Applied Acoustics 74 (2013) 343–355

Contents lists available at SciVerse ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A perspective on multichannel noise reduction in the time domain

Jacob Benesty [a], Mehrez Souden [b,*], Jingdong Chen [c]

[a] INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900 Montreal, QC, Canada H5A 1K6
[b] NTT Communication Science Laboratories, 2-4, Hikaridai, Seika-cho, Kyoto 619-0237, Japan
[c] Northwestern Polytechnical University, 127, Youyi West Rd., Xi'an, Shanxi 710072, China

ARTICLE INFO

Article history:
Received 27 December 2011
Received in revised form 5 August 2012
Accepted 6 August 2012
Available online 24 October 2012

Keywords:
Noise reduction
Speech enhancement
Microphone array
Beamforming
Wiener filter
Minimum variance distortionless response (MVDR) filter
Tradeoff filter
Linearly constrained minimum variance (LCMV) filter
Space–time (ST) prediction filter

ABSTRACT

Conventional multichannel noise reduction techniques are formulated by splitting the processed microphone observations into two terms: filtered noise-free speech and residual additive noise. The first term is treated as desired signal while the second is a nuisance. Then, the objective has typically been to reduce the nuisance while keeping the filtered speech as similar as possible to the clean speech. It turns out that this treatment of the overall filtered speech as the desired signal is inappropriate as will become clear soon. In this paper, we present a new study of the multichannel time-domain noise reduction filters. We decompose the noise-free microphone array observations into two components where the first is correlated with the target signal and perfectly coherent across the sensors while the second consists of residual interference. Then, well-known time-domain filters including the minimum variance distortionless response (MVDR), the space–time (ST) prediction, the maximum signal-to-noise ratio (SNR), the linearly constrained minimum variance (LCMV), the multichannel tradeoff, and Wiener filters are derived. Besides, the analytical performance evaluation of these time-domain filters is provided and new insights into their functioning are presented. Numerical results are finally given to corroborate our study.

1. Introduction

Multichannel noise reduction has been garnering increasing research efforts since the pioneering work of Flanagan et al. in 1985 [1]. In fact, numerous multichannel noise reduction approaches have been recently developed [1–12]. These approaches have a common objective, which is to recover the noise-free signal at the reference microphone by employing the spatial and temporal properties of the observed mixtures of sounds.

Noise reduction can be achieved in either the time or some transform domains that include Fourier, Karhunen-Loève, cosine, and Hadamard [7]. Nevertheless, the transformation to the frequency domain is the most widely adopted since it offers an efficient way of implementation. For instance, in [8] Gannot et al. proposed a channel transfer function ratio (CTFR) based generalized side-lobe canceler (GSC) where the CTFRs are estimated online using the non-stationarity of speech. This approach was then extended to extract multiple target sources using the linearly constrained minimum variance (LCMV) in [9]. To properly design noise reduction filters [e.g., LCMV, minimum variance distortionless response (MVDR), tradeoff or parameterized Wiener filter] some fundamental issues have to be taken into account. First, the parameters affecting the tradeoff of noise reduction versus speech distortion and the tradeoff of interference rejection versus ambient noise reduction have to be determined [6,13]. Second, it is known that, similar to the conventional single-channel processing [14], the knowledge of only noise and noisy-data statistics is sufficient to implement noise reduction filters [2,4–6]. Hence, the accurate estimation of these statistics is paramount to effectively reduce the noise without causing significant speech distortion [6]. In [12], Cornelis et al. analytically studied the robustness of the parameterized multichannel Wiener filter to second-order-statistics estimation errors. Finally, even though frequency-domain noise reduction filters are theoretically equivalent to their time-domain counterparts, approximating the acoustic channel effect in the frequency domain remains a major issue from both practical and theoretical standpoints. Indeed, the time-domain linear convolution is commonly approximated by a scalar multiplication in the frequency domain. This approximation is reasonable provided that the analysis window is larger than the channel impulse responses. However, speech signals are inherently non-stationary, and taking long analysis windows compromises the accurate tracking of noise and speech statistics, thereby increasing the residual distortions. To cope with this issue, Talmon et al. proposed to use convolutive transfer functions in the frequency domain in [10]. However, this

[*] Corresponding author.
E-mail address: souden@emt.inrs.ca (M. Souden).

0003-682X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apacoust.2012.08.002

approach is still based on approximating the channel effect, and its performance cannot be exactly predicted from a theoretical point of view. Alternatively, the problem of noise reduction can be directly investigated in the time domain as in [2,4,11]. The analysis is then more rigorous since no approximation in some transformation domain is involved. However, the performance evaluation of the filtering techniques, including the aforementioned ones, is known to be of a challenge. This issue is addressed in this paper.

In this contribution, we introduce a new study of the time-domain multichannel noise reduction. In contrast to earlier conventional investigations, this study is based on the decomposition of the noise-free observations into two orthogonal components: the desired signal, which is fully coherent across the sensors and some additive interference. This decomposition is optimal in the second-order-statistics sense, and is, consequently, tailored to many widely used filters, including the maximum signal-to-noise ratio (SNR), MVDR, space–time (ST) prediction, LCMV, tradeoff, and Wiener filters. By utilizing this decomposition, we determine new expressions for these filters, and show that the time-domain MVDR, Wiener, tradeoff, and maximum SNR filters are identical up to a scaling factor. Finally, we carry out a simplified yet rigorous performance analysis of all these filters in terms of noise reduction, speech distortion, and output SNR. The concepts investigated in this paper can be easily extended to the transform domains including those mentioned above.

The remainder of this paper is organized as follows: Section 2 describes the signal propagation model. Section 3 outlines the second-order-statistics-based decomposition of the multichannel noise-free speech observations into two orthogonal components. An explicit form of the *time-domain steering vector* is obtained. Section 4 defines the objective performances metrics, namely the speech distortion index, noise reduction factor, and output SNR. These measures are perfectly tailored to the noise reduction formulation in this contribution. Section 5 revisits optimal multichannel noise reduction techniques and provides new expressions for the maximum SNR, MVDR, ST prediction, LCMV, tradeoff, and Wiener filters. Section 6 contains some simulation results that corroborate our study. Finally, Section 7 concludes this work.

2. Signal model

We consider the typical formulation of 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 [2,3,6,8]

$$y_n(k) = g_n(k) * s(k) + v_n(k) = x_n(k) + v_n(k), \quad n = 1, 2, \ldots, N, \tag{1}$$

here $g_n(k)$ is the impulse response from the unknown speech source $s(k)$ location to the nth microphone, $*$ stands for linear convolution, and $v_n(k)$ is the additive noise at microphone n. We assume that the signals $x_n(k)$ and $v_n(k)$ are uncorrelated and zero mean. By definition, $x_n(k) = g_n(k)*s(k)$ is coherent across the array for $n = 1, 2, \ldots, N$. The noise signals $v_n(k)$ are typically either partially coherent or non-coherent across the array. All signals are considered to be real, broadband, and to simplify the development and analysis of the main ideas of this work, we further assume that they are Gaussian. Note here that the signal model in (1) is general and no particular transform will be used in the following. Thus, the results of this contribution can be easily extended to noise reduction in transform domains.

By processing the data by blocks of L samples, the signal model given in (1) can be put into a vector form as

$$\mathbf{y}_n(k) = \mathbf{x}_n(k) + \mathbf{v}_n(k), \quad n = 1, 2, \ldots, N, \tag{2}$$

where

$$\mathbf{y}_n(k) = \begin{bmatrix} y_n(k) & y_n(k-1) & \cdots & y_n(k-L+1) \end{bmatrix}^T, \tag{3}$$

is a vector of length L, superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}_n(k)$ and $\mathbf{v}_n(k)$ are defined in a similar way to $\mathbf{y}_n(k)$. It is more convenient to concatenate the N vectors $\mathbf{y}_n(k)$ together as

$$\mathbf{y}(k) = \begin{bmatrix} \mathbf{y}_1^T(k) & \mathbf{y}_2^T(k) & \cdots & \mathbf{y}_N^T(k) \end{bmatrix}^T = \mathbf{x}(k) + \mathbf{v}(k), \tag{4}$$

where vectors $\mathbf{x}(k)$ and $\mathbf{v}(k)$ of length NL are defined in a similar way to $\mathbf{y}(k)$. Since $x_n(k)$ and $v_n(k)$ are uncorrelated by assumption, the correlation matrix (of size $NL \times NL$) of the microphone signals is

$$\mathbf{R}_\mathbf{y} = E[\mathbf{y}(k)\mathbf{y}^T(k)] = \mathbf{R}_\mathbf{x} + \mathbf{R}_\mathbf{v}, \tag{5}$$

where $E[\cdot]$ denotes mathematical expectation, and $\mathbf{R}_\mathbf{x} = E[\mathbf{x}(k)\mathbf{x}^T(k)]$ and $\mathbf{R}_\mathbf{v} = E[\mathbf{v}(k)\mathbf{v}^T(k)]$ are the correlation matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively.

With the above signal models, the objective of noise reduction is to estimate any one of the signals $x_n(k)$ [2,4,8,11]. Without loss of generality, we choose to estimate the speech signal received at microphone 1, i.e., $x_1(k)$ in this paper. Our problem then may be stated as follows [2]: given the N noisy signals $y_n(k)$, our aim is to estimate $x_1(k)$ and minimize the contribution of the noise terms $v_n(k)$ in the array output.

3. Linear array model

In our linear array model, we estimate the desired signal on a sample basis from the corresponding observation signal vector of length NL. At time k, the signal estimate is obtained as

$$\hat{x}_1(k) = \mathbf{h}^T\mathbf{y}(k), \tag{6}$$

where $\mathbf{h}$ is a finite-impulse-response (FIR) filter of length NL. The linear model in (6) can be rewritten as

$$\hat{x}_1(k) = \mathbf{h}^T[\mathbf{x}(k) + \mathbf{v}(k)] = x_\mathrm{f}(k) + v_\mathrm{rn}(k), \tag{7}$$

where $x_\mathrm{f}(k) = \mathbf{h}^T\mathbf{x}(k)$ is the filtered speech signal and $v_\mathrm{rn}(k) = \mathbf{h}^T\mathbf{v}(k)$ is the residual noise. From (7), we see that $\hat{x}_1(k)$ depends on the vector $\mathbf{x}(k)$; however, our desired signal at time k is only $x_1(k)$ [not the whole vector $\mathbf{x}(k)$]. Therefore, we should decompose $\mathbf{x}(k)$ into two orthogonal vectors: one corresponds to the desired signal at time k and the other corresponds to the interference. Indeed, it is easy to see that this decomposition is

$$\mathbf{x}(k) = x_1(k)\boldsymbol{\gamma}_\mathbf{x} + \mathbf{x}'(k) = \mathbf{x}_\mathrm{d}(k) + \mathbf{x}'(k), \tag{8}$$

where $\mathbf{x}_\mathrm{d}(k) = x_1(k)\boldsymbol{\gamma}_\mathbf{x}$ is the desired signal vector (of length NL), $\mathbf{x}'(k)$ is the interference signal vector (of length NL),

$$\boldsymbol{\gamma}_\mathbf{x} = \begin{bmatrix} \boldsymbol{\gamma}_{\mathbf{x}_1}^T & \boldsymbol{\gamma}_{\mathbf{x}_2}^T & \cdots & \boldsymbol{\gamma}_{\mathbf{x}_N}^T \end{bmatrix}^T \tag{9}$$

is the normalized [with respect to $x_1(k)$] cross-correlation vector (of length NL) between $x_1(k)$ and $\mathbf{x}(k)$,

$$\boldsymbol{\gamma}_{\mathbf{x}_n} = \begin{bmatrix} \gamma_{\mathbf{x}_n,0} & \gamma_{\mathbf{x}_n,1} & \gamma_{\mathbf{x}_n,L-1} \end{bmatrix}^T = \frac{E[x_1(k)\mathbf{x}_n(k)]}{E\left[x_1^2(k)\right]}, \; n = 1, 2, \ldots, N \tag{10}$$

is the normalized cross-correlation vector (of length L) between $x_1(k)$ and $\mathbf{x}_n(k)$,

$$\gamma_{\mathbf{x}_n,l} = \frac{E[x_1(k)x_n(k-l)]}{E\left[x_1^2(k)\right]}, \; l = 0, 1, \ldots, L-1 \tag{11}$$

is the normalized cross-correlation coefficient between $x_1(k)$ and $x_n(k-l)$, and

$$\mathbf{x}'(k) = \mathbf{x}(k) - x_1(k)\boldsymbol{\gamma}_\mathbf{x}, \tag{12}$$

$$E[x_1(k)\mathbf{x}'(k)] = \mathbf{0}. \tag{13}$$

Substituting (8) into (7), we get

$$\hat{x}_1(k) = \mathbf{h}^T[x_1(k)\boldsymbol{\gamma}_\mathbf{x} + \mathbf{x}'(k) + \mathbf{v}(k)], = x_\mathrm{fd}(k) + x'_\mathrm{ri}(k) + v_\mathrm{rn}(k), \tag{14}$$

where $x_{\text{fd}}(k) = x_1(k)\mathbf{h}^T\gamma_{\mathbf{x}}$ is the filtered desired signal and $x'_{\text{ri}}(k) = \mathbf{h}^T\mathbf{x}'(k)$ is the residual interference. We observe that the estimate of the desired signal at time k is the sum of three terms: the first one is clearly the filtered desired signal while the two others are the filtered undesired signals (interference-plus-noise). Since the three terms are mutually uncorrelated, the variance of $\hat{x}_1(k)$ is

$$\sigma^2_{\hat{x}_1} = \sigma^2_{x_{\text{fd}}} + \sigma^2_{x'_{\text{ri}}} + \sigma^2_{v_{\text{rn}}}, \tag{15}$$

where

$$\begin{aligned}\sigma^2_{x_{\text{fd}}} &= \sigma^2_{x_1}(\mathbf{h}^T\gamma_{\mathbf{x}})^2 \\ &= \mathbf{h}^T\mathbf{R}_{\mathbf{x}_\text{d}}\mathbf{h},\end{aligned} \tag{16}$$

$$\begin{aligned}\sigma^2_{x'_{\text{ri}}} &= \mathbf{h}^T\mathbf{R}_{\mathbf{x}'}\mathbf{h} \\ &= \mathbf{h}^T\mathbf{R}_{\mathbf{x}}\mathbf{h} - \sigma^2_{x_1}(\mathbf{h}^T\gamma_{\mathbf{x}})^2,\end{aligned} \tag{17}$$

$$\sigma^2_{v_{\text{rn}}} = \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h}, \tag{18}$$

$\sigma^2_{x_1} = E[x_1^2(k)]$ is the variance of the desired signal, $\mathbf{R}_{\mathbf{x}_\text{d}} = \sigma^2_{x_1}\gamma_{\mathbf{x}}\gamma_{\mathbf{x}}^T$ is the correlation matrix (whose rank is equal to 1) of $\mathbf{x}_\text{d}(k)$, and $\mathbf{R}_{\mathbf{x}'} = E[\mathbf{x}'(k)\mathbf{x}'^T(k)]$ is the correlation matrix of $\mathbf{x}'(k)$. Comparing the decomposition of the filtered microphone observations in (14), and recalling its narrowband counterpart in the frequency domain [6,8] we clearly see that $\gamma_{\mathbf{x}}$ plays the role of the steering vector of the desired source. Thus $\gamma_{\mathbf{x}}$ can be viewed as the *time-domain steering vector* of the desired signal.

4. Performance measures

In this section, we define some useful performance measures that will allow us to study the different multichannel noise reduction algorithms in the time domain developed later in this paper. Since the signal we want to recover is the clean (but convolved) signal received at microphone 1, i.e., $x_1(k)$, this microphone will be serving as the reference sensor. In other words, we define our performance measures by taking the first microphone signal as a reference as we clarify in the following. These definitions slightly differ from traditional (old) definitions that can be found in previous Refs. [7,14] in the sense that the decomposition of the noise-free speech observations described above is taken into account.

The first important measure is the input SNR defined as:

$$\text{iSNR} = \frac{\sigma^2_{x_1}}{\sigma^2_{v_1}}, \tag{19}$$

where $\sigma^2_{v_1} = E[v_1^2(k)]$ is the variance of the noise at microphone 1.

To quantify the level of noise remaining at the output of the filter, we define the output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual interference-plus-noise,[1] i.e.,

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma^2_{x_{\text{fd}}}}{\sigma^2_{x'_{\text{ri}}} + \sigma^2_{v_{\text{rn}}}} = \frac{\sigma^2_{x_1}(\mathbf{h}^T\gamma_{\mathbf{x}})^2}{\mathbf{h}^T\mathbf{R}_{\text{in}}\mathbf{h}}, \tag{20}$$

where

$$\mathbf{R}_{\text{in}} = \mathbf{R}_{\mathbf{x}'} + \mathbf{R}_{\mathbf{v}} \tag{21}$$

is the interference-plus-noise correlation matrix. The objective of the noise reduction filter is to make the output SNR greater than the input SNR.

For the particular filter $\mathbf{h} = \mathbf{i}_1$, where $\mathbf{i}_1$ is the first column of the identity matrix $\mathbf{I}$ (of size $NL \times NL$), we have

$$\text{oSNR}(\mathbf{i}_1) = \text{iSNR}. \tag{22}$$

With the FIR filter $\mathbf{i}_1$, the SNR is not improved.

For any two vectors $\mathbf{h}$ and $\gamma_{\mathbf{x}}$ and a positive definite matrix $\mathbf{R}_{\text{in}}$, we have

$$(\mathbf{h}^T\gamma_{\mathbf{x}})^2 \leqslant (\mathbf{h}^T\mathbf{R}_{\text{in}}\mathbf{h})\left(\gamma_{\mathbf{x}}^T\mathbf{R}_{\text{in}}^{-1}\gamma_{\mathbf{x}}\right). \tag{23}$$

Applying the previous inequality to (20), we deduce an upper bound for the output SNR:

$$\text{oSNR}(\mathbf{h}) \leqslant \sigma^2_{x_1}\gamma_{\mathbf{x}}^T\mathbf{R}_{\text{in}}^{-1}\gamma_{\mathbf{x}}, \quad \forall\mathbf{h}. \tag{24}$$

We define the array gain as the ratio of the output SNR (after beamforming) over the input SNR (at the reference microphone) [15,16], i.e.,

$$\mathcal{A}(\mathbf{h}) = \frac{\text{oSNR}(\mathbf{h})}{\text{iSNR}}. \tag{25}$$

From (24), we deduce that the maximum array gain is

$$\mathcal{A}_{\max} = \sigma^2_{v_1}\gamma_{\mathbf{x}}^T\mathbf{R}_{\text{in}}^{-1}\gamma_{\mathbf{x}}. \tag{26}$$

The noise reduction factor [14,17] quantifies the amount of noise rejected by the filter. This quantity is defined as the ratio of the variance of the noise at the reference microphone over the variance of the interference-plus-noise remaining after the beamforming, i.e.,

$$\xi_{\text{nr}}(\mathbf{h}) = \frac{\sigma^2_{v_1}}{\sigma^2_{x'_{\text{ri}}} + \sigma^2_{v_{\text{rn}}}} = \frac{\sigma^2_{v_1}}{\mathbf{h}^T\mathbf{R}_{\text{in}}\mathbf{h}}. \tag{27}$$

The noise reduction factor is expected to be lower bounded by 1 for optimal filters.

In practice, the FIR filter, $\mathbf{h}$, may distort the desired signal. In order to evaluate the level of this distortion, we define the speech reduction factor [2] as the variance of the desired signal over the variance of the filtered desired signal at the output of the beamformer, i.e.,

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma^2_{x_1}}{\sigma^2_{x_{\text{fd}}}} = \frac{1}{(\mathbf{h}^T\gamma_{\mathbf{x}})^2}. \tag{28}$$

An important observation is that the design of a filter that does not distort the desired signal requires the constraint

$$\mathbf{h}^T\gamma_{\mathbf{x}} = 1. \tag{29}$$

Thus, the speech reduction factor is equal to 1 if there is no distortion and expected to be greater than 1 when distortion occurs.

By making the appropriate substitutions, one can derive the relationship among the previous measures:

$$\mathcal{A}(\mathbf{h}) = \frac{\text{oSNR}(\mathbf{h})}{\text{iSNR}} = \frac{\xi_{\text{nr}}(\mathbf{h})}{\xi_{\text{sr}}(\mathbf{h})}. \tag{30}$$

When no distortion occurs, the array gain coincides with the noise reduction factor.

Another useful performance measure is the speech distortion index [14] defined as

$$\upsilon_{\text{sd}}(\mathbf{h}) = \frac{E\{[x_{\text{fd}}(k) - x_1(k)]^2\}}{\sigma^2_{x_1}} = (\mathbf{h}^T\gamma_{\mathbf{x}} - 1)^2. \tag{31}$$

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

5. Optimal noise reduction filters

In this section, we revisit the most popular multichannel noise reduction filters. Using the new decomposition in Section 3, we provide new simplified expressions for the maximum SNR, Wiener,

[1] In this paper, we consider the interference as part of the noise in the definitions of the performance measures.

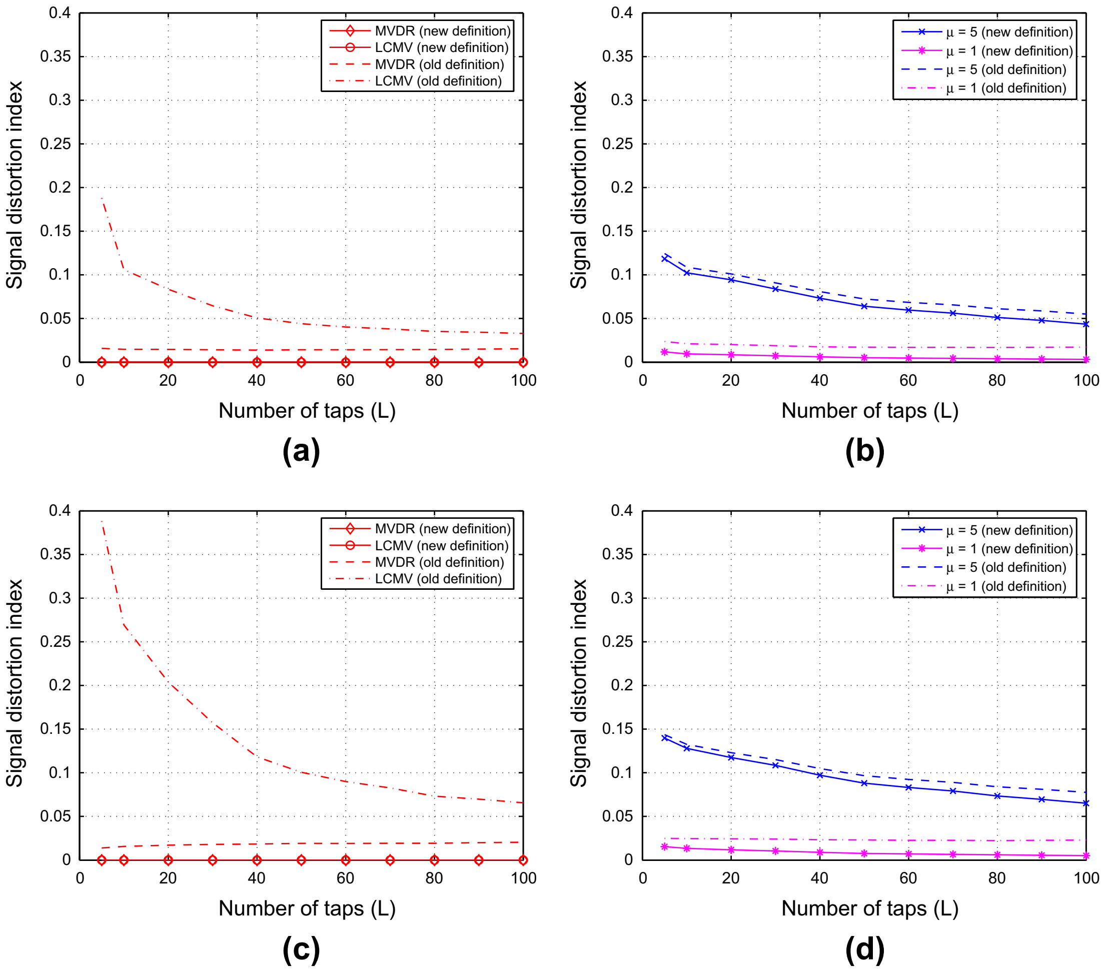

Fig. 1. Effect of the number of taps, L, on the signal distortion index. (a) MVDR and LCMV, T_{60} = 240 ms. (b) Tradeoff filter with $\mu = 5$ and $\mu = 1$, T_{60} = 240 ms. (c) MVDR and LCMV, T_{60} = 580 ms. (d) Tradeoff filter with $\mu = 5$ and $\mu = 1$, T_{60} = 580 ms. Old and new definitions of the signal distortion index are compared. The input speech-to-fan-noise and speech-to-babble-noise ratios are equal to 10 dB.

MVDR, ST prediction, tradeoff, and LCMV filters. A better understanding of the functioning of these filters is then gained thanks to these expressions.

5.1. Maximum SNR filter

The maximum SNR filter, $\mathbf{h}_{\max}$, is obtained by maximizing the output SNR as defined in (20). Therefore, $\mathbf{h}_{\max}$ is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}$. Let us denote this eigenvalue by $\lambda_{\max}$. Since the rank of the matrix $\mathbf{R}_{\mathbf{x}_{\text{d}}}$ is equal to 1, we have

$$\lambda_{\max} = \text{tr}\left(\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}\right) = \sigma_{x_1}^2 \gamma_{\mathbf{x}}^T \mathbf{R}_{\text{in}}^{-1} \gamma_{\mathbf{x}}, \tag{32}$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix. As a result,

$$\text{oSNR}(\mathbf{h}_{\max}) = \sigma_{x_1}^2 \gamma_{\mathbf{x}}^T \mathbf{R}_{\text{in}}^{-1} \gamma_{\mathbf{x}}, \tag{33}$$

which corresponds to the maximum possible SNR according to the inequality in (24). Obviously, we also have

$$\mathbf{h}_{\max} = \alpha \mathbf{R}_{\text{in}}^{-1} \gamma_{\mathbf{x}}, \tag{34}$$

where α is an arbitrary scaling factor different from zero. While this factor has no effect on the output SNR, it may have on the speech distortion. In fact, all filters (except for the LCMV) derived in the rest of this paper are equivalent up to a scaling factor. These filters also try to find the respective scaling factors depending on what we optimize.

5.2. Mean-square error (MSE) criterion

We define the error signal between the estimated and desired signals as

$$e(k) = \hat{x}_1(k) - x_1(k) = \mathbf{h}^T \mathbf{y}(k) - x_1(k), \tag{35}$$

which can be written as the sum of two error signals:

$$e(k) = e_{\text{d}}(k) + e_{\text{r}}(k), \tag{36}$$

where

$$e_{\text{d}}(k) = x_{\text{fd}}(k) - x_1(k) \tag{37}$$

is the signal distortion due to the FIR filter and

$$e_{\text{r}}(k) = x'_{\text{ri}}(k) + v_{\text{rn}}(k) \tag{38}$$

represents the residual interference-plus-noise.

The mean-square error (MSE) is then

$$J(\mathbf{h}) = E[e^2(k)] = J_{\text{d}}(\mathbf{h}) + J_{\text{r}}(\mathbf{h}), \tag{39}$$

where

$$J_{\text{d}}(\mathbf{h}) = E\left[e_{\text{d}}^2(k)\right] = \sigma_{x_1}^2 \left(\mathbf{h}^T \gamma_{\mathbf{x}} - 1\right)^2 \tag{40}$$

and

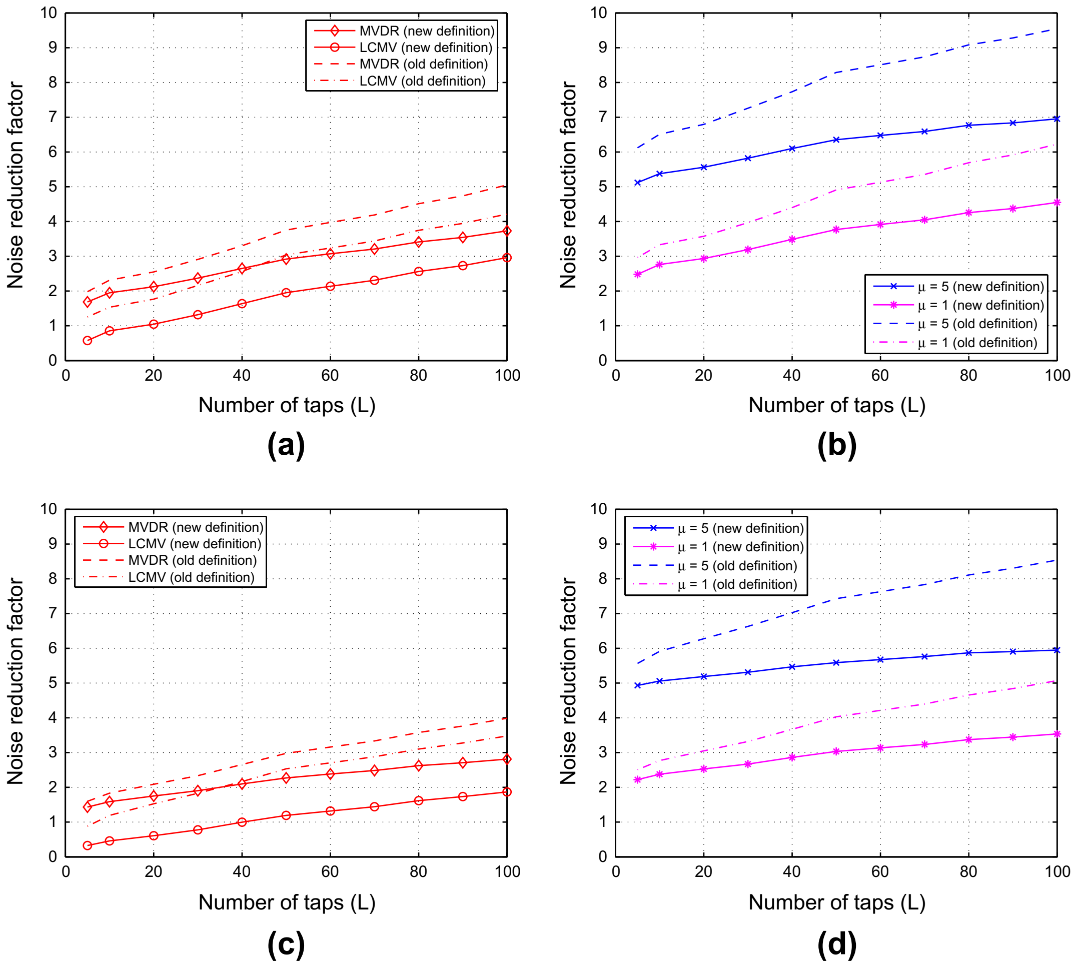

Fig. 2. Effect of the number of taps, L, on the noise reduction factor. (a) MVDR and LCMV, T_{60} = 240 ms. (b) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 240 ms. (c) MVDR and LCMV, T_{60} = 580 ms. (d) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 580 ms. Old and new definitions of the noise reduction factor are compared. The input speech-to-fan-noise and speech-to-babble-noise ratios are equal to 10 dB.

$$J_{\mathrm{r}}(\mathbf{h}) = E\left[e_{\mathrm{r}}^2(k)\right] = \sigma^2_{x'_{\mathrm{ri}}} + \sigma^2_{v_{\mathrm{rn}}}. \tag{41}$$

For the particular filter **h** = $\mathbf{i}_1$, the MSE is

$$J(\mathbf{i}_1) = \sigma^2_{v_1}, \tag{42}$$

so there is neither noise reduction nor speech distortion. We can now define the normalized MSE (NMSE) as

$$\widetilde{J}(\mathbf{h}) = \frac{J(\mathbf{h})}{J(\mathbf{i}_1)} = \mathrm{iSNR}\cdot\upsilon_{\mathrm{sd}}(\mathbf{h}) + \frac{1}{\xi_{\mathrm{nr}}(\mathbf{h})}, \tag{43}$$

where

$$\upsilon_{\mathrm{sd}}(\mathbf{h}) = \frac{J_{\mathrm{d}}(\mathbf{h})}{\sigma^2_{x_1}}, \tag{44}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}) = \frac{\sigma^2_{v_1}}{J_{\mathrm{r}}(\mathbf{h})}. \tag{45}$$

This shows how the MSEs are related to some of the performance measures.

It is clear that the objective of noise reduction with the linear array model is to find optimal FIR filters that would either minimize $J(\mathbf{h})$ or minimize $J_{\mathrm{r}}(\mathbf{h})$ or $J_{\mathrm{d}}(\mathbf{h})$ subject to some constraint.

5.3. Wiener filter

The Wiener filter is easily derived by taking the gradient of the MSE, $J(\mathbf{h})$, with respect to **h** and equating the result to zero:

$$\mathbf{h}_{\mathrm{W}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\mathbf{i}_1 \tag{46}$$

We can use the Woodbury's identity to invert $\mathbf{R}_{\mathbf{y}}$ with the fact that $\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\mathbf{i}_1 = \sigma^2_{x_1}\gamma_{\mathbf{x}}$ to rewrite (46) as

$$\mathbf{h}_{\mathrm{W}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}}{1+\lambda_{\max}}\mathbf{i}_1. \tag{47}$$

Hence, we deduce that the output SNR is

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}) = \lambda_{\max} \tag{48}$$

and the speech distortion index is a clear function of the output SNR:

$$\upsilon_{\mathrm{sd}}(\mathbf{h}_{\mathrm{W}}) = \frac{1}{(1+\lambda_{\max})^2} \leqslant 1. \tag{49}$$

The higher is the value of $\lambda_{\max}$ (and/or the number of microphones), the less the desired signal is distorted. Clearly, we also have:

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}) \geqslant \mathrm{iSNR}, \tag{50}$$

since the Wiener filter maximizes the output SNR.

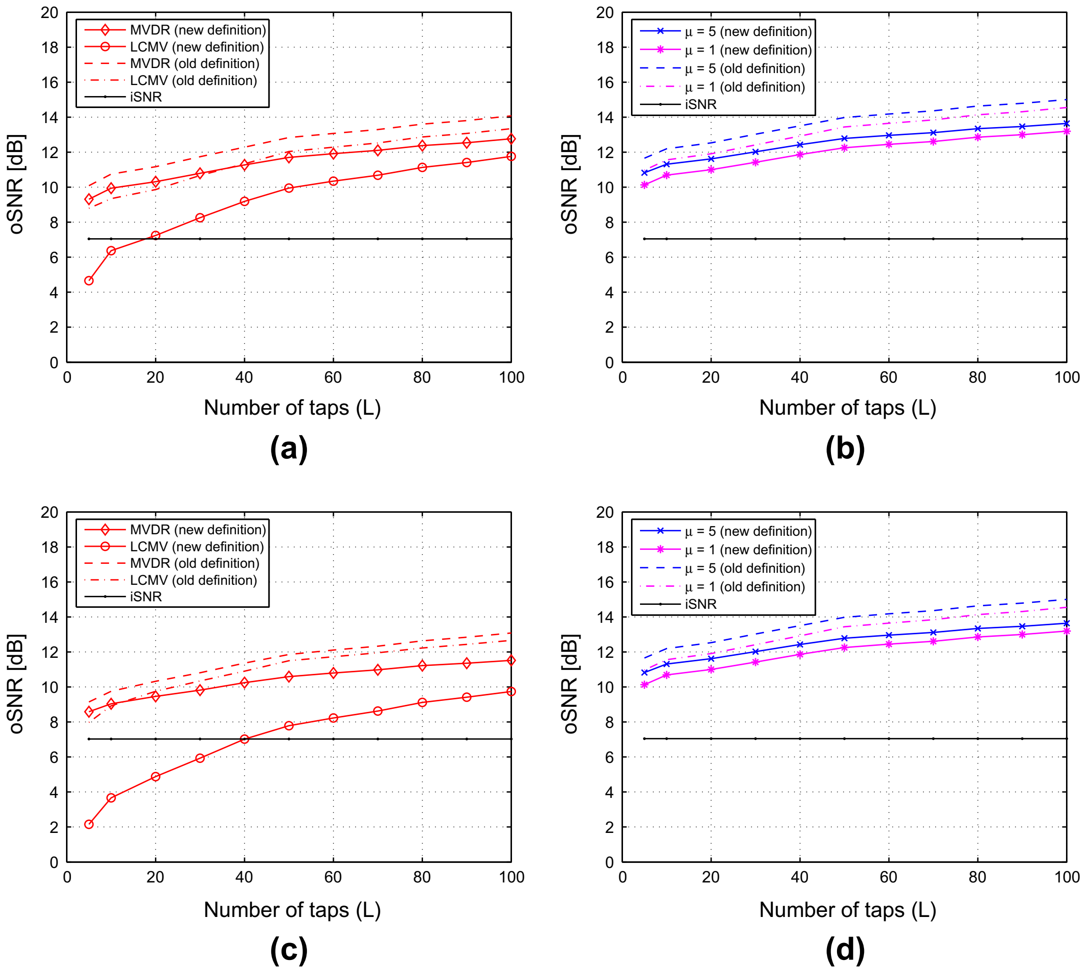

Fig. 3. Effect of the number of taps, L, on the output SNR. (a) MVDR and LCMV, T_{60} = 240 ms. (b) Tradeoff filter with $\mu = 5$ and $\mu = 1$, T_{60} = 240 ms. (c) MVDR and LCMV, T_{60} = 580 ms. (d) Tradeoff filter with $\mu = 5$ and $\mu = 1$, T_{60} = 580 ms. Old and new definitions of the noise reduction factor are compared. The input speech-to-fan-noise and speech-to-babble-noise ratios are equal to 10 dB.

It is of great interest to observe that the two filters $\mathbf{h}_{\max}$ and $\mathbf{h}_{\mathrm{W}}$ are equivalent up to a scaling factor. Indeed, taking

$$\alpha = \frac{\sigma_{x_1}^2}{1+\lambda_{\max}} \tag{51}$$

in (34) (maximum SNR filter), we find (47) (Wiener filter). Finally, with the Wiener filter the noise reduction factor is

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{W}}) = \frac{(1+\lambda_{\max})^2}{\mathrm{iSNR}\cdot\lambda_{\max}} \geqslant \left(1+\frac{1}{\lambda_{\max}}\right)^2. \tag{52}$$

Using (49) and (52) in (43), we find the minimum NMSE (MNMSE):

$$\widetilde{J}(\mathbf{h}_{\mathrm{W}}) = \frac{\mathrm{iSNR}}{1+\mathrm{oSNR}(\mathbf{h}_{\mathrm{W}})} \leqslant 1. \tag{53}$$

5.4. MVDR filter

Another important filter, initially proposed by Capon [18] and delineated in several forms [8,19], is the MVDR beamformer which is obtained by minimizing the variance of the interference-plus-noise at the beamformer output with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\mathrm{in}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^T\boldsymbol{\gamma}_{\mathbf{x}} = 1, \tag{54}$$

for which the solution is

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_{\mathbf{x}}}{\boldsymbol{\gamma}_{\mathbf{x}}^T\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_{\mathbf{x}}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}}{\lambda_{\max}}\mathbf{i}_1. \tag{55}$$

Obviously, we can rewrite the MVDR as

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_{\mathbf{x}}}{\boldsymbol{\gamma}_{\mathbf{x}}^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_{\mathbf{x}}}. \tag{56}$$

Taking

$$\alpha = \frac{\sigma_{x_1}^2}{\lambda_{\max}} \tag{57}$$

in (34) (maximum SNR filter), we find (56) (MVDR filter), showing that the maximum SNR and MVDR filters are equivalent up to a scaling factor. The Wiener and MVDR filters are also simply related as follows:

$$\mathbf{h}_{\mathrm{W}} = \alpha_0\mathbf{h}_{\mathrm{MVDR}}, \tag{58}$$

where

$$\alpha_0 = \mathbf{h}_{\mathrm{W}}^T\boldsymbol{\gamma}_{\mathbf{x}} = \frac{\lambda_x}{1+\lambda_{\max}}. \tag{59}$$

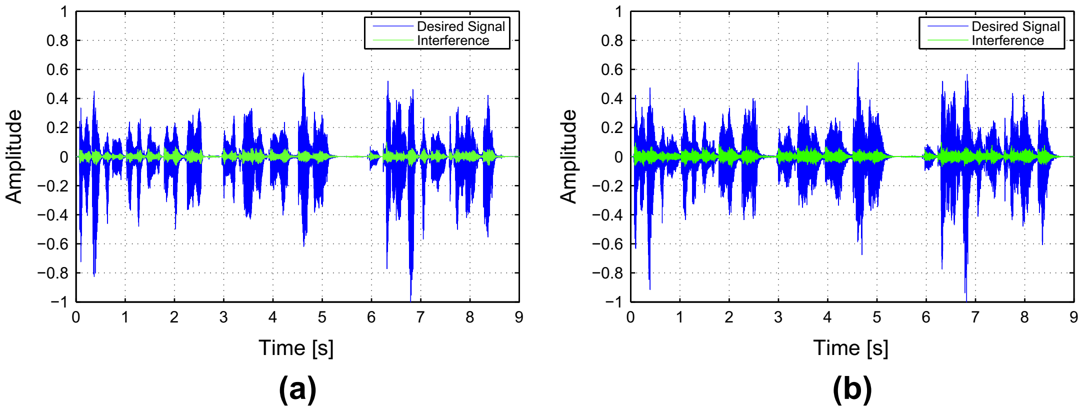

Fig. 4. Desired signal and interference at the output of the multichannel Wiener filter. (a) T_{60} = 240 ms. (b) T_{60} = 580 ms. Four microphones are used. The number of filter taps is L = 50. The input speech-to-fan-noise and speech-to-babble-noise ratios are equal to 10 dB.

Here again the two filters $\mathbf{h}_{\mathrm{W}}$ and $\mathbf{h}_{\mathrm{MVDR}}$ are equivalent up to a scaling factor. From a theoretical point of view, this scaling is not significant but from a practical point of view it can be important. Indeed, the signals are usually non-stationary and the estimation is done frame by frame, so it is essential to have this scaling factor right from one frame to the other in order to avoid large distortions. Therefore, it is recommended to use the MVDR filter rather than the Wiener or maximum SNR filters in speech enhancement applications.

It is clear that we always have

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{MVDR}}) = \mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}), \tag{60}$$

$$\upsilon_{\mathrm{sd}}(\mathbf{h}_{\mathrm{MVDR}}) = 0, \tag{61}$$

$$\xi_{\mathrm{sr}}(\mathbf{h}_{\mathrm{MVDR}}) = 1, \tag{62}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{MVDR}}) = \mathcal{A}(\mathbf{h}_{\mathrm{MVDR}}) \leqslant \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{W}}), \tag{63}$$

and

$$1 \geqslant \tilde{J}(\mathbf{h}_{\mathrm{MVDR}}) = \frac{1}{\mathcal{A}(\mathbf{h}_{\mathrm{MVDR}})} \geqslant \tilde{J}(\mathbf{h}_{\mathrm{W}}). \tag{64}$$

5.5. Link between the MVDR filter and the space–time prediction approach

In the ST prediction approach, we find a distortionless filter in two steps [2,4].

First, we rewrite the error signal as

$$e(k) = \mathbf{h}^T\mathbf{x}(k) - x_1(k) + \mathbf{h}^T\mathbf{v}(k) = e_{\mathrm{d}}(k) + e_{\mathbf{v}}(k), \tag{65}$$

where

$$e_{\mathrm{d}}(k) = \mathbf{h}^T\mathbf{x}(k) - x_1(k), \tag{66}$$

$$e_{\mathbf{v}}(k) = \mathbf{h}^T\mathbf{v}(k). \tag{67}$$

Assume now that we can find an ST filter $\mathbf{g}$ of length NL in such a way that

$$\mathbf{x}(k) = x_1(k)\mathbf{g}. \tag{68}$$

This filter extracts from $\mathbf{x}(k)$ the correlated components to $x_1(k)$. Replacing (68) in (66), we obtain

$$e_{\mathrm{d}}(k) = (\mathbf{h}^T\mathbf{g} - 1)x_1(k). \tag{69}$$

The distortionless filter with the ST approach is then obtained by

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^T\mathbf{g} = 1. \tag{70}$$

We deduce the solution:

$$\mathbf{h}_{\mathrm{ST}} = \frac{\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{g}}{\mathbf{g}^T\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{g}}. \tag{71}$$

The second step consists of finding the optimal $\mathbf{g}$ in the Wiener sense. For that, we need to define the error signal vector

$$\mathbf{e}_{\mathrm{ST}}(k) = \mathbf{x}(k) - x_1(k)\mathbf{g} \tag{72}$$

and form the MSE

$$J(\mathbf{g}) = E\left[\mathbf{e}_{\mathrm{ST}}^T(k)\mathbf{e}_{\mathrm{ST}}(k)\right]. \tag{73}$$

Minimizing $J(\mathbf{g})$ with respect to $\mathbf{g}$, we easily find the optimal ST filter:

$$\mathbf{g}_{\mathrm{o}} = \gamma_{\mathbf{x}}. \tag{74}$$

It is interesting to observe that the error signal vector with the optimal ST filter corresponds to the interference signal, i.e.,

$$\mathbf{e}_{\mathrm{ST,o}}(k) = \mathbf{x}(k) - x_1(k)\mathbf{g}_{\mathrm{o}} = \mathbf{x}'(k). \tag{75}$$

This result is obviously expected because of the orthogonality principle.

Substituting (74) into (71), we finally find that

$$\mathbf{h}_{\mathrm{ST}} = \frac{\mathbf{R}_{\mathbf{v}}^{-1}\gamma_{\mathbf{x}}}{\gamma_{\mathbf{x}}^T\mathbf{R}_{\mathbf{v}}^{-1}\gamma_{\mathbf{x}}}. \tag{76}$$

Comparing $\mathbf{h}_{\mathrm{MVDR}}$ with $\mathbf{h}_{\mathrm{ST}}$, we see that the latter is an approximation of the former. Indeed, in the ST approach the interference signal is neglected: instead of using the correlation matrix of the interference-plus-noise, only the correlation matrix of the noise is used. Nevertheless, this difference between both beamformers is due to the definition of our optimization problem in (70). Identical expressions of the MVDR and ST-prediction filter would have been obtained if we considered minimizing the overall mixture energy subject to the no distortion constraint.

5.6. Tradeoff filter

In the tradeoff approach, we try to compromise between noise reduction and speech distortion. Instead of minimizing the MSE as we already did in finding the Wiener filter, we could minimize the speech distortion index with the constraint that the noise reduction factor is equal to a positive value that is greater than 1. Mathematically, this is equivalent to

$$\begin{aligned} &\min_{\mathbf{h}} J_{\mathrm{d}}(\mathbf{h}) \\ &\text{s.t.} \quad J_{\mathrm{r}}(\mathbf{h}) = \beta\sigma_{v_1}^2, \end{aligned} \tag{77}$$

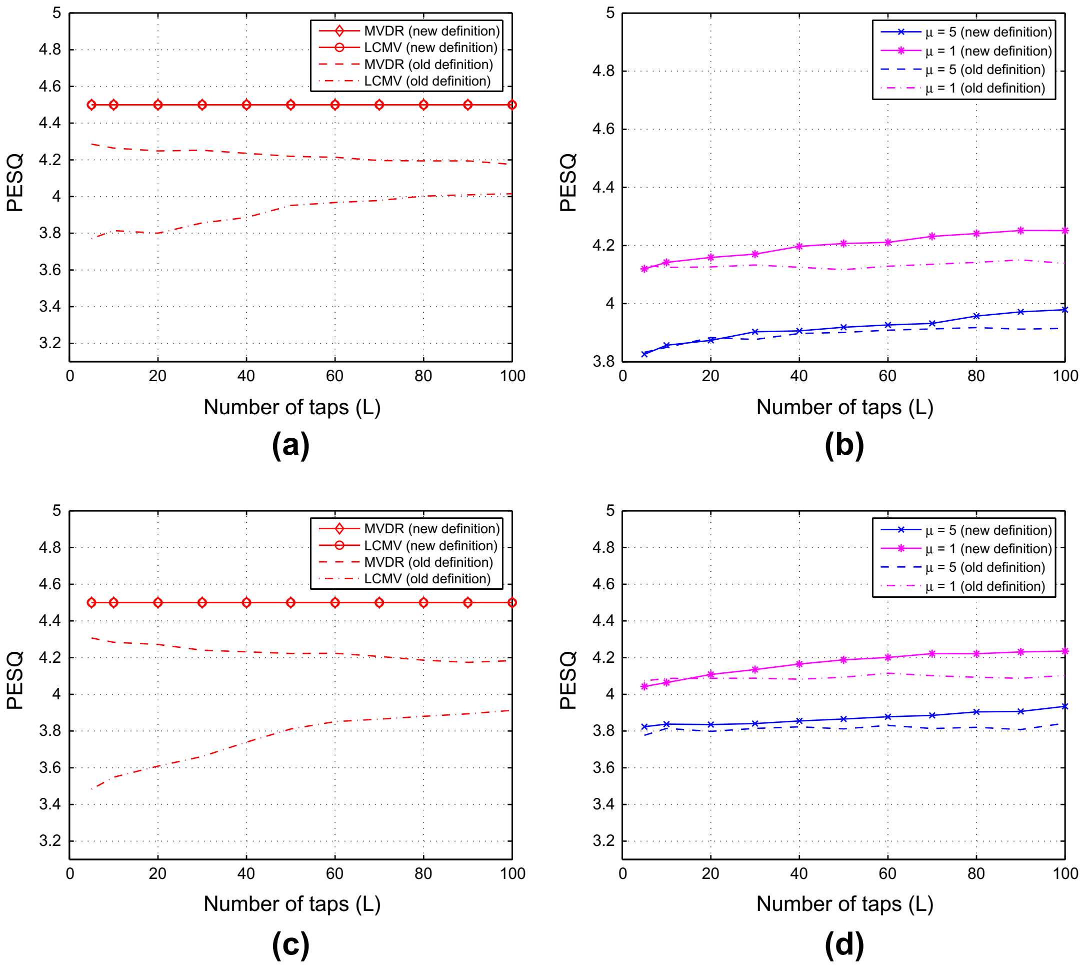

Fig. 5. Effect of the number of taps on the PESQ measure between $x_1(k)$ and the filtered signal. (a) MVDR and LCMV, T_{60} = 240 ms. (b) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 240 ms. (c) MVDR and LCMV, T_{60} = 580 ms. (d) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 580 ms. Old and new definitions of the PESQ measure are compared. The input speech-to-fan-noise and speech-to-babble-noise ratios are equal to 10 dB.

where 0 < β < 1 to insure that we get some noise reduction. By using a Lagrange multiplier, $\mu \geqslant 0$, to adjoin the constraint to the cost function, we easily deduce the tradeoff filter:

$$\mathbf{h}_{\mathrm{T},\mu} = \sigma_{x_1}^2 \left[\sigma_{x_1}^2 \gamma_{\mathbf{x}} \gamma_{\mathbf{x}}^T + \mu \mathbf{R}_{\mathrm{in}}\right]^{-1} \gamma_{\mathbf{x}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1} \gamma_{\mathbf{x}}}{\mu \sigma_{x_1}^{-2} + \gamma_{\mathbf{x}}^T \mathbf{R}_{\mathrm{in}}^{-1} \gamma_{\mathbf{x}}}, \quad (78)$$

where the Lagrange multiplier, μ, satisfies $J_{\mathrm{r}}(\mathbf{h}) = \beta \sigma_{v_1}^2$. By substituting (78) into the constraint in (77), we determine the relationship between the tuning parameter μ and the resulting noise reduction

$$\beta = \frac{\mathrm{iSNR}\ \lambda_{\max}}{(\mu + \lambda_{\max})^2}. \quad (79)$$

In particular, for

- μ = 1, $\mathbf{h}_{\mathrm{T},1} = \mathbf{h}_{\mathrm{W}}$, which is the Wiener filter;
- μ = 0, $\mathbf{h}_{\mathrm{T},0} = \mathbf{h}_{\mathrm{MVDR}}$, which is the MVDR filter;
- μ > 1, results in a filter with low residual noise at the expense of high speech distortion;
- μ < 1, results in a filter with high residual noise and low speech distortion.

Again, we observe here as well that the tradeoff, Wiener, and maximum SNR filters are equivalent up to a scaling factor. As a result, the output SNR of the tradeoff filter is independent of μ and is identical to the output SNR of the Wiener filter, i.e.,

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{T},\mu}) = \mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}), \quad \forall \mu. \quad (80)$$

5.7. LCMV filter

We can derive an LCMV filter [20,21], which can handle more than one linear constraint, by exploiting the structure of the noise signal. In Section 3, we decomposed the vector $\mathbf{x}(k)$ into two orthogonal components to extract the desired signal, $x_1(k)$. We can also decompose (but for a different objective as explained below) the noise signal vector, $\mathbf{v}(k)$, into two orthogonal terms:

$$\mathbf{v}(k) = v_1(k)\gamma_{\mathbf{v}} + \mathbf{v}'(k), \quad (81)$$

where $\gamma_{\mathbf{v}}$ and $\mathbf{v}'(k)$ are defined in a similar way to $\gamma_{\mathbf{x}}$ and $\mathbf{x}'(k)$. Now, our problem is the following. We wish to perfectly recover our desired signal, $x_1(k)$, and completely remove the correlated noise components, $v_1(k)\gamma_{\mathbf{v}}$. Thus, the two constraints can be put together in a matrix form as

$$\mathbf{C}^T \mathbf{h} = \mathbf{i}, \quad (82)$$

where

$$\mathbf{C} = [\gamma_{\mathbf{x}} \quad \gamma_{\mathbf{v}}] \quad (83)$$

is our constraint matrix of size $NL \times 2$ and

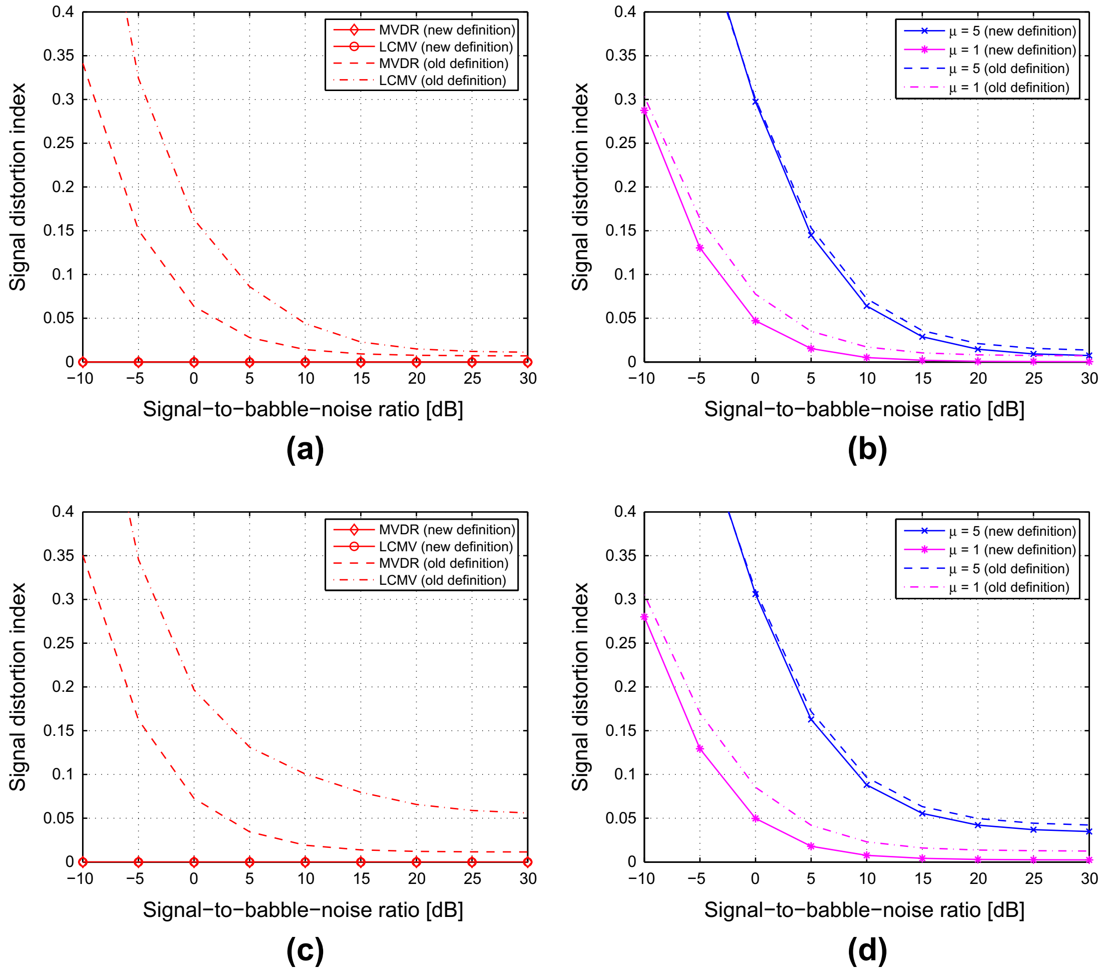

Fig. 6. Effect of the input speech-to-babble noise ratio on the signal distortion index. (a) MVDR and LCMV, T_{60} = 240 ms. (b) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 240 ms. (c) MVDR and LCMV, T_{60} = 580 ms. (d) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 580 ms. Old and new definitions of the signal distortion index are compared. Number of filter taps, L = 50. The input speech-to-fan-noise ratio is 10 dB.

$$\mathbf{i} = [1 \quad 0]^T.$$

Then, our optimal filter is obtained by minimizing the energy at the filter output, with the constraints that the correlated noise components are canceled and the desired speech is preserved, i.e.,

$$\mathbf{h}_{\text{LCMV}} = \arg\min_{\mathbf{h}} \mathbf{h}^T \mathbf{R}_{\mathbf{y}} \mathbf{h} \quad \text{s.t.} \quad \mathbf{C}^T \mathbf{h} = \mathbf{i}. \tag{84}$$

The solution to (84) is given by

$$\mathbf{h}_{\text{LCMV}} = \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{C} \left[\mathbf{C}^T \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{C} \right]^{-1} \mathbf{i}. \tag{85}$$

In a similar way to [13], we demonstrate that the MVDR filter can be written as a linear combination of two beamformers

$$\mathbf{h}_{\text{MVDR}} = \varrho \mathbf{h}_{\text{LCMV}} + (1 - \varrho) \mathbf{h}_{\text{MATCH}} \tag{86}$$

where

$$\mathbf{h}_{\text{MATCH}} = \frac{\mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{x}}}{\boldsymbol{\gamma}_{\mathbf{x}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{x}}}, \tag{87}$$

$$\varrho = \frac{\boldsymbol{\gamma}_{\mathbf{v}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{v}} (1 - \kappa)}{\sigma_{v_1}^{-2} + \boldsymbol{\gamma}_{\mathbf{v}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{v}} (1 - \kappa)}, \tag{88}$$

$$\kappa = \frac{\left(\boldsymbol{\gamma}_{\mathbf{x}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{v}} \right)^2}{\left(\boldsymbol{\gamma}_{\mathbf{v}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{v}} \right) \left(\boldsymbol{\gamma}_{\mathbf{x}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{x}} \right)}, \tag{89}$$

where $\mathbf{R}_{\text{in}'} = \mathbf{R}_{\mathbf{v}'} + \mathbf{R}_{\mathbf{x}'}$ is the correlation matrix of all the incoherent interference-plus-noise components, ϱ is a tradeoff parameter between $\mathbf{h}_{\text{MATCH}}$ and $\mathbf{h}_{\text{LCMV}}$ that are optimal in the absence of interference and non-coherent noise, respectively, and κ measures the collinearity[2] between the vectors $\boldsymbol{\gamma}_{\mathbf{x}}$ and $\boldsymbol{\gamma}_{\mathbf{v}}$ in some transform domain defined by the non-coherent noise whitening matrix $\mathbf{R}_{\text{in}'}^{-1/2}$ [13]. We observe from (86) that when ϱ approaches 0, the MVDR tends to the matched filter while when ϱ approaches 1, it tends to the LCMV, thereby trading off the rejection of the coherent noise and the reduction of the other residual noise components [13]. This tradeoff is determined by the value of the collinearity factor, κ, and the generalized coherent-to-other noise components ratio, $\sigma_{v_1}^2 \boldsymbol{\gamma}_{\mathbf{v}}^T \mathbf{R}_{\text{in}'}^{-1} \boldsymbol{\gamma}_{\mathbf{v}}$.

We always have

$$\text{oSNR}(\mathbf{h}_{\text{LCMV}}) \leqslant \text{oSNR}(\mathbf{h}_{\text{MVDR}}), \tag{90}$$

$$\upsilon_{\text{sd}}(\mathbf{h}_{\text{LCMV}}) = 0, \tag{91}$$

$$\xi_{\text{sr}}(\mathbf{h}_{\text{LCMV}}) = 1, \tag{92}$$

and we can show that

$$\xi_{\text{nr}}(\mathbf{h}_{\text{LCMV}}) \leqslant \xi_{\text{nr}}(\mathbf{h}_{\text{MVDR}}) \leqslant \xi_{\text{nr}}(\mathbf{h}_{\text{W}}). \tag{93}$$

The LCMV filter is able to remove all the correlated noise but at the price of a decreased overall noise reduction factor as compared to the MVDR. Numerous numerical results and discussions are provided next to illustrate our findings.

[2] The larger is κ, the more collinear (or less orthogonal) are $\mathbf{R}_{\text{in}'}^{-1/2}\boldsymbol{\gamma}_{\mathbf{x}}$ and $\mathbf{R}_{\text{in}'}^{-1/2}\boldsymbol{\gamma}_{\mathbf{v}}$.

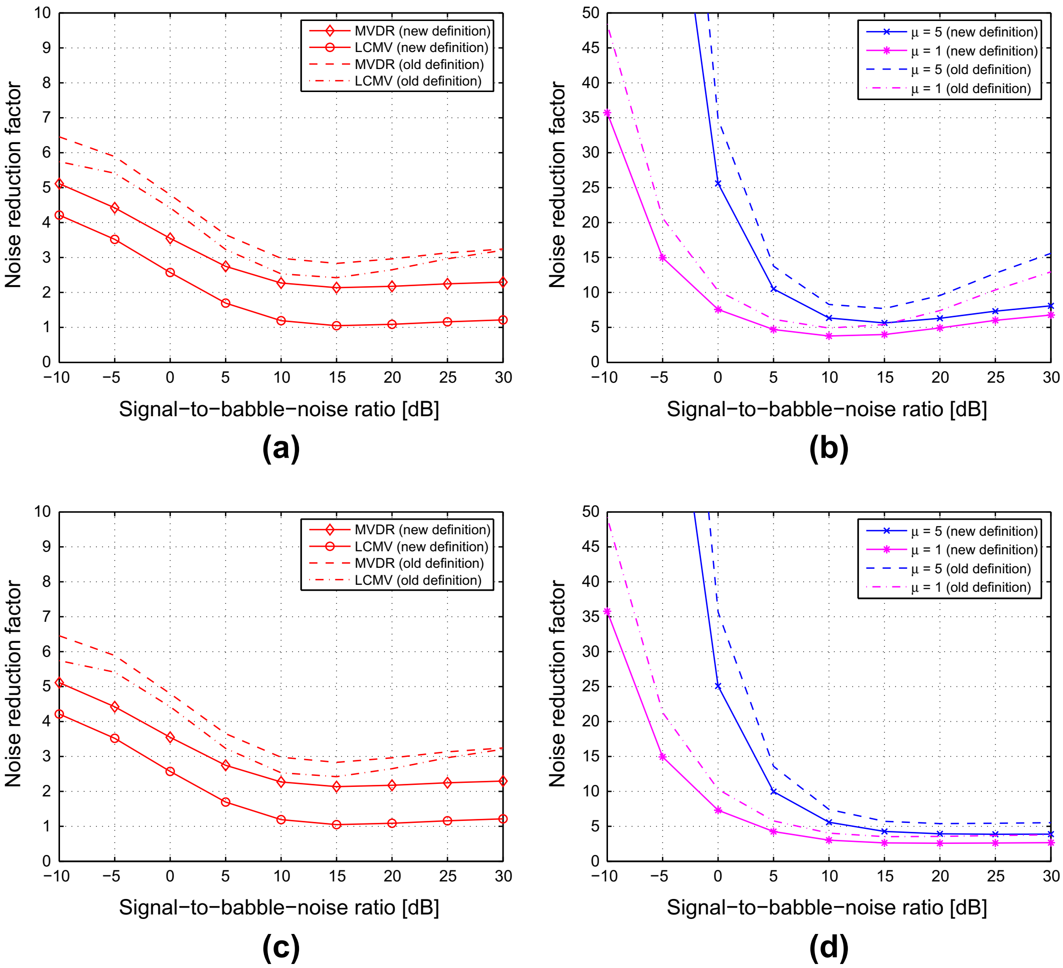

Fig. 7. Effect of the input speech-to-babble noise ratio on the noise reduction factor. (a) MVDR and LCMV, T_{60} = 240 ms. (b) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 240 ms. (c) MVDR and LCMV, T_{60} = 580 ms. (d) Tradeoff filter with μ = 5 and μ = 1, T_{60} = 580 ms. Old and new definitions of the noise reduction factor are compared. The number of filter taps is L = 50. The input speech-to-fan-noise ratio is 10 dB.

6. Experimental results

In our setup, we have several channel impulse responses that were measured at the Bell-Labs varechoic chamber which measures 6700 mm long by 6100 mm wide by 2900 mm high. A linear array of four microphone elements, uniformly located from (2437, 500, 1400) to (2737, 500, 1400), is used. The impulse responses were measured for two source locations: the first one is S_1 = (1337, 1938, 1600) while the second is S_2 = (3337, 1938, 1600). We investigate two reverberation conditions with the reverberation conditions T_{60} = 240 ms and T_{60} = 580 ms, respectively. The target speaker is assumed to be located at S_1 and generating a 12-s-long female speech (8 kHz sampling frequency). A babble noise is added to the noise-free observations. Since we do not have actual multichannel recordings of the babble noise, we take some segments of this signal from the Noisex database [22] and overlap them to each of the noise-free microphone signals. The resulting noise has less spatial coherence than actual recordings and its removal may, consequently, be more challenging. A source located at S_2 and generating a ventilation signal (recorded fan noise from [23]) is also included in the model. We tested several input SNR values as indicated in the figures below.

In order to implement all the filters studied in this paper, accurate estimates of the noise and noisy data correlation matrices $\mathbf{R_y}$ and $\mathbf{R_v}$ are required. The noise-free correlation matrix, $\mathbf{R_x}$, can be retrieved using both matrices and the property of independence between the noise and desired signal. Subsequently, the speech steering vector $\gamma_\mathbf{x}$ is formed using the entries of $\mathbf{R_x}$ as described in Section 3. To implement the LCMV defined in (85), the noise steering vector $\gamma_\mathbf{v}$ needs to be estimated. This is achieved using the definition of $\gamma_\mathbf{v}$ and the estimate of $\mathbf{R_v}$. In contrast to $\mathbf{R_y}$ that can be continuously estimated from the microphone observations, the estimation of $\mathbf{R_v}$ requires a voice activity detector (VAD) in practice. The performance of all investigated filters could vary depending on the performance of the VAD. However, we keep the investigation of the effect of VAD accuracy out of the scope of this paper due to space constraint since we are rather interested in investigating the fundamental limits of the multichannel noise reduction methods described in Section 5. Consequently, we assume the knowledge of the noise samples at every time instant, k, as in some other previous contributions including [4,7]. The statistics estimation and filtering are performed using batch processing with 256 ms-length windows and an overlap rate of 75% between consecutive frames.

For completeness, we investigate the performance of the tradeoff filter for three conditions: μ = 5, 1, and 0. Recall that μ = 0 corresponds to the MVDR and μ = 1 corresponds to the traditional multichannel Wiener filter. We also include comparisons to the LCMV beamformer. All the multichannel filters investigated in this paper are obtained by optimizing second-order-statistics-based criteria. Hence, using second-order-statistics-based metrics to evaluate their performance is the most natural and intuitive way

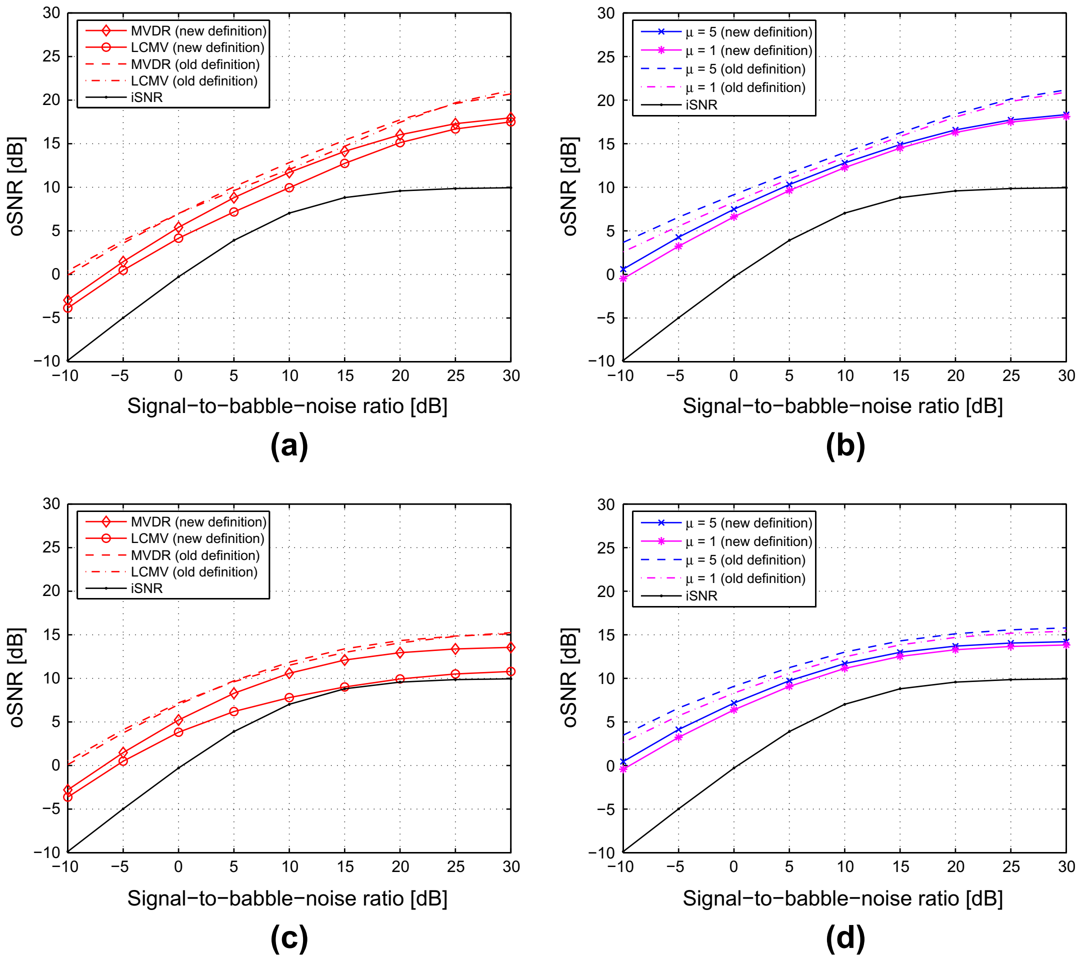

Fig. 8. Effect of the input speech-to-babble noise ratio on the output SNR. (a) MVDR and LCMV, $T_{60} = 240$ ms. (b) Tradeoff filter with $\mu = 5$ and $\mu = 1$, $T_{60} = 240$ ms. (c) MVDR and LCMV, $T_{60} = 580$ ms. (d) Tradeoff filter with $\mu = 5$ and $\mu = 1$, $T_{60} = 580$ ms. Old and new definitions of the output SNR are compared. The number of filter taps is $L = 50$. The input speech-to-fan-noise ratio is 10 dB.

to comprehend their functioning. In particular, the noise reduction factor, signal distortion index and output SNR are used for performance evaluation here. To demonstrate the relevance of the new definitions of our metrics in Section 4, we compare them to the old definitions where the interference (as described in Section 3) is included in the desired signal part (see [14] for details). It is known that the perceptual evaluation of speech quality (PESQ) measure is well correlated to the human perception. Therefore, we include it in our evaluations. Precisely, we measure the PESQ between the reference clean signal $x_1(k)$ and the overall filtered speech signal $x_{\rm f}(k) = \mathbf{h}^T\mathbf{x}(k)$ defined in (7) and we contrast it to the PESQ measure between $x_1(k)$ and the newly defined filtered desired signal $x_{\rm fd}(k) = \mathbf{h}^T\mathbf{x}_{\rm d}(k)$ in (8). The results are first presented for a variable number of filter taps at a constant overall input SNR (the speech-to-babble-noise ratio and the speech-to-fan-noise ratio are both equal to 10 dB). Afterwards, we fix the number of taps and see the effect of the signal-to-babble-noise ratio for a given speech-to-fan-noise ratio.[3]

Figs. 1–3 show the effect of the number of taps on the signal distortion index, noise reduction factor, and SNR at the output of the investigated filter in the two reverberation conditions described above. By observing the newly defined performance measures (denoted as new definitions on the legends of the figures), we first notice that by increasing the number of taps, all filters achieve more noise reduction and higher output SNR. In addition, a decreasing signal distortion is observed at the output of the tradeoff filter with $\mu > 0$. These gains come at the price of an increasing complexity since the noise and noisy data matrices are $NL \times NL$ dimensional. A careful choice of the number of filter taps seems to be required depending on the expected performance and complexity of the noise reduction algorithm. Furthermore, we notice that as the tradeoff parameter increases from 1 to 5, more noise reduction and signal distortion are obtained at the output of the tradeoff filter. Comparing the subplots in (b) and (d) to those in (a) and (c) from Figs. 1 and 2, respectively, we see that the MVDR does not distort the desired signal but achieves less noise reduction than the Wiener filter. These results agree with the theoretical analysis in Section 5. The LCMV does not distort the desired signal as well, but it achieves the lowest values of noise reduction factor; even lower than 1, which means that the total noise at its output was amplified. The latter result is also confirmed by our analysis following the relationship between the MVDR and LCMV in (86). As for the output SNR, we expect from our analysis in Section 5 that its values would be the same for the tradeoff filter regardless of μ. The plots in Fig. 3 do not perfectly agree with our theoretical findings. To explain this mismatch, recall that in all our analysis, we assumed the coexistence of the desired speech signal and noise at every time instant. This assumption is not always valid for speech which is known to be non-stationary and its energy may frequently decay to zero. In noise-only frames, the filters corresponding to $\mu > 0$ attenuate all their outputs to 0 while the MVDR

[3] Similar observations can be made for a varying speech-to-fan-noise ratio and constant signal-to-babble-noise ratio.

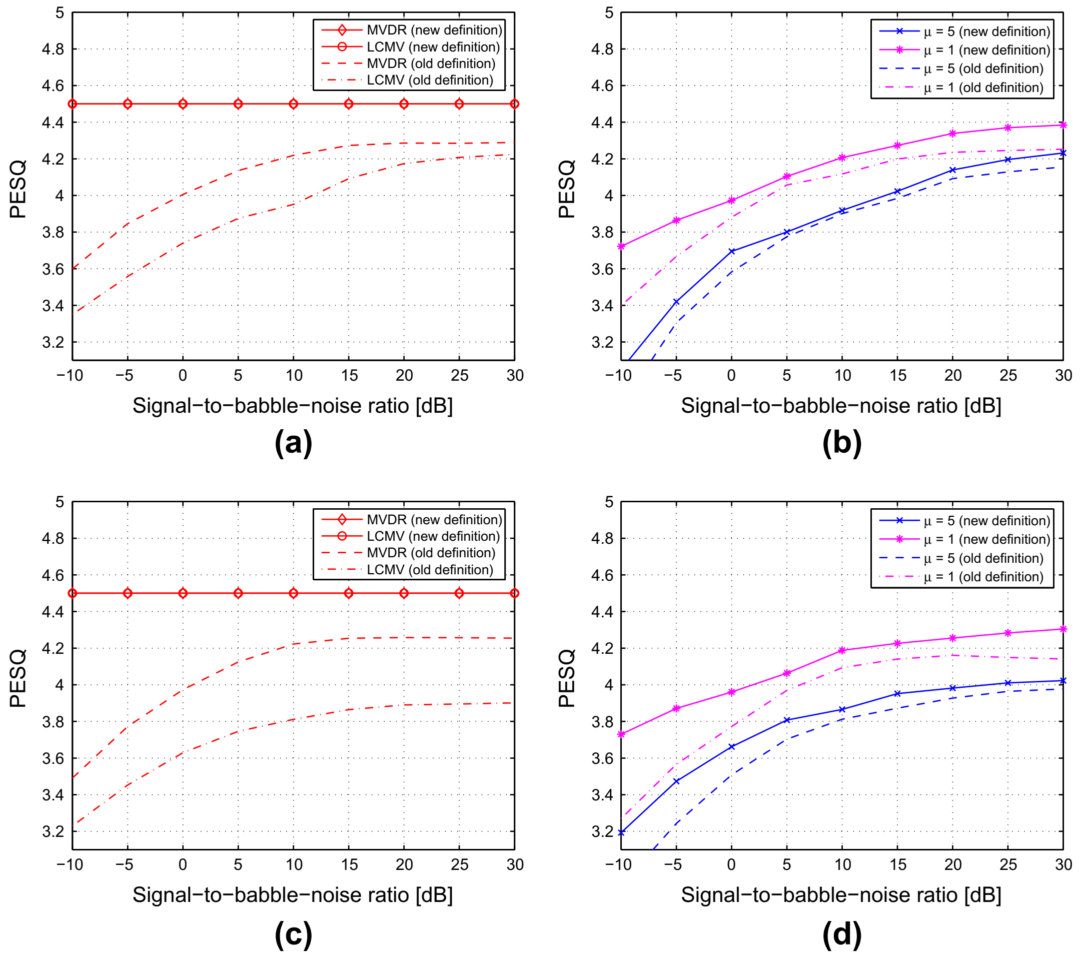

Fig. 9. Effect of the input speech-to-babble noise ratio on the PESQ measure between $x_1(k)$ and the filtered signal. (a) MVDR and LCMV, $T_{60} = 240$ ms. (b) Tradeoff filter with $\mu = 5$ and $\mu = 1$, $T_{60} = 240$ ms. (c) MVDR and LCMV, $T_{60} = 580$ ms. (d) Tradeoff filter with $\mu = 5$ and $\mu = 1$, $T_{60} = 580$ ms. Old and new definitions of the signal distortion index are compared. The number of filter taps is $L = 50$. The input speech-to-fan-noise ratio is 10 dB.

is regularized enough and does not suppress all signals in these frames. Besides, it can be empirically verified that when increasing μ, the tradeoff filter becomes more aggressive in terms of signals suppression when the speech energy decays. Hence, less noise reduction is achieved by the MVDR followed by the Wiener filter in noise-only frames. This translates into larger SNR gains of the tradeoff filter with $\mu = 5$ as compared to the multichannel Wiener filter and MVDR. The LCMV can dramatically deteriorate the SNR especially when the reverberation level is relatively high as shown in Fig. 3c for small number of taps (less than 40).

A fundamental result has to be emphasized in Figs. 1–3. In fact, there is a clear discrepancy between the old and new definitions of the three performance measures. For a given filter, the old metrics indicate larger noise reduction, output SNR, and signal distortion values than their new counterparts. This discrepancy is caused by the interference in the filtered speech signal which is traditionally included in the desired signal part. The behavior of the LCMV evaluated by both types of measures in Figs. 1–3 provides an additional illustration of the relevance of the new definitions of the performance measures: the old performance metrics show that this beamformer significantly distorts the desired signal even though it is defined to be distortionless.

Fig. 4 depicts the residual interference and the desired signal parts at the output of the multichannel Wiener filter when $L = 50$. The interference level is remarkably lower than the desired signal, but taking it into account substantially alters the expected performance of the filters. Including the interference in the definition of the desired signal is certainly not correct since both components are, *by definition*, orthogonal. The relevance of the new decomposition is further illustrated from the perceptional point of view in Fig. 5. Indeed, the fact that the MVDR and LCMV are inherently distortionless is well confirmed by the new definition of the desired signal part (their corresponding PESQ values are equal to 4.5 regardless of the filter length) in contrast to the old definition that considers the interference to be part of the desired signal and results in lower PESQ values. Larger PESQ values are also observed at the output of the tradeoff filter with $\mu = 1$ and 5 when using the new definition of the desired signal part in the filtered speech.

Figs. 6–8 depict the effect of the variations of the babble noise level on the performance of the four filters for both reverberation conditions. We fix the level of the desired signal and the fan noise as in the previous simulations and change the desired signal-to-babble-noise ratio. A clear decrease of the signal distortion index is observed for the tradeoff filter with $\mu = 1$ and 5. The signal distortion at the output of the tradeoff filter with $\mu > 0$ is mainly due to the (spatially) non-coherent noise components associated with the babble noise.[4] The signal distortion at the output of the

[4] The fan noise is more coherent than the babble noise across the sensors. It is known that if only the coherent signals overlap, perfect separation (with no distortion) is theoretically achievable since we have more microphones than sources.

MVDR and LCMV is maintained at its lowest values, by definition. The noise reduction factors of the four filters are decreasing with respect to the speech-to-babble-noise ratio for the range −10 to 10 dB. When the level of babble noise is smaller than that of the fan noise (values of speech-to-babble-noise ratio larger than 10 dB), the filters are more focused on the suppression of the latter. It is known that the suppression of spatially localized signals (fan noise in our case) is easier than the suppression of other types of noise (babble noise in our case). Consequently, we see that the noise reduction factors start to increase when the speech-to-babble-noise ratio is higher than 10 dB, especially at the output of the tradeoff filter with $\mu = 1,2$ when $T_{60} = 240$ ms (i.e., when the fan noise is more coherent across the sensors). An increasing SNR is also observed at the output of all filters as shown in Fig. 8. At relatively high speech-to-babble-noise ratios, the effect of reverberation becomes more noticeable, and we see that as the reverberation time increases from 240 ms to 580 ms, less SNR gain is obtained for all the studied filters. Fig. 9 shows the achieved PESQ values when the old and new definitions of the desired filtered speech are used. When the signal-to-babble-noise ratio increases, the old and new definitions of the PESQ score increase for the tradeoff filter with $\mu > 0$. Finally, the same observations regarding the discrepancy between the old and new definitions of the performance measures can be made, which confirms, again, the relevance of the definitions given in Section 4.

7. Conclusions

In this paper, we presented a new perspective on well-known multichannel time-domain noise reduction approaches. We started by demonstrating that the observed noise-free signals are composed of two main components: the first is obtained by projecting the noise-free observations on the reference microphone speech signal and is perfectly coherent across the sensors, while the second is orthogonal to the desired signal; it is, consequently, termed interference. Thanks to this new decomposition, we introduced the notion of source steering vector in the time domain and exploited it to derive the multichannel Wiener, MVDR, LCMV, maximum SNR, and tradeoff filters. Simplified expressions of these filters were obtained and new insights into their functioning were gained. For instance, it was demonstrated that the main difference between the time-domain MVDR, Wiener, maximum SNR, and tradeoff filters is attributed to a scaling factor that leads to different levels of desired signal distortion and noise reduction, while all filters, theoretically, have the same SNR gain. We applied the same decomposition to the noise observed by the microphones and formulated the time-domain LCMV. We demonstrated a fundamental relationship between the time-domain MVDR and LCMV thanks to this decomposition and showed how the MVDR achieves a tradeoff between the coherent and other residual (incoherent) noise components reduction. Conversely, the LCMV can dramatically amplify the incoherent noise depending on the level of interference and the generalized collinearity factor between the steering vectors of the noise and noise-free data defined in the time domain. We evaluated the performance of all the noise reduction filters considered herein and proved the relevance of our new definitions of performance metrics that are perfectly tailored to our study.

References

[1] Flanagan JL, Johnson JD, Zahn R, Elko GW. Computer-steered microphone arrays for sound transduction in large rooms. J Acoust Soc Am 1985;75:1508–18.

[2] Benesty J, Chen J, Huang Y. Microphone array signal processing. Berlin, Germany: Springer-Verlag; 2008.

[3] Brandstein M, Ward DB, editors. Microphone arrays: signal processing techniques and applications. Berlin, Germany: Springer-Verlag; 2001.

[4] Chen J, Benesty J, Huang Y. A minimum distortion noise reduction algorithm with multiple microphones. IEEE Trans Audio Speech Lang Process 2008;16:481–93.

[5] Doclo S, Spriet A, Wouters J, Moonen M. Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. Speech Commun 2007;49:636–56.

[6] Souden M, Benesty J, Affes S. On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans Audio Speech Lang Process 2010;18:260–76.

[7] Benesty J, Chen J, Huang Y. Noise reduction algorithms in a generalized transform domain. IEEE Trans Audio Speech Lang Process 2009;17:1109–23.

[8] Gannot S, Burshtein D, Weinstein E. Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans Signal Process 2001;49:1614–26.

[9] Markovich S, Gannot S, Cohen I. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Trans Audio Speech Lang Process 2009;17:1071–86.

[10] Talmon R, Cohen I, Gannot S. Convolutive transfer function generalized sidelobe canceler. IEEE Trans Audio Speech Lang Process 2009;17:1420–34.

[11] Doclo S, Moonen M. GSVD-based optimal filtering for single and multimicrophone speech enhancement. IEEE Trans Signal Process 2002;50:2230–44.

[12] Cornelis B, Moonen M, Wouters J. Performance analysis of multichannel Wiener filter based noise reduction in hearing aids under second order statistics estimation errors. IEEE Trans Audio Speech Lang Process 2011;19(5):1368–81.

[13] Souden M, Benesty J, Affes S. A study of the LCMV and MVDR noise reduction filters. IEEE Trans Signal Process 2010;18:4925–35.

[14] Chen J, Benesty J, Huang Y, Doclo S. New insights into the noise reduction Wiener filter. IEEE Trans Audio Speech Lang Process 2006;14:1218–34.

[15] Johnson DH, Dudgeon DE. Array signal processing-concepts and techniques. Englewood Cliffs, NJ: Prentice-Hall; 1993.

[16] Dmochowski JP, Benesty J. Microphone arrays: fundamental concepts. In: Cohen I, Benesty J, Gannot S, editors. Speech processing in modern communication-challenges and perspectives. Berlin, Germany: Springer-Verlag; 2008. p. 199–223 [chapter 8, 2010].

[17] Herbordt W. Combination of robust adaptive beamforming with acoustic echo cancellation for acoustic human/machine interfaces. PhD Thesis, Germany: Erlangen–Nuremberg University; 2004.

[18] Capon J. High resolution frequency-wavenumber spectrum analysis. Proc IEEE 1969;57:1408–18.

[19] Lacoss RT. Data adaptive spectral analysis methods. Geophysics 1971;36:661–75.

[20] Frost O. An algorithm for linearly constrained adaptive array processing. Proc IEEE 1972;60:926–35.

[21] Er M, Cantoni A. Derivative constraints for broad-band element space antenna array processors. IEEE Trans Acoust Speech Signal Process 1983;31:1378–93.

[22] Varga AP, Steenekan HJM, Tomlinson M, Jones D. The noisex-92 study on the effect of additive noise on automatic speech recognition. Tech. Rep., DRA Speech Research Unit; 1992.

[23] http://www.freesound.org/.