefly reviews the state-of-the-art (SOA) DirAC approach. The proposed approach is explained in Sec. 4. Experimental results are shown in Sec. 5. Section 6 concludes the paper.

2. PARAMETRIC SPATIAL SOUND PROCESSING

All processing is carried out in the time-frequency domain with frequency index k and time frame index n. We model the sound field $P(k,n,\mathbf{r})$ at position $\mathbf{r}$ as a sum of a direct sound component and a diffuse sound component, i. e.,

$$P(k,n) = P_{\mathrm{s}}(k,n,\mathbf{r}) + P_{\mathrm{d}}(k,n,\mathbf{r}). \tag{1}$$

The direct sound $P_{\mathrm{s}}(k,n,\mathbf{r})$ and diffuse sound $P_{\mathrm{d}}(k,n,\mathbf{r})$ are assumed to be uncorrelated. The direct sound models the direct sound of the sources while the diffuse sound models the reverberation or ambience. Typically, $P_{\mathrm{s}}(k,n,\mathbf{r})$ is represented by a plane wave with DOA $\varphi(k,n)$ and power $\mathrm{E}\left\{|P_{\mathrm{s}}(k,n,\mathbf{r})|^2\right\} = \Phi_{\mathrm{s}}(k,n)$. The diffuse sound $P_{\mathrm{d}}(k,n,\mathbf{r})$ is modeled as a sum of infinitely many plane waves arriving with random phases from uniformly distributed DOAs [7]. Thus, $P_{\mathrm{d}}(k,n,\mathbf{r})$ can be considered as a complex Gaussian distributed random variable and its expected power is given by $\mathrm{E}\left\{|P_{\mathrm{d}}(k,n,\mathbf{r})|^2\right\} = \Phi_{\mathrm{d}}(k,n)$. The power ratio between the direct sound and diffuse sound is described by the signal-to-diffuse ratio (SDR) given by

$$\mathrm{SDR}(k,n) = \frac{\Phi_{\mathrm{s}}(k,n)}{\Phi_{\mathrm{d}}(k,n)}. \tag{2}$$

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

Note that in practice, the parameters $\varphi(k,n)$, $\Phi_{\mathrm{s}}(k,n)$, $\Phi_{\mathrm{d}}(k,n)$, and $\mathrm{SDR}(k,n)$ are highly time-varying.

We aim at reproducing the sound such that it is *perceptually* equivalent to the original spatial sound at the reference position $\mathbf{r}_1$. This means that the direct sound should be reproduced without distortions from the original direction, while the diffuse sound should be reproduced with the original ambient impression [2]. For the sound field model in (1), the l-th output signal $Z_l(k,n)$ of a spatial sound reproduction system with L channels is given by a weighted sum of the direct sound at $\mathbf{r}_1$ and a diffuse signal, i. e.,

$$Z_l(k,n) = Z_{\mathrm{s},l}(k,n) + Z_{\mathrm{d},l}(k,n) \tag{3a}$$

$$= G_{\mathrm{s},l}(k,\varphi) P_{\mathrm{s}}(k,n,\mathbf{r}_1) + Z_{\mathrm{d},l}(k,n). \tag{3b}$$

The signal $Z_{\mathrm{s},l}(k,n)$ is the target direct signal for the l-th channel and $Z_{\mathrm{d},l}(k,n)$ is the target diffuse signal. The target response $G_{\mathrm{s},l}(k,\varphi)$ depends on the DOA $\varphi(k,n)$ and assures that the direct sound is reproduced from the original direction. When rendering to loudspeakers, $G_{\mathrm{s},l}(k,\varphi)$ corresponds to a panning gain [8].

The target diffuse signal $Z_{\mathrm{d},l}(k,n)$ is proportional to an arbitrary realization of the (random) diffuse field $P_{\mathrm{d}}(k,n,\mathbf{r})$, i. e., the correlation between the actual $Z_{\mathrm{d},l}(k,n)$ and the original $P_{\mathrm{d}}(k,n,\mathbf{r})$ can be arbitrary. However, to recreate the original ambient impression, the signals $Z_{\mathrm{d},l}(k,n)$ should be sufficiently uncorrelated across l, which is achieved by applying decorrelators to $Z_{\mathrm{d},l}(k,n)$ before reproduction. Moreover, $Z_{\mathrm{d},l}(k,n)$ should posses the correct power, which is related to the original diffuse power as

$$\mathrm{E}\left\{|Z_{\mathrm{d},l}(k,n)|^2\right\} = Q_{\mathrm{d},l}(k)\Phi_{\mathrm{d}}(k,n). \tag{4}$$

The target directivity factor $Q_{\mathrm{d},l}(k)$ assures that each channel l reproduces the correct amount of diffuse sound. When rendering to loudspeakers, we typically use $Q_{\mathrm{d},l}(k) = L^{-1}$ [2].

The single-wave sound field model in (1) is violated in practice when multiple direct sounds appear per time and frequency, which can occur e. g. in multi-source scenarios or in the presence of strong early reflections. In this case, the ideal target direct signal $Z_{\mathrm{s},l}(k,n)$ in (3) would be a weighted sum of multiple direct sound components, each direct sound component being weighted with the target response of the corresponding DOA. However, considering multiple direct sound components would significantly increase the complexity of the spatial sound reproduction system.

3. DIRECTIONAL AUDIO CODING

DirAC [2] represents the SOA in perceptually motivated, parametric spatial sound processing and implements the principles explained in Sec. 2. DirAC assumes the single-wave sound field model in (1). The target signals $Z_{\mathrm{s},l}(k,n)$ and $Z_{\mathrm{d},l}(k,n)$ in (3) are estimated using M microphones located at $\mathbf{r}_{1\ldots M}$. Given the sound field model (1), the microphone signals $\mathbf{x}(k,n) = [X_1(k,n),\ldots,X_M(k,n)]^{\mathrm{T}}$ can be written as

$$\mathbf{x}(k,n) = \mathbf{x}_{\mathrm{s}}(k,n) + \mathbf{x}_{\mathrm{d}}(k,n), \tag{5}$$

where $\mathbf{x}_{\mathrm{s}}(k,n)$ are the microphone signals corresponding to the direct sound $P_{\mathrm{s}}(k,n,\mathbf{r}_{1\ldots M})$ and $\mathbf{x}_{\mathrm{d}}(k,n)$ are the microphone signals corresponding to the diffuse sound $P_{\mathrm{d}}(k,n,\mathbf{r}_{1\ldots M})$. Note that DirAC does not assume microphone noise in the signal model.

In the single-channel variant of DirAC [2], the target signals $Z_{\mathrm{s},l}(k,n)$ and $Z_{\mathrm{d},l}(k,n)$ are estimated from a single omnidirectional microphone signal $X_1(k,n)$ with

$$\widehat{Z}_{\mathrm{s},l}(k,n) = H_{\mathrm{s},l}(k,n) X_1(k,n), \tag{6a}$$

$$\widehat{Z}_{\mathrm{d},l}(k,n) = H_{\mathrm{d},l}(k,n) X_1(k,n), \tag{6b}$$

where the single-channel filters are given by

$$H_{\mathrm{s},l}(k,n) = \frac{G_{\mathrm{s},l}(k,\varphi)\mathrm{SDR}(k,n)}{1+\mathrm{SDR}(k,n)}, \tag{7a}$$

$$H_{\mathrm{d},l}(k,n) = \frac{\sqrt{Q_{\mathrm{d},l}(k)}}{1+\mathrm{SDR}(k,n)}. \tag{7b}$$

These filters represent the optimal Wiener filters[1] for estimating $Z_{\mathrm{s},l}(k,n)$ and $Z_{\mathrm{d},l}(k,n)$, as defined in (3) and (4), from $X_1(k,n)$.

In the multi-channel DirAC variant [2], the target signals are estimated from multiple microphone signals $\mathbf{x}(k,n)$ using a spatial filter plus subsequent post filter, i. e.,

$$\widehat{Z}_{\mathrm{s},l}(k,n) = H_{\mathrm{s},l}(k,n) Y_l(k,n), \tag{8a}$$

$$\widehat{Z}_{\mathrm{d},l}(k,n) = H_{\mathrm{d},l}(k,n) Y_l(k,n), \tag{8b}$$

where the output of the l-th multi-channel filter is computed as

$$Y_l(k,n) = \mathbf{w}_l^{\mathrm{H}}(k)\mathbf{x}(k,n). \tag{9}$$

The spatial filter $\mathbf{w}_l(k)$ is computed as a fixed beamformer that is directed towards the l-th loudspeaker position, e. g., a delay-and-sum filter [10]. The post filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ in (8) are computed using (7). This is not optimal since (7) does not represent the Wiener filters for estimating $Z_{\mathrm{s},l}(k,n)$ and $Z_{\mathrm{d},l}(k,n)$ from $Y_l(k,n)$. Using the filter $H_{\mathrm{s},l}(k,n)$ can lead to an incorrect panning of the direct sound since the directivity of the spatial filter $\mathbf{w}_l(k,n)$ is not considered in (7a) [6]. However, using the beamformer signals $Y_l(k,n)$ in (8) can make the direct sound reproduction more stable and reduces the need of decorrelation for the diffuse sound, which reduces decorrelation artifacts [2].

In general, computing the filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ requires to estimate the SDR and DOA $\varphi(k,n)$ for each (k,n) and estimation errors in these parameters can strongly affect the reproduced spatial image. When the signal model in (1) is violated, e. g., when multiple sources are active per time and frequency, the single direct sound component and direct response $G_{\mathrm{s},l}(k,\varphi)$ in the target signal $Z_l(k,n)$ in (3) is not sufficient to represent and recreate the original spatial impression, which can lead to spatial distortions [3].

4. PROPOSED APPROACH

In the following, we propose an approach to estimate the target signals $Z_{\mathrm{s},l}(k,n)$ and $Z_{\mathrm{d},l}(k,n)$ in (3) using multiple microphones. The approach is derived such that the estimation is robust against parameter estimation errors and violations of the sound field model (1).

For this purpose, we estimate the target signals similarly as in the multi-channel DirAC approach in Sec. 3 with (8) and (9). However, we assume a fixed spatial filter $\mathbf{w}_l(k)$ which closely approximates the target response $G_{\mathrm{s},l}(k,\varphi)$ and directivity factor $Q_{\mathrm{d},l}(k)$. If the approximation is accurate, the post filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ [which rely on the signal model (1)] are not required since the output signal $Y_l(k,n)$ of the linear spatial filter $\mathbf{w}_l(k)$ would correspond to the desired target signal $Z_l(k,n)$[2]. Therefore, we aim to compute $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ in an optimal way such that when $\mathbf{w}_l(k)$ is accurate, the post filters have no effect on $Z_l(k,n)$ avoiding any influence of model violations and parameter estimation errors.

[1] The original DirAC approach [2] uses the square-root Wiener filters. Moreover, it defines the filters based on the diffuseness $\Psi(k,n)$, which is related to the SDR as $\Psi(k,n) = [1+\mathrm{SDR}(k,n)]^{-1}$ [9].

[2] This assumes that the diffuse sound at the output of the filter is sufficiently uncorrelated across the channels l.

When $\mathbf{w}_l(k)$ deviates from the desired spatial response, $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ become effective and assure a correct spatial sound reproduction based on the sound field model (1).

4.1. Derivation of the Optimal Post Filters

In the following, we derive optimal single-channel filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ for estimating the target signals $Z_{\mathrm{s},l}(k,n)$ and $Z_{\mathrm{d},l}(k,n)$ from the output signal $Y_l(k,n)$ of the spatial filter in (9). We consider the sound field model in (1) and corresponding microphone signal model (5). Given the model, $Y_l(k,n)$ becomes

$$Y_l(k,n) = \mathbf{w}_l^{\mathrm{H}}(k)\mathbf{x}_{\mathrm{s}}(k,n) + \mathbf{w}_l^{\mathrm{H}}(k)\mathbf{x}_{\mathrm{d}}(k,n) \tag{10a}$$
$$= G_{\mathbf{w},l}(k,\varphi)P_{\mathrm{s}}(k,n,\mathbf{r}_1) + Y_{\mathrm{d},l}(k,n), \tag{10b}$$

where $G_{\mathbf{w},l}(k,\varphi) = \mathbf{w}_l^{\mathrm{H}}(k)\mathbf{a}(k,\varphi)$ is the directivity of the spatial filter $\mathbf{w}_l(k)$, $\mathbf{a}(k,\varphi)$ contains the relative transfer functions between the reference position $\mathbf{r}_1$ and all other microphone positions for a plane wave from direction φ, and $P_{\mathrm{s}}(k,n,\mathbf{r}_1)$ is the direct sound at the reference position. Moreover, $Y_{\mathrm{d},l}(k,n)$ is the filtered diffuse sound. The expected power of $Y_{\mathrm{d},l}(k,n)$ is

$$\mathrm{E}\left\{|Y_{\mathrm{d},l}(k,n)|^2\right\} = \Phi_{\mathrm{d}}(k,n)Q_{\mathbf{w},l}(k), \tag{11}$$

where $Q_{\mathbf{w},l}(k) = \mathbf{w}_l^{\mathrm{H}}(k)\mathbf{\Gamma}_{\mathrm{d}}(k)\mathbf{w}_l(k)$ is the directivity factor of the spatial filter and $\mathbf{\Gamma}_{\mathrm{d}}(k)$ is the diffuse coherence matrix. For omnidirectional microphones and a spherically isotropic diffuse field, $\mathbf{\Gamma}_{\mathrm{d}}(k)$ consists of sinc functions depending on the wavenumber and inter-microphone distances [11]. The optimal post filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ are found by minimizing the mean-square error (MSE) between the true and estimated target signals, i. e.,

$$H_{(\cdot),l}(k,n) = \arg\min_{H} \mathrm{E}\left\{|Z_{(\cdot),l}(k,n) - \widehat{Z}_{(\cdot),l}(k,n)|^2\right\}. \tag{12}$$

The solution to the optimization problem is found by substituting (2) – (4) and (8) – (11) and equating the first derivative of the cost function w. r. t. $H^*_{(\cdot),l}(k,n)$ to zero. This yields the optimal filters

$$H_{\mathrm{s},l}(k,n) = \frac{G_{\mathrm{s},l}(k,\varphi)G^*_{\mathbf{w},l}(k,\varphi)\mathrm{SDR}(k,n)}{|G_{\mathbf{w},l}(k,\varphi)|^2\mathrm{SDR}(k,n) + Q_{\mathbf{w},l}(k)}, \tag{13a}$$

$$H_{\mathrm{d},l}(k,n) = \frac{\sqrt{Q_{\mathrm{d},l}(k)}\mathbf{b}_l^{\mathrm{H}}(k,n)\mathbf{w}_l(k)}{|G_{\mathbf{w},l}(k,\varphi)|^2\mathrm{SDR}(k,n) + Q_{\mathbf{w},l}(k)}. \tag{13b}$$

Note that the filters can be complex-valued. The m-th element of the vector $\mathbf{b}_l(k,n)$ in (13b), given by

$$B_{m,l}(k,n) = \frac{\mathrm{E}\left\{Z^*_{\mathrm{d},l}(k,n)X_{\mathrm{d},m}(k,n)\right\}}{\sqrt{\mathrm{E}\left\{|Z_{\mathrm{d},l}(k,n)|^2\right\}\mathrm{E}\left\{|X_{\mathrm{d},m}(k,n)|^2\right\}}}, \tag{14}$$

is the desired coherence between the target diffuse signal $Z_{\mathrm{d},l}(k,n)$ and diffuse sound $X_{\mathrm{d},m}(k,n)$ captured by the m-th microphone. Note that the denominator in (14) is equal to $\sqrt{Q_{\mathrm{d},l}(k)}\Phi_{\mathrm{d}}(k,n)$ when using omnidirectional microphones.

4.2. Application of the Optimal Post Filters

The post filter $H_{\mathrm{d},l}(k,n)$ in (13b) is the optimal Wiener filter for estimating the target diffuse signal $Z_{\mathrm{d},l}(k,n)$, which is correlated with the diffuse microphone signals $\mathbf{x}_{\mathrm{d}}(k,n)$ as specified by the coherence vector $\mathbf{b}_l(k,n)$, from $Y_l(k,n)$. As discussed in Sec. 2, the desired correlation between $Z_{\mathrm{d},l}(k,n)$ and the true (captured) diffuse sound can be defined arbitrarily in our application. Thus, the coherences $B_{m,l}(k,n)$ can be defined arbitrarily with the restriction $|B_{m,l}(k,n)| \leq 1$, which follows from (14).

As explained in the beginning of this section, we aim at a spatial sound reproduction system where the post filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ in (8) have no effect if the spatial filter $\mathbf{w}_l(k,n)$ in (9) accurately approximates the desired spatial target responses. This means that the estimated target signal should be

$$\widehat{Z}_l(k,n) = \widehat{Z}_{\mathrm{s},l}(k,n) + \widehat{Z}_{\mathrm{d},l}(k,n) \tag{15a}$$
$$= Y_l(k,n), \tag{15b}$$

in case $G_{\mathbf{w},l}(k,\varphi) = G_{\mathrm{s},l}(k,\varphi)\,\forall\varphi$ and $Q_{\mathbf{w},l}(k) = Q_{\mathrm{d},l}(k)$, which means that $H_{\mathrm{s},l}(k,n) + H_{\mathrm{d},l}(k,n) = 1$. It can be shown that this property is obtained when defining the arbitrary coherence vector in (13b) as

$$\mathbf{b}_l^{\mathrm{H}}(k) = \sqrt{Q_{\mathbf{w},l}(k)}\frac{\mathbf{w}_l^{\mathrm{H}}(k)}{\|\mathbf{w}_l(k)\|}. \tag{16}$$

In case the spatial filter $\mathbf{w}_l(k,n)$ deviates from the desired spatial response, i. e., if $G_{\mathbf{w},l}(k,\varphi) \neq G_{\mathrm{s},l}(k,\varphi)$ and/or $Q_{\mathbf{w},l}(k) \neq Q_{\mathrm{d},l}(k)$, we have $H_{\mathrm{s},l}(k,n) + H_{\mathrm{d},l}(k,n) \neq 1$. In this case, the post filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ become effective and assure a correct spatial rendering based on the sound field model (1). When only a single microphone is used, i. e., $\mathbf{w}_l(k) = [1,0,\ldots,0]^{\mathrm{T}}$, $G_{\mathbf{w},l}(k,\varphi) = 1$, and $Q_{\mathbf{w},l}(k) = 1$, the proposed approach becomes equal to the nonlinear, fully parametric single-channel DirAC approach in Sec. 3.

4.3. Computation of the Spatial Filter

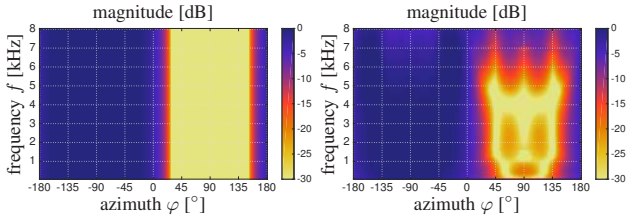
To compute the spatial filter $\mathbf{w}_l(k)$, we consider the well-known least squares (LS) filter derived in [12], which approximates the target direct response $G_{\mathrm{s},l}(k,\varphi)$ for a number of A discrete directions φ in the LS sense. The filter is defined as

$$\mathbf{w}_l(k) = \arg\min_{\mathbf{w}} \sum_{i=1}^{A} |\mathbf{w}^{\mathrm{H}}\mathbf{a}(k,\varphi_i) - G_{\mathrm{s},l}(k,\varphi)|^2. \tag{17}$$

This filter can be computed subject to a white-noise-gain (WNG) constraint as proposed in [13], which assures a specific minimum WNG $\beta(k)$ for sufficient robustness against spatially white noise. By controlling $\beta(k)$, we can control the trade-off between noise robustness and approximation accuracy of the target response $G_{\mathrm{s},l}(k,\varphi)$. In doing so, we can scale the proposed spatial sound reproduction approach between a non-linear, parametric processing scheme (with high noise robustness) and a processing scheme that behaves like a linear system (with high robustness against model violations). In fact, a lower $\beta(k)$ results in a more accurate approximation of $G_{\mathrm{s},l}(k,\varphi)$, and hence, less post filtering is required as discussed in the previous subsection.

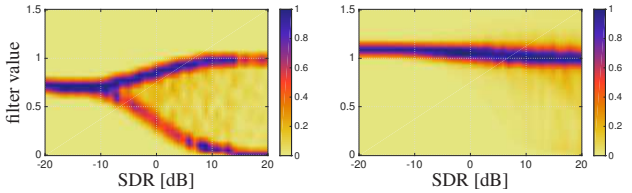
5. SIMULATION RESULTS

We consider a uniform circular array (UCA) with radius $r = 3\,\mathrm{cm}$ and $M = 8$ omnidirectional microphones. We assume a stereo loudspeaker setup ($L = 2$). The target responses $G_{\mathrm{s},l}(k,\varphi)$ are computed using the vector-base amplitude panning (VBAP) scheme [8]. As an example, the target response for the second channel is depicted in Fig. 1(a). For the stereo setup, the target directivity factor is $Q_{\mathrm{d},l}(k) = L^{-1} = 0.5$. To violate the sound field model in (1), we compute the direct microphone signals $\mathbf{x}_{\mathrm{s}}(k,n)$ in (5) by summing two uncorrelated plane waves with specific power and DOAs φ_1 and φ_2. The diffuse microphone signals $\mathbf{x}_{\mathrm{d}}(k,n)$ are found by summing 1000 uncorrelated plane waves with random DOAs [7].



(a) Target response $G_{\mathrm{s},2}(k,\varphi)$ (b) Directivity $G_{\mathbf{w},2}(k,\varphi)$

Fig. 1. Example target response $G_{\mathrm{s},2}(k,\varphi)$ and achieved directivity pattern when using the constraint LS filter with $\beta(k) = -15\,\mathrm{dB}$ (UCA, $M = 8$, $r = 3\,\mathrm{cm}$).



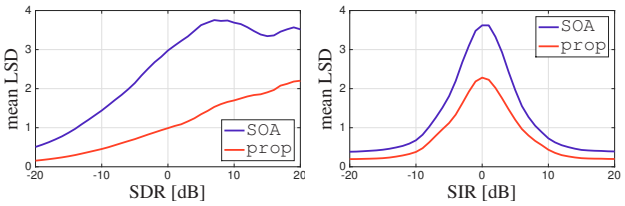
(a) SOA (b) prop

Fig. 2. Distribution of the summed filter $H_{\mathrm{s},l}(k,n) + H_{\mathrm{d},l}(k,n)$ for a scenario with two direct sounds plus diffuse sound. The relative frequency of occurrence of the summed filter value is coded in color. The distribution was computed across frequencies and realizations.

The spatial filter $\mathbf{w}_l(k)$ is computed using the constrained LS filter in Sec. 4.3 with a minimum WNG of $\beta(k) = -3\,\mathrm{dB}$. Figure 1(b) shows the achieved directivity pattern of the spatial filter $\mathbf{w}_2(k)$, which approximates the response in Fig. 1(a). The approximation is relatively accurate for the speech relevant frequency range between $400\,\mathrm{Hz}$ and $4\,\mathrm{kHz}$. However, we can observe undesired side lobes in the directivity pattern for $\varphi \in [45°, 135°]$. Moreover, the approximation becomes inaccurate for very low frequencies (due to the WNG constraint) and towards the spatial aliasing frequency of $7.4\,\mathrm{kHz}$. The directivity factor $Q_{\mathbf{w},l}(k)$ of the spatial filters is computed as discussed below (11). For both filters $\mathbf{w}_1(k)$ and $\mathbf{w}_2(k)$, the average of the directivity factor for the frequency range below $4\,\mathrm{kHz}$ is 0.424, which is close to the desired value $Q_{\mathrm{d},l}(k)$.

Computing the post filters in Sec. 3 and Sec. 4 requires to compute the target responses $G_{\mathrm{s},l}(k,\varphi)$ assuming a single DOA $\varphi(k,n)$ and to estimate the $\mathrm{SDR}(k,n)$. To avoid any specific affect of a practical parameter estimator, we simulate ideal estimators. We estimate $\varphi(k,n)$ by summing the direction vectors corresponding to the two direct sound DOAs φ_1 and φ_2 and then taking the angle. Each direction vector is weighted with the power of the corresponding plane wave. The SDR is computed as the ratio of the summed power of both direct sound plane waves and the diffuse sound power.

First, we study the SOA post filters $H_{\mathrm{s},l}(k,n)$ and $H_{\mathrm{d},l}(k,n)$ in (7) (denoted by SOA) and the proposed post filters in (13) (denoted by prop) for different SDRs for the speech relevant frequency range below $4\,\mathrm{kHz}$. Figure 2 shows the distribution of the summed filters $H_{\mathrm{s},l}(k,n) + H_{\mathrm{d},l}(k,n)$ computed over all frequencies and channels l over 10000 realizations of the sound filed. The direct sound DOAs φ_1 and φ_2 were chosen randomly for each realization. For the SOA filter in Fig. 2(a), the filter values are concentrated around $\sqrt{0.5}$ for low SDRs. Here, the diffuse filter $H_{\mathrm{d},l}(k,n)$ is dominant and assures the desired target directivity $Q_{\mathrm{d}}(k)$. Towards large SDRs, the sum of both filters is mostly either zero or one, depending on the estimated DOA φ. Here, the direct filter $H_{\mathrm{s},l}(k,n)$ is dominant and leads to the panning of the direct sound based on the estimated DOA and the panning function in Fig. 1(a). If the direct filter becomes zero instead of one, which can occur in case of model violations (when multiple sources are active per time and frequency) or DOA estimation errors, the filter would introduce strong distortions to the direct signal. The distribution of the summed filter values for the proposed post filters are depicted in Fig. 2(b). At low SDRs, the diffuse filter $H_{\mathrm{d},l}(k,n)$ is dominant and assures the correct power of the reproduced diffuse sound. Most filter values are slightly larger than one to compensate for the directivity factor $Q_{\mathbf{w}}(k)$ of the spatial filter, which is slightly smaller than the target directivity $Q_{\mathrm{d}}(k)$. For larger SDRs, both proposed filters sum up to one for most realizations. This represents the desired property as discussed in the previous section. The observed slight deviations from this property result from the filter directivity $G_{\mathbf{w},l}(k,\varphi)$, which is slightly different from the target directivity $G_{\mathrm{s},l}(k,\varphi)$ as shown in Fig. 1.



(a) Different SDRs (b) Different SIRs

Fig. 3. Mean LSD for a scenario with two direct sounds with random DOAs plus diffuse sound.

Finally, we study the performance of the entire system. For this purpose, we consider the mean log spectral distortion (LSD) [14] of the estimated target signal $\widehat{Z}_l(k,n)$. The true target signal $Z_l(k,n)$ is computed using (3b), however, considering two direct sound components instead of one. For the target diffuse signal $Z_{\mathrm{d},l}(k,n)$ in (3b), we use the diffuse microphone signal $X_{\mathrm{d},1}(k,n)$ for the SOA approach and the (power adjusted) diffuse signal $Y_{\mathrm{d},l}(k,n)$ at the output of the spatial filter for the proposed approach. In Fig. 3(a), we can see the mean LSD as a function of the SDR. As before, the results were computed over 10000 realizations with random direct sound DOAs φ_1 and φ_2. The power of both direct sounds was equal. As can be seen in Fig. 3(a), the proposed approach outperforms the SOA approach especially for medium and high SDRs. Figure 3(b) shows the same results but for an SDR of 20 dB and for varying power ratios between the two direct sounds [represented by the signal-to-interference ratio (SIR)]. Both the SOA approach and proposed approach lead to low distortions when one of the two direct sounds is dominant. When both direct sounds have similar power, the proposed approach clearly outperforms the SOA approach.

6. CONCLUSIONS

We have proposed an optimal post filter, which, in combination with a spatial filter, can reduce the distortion in parametric spatial sound reproduction. The proposed post filter only becomes effective if the response of the spatial filter is different from the desired target response. In this case, the proposed post filter assures a correct spatial sound reproduction based on a simple sound field model. If the spatial filter accurately approximates the desired target responses, the proposed post filter becomes ineffective such that violations of the sound field model have no effect on the spatial rendering. Simulation results show that the proposed system outperforms the SOA system for a challenging multi-source scenario.

7. REFERENCES

[1] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc*, vol. 33, no. 11, pp. 859–871, 1985.

[2] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, June 2007.

[3] O. Thiergart and E. A. P. Habets, "Sound field model violations in parametric spatial sound processing," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, September 2012.

[4] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *2nd International Symposium on Ambisonics and Spherical Acoustics*, May 2010.

[5] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 2182–2196, December 2014.

[6] O. Thiergart, G. Milano, T. Ascherl, and E. A. P. Habets, "Robust 3D sound capturing with planar microphone arrays using directional audio coding," in *Audio Engineering Society Convention 143*, Oct 2017.

[7] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 204–210, July 2000.

[8] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[9] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2141–2151, March 2012.

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

[11] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr., "Measurement of correlation coefficients in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 27, no. 6, pp. 1072–1077, November 1955.

[12] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Part IV: Optimum Array Processing*. John Wiley & Sons, April 2002, vol. 1.

[13] E. Rasumow, M. Hansen, S. van de Par, D. Püschel, V. Mellert, S. Doclo, and M. Blau, "Regularization approaches for synthesizing hrtf directivity patterns," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 215–225, February 2016.

[14] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, 1st ed. Springer Publishing Company, 2010.