JOINT DEREVERBERATION AND NOISE REDUCTION USING A TWO-STAGE BEAMFORMING APPROACH

Emanuël A.P. Habets

International Audio Laboratories Erlangen[†]
Am Wolfsmantel 33, 91058 Erlangen, Germany

Jacob Benesty

University of Quebec, INRS-EMT
800 de la Gauchetiere Ouest, Suite 6900
Montreal, Quebec, Canada

ABSTRACT

In this paper a two-stage beamforming approach is presented for dereverberation and noise reduction. The first stage comprises a delay-and-sum (DS) beamformer that generates a reference signal that contains a spatially filtered version of the desired speech and interference. In general, the desired speech component at the output of the DS beamformer contains less reverberation compared to reverberant speech signal received at the microphones. The second stage uses the filtered microphone signals and the noisy reference signal to estimate the desired speech component at the output of the DS beamformer. A major advantage over classical approaches is that the proposed approach is able to dereverberate the received desired signal with very low speech distortion. The dereverberation and noise reduction performance is evaluated for a circular microphone array.

Index Terms— Speech enhancement, dereverberation, noise reduction, beamforming, microphone arrays.

1. INTRODUCTION

Distant or hands-free audio acquisition is required in many applications such as audio-bridging and teleconferencing. A microphone array is often used for the acquisition and consists of multiple microphones that are arranged in a specific pattern. The received microphone signals usually consist of a desired speech signal and interference. The received signals are processed in order to extract the desired speech, or in other words to suppress the interference. In the last four decades many algorithms have been proposed to process the received signals (c.f. [1, 2] and the references therein).

The minimum variance distortionless response (MVDR) beamformer, proposed by Capon [3], is aiming at minimizing the output power under the constraint on the response of the array towards a desired source signal. To avoid this constrained optimization, Griffiths and Jim [4] proposed the generalized sidelobe canceller (GSC) structure, which separates the output power minimization and the application of the constraint. Initially, the GSC structure was based on the assumption that the different sensors receive a delayed version of the desired signal, and therefore we refer to it as the delay generalized sidelobe canceller (D-GSC). The D-GSC comprises three blocks: i) a filter-and-sum beamformer (FSB), which aligns the desired direct sound signal components, ii) a blocking matrix (BM), which blocks the desired direct sound components resulting in reference noise signals, and iii) a multichannel adaptive noise canceller (ANC), which eliminates noise components that leak through the sidelobes of the FSB.

The D-GSC suffers from two basic problems [2]. The first problem is caused by the non-ideal FSB, which can lead to a non-coherent filter-and-sum operation that results in a distortion of the desired speech. Doclo and Moonen [5] and Nordholm et al. [6] use spatial and frequency domain constraints to improve the robustness of beamformers. The second problem is caused by non-perfect BM and is known as the *leakage problem*. If the desired speech leaks into the noise reference signals the noise canceller filters will subtract speech components from the FSB output, causing self-cancellation of the desired speech, and hence a severe speech distortion. Even when the ANC filters are estimated or adapted during noise-only periods, the self-cancellation is unavoidable. Although not as intuitively clear, the MVDR also suffers from these two basic problems when the constraint is inadequate (note that the MVDR and GSC are mathematically equivalent).

The techniques to limit the distortion resulting from this leakage can be divided into two classes. Techniques in the first class reduce the leakage components in the noise references. This can be achieved i) by using a more robust fixed blocking matrix design [5], ii) by using an adaptive blocking matrix [7–9], or iii) by constructing a blocking matrix based on estimating the ratios of the acoustic transfer functions from the speech source to the microphone array [10]. Techniques in the second class limit the distorting effect of the leakage components i) by controlling the step-size of the multichannel adaptive algorithm, such that the multichannel ANC is only updated during periods (and for frequencies) where the signal-to-noise ratio (SNR) is low [9], ii) by constraining the update formula for the multichannel adaptive filter, e.g. by imposing a quadratic inequality constraint [11, 12] or by using the leaky least mean square algorithm [13], or iii) by taking speech distortion due to speech leakage into account using a speech distortion weighted multichannel Wiener filter [14, 15].

The MVDR proposed in [1] and GSC proposed in [10] aim at estimating the reverberant signal received by one of the microphones and thereby alleviates the leakage problem. The disadvantage of this approach is that the reverberation of the received speech signal is not reduced. In this paper a novel two-stage beamforming approach is presented that allows both dereverberation and noise reduction with low speech distortion. The first stage comprises of a DS beamformer that generates a reference signal that contains a spatially filtered version of the desired speech signal and interference. In general, the desired speech component at the output of the DS beamformer contains less reverberation compared to reverberant speech signal received at the microphones. The second stage uses the filtered microphone signals and the noisy reference signal to estimate the desired speech component at the output of the DS beamformer.

[†] A joint venture between the Friedrich-Alexander University of Erlangen-Nuremberg and the Fraunhofer-Gesellschaft, Germany.

978-1-4577-0999-9/11/$26.00 ©2011 IEEE

The paper is organized as follows. In Section 2 the signal model is described. In Section 3 the proposed two-stage beamforming approach is presented. In Section 4, performance measures are defined that are subsequently used in Section 5 for the performance evaluation. Finally, conclusions are provided in Section 6.

2. SIGNAL MODEL

We consider the well-accepted room acoustics signal model in which an N-element microphone array captures a convolved source signal in some noise field. The received signals, at the discrete time-index k, are expressed as [1]

$$\begin{aligned} y_n(k) &= g_n(k) * s(k) + v_n(k) \qquad (1) \\ &= x_n(k) + v_n(k),\ n = 1, 2, \ldots, N, \end{aligned}$$

where $g_n(k)$ is the impulse response from the unknown speech source $s(k)$ to the nth microphone, $*$ stands for linear convolution, and $v_n(k)$ is the additive noise at microphone n. We assume that the signals $x_n(k) = g_n(k) * s(k)$ and $v_n(k)$ are uncorrelated and zero mean. By definition, $x_n(k)$ is coherent across the array while $v_n(k)$ may be non-coherent or partially coherent. All previous signals are considered to be real and broadband.

Expression (1) can be rewritten in the frequency domain, at the frequency-index f, as

$$\begin{aligned} Y_n(f) &= G_n(f)S(f) + V_n(f) \qquad (2) \\ &= X_n(f) + V_n(f),\ n = 1, 2, \ldots, N, \end{aligned}$$

where $Y_n(f)$, $G_n(f)$, $S(f)$, $X_n(f) = G_n(f)S(f)$, and $V_n(f)$ are the frequency-domain representations of $y_n(k)$, $g_n(k)$, $s(k)$, $x_n(k)$, and $v_n(k)$, respectively.

3. TWO-STAGE BEAMFORMING APPROACH

3.1. First Stage: Dereverberation

Let $\mathcal{F}_{1n}$ be a function that relates the source position to the relative delay between microphones 1 and n (with $\mathcal{F}_{11} = 0$). The first stage consists of aligning the signals in such a way that the microphone array "looks" in the direction of the source. The geometry of the array as well as the position of the source are implicitly assumed to be known here. Therefore, by delaying the observation signals by $\mathcal{F}_{1n}$, (2) becomes

$$\begin{aligned} \widetilde{Y}_n(f) &= Y_n(f)e^{jf\mathcal{F}_{1n}} \\ &= X_n(f)e^{jf\mathcal{F}_{1n}} + V_n(f)e^{jf\mathcal{F}_{1n}} \\ &= \widetilde{X}_n(f) + \widetilde{V}_n(f),\ n = 1, 2, \ldots, N. \qquad (3) \end{aligned}$$

We then form the classical DS beamformer in the frequency domain:

$$\begin{aligned} \widetilde{Y}_{\mathrm{DS}}(f) &= \frac{1}{N}\sum_{n=1}^{N} \widetilde{Y}_n(f) \\ &= \widetilde{X}_{\mathrm{d}}(f) + \widetilde{V}_{\mathrm{ref}}(f), \qquad (4) \end{aligned}$$

where

$$\widetilde{X}_{\mathrm{d}}(f) = \frac{1}{N}\sum_{n=1}^{N} \widetilde{X}_n(f) = \frac{1}{N}\sum_{n=1}^{N} X_n(f)e^{jf\mathcal{F}_{1n}} \qquad (5)$$

is the desired signal and

$$\widetilde{V}_{\mathrm{ref}}(f) = \frac{1}{N}\sum_{n=1}^{N} \widetilde{V}_n(f) = \frac{1}{N}\sum_{n=1}^{N} V_n(f)e^{jf\mathcal{F}_{1n}} \qquad (6)$$

is the reference noise signal. Clearly, the inverse Fourier transform of $\widetilde{X}_{\mathrm{d}}(f)$ is, in principle, less reverberant than any of the signals $x_n(k),\ n = 1, 2, \ldots, N$. This, basically, concludes the first stage, which consists of dereverberating the reverberant speech signal at the sensors with a DS beamformer[1].

The signal $\widetilde{Y}_{\mathrm{DS}}(f)$ is considered as the reference signal since it contains the desired signal, $\widetilde{X}_{\mathrm{d}}(f)$, that we will try to extract, in the second stage, from the aligned noisy observations, $\widetilde{Y}_n(f),\ n = 1, 2, \ldots, N$. We see that the variance of $\widetilde{Y}_{\mathrm{DS}}(f)$ is

$$\begin{aligned} \phi_{\widetilde{Y}_{\mathrm{DS}}}(f) &= E\left[\left|\widetilde{Y}_{\mathrm{DS}}(f)\right|^2\right] \\ &= \phi_{\widetilde{X}_{\mathrm{d}}}(f) + \phi_{\widetilde{V}_{\mathrm{ref}}}(f), \qquad (7) \end{aligned}$$

where $E[\cdot]$ denotes mathematical expectation, and $\phi_{\widetilde{X}_{\mathrm{d}}}(f) = E\left[\left|\widetilde{X}_{\mathrm{d}}(f)\right|^2\right]$ and $\phi_{\widetilde{V}_{\mathrm{ref}}}(f) = E\left[\left|\widetilde{V}_{\mathrm{ref}}(f)\right|^2\right]$ are the variances of $\widetilde{X}_{\mathrm{d}}(f)$ and $\widetilde{V}_{\mathrm{ref}}(f)$, respectively.

3.2. Second Stage: Noise Reduction

The second stage consists of multichannel noise reduction. The desired signal in the frequency domain is, obviously, $\widetilde{X}_{\mathrm{d}}(f)$ and it is very important to write (3) as a function of this desired signal. Since $\widetilde{X}_{\mathrm{d}}(f)$ and $\widetilde{X}_n(f)$ are coherent (i.e., they come from the same source) and the DS beamformer is a linear processing, it can be shown that

$$\widetilde{X}_n(f) = \gamma_{\widetilde{X}_n\widetilde{X}_{\mathrm{d}}}(f)\widetilde{X}_{\mathrm{d}}(f),\ n = 1, 2, \ldots, N, \qquad (8)$$

where

$$\gamma_{\widetilde{X}_n\widetilde{X}_{\mathrm{d}}}(f) = \frac{E\left[\widetilde{X}_n(f)\widetilde{X}_{\mathrm{d}}^*(f)\right]}{E\left[\left|\widetilde{X}_{\mathrm{d}}(f)\right|^2\right]},\ n = 1, 2, \ldots, N \qquad (9)$$

is the partially normalized [with respect to $\widetilde{X}_{\mathrm{d}}(f)$] coherence function between $\widetilde{X}_{\mathrm{d}}(f)$ and $\widetilde{X}_n(f)$ and the superscript * denotes complex conjugation. Using (8), we can express (3) as

$$\widetilde{Y}_n(f) = \gamma_{\widetilde{X}_n\widetilde{X}_{\mathrm{d}}}(f)\widetilde{X}_{\mathrm{d}}(f) + \widetilde{V}_n(f),\ n = 1, 2, \ldots, N. \qquad (10)$$

It is more convenient to write the N frequency-domain aligned microphone signals in a vector notation as

$$\begin{aligned} \widetilde{\mathbf{y}}(f) &= \widetilde{\mathbf{x}}(f) + \widetilde{\mathbf{v}}(f) \\ &= \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\mathrm{d}}}(f)\widetilde{X}_{\mathrm{d}}(f) + \widetilde{\mathbf{v}}(f), \qquad (11) \end{aligned}$$

where

$$\widetilde{\mathbf{y}}(f) = \begin{bmatrix} \widetilde{Y}_1(f) & \widetilde{Y}_2(f) & \cdots & \widetilde{Y}_N(f) \end{bmatrix}^T,$$

superscript ^{T} denotes transpose of a vector or a matrix, $\widetilde{\mathbf{x}}(f)$ and

[1] A DS beamformer is able to reduce the level of the additive noise as well, but this reduction is often not very significant.

$\widetilde{\mathbf{v}}(f)$ are defined similarly to $\widetilde{\mathbf{y}}(f)$, and

$$\begin{aligned}\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f) &= \begin{bmatrix} \gamma_{\widetilde{X}_1\widetilde{X}_{\rm d}}(f) & \gamma_{\widetilde{X}_2\widetilde{X}_{\rm d}}(f) & \cdots & \gamma_{\widetilde{X}_N\widetilde{X}_{\rm d}}(f) \end{bmatrix}^T \\ &= \frac{E\left[\widetilde{\mathbf{x}}(f)\widetilde{X}_{\rm d}^*(f)\right]}{E\left[\left|\widetilde{X}_{\rm d}(f)\right|^2\right]} \qquad (12)\end{aligned}$$

is the partially normalized [with respect to $\widetilde{X}_{\rm d}(f)$] coherence vector (of length N) between $\widetilde{X}_{\rm d}(f)$ and $\widetilde{\mathbf{x}}(f)$, which can be seen as the steering vector.

We can also express $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)$ as a function of $\tilde{\mathbf{y}}(f)$ and $\tilde{\mathbf{v}}(f)$, i.e.,

$$\begin{aligned}\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f) = &\frac{\phi_{\widetilde{Y}_{\rm DS}}(f)}{\phi_{\widetilde{Y}_{\rm DS}}(f) - \phi_{\widetilde{V}_{\rm ref}}(f)}\boldsymbol{\gamma}_{\widetilde{\mathbf{y}}\widetilde{Y}_{\rm DS}}(f) \\ &- \frac{\phi_{\widetilde{V}_{\rm ref}}(f)}{\phi_{\widetilde{Y}_{\rm DS}}(f) - \phi_{\widetilde{V}_{\rm ref}}(f)}\boldsymbol{\gamma}_{\widetilde{\mathbf{v}}\widetilde{V}_{\rm ref}}(f), \qquad (13)\end{aligned}$$

where $\boldsymbol{\gamma}_{\widetilde{\mathbf{y}}\widetilde{Y}_{\rm DS}}(f)$ and $\boldsymbol{\gamma}_{\widetilde{\mathbf{v}}\widetilde{V}_{\rm ref}}(f)$ are defined similarly to $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)$. Now, $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)$ depends on the second-order statistics of the aligned observation and noise signals. This expression is extremely useful in practice, as the statistics of the noise signals can be computed during silences as in any other noise reduction algorithm.

From (11), we easily deduce the covariance matrix of $\widetilde{\mathbf{y}}(f)$:

$$\begin{aligned}\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}(f) &= E\left[\widetilde{\mathbf{y}}(f)\widetilde{\mathbf{y}}^H(f)\right] \\ &= \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(f) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(f) \\ &= \phi_{\widetilde{X}_{\rm d}}(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\boldsymbol{\gamma}^H_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(f), \qquad (14)\end{aligned}$$

where the superscript ^{H} is the transpose-conjugate operator and $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(f) = \phi_{\widetilde{X}_{\rm d}}(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\boldsymbol{\gamma}^H_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)$ and $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(f) = E\left[\widetilde{\mathbf{v}}(f)\widetilde{\mathbf{v}}^H(f)\right]$ are the covariance matrices of $\widetilde{\mathbf{x}}(f)$ and $\widetilde{\mathbf{v}}(f)$, respectively. The $N \times N$ matrix $\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}(f)$ is the sum of two other matrices: one is of rank equal to 1 and the other one (covariance matrix of the noise) is assumed to be full rank.

Now, multichannel noise reduction is performed by applying a complex weight to the aligned output of each sensor and summing across the array, i.e.,

$$\begin{aligned}Z(f) &= \sum_{n=1}^{N} H_n^*(f)\widetilde{Y}_n(f) \\ &= \mathbf{h}^H(f)\widetilde{\mathbf{y}}(f) \\ &= \widetilde{X}_{\rm fd}(f) + \widetilde{V}_{\rm rn}(f), \qquad (15)\end{aligned}$$

where $Z(f)$ is an estimate of $\widetilde{X}_{\rm d}(f)$,

$$\mathbf{h}(f) = \begin{bmatrix} H_1(f) & H_2(f) & \cdots & H_N(f) \end{bmatrix}^T$$

is a filter of length N containing all the complex gains applied to the aligned microphone outputs at frequency f, $\widetilde{X}_{\rm fd}(f) = \mathbf{h}^H(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\widetilde{X}_{\rm d}(f)$ is the filtered desired signal, and $\widetilde{V}_{\rm rn}(f) = \mathbf{h}^H(f)\widetilde{\mathbf{v}}(f)$ is the residual noise.

We then deduce the variance of $Z(f)$:

$$\begin{aligned}\phi_Z(f) &= \mathbf{h}^H(f)\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}(f)\mathbf{h}(f) \\ &= \phi_{\widetilde{X}_{\rm fd}}(f) + \phi_{\widetilde{V}_{\rm rn}}(f), \qquad (16)\end{aligned}$$

where $\phi_{\widetilde{X}_{\rm fd}}(f) = \phi_{\widetilde{X}_{\rm d}}(f)\left|\mathbf{h}^H(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\right|^2$ and $\phi_{\widetilde{V}_{\rm rn}}(f) = \mathbf{h}^H(f)\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(f)\mathbf{h}(f)$.

The well-known minimum variance distortionless response (MVDR) beamformer proposed by Capon [3] is derived by minimizing the narrowband mean squared error of the residual noise, $\phi_{\widetilde{V}_{\rm rn}}(f)$, with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\arg\min_{\mathbf{h}(f)} \mathbf{h}^H(f)\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(f)\mathbf{h}(f) \quad \text{s.t.} \quad \mathbf{h}^H(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f) = 1, \qquad (17)$$

for which the solution is

$$\mathbf{h}_{\rm MVDR}(f) = \frac{\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)}{\boldsymbol{\gamma}^H_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)}. \qquad (18)$$

Using the fact that $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(f) = \phi_{\widetilde{X}_{\rm d}}(f)\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\boldsymbol{\gamma}^H_{\widetilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)$, the explicit dependence of the above filter on the steering vector is eliminated to obtain the following form:

$$\mathbf{h}_{\rm MVDR}(f) = \frac{\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(f)\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}(f) - \mathbf{I}_N}{\text{tr}\left[\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(f)\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}(f)\right] - N}\mathbf{1}_N, \qquad (19)$$

where tr$[\cdot]$ denotes the trace of a square matrix, $\mathbf{I}_N$ is the identity matrix of size $N \times N$ and $\mathbf{1}_N = \frac{1}{N}\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T$ of length N.

4. PERFORMANCE MEASURES

In this section, we derive some useful performance measures that should fit well with the two-stage beamforming approach explained in the previous section. We can divide these performance measures into three categories. In the first one, we should be able to evaluate dereverberation with the DS beamformer. The second category evaluates the noise reduction performance, while the third one quantifies speech distortion.

4.1. Dereverberation

We define the narrowband input signal-to-reverberation ratio (SRR) as

$$\text{iSRR}(f) = \frac{\phi_{\widetilde{S}}(f)}{\phi_{X_1}(f) - \phi_{\widetilde{S}}(f)}. \qquad (20)$$

where $\phi_{\widetilde{S}}(f) = E\left\{\widetilde{S}(f)\widetilde{S}^*(f)\right\}$, $\widetilde{S}(f) = S(f)e^{if\tau_d}$ and τ_d is the propagation time of the sound from the source position to the position of microphone 1. The broadband input SRR is given by

$$\text{iSRR} = \frac{\int_{-1/2}^{1/2}\phi_{\widetilde{S}}(f)df}{\int_{-1/2}^{1/2}\left[\phi_{X_1}(f) - \phi_{\widetilde{S}}(f)\right]df}. \qquad (21)$$

The narrowband output SRR is given by

$$\text{oSRR}(f) = \frac{\phi_{\widetilde{S}}(f)}{\phi_{\widetilde{X}_{\rm d}}(f) - \phi_{\widetilde{S}}(f)} \qquad (22)$$

and the broadband output SRR is then given by

$$\text{oSRR} = \frac{\int_{-1/2}^{1/2}\phi_{\widetilde{S}}(f)df}{\int_{-1/2}^{1/2}\left[\phi_{\widetilde{X}_{\rm d}}(f) - \phi_{\widetilde{S}}(f)\right]df}. \qquad (23)$$

4.2. Noise Reduction

In this study our objective is to estimate the desired speech component at the output of the DS beamformer, i.e., $\widetilde{X}_{\rm d}(f)$. This signal can be regarded as a filtered version of the desired speech received by one of the microphones. Using the first microphone as a reference the narrowband input SNR can be defined as

$$\text{iSNR}(f) = \frac{\phi_{\widetilde{X}_{\rm d}}(f)}{\phi_{\widetilde{V}_1}(f)}. \tag{24}$$

We define the broadband input SNR as

$$\text{iSNR} \quad = \quad \frac{\int_{-1/2}^{1/2} \phi_{\widetilde{X}_{\rm d}}(f)df}{\int_{-1/2}^{1/2} \phi_{\widetilde{V}_1}(f)df}.$$

To quantify the level of noise remaining at the beamformer output, we define the output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual noise. We have the narrowband output SNR

$$\begin{aligned} \text{oSNR}\left[\mathbf{h}(f)\right] &= \frac{\phi_{\widetilde{X}_{\rm fd}}(f)}{\phi_{\widetilde{V}_{\rm rn}}(f)} \\ &= \frac{\phi_{\widetilde{X}_{\rm d}}(f)\left|\mathbf{h}^H(f)\boldsymbol{\gamma}_{\tilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\right|^2}{\mathbf{h}^H(f)\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(f)\mathbf{h}(f)} \end{aligned} \tag{25}$$

and the broadband output SNR

$$\text{oSNR}\left(\mathbf{h}\right) = \frac{\int_{-1/2}^{1/2} \phi_{\widetilde{X}_{\rm d}}(f)\left|\mathbf{h}^H(f)\boldsymbol{\gamma}_{\tilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\right|^2 df}{\int_{-1/2}^{1/2} \mathbf{h}^H(f)\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(f)\mathbf{h}(f)df}. \tag{26}$$

The noise reduction factor measures the amount of noise being rejected by the beamformer. This quantity is defined as the ratio of the power of the reference noise over the power of the noise remaining at the beamformer output. We define the narrowband and broadband noise reduction factors as

$$\xi_{\rm nr}\left[\mathbf{h}(f)\right] \quad = \quad \frac{\phi_{\widetilde{V}_1}(f)}{\mathbf{h}^H(f)\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(f)\mathbf{h}(f)}, \tag{27}$$

$$\xi_{\rm nr}\left(\mathbf{h}\right) \quad = \quad \frac{\int_{-1/2}^{1/2} \phi_{\widetilde{V}_1}(f)df}{\int_{-1/2}^{1/2} \mathbf{h}^H(f)\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(f)\mathbf{h}(f)df}. \tag{28}$$

The noise reduction factor is expected to be lower bounded by 1; otherwise, the beamformer amplifies the reference noise. The higher the value of the noise reduction factor, the more the noise is rejected.

4.3. Speech Distortion

Since the noise is reduced by the filtering operation, so is, in general, the desired speech. This speech reduction (or cancellation) implies, in general, speech distortion.

The speech reduction factor, which is somewhat similar to the noise reduction factor, is defined as the ratio of the variance of the desired signal over the variance of the filtered desired signal. The narrowband and broadband speech reduction factors are defined as

$$\xi_{\rm sr}\left[\mathbf{h}(f)\right] \quad = \quad \frac{1}{\left|\mathbf{h}^H(f)\boldsymbol{\gamma}_{\tilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\right|^2}, \tag{29}$$

$$\xi_{\rm sr}\left(\mathbf{h}\right) \quad = \quad \frac{\int_{-1/2}^{1/2} \phi_{\widetilde{X}_{\rm d}}(f)df}{\int_{-1/2}^{1/2} \phi_{\widetilde{X}_{\rm d}}(f)\left|\mathbf{h}^H(f)\boldsymbol{\gamma}_{\tilde{\mathbf{x}}\widetilde{X}_{\rm d}}(f)\right|^2 df}. \tag{30}$$

A key observation is that the design of beamformers that do not cancel the desired signal requires the constraint $\xi_{\rm sr} = 1$. Thus, the speech reduction factor is equal to 1 if there is no cancellation and expected to be greater than 1 when cancellation happens.

5. PERFORMANCE EVALUATION

We now evaluate the performance of the two-stage beamfomer in a simulated reverberant environment of size $7.5 \times 9 \times 6$ m (length × width × height) and a reverberation time of 600 ms. All room impulse responses were computed using an efficient implementation of the source-image method [16]. Circular arrays with a radius of 12.5 cm were used with either 4 or 8 microphones. The source was placed on the rotation axis of the array and the source-array distance r_d ranges from 0.5 until 4 meters. The source signal consisted of male and female speech with a total length of 10 minutes. The additive noise was a mixture of spatially white noise and spherically isotropic (diffuse) noise (generated using [17]) with a constant input SNR of 35 and 5 dB, respectively. Monte Carlo simulations were conducted (in total 50 trials per source-array distance) by translating and rotating the source-array configuration in the room.

The processing was done at 8 kHz in the short-time Fourier transform domain with a window length of 256 ms and the overlap between successive time frames was 50%. In practice the covariance matrix $\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}$ can be estimated for each time-frequency bin when the desired source is inactive. In this study we have used the signal vector $\tilde{\mathbf{v}}$ directly in order to put aside the influence of a voice activity detector. The covariance matrixes $\boldsymbol{\Phi}_{\tilde{\mathbf{y}}}$ and $\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}$ were estimated recursively using $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{v}}$ with a weighting factor of 0.98 that was determined experimentally. In this study we assume the direction of arrival is known such that the signals can be properly time-aligned. It is important to note that the proposed approach will not suffer from any signal cancellation if the microphone signals are not correctly time aligned. Instead, only the desired signal will be defined differently. For the second stage we have used the beamformer given in (18). In addition, (13) was used to compute $\boldsymbol{\gamma}_{\tilde{\mathbf{x}}\widetilde{X}_{\rm d}}$, where $\boldsymbol{\gamma}_{\tilde{\mathbf{y}}\widetilde{Y}_{\rm DS}}$ and $\boldsymbol{\gamma}_{\tilde{\mathbf{v}}\widetilde{V}_{\rm ref}}$ were computed using $\boldsymbol{\Phi}_{\tilde{\mathbf{y}}}$, $\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}$, $\phi_{\widetilde{V}_{\rm ref}}$ and $\phi_{\widetilde{Y}_{\rm DS}}$.

In Fig. 1 the results for the broadband performance measures defined in Section 4 are given. The performance measures were computed per frame and then averaged across the frames and subsequently averaged across the Monte Carlo trials. The SRR and SNR measures were averaged in the log domain while the speech reduction factor was averaged in the linear domain. In Fig. 1a the input and output SRRs are depicted. As expected both values decrease monotonically with increasing source-array distance. While the reverberation reduction in terms of the SRR for $N = 8$ is only around 2 dB larger compared to $N = 4$, there is a noticeable perceptual improvement that is not illustrated by this measure. It should be noted that when the source-array distance is smaller than the critical distance the variance of the desired source increases when the source-array distance decreases. Because the input SNR is kept constant the noise variance increases when the variance of the desired source increases. In Fig. 1b the SNR improvement (oSNR - iSNR) is shown for the DS beamformer (dashed line) and for the MVDR beamformer that jointly reduces reverberation and noise (solid line). The SNR at the output of the DS is shown to be independent of the source-array distance. It is extremely interesting to see that SNR improvement of the MVDR beamformer increases when the SRR

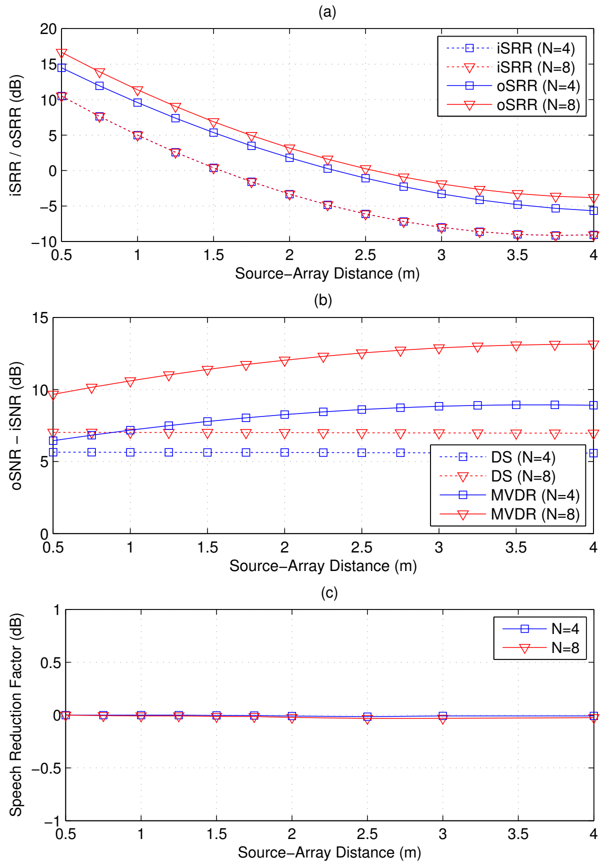

Fig. 1. Performance measures for different source-array distances r_d and a constant input SNR. The results for $N = 4$ are denoted by a square and the results for $N = 8$ by a triangle. In the center the SNR improvement is shown for the output of the DS beamformer (dashed line) and for the output of the MVDR beamformer (solid line).

decreases, and that the output SNR reaches its asymptotic maximum when the SRR reaches its asymptotic minimum. In Fig. 1c it is shown that the speech reduction factor is close to zero dB for all distances and number of microphones. Hence, the desired speech at the output of the noise reduction stage is equal to the desired speech component at the output of the DS beamformer.

6. CONCLUSIONS

Recently, various beamformers were designed to estimate the desired speech as received by one of the microphones. While these allow for a high noise reduction and low speech distortion they do not provide the possibility to dereverberate the speech signal. In this work, a two-stage beamformer was proposed that allows for joint dereverberation and noise reduction. The performance evaluation showed that both dereverberation and noise reduction can be achieved with little speech distortion.

7. REFERENCES

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

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

[3] J. Capon, “High resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[4] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[5] S. Doclo and M. Moonen, “Design of far-field and near-field broadband beamformers using eigenfilters,” *Signal Processing*, vol. 83, no. 12, pp. 2641–2673, Dec. 2003.

[6] S. Nordholm, H. Q. Dam, N. Grbić, and S. Y. Low, “Adaptive microphone array employing spatial quadratic soft constraints and spectral shaping,” in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., Signals and Communication Technology, pp. 229–246. Springer, Berlin, Germany, 2005.

[7] D. van Compernolle, “Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1990, vol. 2, pp. 833–836.

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

[9] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, “Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1340–1351, May 2007.

[10] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[11] N. Jablon, “Adaptive beamforming with the generalized sidelobe canceller in the presence of array imperfections,” *IEEE Trans. Antennas Propag.*, vol. 34, no. 8, pp. 996–1012, Aug. 1986.

[12] H. Cox, R. M. Zeskind, and M. M. Owen, “Robust adaptive beamforming,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[13] I. Claesson and S. Nordholm, “A spatial filtering approach to robust adaptive beamforming,” *IEEE Trans. Antennas Propag.*, vol. 40, no. 1, pp. 1093–1096, Sept. 1992.

[14] A. Spriet, M. Moonen, and J. Wouters, “Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids,” *IEEE Trans. Signal Process.*, vol. 53, no. 3, pp. 911–925, Mar. 2005.

[15] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction,” *Speech Communication*, vol. 49, no. 7–8, pp. 636–656, Aug. 2007.

[16] E. A. P. Habets, “Room impulse response (RIR) generator,” May 2008.

[17] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.