COHERENT AND INCOHERENT INTERFERENCE REDUCTION USING A SUBBAND TRADEOFF BEAMFORMER

Emanuël A. P. Habets[1] *and Jacob Benesty*[2]

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

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

ABSTRACT

Signals captured by a set of microphones in a speech communication system are mixtures of desired and undesired signals. In this paper another perspective on subband beamformers in room acoustics is provided. Specifically, the observed undesired signals are divided into coherent and incoherent additive components while no assumption is being made regarding the number of coherent undesired sources. From this perspective a general tradeoff beamformer is proposed that enables a compromise between noise reduction and speech distortion on the one hand, and coherent noise versus incoherent noise reductions on the other hand. The presented performance evaluation shows how existing beamformers and the tradeoff beamformer perform in a particular scenario.

1. INTRODUCTION

Distant or hands-free speech capture is required in many applications such as hearing aids and teleconferencing. Microphone arrays are often used for the acquisition and consist of sets of microphones that are arranged in specific topologies. The received microphone signals usually consist of a mixture of signals of the desired source, signals of the undesired sources, and ambient noise. As the acoustic interference degrades the quality and intelligibility of the desired source, the received signals are processed (i.e., filtered and summed) in order to extract the desired source signals or in other words, reduce the interference (i.e., signals of the undesired sources plus ambient noise).

In the last four decades numerous spatio-temporal filters have been proposed to process the received microphone signals (see [1,2] and the references therein). Many filters were originally developed for, and used in, wireless communication systems. More recently filters were developed specifically for speech communication systems. Existing filters can be divided into those that preserve or distort the desired signal. Filters that preserve the desired signal are, for example, the minimum variance distortionless response (MVDR) beamformer (also known as Capon's beamformer) [3] that reduces the interference-plus-noise power, and the linearly constrained minimum variance (LCMV) beamformer [4]. The LCMV is a generalization of the MVDR and commonly aims at minimizing the beamformer's output power while satisfying multiple constraints such as rejecting the signals of the undesired sources and passing the desired signal through undistorted. Another beamformer that consists of a weighted sum of the LCMV and a matched filter (i.e., an MVDR that reduces ambient noise only) has been recently proposed by Souden et al. [5]. The proposed beamformer allows a tradeoff between the undesired signal and ambient noise reductions. The multichannel Wiener filter (MWF), on the other hand, reduces the interference-plus-noise power without exactly preserving the desired signal. In order to control the amount of distortion, the parameterized multichannel Wiener filter (also known as speech-distortion weighted multichannel Wiener filter) has been proposed [6, 7].

To design an effective speech acquisition system, a clear understanding of the functioning of noise reduction filters as well as the ability to control various tradeoffs is paramount. In previous work the undesired signals are commonly composed of coherent noise source components (one for each noise source) and ambient noise. The ambient noise is often considered to be a mixture of spatially-white and diffuse noise. In this paper we express the undesired signals in terms of a coherent signal component and an incoherent signal component. From this perspective, we deduce a new and general beamformer that allows a tradeoff between speech distortion and noise reduction on the one hand, and coherent and incoherent noise reductions on the other hand. Using the proposed signal model, we provide additional insight into the behavior of subband beamformers, which can aid the design process of beamformers in which various tradeoffs need to be realized.

This paper is organized as follows. Section 2 provides a new signal model, linear array model, definitions, and fundamental assumptions made in this paper. In Section 3 the performance measures are defined. In Section 4 a new general tradeoff beamformer is deduced using the proposed signal model. Section 5 investigates the performance of the proposed beamformer. Finally, Section 6 provides some concluding remarks.

2. PROPOSED 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. In the short-term Fourier transform domain we can express the spectral coefficients of the received signals at time frame m and discrete frequency k as[1]

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

where $G_n(k)$ is the transfer function from the unknown speech source $S(k,m)$ to the nth microphone that is assumed to be time-invariant, and $V_n(k,m)$ is the additive noise at microphone n. We assume that the spectral coefficients $X_n(k,m) = G_n(k)S(k,m)$ and $V_n(k,m)$ are uncorrelated and zero-mean complex random variables. By definition, $X_n(k,m)$ is coherent across the array. We will get back to the noise components, $V_n(k,m)$, at the end of this section.

It is more convenient to write the N microphone signals in a vector notation:

$$\begin{aligned} \mathbf{y}(k,m) &= \mathbf{g}(k)S(k,m) + \mathbf{v}(k,m) \\ &= \mathbf{x}(k,m) + \mathbf{v}(k,m) \\ &= \mathbf{d}(k)X_1(k,m) + \mathbf{v}(k,m), \end{aligned} \tag{2}$$

† A joint institution of Fraunhofer IIS and the Friedrich-Alexander University of Erlangen-Nuremberg, Germany.

[1] In this work we assume that the analysis window is sufficiently long such that the multiplicative transfer function approximation [8] holds.

where

$$\begin{aligned}
\mathbf{y}(k,m) &= [Y_1(k,m)\ Y_2(k,m)\ \cdots\ Y_N(k,m)]^T,\\
\mathbf{x}(k,m) &= [X_1(k,m)\ X_2(k,m)\ \cdots\ X_N(k,m)]^T,\\
&= S(k,m)\,[G_1(k)\ G_2(k)\ \cdots\ G_N(k)]^T\\
&= S(k,m)\mathbf{g}(k),\\
\mathbf{v}(k,m) &= [V_1(k,m)\ V_2(k,m)\ \cdots\ V_N(k,m)]^T,\\
\mathbf{d}(k) &= \left[1\ \frac{G_2(k)}{G_1(k)}\ \cdots\ \frac{G_N(k)}{G_1(k)}\right]^T\\
&= \frac{\mathbf{g}(k)}{G_1(k)},
\end{aligned}$$

and superscript ^{T} denotes transpose of a vector or a matrix. The vector $\mathbf{d}(k)$ is termed the steering vector since it determines the direction of the desired signal $X_1(k,m)$. This definition is a generalization of the classical steering vector to a reverberant (multipath) environment.

The covariance matrix of $\mathbf{y}(k,m)$ is given by

$$\begin{aligned}
\mathbf{\Phi}_{\mathbf{y}}(k,m) &= E\left[\mathbf{y}(k,m)\mathbf{y}^H(k,m)\right]\\
&= \mathbf{\Phi}_{\mathbf{x}}(k,m) + \mathbf{\Phi}_{\mathbf{v}}(k,m) \qquad (3a)\\
&= \phi_{X_1}(k,m)\mathbf{d}(k)\mathbf{d}^H(k) + \mathbf{\Phi}_{\mathbf{v}}(k,m), \qquad (3b)
\end{aligned}$$

where $E[\cdot]$ is the mathematical expectation, the superscript ^{H} denotes the transpose-conjugate operator, $\phi_{X_1}(k,m) = E\left[|X_1(k,m)|^2\right]$, $\mathbf{\Phi}_{\mathbf{x}}(k,m) = E\left[\mathbf{x}(k,m)\mathbf{x}^H(k,m)\right]$ and $\mathbf{\Phi}_{\mathbf{v}}(k,m) = E\left[\mathbf{v}(k,m)\mathbf{v}^H(k,m)\right]$. The $N\times N$ matrix $\mathbf{\Phi}_{\mathbf{y}}(k,m)$ 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.

Expressions (2) and (3b) exploit only the fact that the desired signal, $X_1(k,m)$, is completely coherent across all microphones but, so far, nothing is said about the reference noise term $V_1(k,m)$. In practice, however, $V_1(k,m)$ is likely at least partially coherent with the noise components, $V_n(k,m)$, at the other microphones. Therefore, any noise term $V_n(k,m)$ can be easily decomposed into two orthogonal components, i.e.,

$$V_n(k,m) = \gamma_{V_nV_1}(k,m)V_1(k,m) + V_n'(k,m),\ n = 1,2,\ldots,N, \qquad (4)$$

where

$$\gamma_{V_nV_1}(k,m) = \frac{E\left[V_n(k,m)V_1^*(k,m)\right]}{E\left[|V_1(k,m)|^2\right]} \qquad (5)$$

is the partially normalized [with respect to $V_1(k,m)$] coherence function between $V_n(k,m)$ and $V_1(k,m)$, the superscript * denotes complex conjugation, and

$$E\left[V_1^*(k,m)V_n'(k,m)\right] = 0,\ n = 1,2,\ldots,N. \qquad (6)$$

In contrast to previous works we do not make any assumptions regarding the number of coherent noise sources. The vector $\mathbf{v}(k,m)$ can then be written as the sum of two other vectors: one that is coherent with $V_1(k,m)$ and the other one that is incoherent with $V_1(k,m)$, i.e.,

$$\mathbf{v}(k,m) = \boldsymbol{\gamma}(k,m)V_1(k,m) + \mathbf{v}'(k,m), \qquad (7)$$

where

$$\boldsymbol{\gamma}(k,m) = [1\ \gamma_{V_2V_1}(k,m)\ \cdots\ \gamma_{V_NV_1}(k,m)]^T$$

is the partially normalized [with respect to $V_1(k,m)$] coherence vector and

$$\mathbf{v}'(k,m) = \left[0\ V_2'(k,m)\ \cdots\ V_N'(k,m)\right]^T.$$

Substituting (7) into (2), we get

$$\mathbf{y}(k,m) = \mathbf{d}(k)X_1(k,m) + \boldsymbol{\gamma}(k,m)V_1(k,m) + \mathbf{v}'(k,m). \qquad (8)$$

We see that the microphone signal vector is the sum of three other vectors that are mutually incoherent and from which the first two depend explicitly on the desired and noise signals at the reference microphone.

Using (8), we can now decompose the covariance matrix of $\mathbf{y}(k,m)$ as

$$\begin{aligned}
\mathbf{\Phi}_{\mathbf{y}}(k,m) &= \phi_{X_1}(k,m)\mathbf{d}(k)\mathbf{d}^H(k)\\
&\quad + \phi_{V_1}(k,m)\boldsymbol{\gamma}(k,m)\boldsymbol{\gamma}^H(k,m) + \mathbf{\Phi}_{\mathbf{v}'}(k,m), \qquad (9)
\end{aligned}$$

where $\phi_{V_1}(k,m) = E\left[|V_1(k,m)|^2\right]$ is the variance of $V_1(k,m)$ and $\mathbf{\Phi}_{\mathbf{v}'}(k,m)$ is the covariance matrix of $\mathbf{v}'(k,m)$. The matrix $\mathbf{\Phi}_{\mathbf{y}}(k,m)$ is the sum of three other matrices: the first two are of rank equal to 1 and the last one (covariance matrix of the incoherent noise) is assumed to be of rank equal to $N-1$.

The conventional linear beamforming is performed by applying a complex weight to the output of each microphone and summing across the array, i.e.,

$$\begin{aligned}
Z(k,m) &= \mathbf{h}^H(k,m)\mathbf{y}(k,m)\\
&= X_1(k,m)\mathbf{h}^H(k,m)\mathbf{d}(k)\\
&\quad + V_1(k,m)\mathbf{h}^H(k,m)\boldsymbol{\gamma}(k,m) + \mathbf{h}^H(k,m)\mathbf{v}'(k,m)\\
&= X_{1,\mathrm{f}}(k,m) + V_{1,\mathrm{f}}(k,m) + V_{\mathrm{rn}}(k,m), \qquad (10)
\end{aligned}$$

where

$$\mathbf{h}(k,m) = [H_1(k,m)\ H_2(k,m)\ \cdots\ H_N(k,m)]^T \qquad (11)$$

is a filter of length N containing all the complex gains applied to the microphone outputs, $X_{1,\mathrm{f}}(k,m) = X_1(k,m)\mathbf{h}^H(k,m)\mathbf{d}(k)$ is the filtered desired signal, $V_{1,\mathrm{f}}(k,m) = V_1(k,m)\mathbf{h}^H(k,m)\boldsymbol{\gamma}(k,m)$ is the filtered coherent noise, and $V_{\mathrm{rn}}(k,m) = \mathbf{h}^H(k,m)\mathbf{v}'(k,m)$ is the residual incoherent noise. Therefore, with an appropriate choice of $\mathbf{h}(k,m)$, $Z(k,m)$ can be a good estimate of $X_1(k,m)$.

The three terms on the right-hand side of (10) are mutually incoherent. Hence, the variance of $Z(k,m)$ is also the sum of three variances:

$$\begin{aligned}
\phi_Z(k,m) &= \mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{y}}(k,m)\mathbf{h}(k,m)\\
&= \phi_{X_{1,\mathrm{f}}}(k,m) + \phi_{V_{1,\mathrm{f}}}(k,m) + \phi_{V_{\mathrm{rn}}}(k,m), \qquad (12)
\end{aligned}$$

where $\phi_{X_{1,\mathrm{f}}}(k,m) = \phi_{X_1}(k,m)\left|\mathbf{h}^H(k,m)\mathbf{d}(k)\right|^2$, $\phi_{V_{1,\mathrm{f}}}(k,m) = \phi_{V_1}(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}(k,m)\right|^2$, and $\phi_{V_{\mathrm{rn}}}(k,m) = \mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}'}(k,m)\mathbf{h}(k,m)$. The different variances on the right-hand side of (12) are important in the definitions of the performance measures.

3. PERFORMANCE MEASURES

Many performance measures exist in the literature but in this part we only discuss the ones that are the most useful for studying and evaluating the proposed beamformer. Since the signal we want to recover is the clean (but convolved) signal received at microphone 1, i.e., $X_1(k,m)$, therefore microphone 1 is serving as the reference microphone.

We define the subband input SNR as

$$\mathrm{iSNR}(k,m) = \frac{\phi_{X_1}(k,m)}{\phi_{V_1}(k,m)}. \qquad (13)$$

To quantify the level of noise remaining in the output signal of the beamformer, $Z(k,m)$, we define the subband output SNR as the

ratio of the power of the filtered desired signal over the power of the residual noise, i.e.,

$$\text{oSNR}\left[\mathbf{h}(k,m)\right] = \frac{\phi_{X_{1,\text{f}}}(k,m)}{\phi_{V_{1,\text{f}}}(k,m)+\phi_{V_{\text{rn}}}(k,m)} = \frac{\phi_{X_1}(k,m)\left|\mathbf{h}^H(k,m)\mathbf{d}(k)\right|^2}{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}(k,m)\mathbf{h}(k,m)}. \quad (14)$$

In this study it is desired to know the decomposition of the residual noise at the output of the beamformer, i.e., $\phi_{V_{1,\text{f}}}(k,m) + \phi_{V_{\text{rn}}}(k,m)$. We therefore define the subband coherent noise fraction (CNF) as the variance of the residual coherent noise over the variance of the total residual noise and subband incoherent noise fraction (INF) as the variance of the residual incoherent noise over the variance of the total residual noise, i.e.,

$$\text{CNF}[\mathbf{h}(k,m)] = \frac{\phi_{V_{1,\text{f}}}(k,m)}{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}(k,m)\mathbf{h}(k,m)} = \frac{\phi_{V_1}(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}(k,m)\right|^2}{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}(k,m)\mathbf{h}(k,m)}. \quad (15)$$

and

$$\text{INF}[\mathbf{h}(k,m)] = \frac{\phi_{V_{\text{rn}}}(k,m)}{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}(k,m)\mathbf{h}(k,m)} = \frac{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}'}(k,m)\mathbf{h}(k,m)}{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}(k,m)\mathbf{h}(k,m)}. \quad (16)$$

Therefore we always have $0 \leq \text{CNF}[\mathbf{h}(k,m)] \leq 1$, $0 \leq \text{INF}[\mathbf{h}(k,m)] \leq 1$, and $\text{CNF}[\mathbf{h}(k,m)] + \text{INF}[\mathbf{h}(k,m)] = 1$. When the coherent noise is completely reduced the subband CNF is 0.

In general beamforming algorithms distort the desired signal. In order to quantify the level of this distortion, we define the subband speech-distortion index as

$$\upsilon_{\text{sd}}\left[\mathbf{h}(k,m)\right] = \frac{E\left[\left|\mathbf{h}^H(k,m)\mathbf{d}(k) - X_1(k,m)\right|^2\right]}{\phi_{X_1}(k,m)} = \left|\mathbf{h}^H(k,m)\mathbf{d}(k) - 1\right|^2. \quad (17)$$

The speech-distortion index is always greater than or equal to 0 and should be upper bounded by 1; so the higher is the value of $\upsilon_{\text{sd}}\left[\mathbf{h}(k,m)\right]$, the more the desired signal is distorted.

Corresponding segmental performance measures are obtained by averaging the numerator and denominator in (14)-(17) across all discrete frequencies.

4. MULTICHANNEL TRADEOFF BEAMFORMER

In this section, we derive a general tradeoff filter that is able to compromise between noise reduction and speech distortion on the one hand, and coherent noise versus incoherent noise reductions on the other hand.

In order to control the speech distortion we can use the constraint

$$\mathbf{h}^H(k,m)\mathbf{d}(k) = \alpha(k,m), \quad (18)$$

where $\alpha(k,m)$ is a complex number. The closer the value of $|\alpha(k,m)|^2$ is to one, the less the amplitude response of the desired signal is distorted; and for $\alpha(k,m) = 1$, there is no distortion. We can introduce a second constraint in order to compromise between reduction of the coherent and incoherent noises. Thus, we have

$$\mathbf{h}^H(k,m)\boldsymbol{\gamma}(k,m) = \beta(k,m), \quad (19)$$

where $\beta(k,m)$ is a complex number. The closer the value of $|\beta(k,m)|^2$ is to zero, the more the coherent noise is reduced; and for $\beta(k,m) = 0$, the coherent noise is completely removed.

Putting the two previous constraints together

$$\mathbf{M}_{\text{C}}^H(k,m)\mathbf{h}(k,m) = \mathbf{i}_{\text{T}}(k,m), \quad (20)$$

where

$$\mathbf{i}_{\text{T}}(k,m) = \left[\begin{array}{cc} \alpha(k,m) & \beta(k,m) \end{array}\right]^H,$$

and minimizing the energy of the residual noise at the beamformer output subject to (20), we obtain the tradeoff beamformer

$$\mathbf{h}_{\text{T}}(k,m) = \mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{M}_{\text{C}}(k,m)\left[\mathbf{M}_{\text{C}}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{M}_{\text{C}}(k,m)\right]^{-1}\mathbf{i}_{\text{T}}(k,m). \quad (21)$$

For convenience, we can rewrite (21) as

$$\mathbf{h}_{\text{T}}(k,m) = \alpha(k,m)\mathbf{m}_1(k,m) + \beta(k,m)\mathbf{m}_2(k,m), \quad (22)$$

where

$$\left[\begin{array}{cc} \mathbf{m}_1(k,m) & \mathbf{m}_2(k,m) \end{array}\right] = \mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{M}_{\text{C}}(k,m)\left[\mathbf{M}_{\text{C}}^H(k,m)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{M}_{\text{C}}(k,m)\right]^{-1}. \quad (23)$$

All performance measures can be written as a function of $\alpha(k,m)$ and/or $\beta(k,m)$:

$$\text{oSNR}\left[\mathbf{h}_{\text{T}}(k,m)\right] = \frac{\phi_{X_1}(k,m)|\alpha(k,m)|^2}{\phi_{V_1}(k,m)|\beta(k,m)|^2 + \mathbf{h}_{\text{T}}^H(k,m)\mathbf{\Phi}_{\mathbf{v}'}(k,m)\mathbf{h}_{\text{T}}(k,m)}, \quad (24)$$

$$\upsilon_{\text{sd}}\left[\mathbf{h}_{\text{T}}(k,m)\right] = |\alpha(k,m) - 1|^2, \quad (25)$$

and

$$\text{CNF}\left[\mathbf{h}_{\text{T}}(k,m)\right] = \phi_{V_1}(k,m)|\beta(k,m)|^2\left[m_1(k,m)|\alpha(k,m)|^2 + \left[\phi_{V_1}(k,m) + m_2(k,m)\right]|\beta(k,m)|^2 + 2\,\text{Re}\{m_{12}(k,m)\alpha(k,m)\beta(k,m)\}\right]^{-1}, \quad (26)$$

where

$$m_1(k,m) = \mathbf{m}_1^H(k,m)\mathbf{\Phi}_{\mathbf{v}'}(k,m)\mathbf{m}_1(k,m),$$
$$m_2(k,m) = \mathbf{m}_2^H(k,m)\mathbf{\Phi}_{\mathbf{v}'}(k,m)\mathbf{m}_2(k,m),$$
$$m_{12}(k,m) = \mathbf{m}_1^H(k,m)\mathbf{\Phi}_{\mathbf{v}'}(k,m)\mathbf{m}_2(k,m).$$

We will now show that the LCMV, MVDR and MWF are special cases of the proposed beamformer. The conventional LCMV is obtained for

$$\mathbf{h}_{\text{LCMV}}(k,m) = \mathbf{m}_1(k,m), \quad (27)$$

such that

$$\alpha_{\text{LCMV}}(k,m) = 1,$$
$$\beta_{\text{LCMV}}(k,m) = 0.$$

We are now going to derive the MVDR from (22). Because the MVDR does not distort the desired signal we have

$$\mathbf{h}_{\text{MVDR}}(k,m) = \mathbf{m}_1(k,m) + \beta_{\text{MVDR}}(k,m)\,\mathbf{m}_2(k,m). \quad (28)$$

Since the output SNR is upper bounded and its maximum is obtained with the MVDR, we can derive $\beta_{\text{MVDR}}(k,m)$ from (24). Indeed, taking $\alpha(k,m) = 1$ and maximizing $\text{oSNR}\left[\mathbf{h}_{\text{T}}(k,m)\right]$ with respect to $\beta(k,m)$, we find that

$$\beta_{\text{MVDR}}(k,m) = -\frac{m_{12}^*(k,m)}{\phi_{V_1}(k,m) + m_2(k,m)}. \quad (29)$$

Finally, the multichannel Wiener filter can be expressed as [1]

$$\mathbf{h}_{\mathrm{W}}(k,m) = C_{\mathrm{W}}(k,m)\mathbf{h}_{\mathrm{MVDR}}(k,m), \tag{30}$$

where

$$\begin{aligned} C_{\mathrm{W}}(k,m) &= \mathbf{h}_{\mathrm{W}}^{H}(k,m)\mathbf{d}(k) \\ &= \frac{\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{\Phi}_{\mathbf{x}}(k,m)\right]}{1+\mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{\Phi}_{\mathbf{x}}(k,m)\right]} \end{aligned} \tag{31}$$

can be seen as a single-channel Wiener gain. Therefore, we can also deduce another form for the Wiener filter:

$$\begin{aligned} \mathbf{h}_{\mathrm{W}}(k,m) &= C_{\mathrm{W}}(k,m)\mathbf{m}_1(k,m) \\ &\quad - \frac{C_{\mathrm{W}}(k,m)\, m_{12}^{*}(k,m)}{\phi_{V_1}(k,m)+m_2(k,m)}\mathbf{m}_2(k,m), \end{aligned} \tag{32}$$

such that

$$\begin{aligned} \alpha_{\mathrm{W}}(k,m) &= C_{\mathrm{W}}(k,m), \\ \beta_{\mathrm{W}}(k,m) &= -\frac{C_{\mathrm{W}}(k,m)\, m_{12}^{*}(k,m)}{\phi_{V_1}(k,m)+m_2(k,m)}. \end{aligned}$$

5. PERFORMANCE EVALUATION

We will now evaluate the performance of the proposed subband beamformer in a reverberant environment. Here we focus in particular on the tradeoff between noise reduction and speech distortion on the one hand, and coherent noise versus incoherent noise reductions on the other hand.

The results of our simulations are presented in terms of the segmental performance measures described in Section 3. Because the filters are computed and applied on a frame-by-frame basis the segmental performance measures are evaluated per frame and subsequently averaged over all frames. The SNRs are averaged in the log domain while other performance measures are averaged in the linear domain.

5.1 Experimental Setup and Implementation

A linear microphone array was used with 4 microphones and an inter-microphone distance of 2.5 cm. The distance between the floor and the microphone array was 1.6 m. The acoustic impulse responses (each with a duration of 500 ms) were generated using an efficient implementation of the image-method [9]. The room size was $6.4 \times 5 \times 4$ m (length$\times$width$\times$height) and the reflection coefficients of the walls, ceiling, and floor were set to achieve a reverberation time $T_{60} = 0.4$ s. The distance between the first microphone and the desired and undesired sources are denoted by r_{d} and r_{u}, respectively. Furthermore, θ_{d} and θ_{u} determine the azimuth angle of the desired and undesired sources as shown in Figure 1.

The desired source consists of 10 minutes of male and female speech composed of data from the APLAWD speech database [10] sampled at 8 kHz and is located at ($\theta_{\mathrm{d}} = 90$ degrees, $r_{\mathrm{d}} = 1$ m). The undesired source is located at ($\theta_{\mathrm{u}} = 130$ degrees, $r_{\mathrm{u}} = 2$ m) and consists of an USASI noise sequence that exhibits the same spectral properties as speech. The input signal to coherent noise ratio (iSCNR) was set to 10 dB. Finally, spatio-temporal white Gaussian noise with a long-term input signal to incoherent noise ratio (iSINR) of 20 dB (evaluated at the first microphone) was added to the microphone signals. 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 noise statistics can be estimated when the desired source is inactive. Here we have put aside the problem of detecting when the desired source is active as we are interested in assessing the different tradeoffs of the proposed beamformer. Therefore, the covariance matrix $\mathbf{\Phi}_{\mathbf{v}}(k,m)$ is recursively estimated from $\mathbf{v}(k,m)$ using

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(k,m) = \eta\hat{\mathbf{\Phi}}_{\mathbf{v}}(k,m-1) + (1-\eta)\mathbf{v}(k,m)\mathbf{v}^{H}(k,m), \tag{33}$$

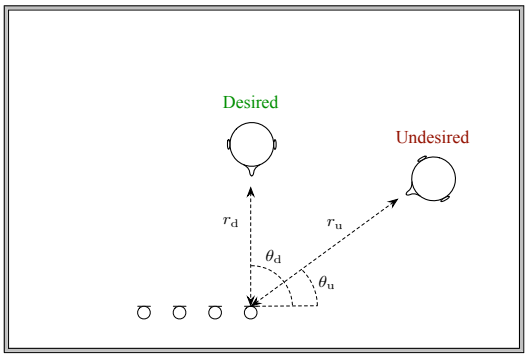


Figure 1: Schematic drawing showing the array configuration and the location of the desired and undesired source.

where η ($0 \leq \eta < 1$) denotes a weighting factor that was set to 0.98. In addition, $\mathbf{\Phi}_{\mathbf{y}}(k,m)$ was estimated similarly to $\mathbf{\Phi}_{\mathbf{v}}(k,m)$ using $\mathbf{y}(k,m)$. Using (3a) we can then obtain an estimate of $\mathbf{\Phi}_{\mathbf{x}}(k,m)$. The vector $\boldsymbol{\gamma}(k,m)$ is obtained using

$$\hat{\boldsymbol{\gamma}}(k,m) = \eta\,\hat{\boldsymbol{\gamma}}(k,m-1) + (1-\eta)\,\frac{\hat{\mathbf{\Phi}}_{\mathbf{v}}(k,m)}{\hat{\phi}_{V_1}(k,m)}\mathbf{u}, \tag{34}$$

where $\mathbf{u} = [1,0,\ldots,0]^{T}$. An estimate of the steering vector $\mathbf{d}(k)$ is estimated similarly to $\boldsymbol{\gamma}$ in (34) using $\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,m)$. Finally, to compute (21) we require the inverse of $\mathbf{\Phi}_{\mathbf{v}}(k,m)$. In this study a regularization technique is used to compute the inverse such that $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)$ is replaced by

$$\left[\mathbf{\Phi}_{\mathbf{v}}(k,m) + \frac{\delta\,\mathrm{tr}\{\mathbf{\Phi}_{\mathbf{v}}(k,m)\}}{N}\mathbf{I}_{N\times N}\right]^{-1}, \tag{35}$$

where δ is the regularization parameter that was determined experimentally and set to 10^{-6} and $\mathbf{I}_{N\times N}$ is an identify matrix of size $N \times N$.

For the proposed beamformer we have two design parameters, viz., $\alpha(k,m)$ and $\beta(k,m)$. These parameters can be chosen in many ways, some of which perceptually motivated. Rather than evaluating the entire parameter space of the beamformer, we propose to investigate another parameter space such that it includes the MVDR, LCMV, and parameterized multichannel Wiener filter (PMWF) [1, 6, 7]. Specifically, we define $\alpha(k,m)$ as

$$\alpha(k,m;\lambda) = H_{\mathrm{PW}}(k,m;\lambda), \tag{36}$$

with

$$H_{\mathrm{PW}}(k,m;\lambda) = \frac{\mathrm{tr}\left[\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(k,m)\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,m)\right]}{\lambda+\mathrm{tr}\left[\hat{\mathbf{\Phi}}_{\mathbf{v}}^{-1}(k,m)\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,m)\right]} \tag{37}$$

such that we can control the total noise reduction and speech distortion using λ. Furthermore, we define $\beta(k,m)$ as

$$\beta(k,m;\beta',\lambda) = -\beta'\,\frac{\alpha(k,m;\lambda)\,\hat{m}_{12}^{*}(k,m)}{\hat{\phi}_{V_1}(k,m)+\hat{m}_2(k,m)}, \tag{38}$$

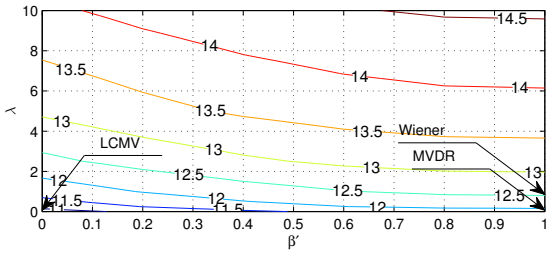
such that we can investigate the behavior of the beamformer as a function of β'. For ($\lambda = 0, \beta' = 0$), we then obtain the LCMV and for ($\lambda = 0, \beta' = 1$), we then obtain the MVDR. For ($\lambda \geq 0, \beta' = 1$), we obtain solutions of the PMWF. Previously unexplored solutions that allow a tradeoff between speech distortion and noise reduction on the one hand, and coherent and incoherent noise reductions on the other hand are given by ($\lambda \geq 0, 0 < \beta' < 1$).

5.2 Results

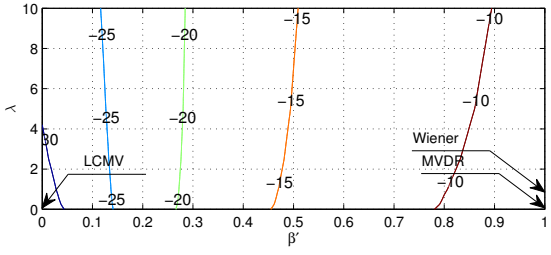
We investigated the beamformer's performance in terms of the average segmental SNR improvement (i.e., oSNR - iSNR in dB), coherent noise fraction, and speech-distortion index as a function of λ and β'. The contour plots of the obtained results are shown in Fig. 2. We observe that the SNR improvement depicted in Fig. 2(a) increases monotonically with decreasing β' and increasing λ. We also observe that for the considered scenario and compared to λ, β' has only a minor influence on the SNR improvement. In Fig. 2(b) the CNF is depicted as a function of λ and β'. We observe that this power ratio is mainly controlled by β'. For the considered scenario the range in which we can vary the coherent noise fraction using β' is about 20 dB. Hence, for any given value λ we are able to control the amount of residual coherent and residual incoherent noise at the output of the beamformer. In Fig. 2(c) the speech-distortion index is shown in dB. Firstly, we observe that there is minimal distortion of the desired signal for small values of λ. Secondly, we observe that β' has negligible influence on the speech-distortion index. As stated in previous works (e.g., [6,7]) we see that the output SNR can be increased at the expense of introducing more speech distortion.

In Fig. 3 the coherent and incoherent noise fractions averaged across the time frames are shown for $\beta' = \{0, 0.5, 1\}$ and $\lambda = 0$. Clearly the largest amount of coherent noise is reduced when $\beta' = 0$ (i.e., LCMV beamformer) and the least amount of coherent noise is reduced when $\beta' = 1$ (i.e., MVDR beamformer). Especially at low frequencies the INF is very close to 0 dB, indicating that the residual noise consists mainly of non-coherent noise.

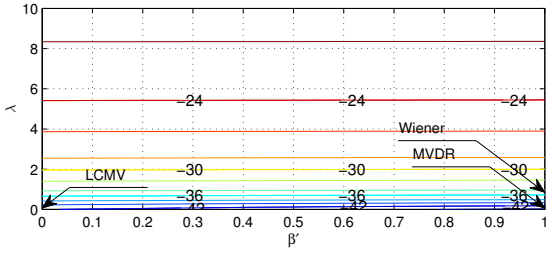
Hence, we see that the tradeoff beamformer enables a wide range of alternative solutions that allow a tradeoff between noise reduction and speech distortion on the one hand, and coherent noise versus incoherent noise reductions on the other hand.



(a) SNR improvement in dB.



(b) Coherent noise fraction in dB.



(c) Speech-distortion index in dB.

Figure 2: Contour plots of different performance measures as a function of β' and λ.

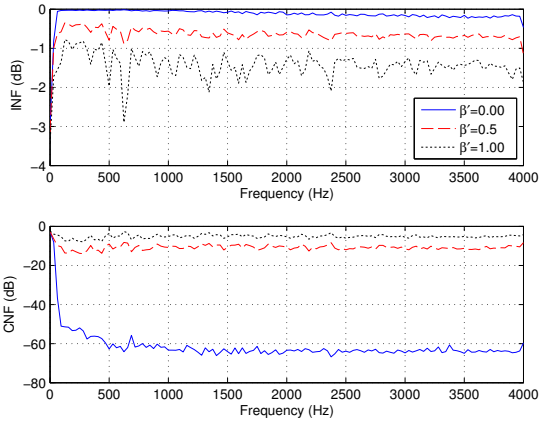


Figure 3: Average incoherent and coherent noise fractions.

6. CONCLUSIONS

In this contribution, the observed undesired signals are divided into spatially coherent and incoherent additive components while no assumption is being made regarding the number of coherent undesired sources. From this perspective performance measures are defined, and a new and general tradeoff beamformer is proposed that enables a compromise between noise reduction and speech distortion on the one hand, and coherent noise versus incoherent noise reductions on the other hand. It has been shown that existing beamformers are special cases of the proposed beamformer. The presented performance evaluation shows how existing beamformers and the tradeoff beamformer perform in a specific noise field. The tradeoffs facilitated by the proposed beamformer can be used in the development of effective speech acquisition systems and in the design of perceptually motivated beamformers.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 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. Springer-Verlag, 2008, ch. 47.

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

[4] M. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1378–1393, Dec. 1983.

[5] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

[6] A. Spriet, M. Moonen, and J. Wouters, "Spatially preprocessed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[7] S. Mehrez, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[8] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[9] E. A. P. Habets. (2008, May) Room impulse response (RIR) generator. [Online]. Available: http://home.tiscali.nl/ehabets/rirgenerator.html

[10] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Technical Report, Jun. 1987.