SPHERICAL HARMONIC DOMAIN NOISE REDUCTION USING AN MVDR BEAMFORMER AND DOA-BASED SECOND-ORDER STATISTICS ESTIMATION

Daniel P. Jarrett[†]*, Emanuël A. P. Habets*[⋆]*, Patrick A. Naylor* [†]

[†] Dept. of Electrical & Electronic Engineering, Imperial College London, UK
email: {daniel.jarrett05,p.naylor}@imperial.ac.uk
[⋆] International Audio Laboratories Erlangen, Germany
email: emanuel.habets@audiolabs-erlangen.de

ABSTRACT

Most beamformers used for noise reduction rely on the accurate estimation of the second-order statistics of the noise, and in some cases, of the desired signal. Speech presence probability (SPP) based statistics estimators seek to update the estimates only when speech is absent/present, however, when used with a fixed *a priori* SPP, they cannot distinguish between a coherent desired source and coherent noise sources. We propose to distinguish between desired and noise sources by estimating the second-order statistics with a direction of arrival dependent *a priori* SPP, which we then use to compute the weights of a spherical harmonic domain minimum variance distortionless response filter.

Index Terms— Speech enhancement, noise reduction, speech presence probability, spherical harmonic domain

1. INTRODUCTION

In many distant speech acquisition scenarios, such as hands-free telephony, hearing aids, or teleconferencing, the acquired speech signal is corrupted by noise, such as sensor noise, diffuse noise or interfering speech. This noise degrades both the speech quality and intelligibility, making communication difficult or even impossible. Noise reduction algorithms seek to mitigate these effects and extract the desired speech signal.

This objective is commonly achieved through the use of microphone arrays [1], which allow us to take advantage of the spatial properties of the sound field in order to improve the noise reduction performance. These microphone arrays are mostly two dimensional (planar). Spherical microphone arrays, where the microphones are arranged in a spherical configuration, either suspended in free space (an *open* array) or mounted on a rigid spherical baffle (a *rigid* array), are advantageous due to their ability to analyze the sound field in three dimensions [2–4]; the captured sound field can then be efficiently described in the spherical harmonic domain (SHD) [5].

Over the past few decades, many spatio-temporal filters or *beamformers* have been proposed to process the signals received by microphone arrays in the spatial domain [1, 6]. SHD beamformers, where instead of filtering and combining the individual microphone signals, we filter and sum the SHD signals (the *eigenbeams*), have more recently been proposed [7, 8].

The weights of these filters are most often a function of the noise power spectral density (PSD) matrix. Unfortunately, in practice the noise signals are not observable and the noise PSD must be estimated from the noisy signals. Previously proposed noise estimators based on the speech presence probability (SPP) [9, 10] seek to update the noise PSD estimate only in time-frequency bins where speech is absent. A recent contribution by Souden et al. [11] proposes a Gaussian model based multichannel SPP estimator with a fixed *a priori* SPP. However, this estimator does not work when coherent noise sources, such as interfering talkers, are present, as it cannot distinguish between desired and undesired coherent sources.

In this work, we seek to differentiate between these two types of coherent sources. We propose to estimate the noise and desired PSD matrices using a desired speech presence probability (DSPP) estimator based on a direction of arrival (DOA) dependent *a priori* DSPP. The *a priori* DSPP is estimated by comparing the instantaneous DOA estimate for each time-frequency bin to a given steering direction: the closer the DOA to the steering direction, the more likely it is that desired speech is present in this bin. We then use the signal statistics to compute the weights of a SHD minimum variance distortionless response (MVDR) filter.

The estimation of the noise PSD is performed in a similar way to [12], with two key differences: we work in the SHD instead of the spatial domain, and we use a DOA dependent *a priori* DSPP instead of a direct-to-diffuse ratio dependent *a priori* SPP, thus allowing us to suppress coherent sources that do not originate from the desired look direction. Instantaneous DOA estimates are obtained using a pseudointensity vector based method [13]. The MVDR beamformer used is a special case of the tradeoff beamformer in [8, 14].

2. PROBLEM FORMULATION

2.1. Signal Model

In this paper, we consider a scenario in which we receive a mixture of desired speech X originating from a source S, coherent noise V_{c} (e.g., interfering speech), and incoherent noise (e.g., sensor noise) V_{i}. The signal model can be expressed in the short-time Fourier transform (STFT) and spherical harmonic domains as[1]:

$$\begin{aligned} P_{lm}(k) &= G_{lm}(k)S(k) + V_{lm,\mathrm{c}}(k) + V_{lm,\mathrm{i}}(k) \\ &= X_{lm}(k) + V_{lm,\mathrm{c}}(k) + V_{lm,\mathrm{i}}(k), \end{aligned} \tag{1}$$

where P_{lm} denotes the received SHD signal (or *eigenbeam*) of order l and degree m, k denotes the discrete frequency index, and X_{lm}, $V_{lm,\mathrm{c}}$ and $V_{lm,\mathrm{i}}$ respectively denote the desired speech, coherent noise and incoherent noise components of the eigenbeam P_{lm}. The acoustic transfer function is represented by G_{lm} in the SHD.

The eigenbeams P_{lm}, G_{lm}, X_{lm}, $V_{lm,\mathrm{c}}$ and $V_{lm,\mathrm{i}}$ are dependent on the mode strength $B_l(k)$, which is a function of the array prop-

[1] For brevity the time index t is omitted in this section.

erties (radius, configuration, microphone type). For example, for a rigid array of radius r, $B_l(k)$ is given by [15, p. 228]:

$$B_l(k) = i^l \left(j_l(kr) - \frac{j_l'(kr)}{h_l^{(2)'}(kr)} h_l^{(2)}(kr) \right), \tag{2}$$

where $h^{(2)}$ is the spherical Hankel function of the second kind, and $h^{(2)'}$ is its first derivative with respect to kr. To cancel this dependence, we divide our eigenbeams by the mode strength to yield mode strength compensated eigenbeams:

$$\begin{aligned} \widetilde{P}_{lm}(k) &= \left[\sqrt{4\pi} B_l^{-1}(k)\right] P_{lm}(k) \\ &= \widetilde{G}_{lm}(k) S(k) + \widetilde{V}_{lm,\mathrm{c}}(k) + \widetilde{V}_{lm,\mathrm{i}}(k) \\ &= \widetilde{X}_{lm}(k) + \widetilde{V}_{lm,\mathrm{c}}(k) + \widetilde{V}_{lm,\mathrm{i}}(k), \end{aligned} \tag{3}$$

where $\widetilde{P}_{lm}$, $\widetilde{G}_{lm}$, $\widetilde{X}_{lm}$, $\widetilde{V}_{lm,\mathrm{c}}$ and $\widetilde{V}_{lm,\mathrm{i}}$ respectively denote the eigenbeams P_{lm}, G_{lm}, X_{lm}, $V_{lm,\mathrm{c}}$ and $V_{lm,\mathrm{i}}$ after mode strength compensation. With the addition of the $\sqrt{4\pi}$ scaling factor, $\widetilde{P}_{00}(k)$ is equal to the signal which would be received at an omnidirectional microphone $\mathcal{M}_{\mathrm{ref}}$ placed at the center of the sphere (in the absence of the sphere) [8, 14]. Our aim is to estimate the desired speech component $\widetilde{X}_{00}$ of this signal.

2.2. MVDR Beamformer

For convenience, we choose to rewrite (3) in vector notation, where each of the vectors is of length $N = (L+1)^2$, the total number of eigenbeams up to order L:

$$\begin{aligned} \widetilde{\mathbf{p}}(k) &= \widetilde{\mathbf{g}}(k) S(k) + \widetilde{\mathbf{v}}_{\mathrm{c}}(k) + \widetilde{\mathbf{v}}_{\mathrm{i}}(k) \\ &= \widetilde{\mathbf{x}}(k) + \widetilde{\mathbf{v}}_{\mathrm{c}}(k) + \widetilde{\mathbf{v}}_{\mathrm{i}}(k) \\ &= \mathbf{d}(k) \widetilde{X}_{00}(k) + \widetilde{\mathbf{v}}_{\mathrm{c}}(k) + \widetilde{\mathbf{v}}_{\mathrm{i}}(k), \end{aligned} \tag{4}$$

where

$$\widetilde{\mathbf{p}}(k) = \left[\widetilde{P}_{00}(k)\, \widetilde{P}_{1(-1)}(k)\, \widetilde{P}_{10}(k)\, \widetilde{P}_{11}(k) \cdots \widetilde{P}_{LL}(k)\right]^T,$$

$$\mathbf{d}(k) = \left[1\, \frac{\widetilde{G}_{1(-1)}(k)}{\widetilde{G}_{00}(k)}\, \frac{\widetilde{G}_{10}(k)}{\widetilde{G}_{00}(k)}\, \frac{\widetilde{G}_{11}(k)}{\widetilde{G}_{00}(k)} \cdots \frac{\widetilde{G}_{LL}(k)}{\widetilde{G}_{00}(k)}\right]^T,$$

and $\widetilde{\mathbf{x}}(k)$, $\widetilde{\mathbf{g}}(k)$, $\widetilde{\mathbf{v}}_{\mathrm{c}}(k)$ and $\widetilde{\mathbf{v}}_{\mathrm{i}}(k)$ are defined similarly to $\widetilde{\mathbf{p}}(k)$.

The desired speech eigenbeams $\widetilde{X}_{lm}$ are coherent across l and m [8, 14], therefore the desired signal vector $\widetilde{\mathbf{x}}(k)$ can be expressed as $\widetilde{\mathbf{x}}(k) = \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \widetilde{X}_{00}(k)$, where

$$\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) = \frac{E\left[\widetilde{\mathbf{x}}(k) \widetilde{X}_{00}^*(k)\right]}{E\left[|\widetilde{X}_{00}(k)|^2\right]} \tag{5}$$

is the partially normalized [with respect to $\widetilde{X}_{00}(k)$] coherence vector between $\widetilde{\mathbf{x}}(k)$ and $\widetilde{X}_{00}(k)$, and $E[\cdot]$ denotes mathematical expectation. Using (5), (4) can be expressed as

$$\begin{aligned} \widetilde{\mathbf{p}}(k) &= \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \widetilde{X}_{00}(k) + \widetilde{\mathbf{v}}_{\mathrm{c}}(k) + \widetilde{\mathbf{v}}_{\mathrm{i}}(k) \\ &= \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \widetilde{X}_{00}(k) + \widetilde{\mathbf{v}}(k), \end{aligned} \tag{6}$$

where we have defined the noise signal vector $\widetilde{\mathbf{v}}(k) = \widetilde{\mathbf{v}}_{\mathrm{c}}(k) + \widetilde{\mathbf{v}}_{\mathrm{i}}(k)$.

We assume that $\widetilde{\mathbf{x}}(k)$, $\widetilde{\mathbf{v}}_{\mathrm{c}}(k)$ and $\widetilde{\mathbf{v}}_{\mathrm{i}}(k)$ are mutually uncorrelated, therefore the PSD matrix $\boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}$ of $\widetilde{\mathbf{p}}$ can be expressed as

$$\begin{aligned} \boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}(k) &= E\left[\widetilde{\mathbf{p}}(k) \widetilde{\mathbf{p}}^H(k)\right] = \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(k) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(k) \\ &= \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(k) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{c}}}(k) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{i}}}(k), \end{aligned} \tag{7}$$

where $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(k) = \phi_{\widetilde{X}_{00}}(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}^H(k)$, $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(k) = E\left[\widetilde{\mathbf{v}}(k) \widetilde{\mathbf{v}}^H(k)\right]$, $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{c}}}(k) = E\left[\widetilde{\mathbf{v}}_{\mathrm{c}}(k) \widetilde{\mathbf{v}}_{\mathrm{c}}^H(k)\right]$ and $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{i}}}(k) = E\left[\widetilde{\mathbf{v}}_{\mathrm{i}}(k) \widetilde{\mathbf{v}}_{\mathrm{i}}^H(k)\right]$ are respectively the PSD matrices of $\widetilde{\mathbf{x}}(k)$, $\widetilde{\mathbf{v}}(k)$, $\widetilde{\mathbf{v}}_{\mathrm{c}}(k)$ and $\widetilde{\mathbf{v}}_{\mathrm{i}}(k)$, and $\phi_{\widetilde{X}_{00}}(k) = E\left[|\widetilde{X}_{00}(k)|^2\right]$ is the variance of $\widetilde{X}_{00}(k)$.

The output of our beamformer is obtained by applying a complex weight to each eigenbeam, and summing over all eigenbeams:

$$\begin{aligned} Z(k) &= \mathbf{h}^H(k) \widetilde{\mathbf{x}}(k) + \mathbf{h}^H(k) \widetilde{\mathbf{v}}_{\mathrm{c}}(k) + \mathbf{h}^H(k) \widetilde{\mathbf{v}}_{\mathrm{i}}(k) \\ &= \widetilde{X}_{\mathrm{fd}}(k) + \widetilde{V}_{\mathrm{rcn}}(k) + \widetilde{V}_{\mathrm{rin}}(k), \end{aligned} \tag{8}$$

where $\widetilde{X}_{\mathrm{fd}}(k) = \mathbf{h}^H(k) \widetilde{\mathbf{x}}(k) = \mathbf{h}^H(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \widetilde{X}_{00}(k)$ is the filtered desired signal, $\widetilde{V}_{\mathrm{rcn}}(k) = \mathbf{h}^H(k) \widetilde{\mathbf{v}}_{\mathrm{c}}(k)$ is the residual coherent noise and $\widetilde{V}_{\mathrm{rin}}(k) = \mathbf{h}^H(k) \widetilde{\mathbf{v}}_{\mathrm{i}}(k)$ is the residual incoherent noise.

We design a minimum variance distortionless response (MVDR) beamformer which seeks to minimize the residual (coherent and incoherent) noise with the constraint that the desired signal is not distorted, i.e.,

$$\min_{\mathbf{h}(k)} \mathbf{h}^H(k) \Phi_{\widetilde{\mathbf{v}}}(k) \mathbf{h}(k) \text{ s.t. } \mathbf{h}^H(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) = 1,$$

for which the filter weights are given by [8, 16]

$$\mathbf{h}_{\mathrm{MVDR}}(k) = \frac{\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)}{\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}^H(k) \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)}. \tag{9}$$

3. SIGNAL STATISTICS ESTIMATION

In order to compute the MVDR filter in (9), we must estimate the noise PSD matrix $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}$, as well as the propagation vector $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$. In this section, we propose a method for this based on SPPs.

We introduce the following two hypotheses regarding the presence of desired speech in each time-frequency bin:

$$\begin{aligned} &\mathcal{H}_0(k,t) : \widetilde{\mathbf{p}}(k,t) = \widetilde{\mathbf{v}}(k,t) \\ &\qquad\qquad \text{indicating desired speech absence} \\ &\mathcal{H}_1(k,t) : \widetilde{\mathbf{p}}(k,t) = \widetilde{\mathbf{x}}(k,t) + \widetilde{\mathbf{v}}(k,t) \\ &\qquad\qquad \text{indicating desired speech presence} \end{aligned}$$

A minimum mean square error estimate of the noise PSD matrix taking into account the probability of these two hypotheses is given by[2]

$$\begin{aligned} E\left[\widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^H | \widetilde{\mathbf{p}}\right] =& f(\mathcal{H}_0|\widetilde{\mathbf{p}}) E\left[\widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^H | \widetilde{\mathbf{p}}, \mathcal{H}_0\right] \\ &+ f(\mathcal{H}_1|\widetilde{\mathbf{p}}) E\left[\widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^H | \widetilde{\mathbf{p}}, \mathcal{H}_1\right], \end{aligned} \tag{10}$$

where $f(\mathcal{H}_1|\widetilde{\mathbf{p}})$ is the multichannel *a posteriori* desired speech presence probability (DSPP) and $f(\mathcal{H}_0|\widetilde{\mathbf{p}}) = 1 - f(\mathcal{H}_1|\widetilde{\mathbf{p}})$. A common

[2] For brevity, the dependencies on the discrete frequency and time indices k and t are omitted where possible in this section.

way of approximating (10) is to recursively estimate the PSD matrix with a smoothing factor that depends on the SPP, as in [11, 12], such that the estimate is updated most rapidly when desired speech is absent, i.e.,

$$\begin{aligned}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t) &= f(\mathcal{H}_0|\widetilde{\mathbf{p}})\left(\alpha_{\mathrm{v}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t-1) + (1-\alpha_{\mathrm{v}})\widetilde{\mathbf{p}}\widetilde{\mathbf{p}}^H\right) \\ &\quad + f(\mathcal{H}_1|\widetilde{\mathbf{p}})\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t-1) \\ &= \alpha'\widetilde{\mathbf{p}}\widetilde{\mathbf{p}}^H + [1-\alpha']\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t-1) \end{aligned} \tag{11}$$

where $\alpha' = f(\mathcal{H}_0|\widetilde{\mathbf{p}})(1-\alpha_{\mathrm{v}})$ and $0 < \alpha_{\mathrm{v}} \leq 1$ is a smoothing factor.

The propagation vector $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ is given by the first column of $\mathbf{\Phi}_{\widetilde{\mathbf{x}}}$ divided by the first element $\phi_{\widetilde{X}_{00}}$, and is estimated by

$$\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}} = \hat{\phi}^{-1}_{\widetilde{X}_{00}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}\,\mathbf{i}_N, \tag{12}$$

where $\mathbf{i}_N = [1\ 0\ \cdots\ 0]^T$ is a vector of length N. Since the noise is always present, the desired signal is not directly observable. Therefore, we first compute an estimate of the desired speech plus incoherent noise PSD $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{i}}}$ in a similar way to $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}$, i.e.,

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{i}}}(t) = \alpha''\widetilde{\mathbf{p}}\widetilde{\mathbf{p}}^H + [1-\alpha'']\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{i}}}(t-1), \tag{13}$$

where $\alpha'' = f(\mathcal{H}_1|\widetilde{\mathbf{p}})(1-\alpha_{\mathrm{xv_i}})$ and $0 < \alpha_{\mathrm{xv_i}} \leq 1$ is a smoothing factor. We can now obtain an estimate $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}$ of the desired speech PSD matrix using

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}} = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{i}}} - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{i}}}. \tag{14}$$

The propagation vector estimate $\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ is therefore updated most rapidly when desired speech is present. We assume that the incoherent noise $\widetilde{\mathbf{v}}_{\mathrm{i}}$ is stationary; its PSD matrix can therefore be estimated during initial noise only frames.

3.1. Multichannel Desired Speech Presence Probability

Assuming the desired speech, coherent noise, and incoherent noise can be modeled as complex multivariate Gaussian random variables, a multichannel DSPP estimate is given by [11]:

$$\hat{f}(\mathcal{H}_1|\widetilde{\mathbf{p}}) = \left\{1 + \frac{1-q}{q}(1+\xi)e^{-\frac{\beta}{1+\xi}}\right\}^{-1}, \tag{15}$$

where $q = f(\mathcal{H}_1)$ denotes the *a priori* DSPP, β is defined as

$$\beta = \widetilde{\mathbf{p}}^H\hat{\mathbf{\Phi}}^{-1}_{\widetilde{\mathbf{v}}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}\hat{\mathbf{\Phi}}^{-1}_{\widetilde{\mathbf{v}}}\widetilde{\mathbf{p}}, \tag{16}$$

and $\xi = \mathrm{tr}\left(\hat{\mathbf{\Phi}}^{-1}_{\widetilde{\mathbf{v}}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}\right)$. The PSD matrix $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}$ is given by

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}} = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}} - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}, \tag{17}$$

and represents the desired signal plus residual noise (i.e., the noise that has not yet been captured by the noise PSD matrix). The PSD matrix $\mathbf{\Phi}_{\widetilde{\mathbf{p}}}$ is recursively estimated as

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(t) = \alpha_{\mathrm{p}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(t-1) + (1-\alpha_{\mathrm{p}})\widetilde{\mathbf{p}}\widetilde{\mathbf{p}}^H, \tag{18}$$

where $0 < \alpha_{\mathrm{p}} \leq 1$ is a smoothing factor.

The *a priori* DSPP represents our prior knowledge of the probability of desired speech presence. In previous approaches, the SPP has been fixed [11, 17], or signal dependent [9, 10, 12]. In order for our DSPP estimator to be able to distinguish between desired and interfering speech, we make the *a priori* DSPP signal dependent.

3.2. DOA-based *A Priori* Desired Speech Presence Probability

In this work, we estimate the *a priori* DSPP using instantaneous time and frequency dependent DOA estimates $\hat{\Omega}_{\mathrm{DOA}}$. The closer the instantaneous DOA is to the steering direction Ω_{steer}, the more likely it is that the desired source is present in that time-frequency bin. We define the *opening angle* $\Theta_{\Omega_{\mathrm{steer}},\hat{\Omega}_{\mathrm{DOA}}}$ as the angle between Ω_{steer} and $\hat{\Omega}_{\mathrm{DOA}}$. We can then express the *a priori* DSPP as

$$q(k,t) = w\left(\Theta_{\Omega_{\mathrm{steer}},\hat{\Omega}_{\mathrm{DOA}}}(k,t)\right), \tag{19}$$

where $w(\Theta)$ is a windowing function (e.g., Hamming, Gaussian) centered around $\Theta = 0$. The width of the main lobe determines the region of interest around Ω_{steer}; e.g., with a Gaussian window this region is determined by the standard deviation of the Gaussian.

As previously shown in [13], DOA estimates can be obtained for each time-frequency bin using a combination of zero- and first-order eigenbeams obtained with a spherical microphone array. The reader is referred to [13] for details of the computation of the DOA estimates from X_{00}, $X_{1(-1)}$, X_{10} and X_{11}.

3.3. Algorithm Summary

The noise PSD matrix $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}$ and propagation vector $\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ are recursively estimated according to the following steps:

1. Estimate the *a priori* DSPP $q(t)$ for the current frame using the instantaneous DOA estimate $\hat{\Omega}_{\mathrm{DOA}}(t)$ in (19).
2. Update $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(t)$ using (18).
3. Estimate $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}(t)$ as $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}(t) = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(t) - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t-1)$.
4. Estimate the (*a posteriori*) multichannel DSPP according to (15), using $q(t)$, $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}(t)$ and $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t-1)$.
5. Compute a recursively smoothed DSPP:

$$\bar{f}(t) = \alpha_{\mathrm{f}}\bar{f}(t-1) + (1-\alpha_{\mathrm{f}})f(\mathcal{H}_1(t)|\widetilde{\mathbf{p}}(t)), \tag{20}$$

   where $0 < \alpha_{\mathrm{f}} \leq 1$ denotes a smoothing parameter.
6. Avoid stagnation of the noise PSD matrix by setting the multichannel DSPP to $\min\left(f_{\max}, f(\mathcal{H}_1(t)|\widetilde{\mathbf{p}}(t))\right)$ whenever $\bar{f}(t) > f_{\max}$.
7. Update $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t)$ according to (11) by using $f(\mathcal{H}_0(t)|\widetilde{\mathbf{p}}(t))$.
8. Update $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{i}}}(t)$ by using $f(\mathcal{H}_1(t)|\widetilde{\mathbf{p}}(t))$ in (13), and estimate $\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(t)$ according to (12).

4. PERFORMANCE EVALUATION

4.1. Experimental Setup

We evaluated the performance of the proposed DOA-based second-order statistics estimation algorithm by using the estimated statistics to compute the weights of the MVDR filter defined in (9).

We simulated a rigid $Q = 32$ microphone array with radius 4.2 cm placed approximately in the center of a room with dimensions $5 \times 7 \times 4$ m and a reverberation time of 300 ms, using SMIRgen, a room impulse response generator for spherical microphone arrays [18, 19]. We applied the MVDR beamformer to eigenbeams up to order $L = 3$, of which there are $N = (L+1)^2 = 16$ in total.

A desired talker was placed at an azimuth of 0°, and two interfering talkers at azimuths of 130° and 210°; all three talkers were placed at an elevation of 0° and a distance of 1 m from the center of the array. The desired and interfering speech signals consisted of male and female speech from the EBU SQAM dataset [20].

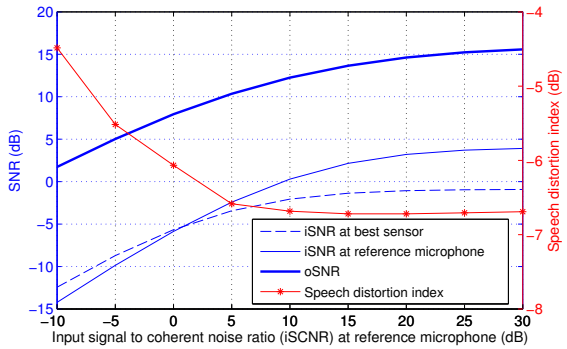

Fig. 1. Speech distortion index, input SNRs and output SNR as a function of the input signal to coherent noise ratio

The incoherent noise consisted of spatio-temporal white Gaussian noise with a constant input signal to incoherent noise ratio (iSINR) of 25 dB at the reference microphone $\mathcal{M}_{\text{ref}}$. It should be noted that the incoherent noise power at $\mathcal{M}_{\text{ref}}$ is reduced by a factor of $Q|B_0(k)|^2$ with respect to the sensors [14, 21]; with $Q = 32$ microphones, at low frequencies an iSINR of 25 dB at $\mathcal{M}_{\text{ref}}$ corresponds to an iSINR of around 10 dB based on the noise power at the sensors. The coherent noise had an input signal to coherent noise ratio (iSCNR) between -10 and 30 dB at $\mathcal{M}_{\text{ref}}$. The coherent and incoherent noise levels were set based on active speech levels, computed according to ITU-T Rec. P.56 [22].

Processing was performed in the STFT domain at a sampling frequency of 8 kHz, with a frame length of 64 ms and an overlap of 50% between successive frames. In order to obtain the *a priori* DSPP, we applied a Gaussian window with a standard deviation of 6° to the opening angles Θ. The smoothing factors were empirically chosen as $\alpha_{\text{v}} = 0.93$, $\alpha_{\text{xv}_\text{i}} = 0.7$, $\alpha_{\text{p}} = 0.7$, and $\alpha_{\text{f}} = 0.8$; the maximum long-term DSPP was chosen as $f_{\max} = 0.99$.

4.2. Results

In Fig. 1 we plot (as a function of the iSCNR) the speech distortion index at $\mathcal{M}_{\text{ref}}$, as well as the signal to (coherent plus incoherent) noise ratio (SNR) at the sensor with the highest SNR (the '*best sensor*'), at the reference microphone $\mathcal{M}_{\text{ref}}$, and at the output of the beamformer. The speech distortion index was computed by averaging the fullband speech distortion index defined in [16, eqn. 4.44] over 16 ms frames. The input and output SNRs were computed by taking the segmental SNR over 16 ms frames, after applying the frequency weighting defined in ITU-R 468, and discarding silent frames determined according to ITU-T Rec. P.56. Both the speech distortion index and the SNRs were computed over a 40 s multi-talk segment, with one or two interfering talkers active at all times.

As expected, we find that as the iSCNR increases, our *a priori* DSPP and *a posteriori* DSPP estimates become more accurate, and the speech distortion index decreases. We also find that the array gain with respect to $\mathcal{M}_{\text{ref}}$ (i.e., oSNR - iSNR at $\mathcal{M}_{\text{ref}}$ in dB) decreases as the iSCNR increases; this is due to the fact that at high iSCNRs, we are limited by the MVDR filter's maximum incoherent noise reduction factor of $10\log_{10}(Q) = 15$ dB [14, 23] with respect to the sensors.

Finally we note that at low iSCNRs where the coherent noise has high power, the input SNR at $\mathcal{M}_{\text{ref}}$ is *lower* than at the best sensor, since the reference microphone $\mathcal{M}_{\text{ref}}$ is omnidirectional, while the best sensor has some directionality due to the scattering introduced by the rigid sphere. At high iSCNRs where the incoherent noise is more significant, the input SNR at $\mathcal{M}_{\text{ref}}$ is *higher* than at the best sensor, due to the fact that the incoherent noise power is reduced by a factor of $Q|B_0(k)|^2$ (see Sec. 4.1).

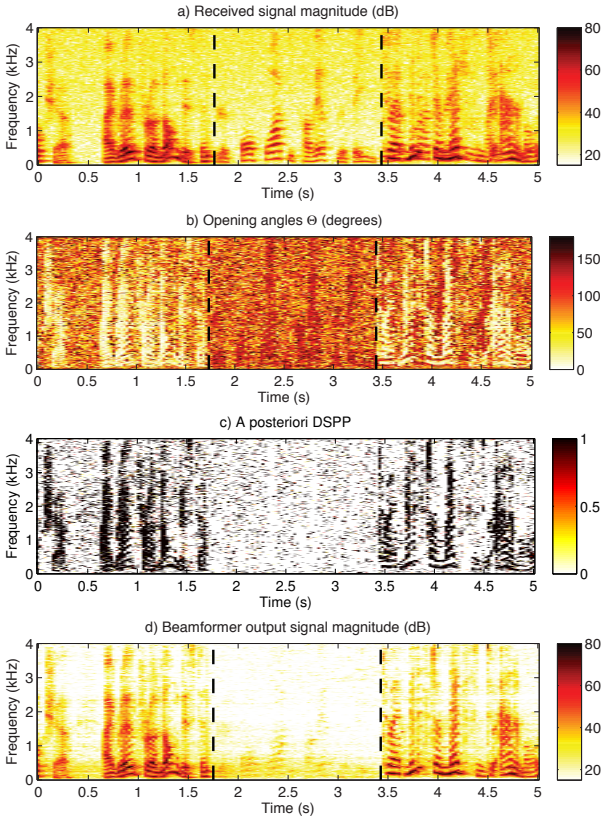

Fig. 2. Time-frequency plots of the received signal $\tilde{P}_{00}$ (a), opening angles Θ (b), *a posteriori* DSPP (c), and output signal Z (d).

In Fig. 2 we show some example time-frequency plots of the received signal $\tilde{P}_{00}$, the opening angles Θ, the *a posteriori* DSPP, and the beamformer output signal Z, for an iSCNR of 10 dB and three time segments: single-talk (desired speech only), single-talk (interfering speech only), and double-talk. It can be seen that during the interfering speech only segment, the *a posteriori* DSPP remains low for most time-frequency bins, as desired, thanks to the DOA-based *a priori* DSPP. In Fig. 2 (d) we see that, as a result, the interfering speech is suppressed in addition to the incoherent noise.

5. CONCLUSIONS

In this paper, we proposed an algorithm to suppress both coherent and incoherent noise sources. We first proposed a desired and noise PSD matrix estimator based on the DSPP, with a DOA-dependent *a priori* DSPP. We then used the signal statistics to compute the weights of an MVDR filter, which we applied to the received eigenbeams. Finally, we showed that the proposed algorithm achieved good noise suppression (with an array gain of around 15 dB) with only moderate speech distortion (with a speech distortion index of approximately 6.5 dB). These results were confirmed by informal listening tests.

6. REFERENCES

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

[2] G. W. Elko and J. Meyer, "Spherical microphone arrays for 3D sound recordings," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., chapter 3, pp. 67–89.

[3] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[4] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1949–1952.

[5] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2002, vol. 2, pp. 1781–1784.

[6] E. A. P Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 854–867, Mar. 2012.

[7] H. Sun, S. Yan, and U. P. Svensson, "Robust minimum sidelobe beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 1045–1051, May 2011.

[8] D. P. Jarrett, E. A. P. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[10] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[11] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, July 2010.

[12] 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.

[13] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[14] D. P. Jarrett and E. A. P. Habets, "On the noise reduction performance of a spherical harmonic domain tradeoff beamformer," *IEEE Signal Process. Lett.*, vol. 19, no. 11, pp. 773–776, Nov. 2012.

[15] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, first edition, 1999.

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

[17] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2011, pp. 145–148.

[18] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sept. 2012.

[19] D. P. Jarrett, "Spherical microphone array impulse response (SMIR) generator," http://www.ee.ic.ac.uk/sap/smirgen/.

[20] European Broadcasting Union, "Sound quality assessment material recordings for subjective tests," 1988, http://tech.ebu.ch/publications/sqamcd.

[21] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[22] ITU-T, "Objective measurement of active speech level," Mar. 1993.

[23] H. L. van Trees, *Detection, Estimation, and Modulation Theory*, vol. IV, Optimum Array Processing, Wiley, New York, USA, Apr. 2002.