A TRADEOFF BEAMFORMER FOR NOISE REDUCTION IN THE SPHERICAL HARMONIC DOMAIN

Daniel P. Jarrett[1,2], *Emanuël A. P. Habets*[2], *Jacob Benesty*[3], *Patrick A. Naylor*[1]

[1]: Dept. of Electrical & Electronic Engineering, Imperial College London, London, SW7 2AZ, United Kingdom
[2]: International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany
[3]: INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, Quebec, Canada

ABSTRACT

Spherical microphone arrays present the advantage of enabling a three dimensional analysis of the sound field that can be described efficiently in the spherical harmonic domain. In this paper, we present a tradeoff beamformer which operates in this domain and enables a compromise between noise reduction and speech distortion. The experimental results obtained using simulated data demonstrate the beamformer's ability to reduce high levels of coherent noise with low speech distortion. With a 32 microphone array in the presence of one interfering talker and signal to coherent noise ratios as low as -40 dB, the beamformer is able to achieve an array gain of up to 72 dB while retaining good speech quality.

Index Terms— Speech enhancement, noise reduction, spherical harmonic domain, spherical microphone arrays.

1. INTRODUCTION

Distant speech acquisition is required in many applications such as hands-free telephony and hearing aids, where in addition to the desired speech signal, the received signal is corrupted by sensor noise and other interfering sound sources. This unfortunately degrades the quality and intelligibility of the desired speech. Microphone arrays, where a set of microphones is arranged in a specific configuration, have frequently been used to mitigate these effects [1], exploiting the spatial properties of the sound field.

Spherical microphone arrays present the advantage of enabling a three dimensional analysis of the sound field that can be described efficiently in the spherical harmonic domain. Numerous spatio-temporal filters, called beamformers, have been proposed to process the received microphone signals in the spatial domain (see [1] and the references therein). More recently, spherical harmonic domain (SHD) beamformers have become a topic of interest [2], where we filter the SHD signals (eigenbeams) instead of the individual microphone signals. In optimal beamforming the filter weights are chosen in a statistically optimal way: a cost function (typically the mean square error between the filtered and desired signals) is minimized, often subject to some additional constraints.

Fixed beamformers apply a constraint to a specific look direction [3, 4] and optimize the beamformer weights with respect to array performance measures such as the directivity index, white noise gain or sidelobe levels. Signal-dependent beamformers (e.g. [5, 6]) optimize the filter weights taking into account characteristics of the desired signal and noise, although they usually still assume knowledge of the source direction of arrival (DOA). In order to perform dereverberation in addition to noise reduction, in [6] the desired signal consists only of the direct path and some early reflections, requiring an estimate of the DOAs of all plane waves of interest. In this contribution our aim is to concentrate only on noise reduction and we therefore do not perform any dereverberation, keeping in mind the fact that there is a tradeoff between noise reduction and speech dereverberation [7].

In this work, we propose a tradeoff beamformer in the spherical harmonic domain, where the tradeoff is between noise reduction on the one hand, and speech distortion on the other. This tradeoff beamformer includes the minimum variance distortionless response (MVDR) and Wiener filters as special cases. The proposed beamformer depends on the covariance matrices of the desired plus noise signals and of the noise only signals, but does not explicitly require knowledge of the desired source DOA. We evaluate the performance of our beamformer using signal-based measures.

This paper is organized as follows. In Section 2, we describe the signal model and formulate the problem. In Section 3, we define a number of performance measures which we will use to evaluate the performance of our beamformer, and which we use to derive the proposed tradeoff beamformer in Section 4. In Section 5, we evaluate its performance as a function of the tradeoff parameter. Finally we provide conclusions in Section 6.

2. PROBLEM FORMULATION

We consider a conventional time-domain signal model in which a spherical microphone array captures Q noisy signals $p(n, \mathbf{r}_q)$ at a discrete time index n and microphone positions $\mathbf{r}_q = (r, \Omega_q)$ (in spherical coordinates), where r is the radius of the sphere and $q \in \{1, \ldots, Q\}$. The q^{th} microphone signal, $p(n, \mathbf{r}_q)$, consists of a convolved speech signal $x(n, \mathbf{r}_q)$ and a noise signal $v(n, \mathbf{r}_q)$:

$$p(n, \mathbf{r}_q) = g(n, \mathbf{r}_q) * s(n) + v(n, \mathbf{r}_q) = x(n, \mathbf{r}_q) + v(n, \mathbf{r}_q) \quad (1)$$

where $g(n, \mathbf{r}_q)$ is the acoustic impulse response from the unknown speech source $s(n)$ to the microphone at angle Ω_q and $*$ is the linear convolution operator.

The acoustic impulse responses are assumed to be time-invariant. We also assume that the received speech signals $x(n, \mathbf{r}_q)$ and the received noise signals $v(n, \mathbf{r}_q)$ are uncorrelated. The received speech signals originate from a single source and are therefore, by definition, coherent across the array. The noise signals, on the other hand, are usually only partially coherent across the array.

We can rewrite (1) in the STFT (short-time Fourier transform) domain as[1]

$$\begin{aligned} P(k, \mathbf{r}_q) &= G(k, \mathbf{r}_q) S(k) + V(k, \mathbf{r}_q) \\ &= X(k, \mathbf{r}_q) + V(k, \mathbf{r}_q) \end{aligned} \quad (2)$$

[1] For brevity, the dependency on the time frame index is omitted.

where k is the discrete frequency index and $P(k, \mathbf{r}_q)$, $G(k, \mathbf{r}_q)$, $S(k)$, and $V(k, \mathbf{r}_q)$ are STFT-domain representations of $p(n, \mathbf{r}_q)$, $g(n, \mathbf{r}_q)$, $s(n)$ and $v(n, \mathbf{r}_q)$, respectively.

2.1. Spherical harmonic domain signal model

When dealing with spherical microphone arrays, it is convenient to work in the spherical harmonic domain. The spherical Fourier transform $F_{lm}(k)$ of a spatial domain signal $F(k, \mathbf{r}_q)$ involves an integral over all angles Ω, however it can be approximated for a discretely sampled sound field using the expression [8]

$$F_{lm}(k) \approx \sum_{q=1}^{Q} c_q F(k, \mathbf{r}_q) Y_{lm}^*(\Omega_q), \tag{3}$$

where Y_{lm} is the spherical harmonic of order l and degree m, and $(\cdot)^*$ denotes the complex conjugate. The spherical Fourier transform coefficient $F_{lm}(k)$ for all values of k is often called the *eigenbeam* of order l and degree m, due to the fact that the spherical harmonics are the eigensolutions of the acoustic wave equation in spherical coordinates. The weights c_q are chosen such that the approximation in (3) is as accurate as possible (c.f. [8] for examples); with a sufficient number of microphones and appropriate positioning, the error involved in this approximation can be eliminated entirely. All spatial sampling schemes require at least $Q = (L+1)^2$ microphones to sample a sound field up to order $l = L$.

We can now express our signal model in the SHD as:

$$\begin{aligned} P_{lm}(k) &= G_{lm}(k) S(k) + V_{lm}(k) \\ &= X_{lm}(k) + V_{lm}(k) \end{aligned} \tag{4}$$

where $P_{lm}(k)$, $G_{lm}(k)$, $X_{lm}(k)$ and $V_{lm}(k)$ denote the SHD representations of $P(k, \mathbf{r}_q)$, $G(k, \mathbf{r}_q)$, $X(k, \mathbf{r}_q)$ and $V(k, \mathbf{r}_q)$, respectively.

The SHD signals $P_{lm}(k)$, $G_{lm}(k)$, $X_{lm}(k)$ and $V_{lm}(k)$ are dependent on the mode strength $B_l(k)$, which is a function of the array configuration. Expressions for mode strength in various configurations (open, rigid, dual-sphere, etc.) can be found in [2]. In order to cancel the dependence on the array configuration, we divide our eigenbeams by the mode strength to yield mode strength compensated signals:

$$\begin{aligned} \widetilde{P}_{lm}(k) &= B_l^{-1}(k) P_{lm}(k) \\ &= B_l^{-1}(k) \left[G_{lm}(k) S(k) + V_{lm}(k) \right] \\ &= \widetilde{G}_{lm}(k) S(k) + \widetilde{V}_{lm}(k) \\ &= \widetilde{X}_{lm}(k) + \widetilde{V}_{lm}(k) \end{aligned} \tag{5}$$

where $\widetilde{P}_{lm}(k)$, $\widetilde{G}_{lm}(k)$, $\widetilde{X}_{lm(k)}$ and $\widetilde{V}_{lm}(k)$ respectively denote the signals $P_{lm}(k)$, $G_{lm}(k)$, $X_{lm}(k)$ and $V_{lm}(k)$ after mode strength compensation.

As $X(k, \mathbf{r}_q)$ and $V(k, \mathbf{r}_q)$ are uncorrelated and the STFT, spherical Fourier transform and division by the mode strength are linear operations, $\widetilde{X}_{lm}(k)$ and $\widetilde{V}_{lm}(k)$ are also uncorrelated. It can be shown that as the spatial domain signals $X(k, \mathbf{r}_q)$ are coherent across $\mathbf{r}_q$, the eigenbeams $\widetilde{X}_{lm}(k)$ are also coherent across l and m.

2.2. Beamforming in the spherical harmonic domain

In this work, our desired signal is chosen to be $\widetilde{X}_{00}(k)$. This signal is proportional to the speech signal which would be received by an omnidirectional microphone placed at the center of the sphere, and will therefore be referred to as the *omnidirectional speech signal*. For convenience, we rewrite the SHD signals in a vector notation, where each of the vectors has length $N = (L+1)^2$, the total number of eigenbeams up to order L:

$$\begin{aligned} \widetilde{\mathbf{p}}(k) &= \widetilde{\mathbf{g}}(k) S(k) + \widetilde{\mathbf{v}}(k) \\ &= \widetilde{\mathbf{x}}(k) + \widetilde{\mathbf{v}}(k) \\ &= \mathbf{d}(k) \widetilde{X}_{00}(k) + \widetilde{\mathbf{v}}(k), \end{aligned} \tag{6}$$

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 $\widetilde{\mathbf{g}}(k)$ and $\widetilde{\mathbf{v}}(k)$ are defined similarly to $\widetilde{\mathbf{p}}(k)$. Note that we have implicitly assumed here that $G_{00}(k) \neq 0 \ \forall k$.

The eigenbeams $\widetilde{X}_{lm}(k)$ are coherent, therefore the signal vector $\widetilde{\mathbf{x}}(k)$ can also be written as

$$\widetilde{\mathbf{x}}(k) = \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \widetilde{X}_{00}(k), \tag{7}$$

where

$$\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) = \frac{E\left[\widetilde{\mathbf{x}}(k) \widetilde{X}_{00}^*(k) \right]}{E\left[|\widetilde{X}_{00}(k)|^2 \right]} = \mathbf{d}(k) \tag{8}$$

is the partially normalized [with respect to $\widetilde{X}_{00}(k)$] coherence vector between $\widetilde{\mathbf{x}}(k)$ and $\widetilde{X}_{00}(k)$ and $E\left[\cdot\right]$ denotes mathematical expectation. Using (8), (6) can be expressed as

$$\widetilde{\mathbf{p}}(k) = \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \widetilde{X}_{00}(k) + \widetilde{\mathbf{v}}(k). \tag{9}$$

As $\widetilde{\mathbf{p}}(k)$ is the sum of two uncorrelated components $\widetilde{\mathbf{x}}(k)$ and $\widetilde{\mathbf{v}}(k)$, the correlation matrix of $\widetilde{\mathbf{p}}(k)$ is given by

$$\begin{aligned} \boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}(k) &= E\left[\widetilde{\mathbf{p}}(k) \widetilde{\mathbf{p}}^H(k) \right] \\ &= \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(k) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(k), \end{aligned} \tag{10}$$

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

Equation (9) contains our desired signal $\widetilde{X}_{00}(k)$ and is the basis for the design of our noise reduction beamformer. The output $Z(k)$ of our beamformer is obtained by applying a complex weight $\mathbf{h}^H(k)$ to each eigenbeam, and summing across all eigenbeams:

$$\begin{aligned} Z(k) &= \mathbf{h}^H(k) \widetilde{\mathbf{p}}(k) \\ &= \mathbf{h}^H(k) \widetilde{\mathbf{x}}(k) + \mathbf{h}^H(k) \widetilde{\mathbf{v}}(k) \\ &= \widetilde{X}_{\mathrm{fd}}(k) + \widetilde{V}_{\mathrm{rn}}(k), \end{aligned} \tag{11}$$

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 and $\widetilde{V}_{\mathrm{rn}}(k) = \mathbf{h}^H(k) \widetilde{\mathbf{v}}(k)$ is the residual noise. The variance of $Z(k)$ is then given by

$$\begin{aligned} \phi_Z(k) &= \mathbf{h}^H(k) \Phi_{\widetilde{\mathbf{p}}}(k) \mathbf{h}(k) \\ &= \phi_{\widetilde{X}_{\mathrm{fd}}}(k) + \phi_{\widetilde{V}_{\mathrm{rn}}}(k), \end{aligned} \tag{12}$$

where $\phi_{\widetilde{X}_{\mathrm{fd}}}(k) = \phi_{\widetilde{X}_{00}}(k) |\mathbf{h}^H(k) \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)|^2$ and $\phi_{\widetilde{V}_{\mathrm{rn}}}(k) = \mathbf{h}^H(k) \Phi_{\widetilde{\mathbf{v}}}(k) \mathbf{h}(k)$.

3. PERFORMANCE MEASURES

In the following section, we define the performance measures which we will be using to design and evaluate our beamformer.

Array gain: The array gain is defined as the ratio of the output signal to noise ratio (SNR) to the input SNR [9], where here noise refers to both incoherent (sensor) noise and coherent noise (interference). Our objective is to estimate the desired speech component $\widetilde{X}_{00}(k)$, and the input SNR is therefore defined as the ratio of the power of $\widetilde{X}_{00}(k)$ to the power of the noise $\widetilde{V}_{00}(k)$. The output SNR quantifies the amount of noise remaining at the output of our beamformer, and is defined as the ratio of the power of the filtered desired signal $\widetilde{X}_{\mathrm{fd}}(k)$ over the power of the residual noise $\widetilde{V}_{\mathrm{rn}}(k)$. The narrowband array gain is therefore given by

$$\mathcal{A}\left[\mathbf{h}(k)\right] = \frac{\mathrm{oSNR}\left[\mathbf{h}(k)\right]}{\mathrm{iSNR}(k)} = \frac{\phi_{\widetilde{X}_{\mathrm{fd}}}(k)}{\phi_{\widetilde{V}_{\mathrm{rn}}}(k)}\,\frac{\phi_{\widetilde{V}_{00}}(k)}{\phi_{\widetilde{X}_{00}}(k)} \tag{13a}$$

$$= \frac{\phi_{\widetilde{V}_{00}}(k)|\mathbf{h}^H(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)|^2}{\mathbf{h}^H(k)\Phi_{\widetilde{\mathbf{v}}}(k)\mathbf{h}(k)}. \tag{13b}$$

where $\phi_{\widetilde{V}_{00}}(k) = E\left[|\widetilde{V}_{00}(k)|^2\right]$ is the variance of $\widetilde{V}_{00}(k)$.

Noise reduction factor: The noise reduction factor measures the amount of noise being attenuated by the beamformer [10], and is defined as the ratio of the power of the noise in the omnidirectional signal $\widetilde{V}_{00}(k)$ over the power of the residual noise at the beamformer output $\widetilde{V}_{\mathrm{rn}}(k)$. We define the narrowband noise reduction factor as

$$\xi_{\mathrm{nr}}\left[\mathbf{h}(k)\right] = \frac{\phi_{\widetilde{V}_{00}}(k)}{\phi_{\widetilde{V}_{\mathrm{rn}}}(k)} = \frac{\phi_{\widetilde{V}_{00}}(k)}{\mathbf{h}^H(k)\Phi_{\widetilde{\mathbf{v}}}(k)\mathbf{h}(k)} \tag{14}$$

Speech distortion index: The filtering operation unfortunately introduces distortion to the desired speech signal $\widetilde{X}_{00}(k)$. The narrowband speech distortion index [10] is defined as

$$\upsilon_{\mathrm{sd}}\left[\mathbf{h}(k)\right] = \frac{E\left[|\widetilde{X}_{\mathrm{fd}}(k) - \widetilde{X}_{00}(k)|^2\right]}{\phi_{\widetilde{X}_{00}}(k)} \tag{15a}$$

$$= \left|\mathbf{h}^H(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) - 1\right|^2. \tag{15b}$$

Speech quality: We evaluate the speech quality using the Perceptual Evaluation of Speech Quality (PESQ) measure, standardized as ITU-T Recommendation P. 862, which normally yields a mean opinion score (MOS) between 1 (bad) and 4.5 (excellent) [11]. In cases of high distortion, the PESQ score may drop below 1.

4. SPHERICAL HARMONIC DOMAIN TRADEOFF BEAMFORMER

In this section, we derive a signal-dependent tradeoff beamformer. Our aim is to minimize the narrowband speech distortion index with the constraint that the narrowband noise reduction factor is greater than one. Mathematically, this is equivalent to

$$\min_{\mathbf{h}(k)} \upsilon_{\mathrm{sd}}\left[\mathbf{h}(k)\right] \quad \text{s.t.} \quad \xi_{\mathrm{nr}}\left[\mathbf{h}(k)\right] = \beta^{-1} \tag{16}$$

$$\min_{\mathbf{h}(k)}\left|\mathbf{h}^H(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) - 1\right|^2 \quad \text{s.t.} \quad \mathbf{h}^H(k)\Phi_{\widetilde{\mathbf{v}}}(k)\mathbf{h}(k) = \beta\phi_{\widetilde{V}_{00}}(k),$$

where $0 < \beta < 1$ to ensure that we get some noise reduction. Using the Woodbury matrix identity and a Lagrange multiplier, $\mu \geq 0$, to adjoin the constraint to the cost function, we deduce the tradeoff filter:

$$\mathbf{h}_{\mathrm{T},\mu}(k) = \phi_{\widetilde{X}_{00}}(k)\left[\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k) + \mu\mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k)\right]^{-1}\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k) \tag{17a}$$

$$= \frac{\phi_{\widetilde{X}_{00}}(k)\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)}{\mu + \phi_{\widetilde{X}_{00}}(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}^H(k)\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)} \tag{17b}$$

$$= \frac{\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\mathbf{\Phi}_{\widetilde{\mathbf{p}}}(k) - \mathbf{I}_N}{\mu + \mathrm{tr}\left[\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\mathbf{\Phi}_{\widetilde{\mathbf{p}}}(k)\right] - N}\mathbf{i}_{1,N}, \tag{17c}$$

where the Lagrange multiplier, μ, satisfies

$$\mathbf{h}_{\mathrm{T},\mu}(k)\Phi_{\widetilde{\mathbf{v}}}(k)\mathbf{h}_{\mathrm{T},\mu}(k) = \beta\phi_{\widetilde{V}_{00}}(k), \tag{18}$$

$\mathbf{I}_N$ is the $N \times N$ identity matrix, and $\mathbf{i}_{1,N} = [1\ 0\ \cdots\ 0]^T$ is the first column of $\mathbf{I}_N$.

However, in practice it is not easy to determine the optimal μ. Nevertheless, when this parameter is chosen in an ad-hoc way, it has been shown [12] that the $\mu = 0$ and $\mu = 1$ cases respectively correspond to the MVDR and Wiener filters, that $\mu > 1$ results in low residual noise at the expense of high speech distortion, and that $\mu < 1$ results in high residual noise and low speech distortion. While in principle μ could be frequency-dependent, in this work for simplicity μ is chosen to be a frequency-independent constant.

In some cases it can be advantageous to separate the tradeoff filter into two parts, using (17b) and (17c):

$$\mathbf{h}_{\mathrm{T},\mu}(k) = \frac{\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)}{\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}^H(k)\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(k)}\,\frac{\mathrm{tr}\left[\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\mathbf{\Phi}_{\widetilde{\mathbf{p}}}(k)\right]-N}{\mu+\mathrm{tr}\left[\mathbf{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(k)\mathbf{\Phi}_{\widetilde{\mathbf{p}}}(k)\right]-N}. \tag{19}$$

The first term represents an MVDR beamformer in the SHD, while the second term is a single-channel parametric Wiener filter which is always smaller than or equal to one. The behaviour of the tradeoff filter as a function of μ is clearly seen here: as μ increases, the second term term becomes smaller and the beamformer attenuates more noise but also causes more speech distortion.

5. PERFORMANCE EVALUATION

5.1. Experimental setup

We evaluated the performance of our beamformer in a simulated room with dimensions 5 × 7 × 4 m and a reverberation time of 300 ms. We computed room impulse responses (RIRs) using SMIRgen, a RIR generator for spherical microphone arrays [13]. We simulated a rigid 32 microphone spherical array with radius 4.2 cm placed approximately in the center of the room, and applied our 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 $0°$, while an interfering talker was placed at an azimuth of $140°$. For both talkers the elevation was $0°$ and the source-array distance was 1 m.

The desired source and coherent noise signals consisted of 60 s of male and female speech from the EBU SQAM disc [14]. The incoherent sensor noise consisted of spatially white noise with a constant input signal to incoherent noise ratio (iSINR) of 20 dB, while the coherent interference had an input signal to coherent noise ratio (iSCNR) between -40 and 40 dB. Noise levels were set based on active speech levels, computed according to ITU-T Rec. P. 56, at a microphone located at $(r, 0°, 69°)$.

Processing was done at a sampling frequency of 8 kHz in the STFT domain with a block length of 256 ms and a 50% overlap between successive frames. In practice the covariance matrix $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}$

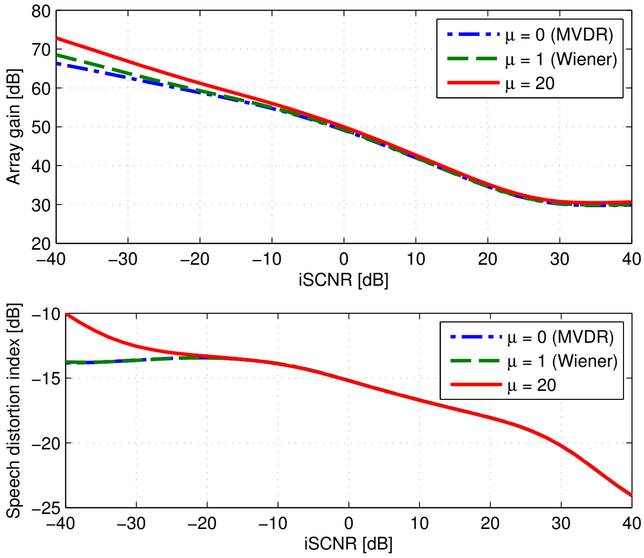

Fig. 1. Array gain and speech distortion index as a function of iSCNR, for various values of μ. The input SINR was fixed at 20 dB.

and coherence vector $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ can be computed using (10) from time-frequency bins where the coherent noise source is inactive, where the desired source is inactive, and where the desired and coherent noise sources are inactive. In this study, we used the signal vectors $\widetilde{\mathbf{p}}$ and $\widetilde{\mathbf{v}}$ directly, in order to neglect the influence of a multi-speaker voice activity detector. We estimated the covariance matrices recursively as $\boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}(k,t) = \alpha\boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}(k,t-1) + (1-\alpha)\boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}(k,t)$, where t denotes the time frame index and the weighting factor α was set to 0.8.

The broadband performance measures were computed per frame and then averaged over all frames, based on frequencies from 100 Hz to 4 kHz. The broadband array gain was computed as the ratio of the broadband equivalents of the output and input SNRs, as in [12, eqn. 4.38]. The broadband speech distortion index was computed by taking the sum over the above frequencies of the numerator and denominator of (15a), the narrowband speech distortion index.

5.2. Results

In Fig. 1 we plot the broadband array gain and speech distortion index as a function of iSCNR, for three values of μ. We obtain array gains of up to 72 dB for low iSCNR values (high levels of coherent noise), and speech distortion as low as -24 dB for high iSCNR values. As expected, both the array gain and speech distortion index increase with μ, reflecting the tradeoff between noise reduction and speech distortion. The difference is most noticeable for low iSCNR values; for higher iSCNRs it is worth choosing a high μ value as this adds little distortion but still gives higher noise reduction.

We also plot the subjective speech quality as given by the PESQ MOS in Fig. 2, using the omnidirectional speech signal as a reference. It can be seen that after being filtered the speech quality is greatly improved, especially for low iSCNRs. The input scores should be disregarded below -30 dB due to the fact that the desired speech is essentially inaudible at these iSCNRs.

6. CONCLUSIONS

The proposed tradeoff beamformer achieves high performance even in high levels of coherent noise: the noise is substantially reduced while keeping the distortion of the desired speech low. The tradeoff

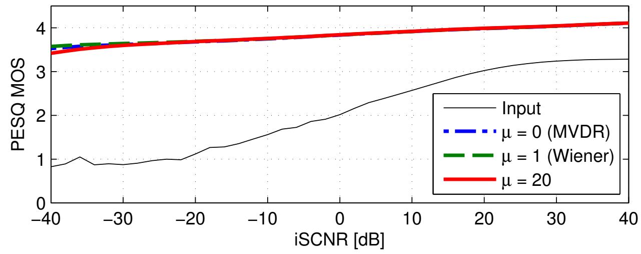

Fig. 2. PESQ MOS as a function of iSCNR, before and after applying the tradeoff beamformer, for various values of μ.

between noise reduction and speech distortion is illustrated, and it is shown that the tradeoff parameter μ should be kept small in high levels of coherent noise. The speech quality results obtained using PESQ are in agreement with the objective performance measures we computed, as well as the results of informal listening tests[2].

7. REFERENCES

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

[2] B. Rafaely, "Spatial sampling and beamforming for spherical microphone arrays," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2008, pp. 5–8.

[3] J. Meyer and T. Agnello, "Spherical microphone array for spatial sound recording," in *Proc. Audio Eng. Soc. Convention*, New York, NY, USA, Oct. 2003, pp. 1–9.

[4] 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.

[5] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.

[6] Y. Peled and B. Rafaely, "Linearly constrained minimum variance method for spherical microphone arrays in a coherent environment," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, June 2011, pp. 86–91.

[7] E.A.P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 158–170, Jan. 2010.

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

[9] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, 1993.

[10] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

[11] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

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

[13] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 129–132.

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

[2] A number of audio examples are available at http://www.ee.ic.ac.uk/sap/sphtradeoff/