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 $\mathbf{x}(k)$,

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

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

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

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

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