Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction

Jacob Benesty, Jingdong Chen, Yiteng (Arden) Huang, and Tomas Gaensler

Citation: The Journal of the Acoustical Society of America **132**, 452 (2012); doi: 10.1121/1.4726071
View online: https://doi.org/10.1121/1.4726071
View Table of Contents: https://asa.scitation.org/toc/jas/132/1
Published by the Acoustical Society of America

Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction

Jacob Benesty
INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, Quebec H5A 1K6, Canada

Jingdong Chen
Northwestern Polytechnical University, 127 Youyi West Road, Xi'an, Shaanxi 710072, China

Yiteng (Arden) Huang
WeVoice, Inc., 1065 Route 22 West, Suite 2E, Bridgewater, New Jersey 08807

Tomas Gaensler
mh acoustics LLC, 25A Summit Avenue, Summit, New Jersey 07901

(Received 22 December 2011; revised 13 April 2012; accepted 16 May 2012)

This paper addresses the problem of noise reduction in the time domain where the clean speech sample at every time instant is estimated by filtering a vector of the noisy speech signal. Such a clean speech estimate consists of both the filtered speech and residual noise (filtered noise) as the noisy vector is the sum of the clean speech and noise vectors. Traditionally, the filtered speech is treated as the desired signal after noise reduction. This paper proposes to decompose the clean speech vector into two orthogonal components: one is correlated and the other is uncorrelated with the current clean speech sample. While the correlated component helps estimate the clean speech, it is shown that the uncorrelated component interferes with the estimation, just as the additive noise. Based on this orthogonal decomposition, the paper presents a way to define the error signal and cost functions and addresses the issue of how to design different optimal noise reduction filters by optimizing these cost functions. Specifically, it discusses how to design the maximum SNR filter, the Wiener filter, the minimum variance distortionless response (MVDR) filter, the tradeoff filter, and the linearly constrained minimum variance (LCMV) filter. It demonstrates that the maximum SNR, Wiener, MVDR, and tradeoff filters are identical up to a scaling factor. It also shows from the orthogonal decomposition that many performance measures can be defined, which seem to be more appropriate than the traditional ones for the evaluation of the noise reduction filters.
 [http://dx.doi.org/10.1121/1.4726071]

PACS number(s): 43.72.Dv, 43.60.Fg [CYE] Pages: 452–464

I. INTRODUCTION

In applications related to speech, sound recording, telecommunications, teleconferencing, telecollaboration, and human-machine interfaces, the signal of interest (usually speech) that is picked up by a microphone is always contaminated by noise. Such a contamination can dramatically change the statistics of the speech signal and can degrade the speech quality and intelligibility, thereby causing significant performance degradation to human-to-human and human-to-machine communication systems. In order to mitigate the detrimental effect of noise, it is indispensable to develop digital signal processing techniques to "clean" the noisy speech before it is stored, transmitted, or rendered. This cleaning process, which is referred to as noise reduction, has been a major challenge for many researchers and engineers over the past few decades (Boll, 1979; Vary, 1985; Martin, 2001; Ephraim and Malah, 1984; Chen *et al*., 2009; Benesty *et al*., 2009; Vary and Martin, 2006; Loizou, 2007; Benesty *et al*., 2005).

Typically, noise reduction is formulated as a filtering problem where the clean speech estimate is obtained by passing the noisy speech through a digital filter. With such a formulation, the core issue of noise reduction is to construct an optimal filter that can fully exploit the speech and noise statistics to achieve maximum noise suppression without introducing perceptually noticeable speech distortion. The design of optimal noise reduction filters can be accomplished either directly in the time domain or in a transform space. Practically, working in a transform space such as the frequency (Boll, 1979; Vary, 1985; Martin, 2001; Ephraim and Malah, 1984) or Karhunen-Loève expansion (KLE) domains (Chen *et al*., 2009) may offer some advantages in terms of real-time implementation and flexibility. But the filter design process in different domains remains the same and any noise reduction filter designed in a transform space can be equivalently constructed in the time domain from a theoretical point of view. So, in this paper we will focus our discussion on the time-domain formulation. However, any approach developed here should not be limited to the time domain and can be extended to other domains.

In the time domain, noise reduction is generally achieved on a sample-by-sample basis where the clean speech sample at every time instant is estimated by filtering a vector of the noisy speech signal. Since the noisy speech vector is the sum of the clean speech and noise signal vectors, this estimate consists of both the filtered speech and

residual noise (filtered noise). Traditionally, the filtered speech is treated as the desired signal after noise reduction. This definition of the desired speech, however, can cause many problems for both the design and evaluation of the noise reduction filters. For example, with this definition, the output signal-to-noise ratio (SNR) would be the ratio of the power of the filtered speech over the power of the residual noise. We should expect then that the filter that maximizes the output SNR should be a good optimal noise reduction filter. It has been found, however, that such a filter causes so much speech distortion that it is not useful in practice. In this paper, we propose to decompose the clean speech vector into two orthogonal components: one is correlated and the other is uncorrelated with the current clean speech sample. While the correlated component helps estimate the clean speech, we show that the uncorrelated component interferes with the estimation, just as the additive noise. Therefore, we introduce a new term, interference, in noise reduction. Based on this orthogonal decomposition and the new interference term, we present a way to redefine the error signal and cost functions. By optimizing these cost functions, we can derive many new noise reduction filters such as the minimum variance distortionless response (MVDR) filter and the linearly constrained minimum variance (LCMV) filter that are impossible to obtain with the traditional approaches. We show that the maximum SNR filter derived from the new form of the error signal is identical, up to a scaling factor, to the Wiener and MVDR filters. This, on one hand, proves that the new decomposition makes sense, and on the other hand, demonstrates that the Wiener filter is an optimal filter not only from the minimum mean-square error (MMSE) sense but also from the maximum SNR standpoint. Based on the decomposition of the filtered speech, we also show that many performance measures should be redefined and the new measures are more appropriate to quantify the noise reduction performance than the traditional ones.

The rest of this paper is organized as follows. In Sec. II, we formulate the single-channel noise reduction problem in the time domain. We briefly review the classical approaches in Sec. III. Section IV presents a new way to decompose the error signal based on the decomposition of the filtered speech into the filtered desired speech and interference, and we explain the difference between the new error signals and the traditional ones. Section V discusses different performance measures. In Sec. VI, we derive several optimal noise reduction filters. Section VII deals with the linearly constrained minimum variance (LCMV) filter. Section VIII presents some experiments confirming the theoretical derivations. Finally, we give our conclusions in Sec. IX.

II. SIGNAL MODEL

The noise reduction problem considered in this paper is one of recovering the desired signal (or clean speech) $x(k)$, k being the discrete-time index, of zero mean from the noisy observation (microphone signal) (Benesty *et al.*, 2009; Vary and Martin, 2006; Loizou, 2007)

$$y(k) = x(k) + v(k), \tag{1}$$

where $v(k)$, assumed to be a zero-mean random process, is the unwanted additive noise that can be either white or colored but is uncorrelated with $x(k)$. All signals are considered to be real and broadband, and $x(k)$ is assumed to be quasi-stationary so that its statistics can be estimated on a short-time basis.

The signal model given in (1) can also be written into a vector form as

$$\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k), \tag{2}$$

where

$$\mathbf{y}(k) \triangleq \left[y(k)\, y(k-1)\, \cdots\, y(k-L+1)\right]^T \tag{3}$$

is a vector of length L, superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}(k)$ and $\mathbf{v}(k)$ are defined in a similar way to $\mathbf{y}(k)$. Since $x(k)$ and $v(k)$ are uncorrelated by assumption, the correlation matrix (of size $L \times L$) of the noisy signal can be written as

$$\mathbf{R}_\mathbf{y} \triangleq E[\mathbf{y}(k)\mathbf{y}^T(k)] = \mathbf{R}_\mathbf{x} + \mathbf{R}_\mathbf{v}, \tag{4}$$

where $E[\cdot]$ denotes mathematical expectation, and $\mathbf{R}_\mathbf{x} \triangleq E[\mathbf{x}(k)\mathbf{x}^T(k)]$ and $\mathbf{R}_\mathbf{v} \triangleq E[\mathbf{v}(k)\mathbf{v}^T(k)]$ are the correlation matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively. The objective of noise reduction is then to find a "good" estimate of either $x(k)$ or $\mathbf{x}(k)$ in the sense that the additive noise is significantly reduced while the desired signal is not much distorted. In this paper, we focus only on the estimation of $x(k)$ to make the presentation concise. In other words, we only consider to estimate the desired speech on a sample-by-sample basis and, at each time instant k, the signal sample $x(k)$ is estimated from the corresponding observation signal vector $\mathbf{y}(k)$ of length L.

III. CLASSICAL LINEAR FILTERING APPROACH

In the classical approach, the estimate of the desired signal $x(k)$ is obtained by applying a finite-impulse-response (FIR) filter to the observation signal vector $\mathbf{y}(k)$, (Chen *et al.*, 2006) i.e.,

$$\hat{x}(k) = \mathbf{h}^T\mathbf{y}(k) = x_\mathrm{fd}(k) + v_\mathrm{rn}(k), \tag{5}$$

where

$$\mathbf{h} \triangleq [h_0\ h_1 \cdots h_{L-1}]^T \tag{6}$$

is an FIR filter of length L, $x_\mathrm{fd}(k) \triangleq \mathbf{h}^T\mathbf{x}(k)$ is the filtered speech, which is treated as the desired signal component after noise reduction, and $v_\mathrm{rn}(k) \triangleq \mathbf{h}^T\mathbf{v}(k)$ is the residual noise that is uncorrelated with the filtered speech.

The error signal for this estimation problem is

$$e(k) \triangleq \hat{x}(k) - x(k) = e_\mathrm{d}^\mathrm{C}(k) + e_\mathrm{r}^\mathrm{C}(k), \tag{7}$$

where

$$e_{\mathrm{d}}^{\mathrm{C}}(k) \triangleq x_{\mathrm{f}}(k) - x(k) = \mathbf{h}^T\mathbf{x}(k) - x(k) \tag{8}$$

is the signal distortion due to the FIR filter,

$$e_{\mathrm{r}}^{\mathrm{C}}(k) \triangleq v_{\mathrm{rn}}(k) \tag{9}$$

represents the residual noise, and we use the superscript $^{\mathrm{C}}$ to denote the classical model.

The mean-square error (MSE) is then

$$J(\mathbf{h}) \triangleq E[e^2(k)]. \tag{10}$$

Since $x(k)$ and $v(k)$ are uncorrelated, the MSE can be decomposed into two terms as

$$J(\mathbf{h}) = J_{\mathrm{d}}^{\mathrm{C}}(\mathbf{h}) + J_{\mathrm{r}}^{\mathrm{C}}(\mathbf{h}), \tag{11}$$

where

$$J_{\mathrm{d}}^{\mathrm{C}}(\mathbf{h}) \triangleq E\{[e_{\mathrm{d}}^{\mathrm{C}}(k)]^2\}. \tag{12}$$

and

$$J_{\mathrm{r}}^{\mathrm{C}}(\mathbf{h}) \triangleq E\{[e_{\mathrm{r}}^{\mathrm{C}}(k)]^2\}. \tag{13}$$

Given the definition of the MSE, the optimal noise reduction filters can be obtained by directly minimizing $J(\mathbf{h})$, or by minimizing either $J_{\mathrm{d}}^{\mathrm{C}}(\mathbf{h})$ or $J_{\mathrm{r}}^{\mathrm{C}}(\mathbf{h})$ with some constraint.

IV. A LINEAR MODEL BASED ON AN ORTHOGONAL DECOMPOSITION FOR DESIRED SIGNAL EXTRACTION

From the filtering model given in (5), we see that $\hat{x}(k)$ depends on the vector $\mathbf{x}(k)$. However, not all the components in $\mathbf{x}(k)$ contribute to the estimation of the desired signal sample $x(k)$; therefore, treating the filtered speech, i.e., $x_{\mathrm{f}}(k) = \mathbf{h}^T\mathbf{x}(k)$, as the desired signal after noise reduction seems inappropriate in the derivation and evaluation of noise reduction filters. To see this clearly, let us decompose the vector $\mathbf{x}(k)$ into the following form:

$$\mathbf{x}(k) = x(k)\boldsymbol{\gamma}_x + \mathbf{x}'(k) = \mathbf{x}_{\mathrm{d}}(k) + \mathbf{x}'(k), \tag{14}$$

where

$$\mathbf{x}_{\mathrm{d}}(k) = [\,x_{\mathrm{d},0}(k) \quad x_{\mathrm{d},1}(k) \quad \cdots \quad x_{\mathrm{d},L-1}(k)\,]^T = x(k)\boldsymbol{\gamma}_x, \tag{15}$$

$$\begin{aligned}\mathbf{x}'(k) &= [\,x'_0(k) \quad x'_1(k) \quad \cdots \quad x'_{L-1}(k)\,]^T \\ &= \mathbf{x}(k) - x(k)\boldsymbol{\gamma}_x,\end{aligned} \tag{16}$$

$$\begin{aligned}\boldsymbol{\gamma}_x &= [\gamma_{x,0} \quad \gamma_{x,1} \quad \cdots \quad \gamma_{x,L-1}]^T = [1 \quad \gamma_{x,1} \quad \cdots \quad \gamma_{x,L-1}]^T \\ &= \frac{E[x(k)\mathbf{x}(k)]}{E[x^2(k)]}\end{aligned} \tag{17}$$

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

$$\gamma_{x,l} = \frac{E[x(k)x(k-l)]}{E[x^2(k)]} \tag{18}$$

is the correlation coefficient between $x(k)$ and $x(k-l)$ with $-1 \le \gamma_{x,l} \le 1$.

It is easy to check that $\mathbf{x}_{\mathrm{d}}(k)$ is correlated with the desired signal sample $x(k)$, while $\mathbf{x}'(k)$ is uncorrelated with $x(k)$, i.e., $E[x(k)\mathbf{x}'(k)] = \mathbf{0}$. To illustrate this decomposition, we took a frame (400 samples) of an /i:/ sound signal recorded from a female speaker and computed its correlation coefficients $\gamma_{x,l}$, $l = 0, 1, \ldots, 20$, using a short-time average. Both the waveform and correlation coefficients are plotted in Fig. 1. Using these estimated correlation coefficients and setting the parameter L to 20, we performed the orthogonal decomposition of the /i:/ sound signal $x(k)$. The first and second coefficients of $\mathbf{x}_{\mathrm{d}}(k)$ and $\mathbf{x}'(k)$ as a function of time k are shown in Fig. 2. Since $\gamma_{x,0} = 1$, we have $x_{\mathrm{d},0}(k) = x(k)$ and $x'_0(k) = 0$, which can be seen from Figs. 2(a) and 2(b). But as l increases, the correlation between $x(k)$ and $x(k-l)$ decreases, and as a result, the level of $x_{\mathrm{d},l}(k)$ decreases while that of $x'_l(k)$ increases, which can be seen by comparing Figs. 2(a) with 2(c) and 2(b) with 2(d). Figures 2(c) and 2(d) show $x_{\mathrm{d},1}(k)$ and $x'_1(k)$. Note that in practice $\boldsymbol{\gamma}_x$ cannot be computed directly since $x(k)$ is not accessible. However, slightly rearranging (17), we get

$$\boldsymbol{\gamma}_x = \frac{E[\mathbf{y}(k)y(k)] - E[\mathbf{v}(k)v(k)]}{E[y^2(k)] - E[v^2(k)]} = \frac{\sigma_y^2\boldsymbol{\gamma}_y - \sigma_v^2\boldsymbol{\gamma}_v}{\sigma_y^2 - \sigma_v^2}, \tag{19}$$

where $\sigma_y^2 \triangleq E[y^2(k)]$ and $\sigma_v^2 \triangleq E[v^2(k)]$ are the variances of $y(k)$ and $v(k)$, respectively. One can see that now $\boldsymbol{\gamma}_x$ depends on the statistics of $y(k)$ and $v(k)$. The statistics of $y(k)$ can be computed directly since $y(k)$ is accessible while the statistics of $v(k)$ can be estimated based on the use of a voice activity detector (VAD) (Cohen *et al.*, 2010).

Now substituting (14) into (5), we get

$$\hat{x}(k) = \mathbf{h}^T\mathbf{x}_{\mathrm{d}}(k) + \mathbf{h}^T\mathbf{x}'(k) + \mathbf{h}^T\mathbf{v}(k). \tag{20}$$

Since it is correlated with the desired signal sample $x(k)$, the vector $\mathbf{x}_{\mathrm{d}}(k)$ will help estimate $x(k)$. So, the first term on the

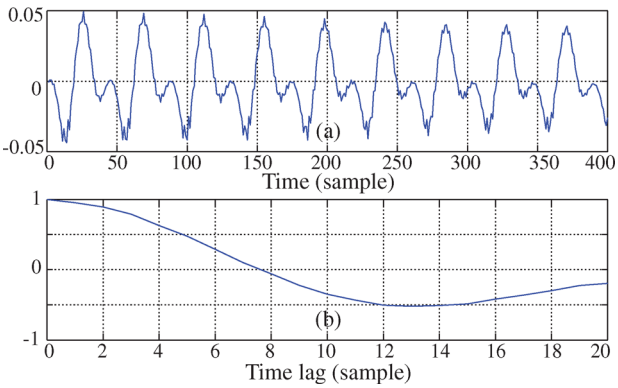

FIG. 1. (Color online) A frame of an /i:/ sound signal recorded from a female talker: (a) waveform and (b) its normalized correlation coefficients estimated using a short-time average.

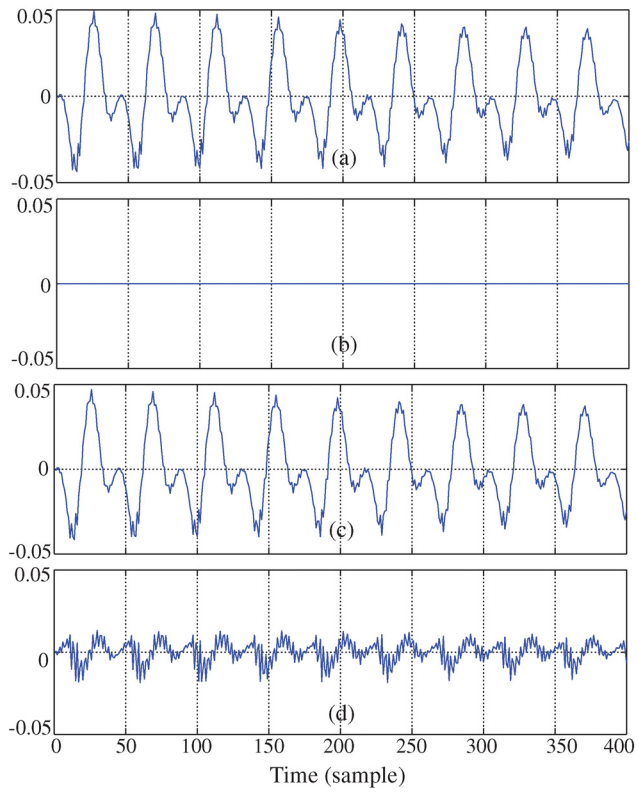

FIG. 2. (Color online) The orthogonal decomposition of the signal in Fig. 1: (a) $x_{\mathrm{d},0}(k) = x(k)$, (b) $x'_0(k)$, (c) $x_{\mathrm{d},1}(k)$, and (d) $x'_1(k)$.

right-hand side of (20) is clearly the filtered desired signal and we denote it as $x_{\mathrm{fd}}(k) \triangleq \mathbf{h}^T\mathbf{x}_{\mathrm{d}}(k) = x(k)\mathbf{h}^T\boldsymbol{\gamma}_x$. In comparison, $\mathbf{x}'(k)$ is orthogonal to $x(k)$; so this vector would interfere with the estimation. Therefore, we introduce the term "interference," defined as $x'_{\mathrm{ri}}(k) \triangleq \mathbf{h}^T\mathbf{x}'(k)$. The third term on the right-hand side of (20) is the residual noise, as in the classical approaches, i.e., $v_{\mathrm{rn}}(k) \triangleq \mathbf{h}^T\mathbf{v}(k)$. So, the signal estimate can now be written as

$$\hat{x}(k) = x_{\mathrm{fd}}(k) + x'_{\mathrm{ri}}(k) + v_{\mathrm{rn}}(k). \tag{21}$$

It can be checked that the three terms $x_{\mathrm{fd}}(k)$, $x'_{\mathrm{ri}}(k)$, and $v_{\mathrm{rn}}(k)$ are mutually uncorrelated. Therefore, the variance of $\hat{x}(k)$ is

$$\sigma^2_{\hat{x}} = \sigma^2_{x_{\mathrm{fd}}} + \sigma^2_{x'_{\mathrm{ri}}} + \sigma^2_{v_{\mathrm{rn}}}, \tag{22}$$

where

$$\sigma^2_{x_{\mathrm{fd}}} = \sigma^2_x(\mathbf{h}^T\boldsymbol{\gamma}_x)^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\mathbf{h}, \tag{23}$$

$$\sigma^2_{x'_{\mathrm{ri}}} = \mathbf{h}^T\mathbf{R}_{\mathbf{x}'}\mathbf{h} = \mathbf{h}^T\mathbf{R}_{\mathbf{x}}\mathbf{h} - \sigma^2_x(\mathbf{h}^T\boldsymbol{\gamma}_x)^2, \tag{24}$$

$$\sigma^2_{v_{\mathrm{rn}}} = \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h}, \tag{25}$$

$\sigma^2_x \triangleq E[x^2(k)]$ is the variance of $x(k)$, $\mathbf{R}_{\mathbf{x}_{\mathrm{d}}} = \sigma^2_x\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T$ is the correlation matrix (whose rank is equal to 1) of $\mathbf{x}_{\mathrm{d}}(k)$, and $\mathbf{R}_{\mathbf{x}'} \triangleq E[\mathbf{x}'(k)\mathbf{x}'^T(k)]$ is the correlation matrix of $\mathbf{x}'(k)$.

With the above decomposition, one can see that the objective of noise reduction is to find a good filter that makes $x_{\mathrm{fd}}(k)$ as close as possible to $x(k)$ and meanwhile minimizes the effect of both $x'_{\mathrm{ri}}(k)$ and $v_{\mathrm{rn}}(k)$. To find such a filter, we first define the error signal between the estimated and desired signals as

$$e(k) \triangleq \hat{x}(k) - x(k) = e_{\mathrm{d}}(k) + e_{\mathrm{r}}(k), \tag{26}$$

where

$$e_{\mathrm{d}}(k) \triangleq x_{\mathrm{fd}}(k) - x(k) \tag{27}$$

is the signal distortion due to the FIR filter and

$$e_{\mathrm{r}}(k) \triangleq x'_{\mathrm{ri}}(k) + v_{\mathrm{rn}}(k) \tag{28}$$

represents the residual interference-plus-noise.

The MSE is then

$$J(\mathbf{h}) = E[e^2(k)] = J_{\mathrm{d}}(\mathbf{h}) + J_{\mathrm{r}}(\mathbf{h}), \tag{29}$$

where

$$J_{\mathrm{d}}(\mathbf{h}) = E[e^2_{\mathrm{d}}(k)] = \sigma^2_x(\mathbf{h}^T\boldsymbol{\gamma}_x - 1)^2 \tag{30}$$

and

$$J_{\mathrm{r}}(\mathbf{h}) = E[e^2_{\mathrm{r}}(k)] = \sigma^2_{x'_{\mathrm{ri}}} + \sigma^2_{v_{\mathrm{rn}}}. \tag{31}$$

Comparing (26) with (7) and (29) with (11), one can clearly see the difference between the new definitions of the error signal and MSE in our new model and the traditional definitions. It is clear that the objective of noise reduction with the new linear 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. But before deriving the optimal filters, we first give some very useful measures that fit well with the new linear model.

V. PERFORMANCE MEASURES

Many distance measures have been developed to evaluate noise reduction, such as the Itakura distance, the Itakura-Saito distance (ISD) [that performs a comparison of spectral envelopes (AR parameters) between the clean and the processed speech] (Itakura and Saito, 1970; Quackenbush *et al.*, 1988; Chen *et al.*, 2003), SNR, speech distortion index, noise reduction factor (Benesty *et al.*, 2009; Benesty *et al.*, 2005; Chen *et al.*, 2006), etc. Many of these measures are defined based on the classical linear filter model. As it was shown in the previous section, the filtered speech (where the desired signal does not appear explicitly) should be separated into the filtered desired speech and interference. The interference part interferes with the estimation of the desired speech signal and it should be treated as part of the noise. Therefore, it is necessary to redefine some of the performance measures originally established for the classical model.

The first measure is the input SNR defined as

$$\mathrm{iSNR} = \frac{\sigma^2_x}{\sigma^2_v}. \tag{32}$$

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 (in this paper, we consider the interference as part of the noise in the definitions of the performance measures since it is uncorrelated with the desired signal), i.e.,

$$\mathrm{oSNR}(\mathbf{h}) = \frac{\sigma_{x_{\mathrm{fd}}}^2}{\sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2} = \frac{\sigma_x^2(\mathbf{h}^T\boldsymbol{\gamma}_x)^2}{\mathbf{h}^T\mathbf{R}_{\mathrm{in}}\mathbf{h}}, \tag{33}$$

where

$$\mathbf{R}_{\mathrm{in}} = \mathbf{R}_{\mathbf{x}'} + \mathbf{R}_{\mathbf{v}} \tag{34}$$

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 so that the quality of the noisy signal will be enhanced. For the particular filter $\mathbf{h} = \mathbf{i}_0$, where $\mathbf{i}_0$ is the first column of the identity matrix $\mathbf{I}$ (of size $L \times L$), we have

$$\mathrm{oSNR}(\mathbf{i}_0) = \mathrm{iSNR}. \tag{35}$$

Now, let us define the quantity

$$\mathrm{oSNR}_{\mathrm{max}} \triangleq \lambda_{\mathrm{max}}(\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x_d}}), \tag{36}$$

where $\lambda_{\mathrm{max}}(\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}})$ denotes the maximum eigenvalue of the matrix $\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}$. Since the rank of the matrix $\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}$ is equal to 1, we also have

$$\mathrm{oSNR}_{\mathrm{max}} = \mathrm{tr}[\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}] = \sigma_x^2\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x, \tag{37}$$

where $\mathrm{tr}[\cdot]$ denotes the trace of a square matrix. It can be checked that the quantity $\mathrm{oSNR}_{\mathrm{max}}$ corresponds to the maximum SNR that can be achieved through filtering since the filter, $\mathbf{h}_{\mathrm{max}}$, that maximizes $\mathrm{oSNR}(\mathbf{h})$ [Eq. (33)] is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x_d}}$. As a result, we have

$$\mathrm{oSNR}(\mathbf{h}) \leq \mathrm{oSNR}_{\mathrm{max}}, \quad \forall \mathbf{h} \tag{38}$$

and

$$\mathrm{oSNR}_{\mathrm{max}} = \mathrm{oSNR}(\mathbf{h}_{\mathrm{max}}) \geq \mathrm{oSNR}(\mathbf{i}_0) = \mathrm{iSNR}. \tag{39}$$

The noise reduction factor (Benesty *et al.*, 2005; Chen *et al.*, 2006) quantifies the amount of noise that is rejected by the filter. This quantity is defined as the ratio of the variance of the noise at the microphone over the variance of the interference-plus-noise remaining after the filtering operation, i.e.,

$$\xi_{\mathrm{nr}}(\mathbf{h}) \triangleq \frac{\sigma_v^2}{\sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2} = \frac{\sigma_v^2}{\mathbf{h}^T\mathbf{R}_{\mathrm{in}}\mathbf{h}}. \tag{40}$$

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

In practice, the FIR filter, $\mathbf{h}$, distorts the desired signal. In order to evaluate the level of this distortion, we define the speech reduction factor (Benesty *et al.*, 2009) as the variance of the desired signal over the variance of the filtered desired signal at the output of the filter, i.e.,

$$\xi_{\mathrm{sr}}(\mathbf{h}) \triangleq \frac{\sigma_x^2}{\sigma_{x_{\mathrm{fd}}}^2} = \frac{1}{(\mathbf{h}^T\boldsymbol{\gamma}_x)^2}. \tag{41}$$

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

$$\mathbf{h}^T\boldsymbol{\gamma}_x = 1. \tag{42}$$

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 four previous measures:

$$\frac{\mathrm{oSNR}(\mathbf{h})}{\mathrm{iSNR}} = \frac{\xi_{\mathrm{nr}}(\mathbf{h})}{\xi_{\mathrm{sr}}(\mathbf{h})}. \tag{43}$$

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

Another useful performance measure is the speech distortion index (Benesty *et al.*, 2005; Chen *et al.*, 2006) defined as

$$v_{\mathrm{sd}}(\mathbf{h}) = \frac{E\{[x_{\mathrm{fd}}(k) - x(k)]^2\}}{\sigma_x^2} = (\mathbf{h}^T\boldsymbol{\gamma}_x - 1)^2. \tag{44}$$

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 $v_{\mathrm{sd}}(\mathbf{h})$, the more the desired signal is distorted.

VI. OPTIMAL FILTERS

We have defined the MSE criterion with the new linear model in Sec. IV. For the particular filter $\mathbf{h} = \mathbf{i}_0$, the MSE is

$$J(\mathbf{i}_0) = \sigma_v^2. \tag{45}$$

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

$$\tilde{J}(\mathbf{h}) = \frac{J(\mathbf{h})}{J(\mathbf{i}_0)} = \mathrm{iSNR} \cdot v_{\mathrm{sd}}(\mathbf{h}) + \frac{1}{\xi_{\mathrm{nr}}(\mathbf{h})}, \tag{46}$$

where

$$v_{\mathrm{sd}}(\mathbf{h}) = \frac{J_{\mathrm{d}}(\mathbf{h})}{\sigma_x^2}, \tag{47}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}) = \frac{\sigma_v^2}{J_{\mathrm{r}}(\mathbf{h})}. \tag{48}$$

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

It is clear that the objective of noise reduction with the new linear 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. In this section, we derive three fundamental filters with the revisited linear model and show that they are fundamentally equivalent. We also show their equivalence with $\mathbf{h}_{\mathrm{max}}$ (i.e., the maximum SNR filter).

A. Wiener

The Wiener filter is easily derived by taking the gradient of the MSE, i.e., $J(\mathbf{h})$ defined in (29), with respect to $\mathbf{h}$ and equating the result to zero:

$$\mathbf{h}_{\mathrm{W}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}\mathbf{i}_0 = [\mathbf{I} - \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}]\mathbf{i}_0. \tag{49}$$

Since

$$\mathbf{R}_{\mathbf{x}}\mathbf{i}_0 = \sigma_x^2\boldsymbol{\gamma}_x, \tag{50}$$

we can rewrite (49) as

$$\mathbf{h}_{\mathrm{W}} = \sigma_x^2\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x. \tag{51}$$

From Sec. IV, it is easy to verify that

$$\mathbf{R}_{\mathbf{y}} = \sigma_x^2\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T + \mathbf{R}_{\mathrm{in}}. \tag{52}$$

Determining the inverse of $\mathbf{R}_{\mathbf{y}}$ from (52) with Woodbury's identity

$$\mathbf{R}_{\mathbf{y}}^{-1} = \mathbf{R}_{\mathrm{in}}^{-1} - \frac{\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathrm{in}}^{-1}}{\sigma_x^{-2} + \boldsymbol{\gamma}_x^T\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x} \tag{53}$$

and substituting the result into (51), we get another interesting formulation of the Wiener filter:

$$\mathbf{h}_{\mathrm{W}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x}{\sigma_x^{-2} + \boldsymbol{\gamma}_x^T\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x}, \tag{54}$$

that we can rewrite as

$$\mathbf{h}_{\mathrm{W}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{y}} - \mathbf{I}}{1 - L + \mathrm{tr}[\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{y}}]}\mathbf{i}_0 = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}}{1 + \mathrm{oSNR}_{\mathrm{max}}}\mathbf{i}_0. \tag{55}$$

Using (54), we deduce that the output SNR is

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}) = \mathrm{oSNR}_{\mathrm{max}} = \mathrm{tr}[\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{y}}] - L, \tag{56}$$

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

$$v_{\mathrm{sd}}(\mathbf{h}_{\mathrm{W}}) = \frac{1}{(1 + \mathrm{oSNR}_{\mathrm{max}})^2}. \tag{57}$$

The higher is the value of $\mathrm{oSNR}_{\mathrm{max}}$, the less the desired signal is distorted.

Since the Wiener filter maximizes the output SNR according to (56), we have

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}) \geq \mathrm{oSNR}(\mathbf{i}_0) = \mathrm{iSNR}. \tag{58}$$

It is interesting to see that the two filters $\mathbf{h}_{\mathrm{W}}$ and $\mathbf{h}_{\mathrm{max}}$ both maximize the output SNR. So, they are equivalent (different only by a scaling factor).

With the Wiener filter the noise reduction factor is

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{W}}) = \frac{(1 + \mathrm{oSNR}_{\mathrm{max}})^2}{\mathrm{iSNR} \cdot \mathrm{oSNR}_{\mathrm{max}}} \geq \left(1 + \frac{1}{\mathrm{oSNR}_{\mathrm{max}}}\right)^2. \tag{59}$$

Using (57) and (59) in (46), we find the minimum NMSE (MNMSE):

$$\tilde{J}(\mathbf{h}_{\mathrm{W}}) = \frac{\mathrm{iSNR}}{1 + \mathrm{oSNR}(\mathbf{h}_{\mathrm{W}})}. \tag{60}$$

B. MVDR

The celebrated MVDR filter proposed by Capon (Capon, 1969) is usually derived in a context where we have at least two sensors (or microphones) available. Interestingly, with the new linear model, we can also derive the MVDR (with one sensor only) by minimizing the MSE of the residual interference-plus-noise, $J_{\mathrm{r}}(\mathbf{h})$, with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\min_{\mathbf{h}} \quad \mathbf{h}^T\mathbf{R}_{\mathrm{in}}\mathbf{h} \quad \text{subject to} \quad \mathbf{h}^T\boldsymbol{\gamma}_x = 1. \tag{61}$$

The solution to the above optimization problem is

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x}, \tag{62}$$

which can also be written as

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{y}} - \mathbf{I}}{\mathrm{tr}[\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{y}}] - L}\mathbf{i}_0 = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}}{\mathrm{oSNR}_{\mathrm{max}}}\mathbf{i}_0. \tag{63}$$

Obviously, we can rewrite the MVDR as

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x}. \tag{64}$$

The Wiener and MVDR filters are simply related as follows

$$\mathbf{h}_{\mathrm{W}} = \alpha\mathbf{h}_{\mathrm{MVDR}}, \tag{65}$$

where

$$\alpha = \mathbf{h}_{\mathrm{W}}^T\boldsymbol{\gamma}_x = \frac{\mathrm{oSNR}_{\mathrm{max}}}{1 + \mathrm{oSNR}_{\mathrm{max}}}. \tag{66}$$

So, 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 nonstationary and the estimations are done on a frame-by-frame basis, so it is essential to have this scaling factor right from one frame to another in order to avoid large distortions. Therefore, it is recommended to use the MVDR filter rather than the Wiener filter in speech enhancement applications.

Locally, a scaling factor should not affect the SNR, but it can change the level of speech distortion and noise reduction. We should have

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{MVDR}}) = \mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}), \tag{67}$$

$$v_{\mathrm{sd}}(\mathbf{h}_{\mathrm{MVDR}}) = 0, \tag{68}$$

$$\xi_{\mathrm{sr}}(\mathbf{h}_{\mathrm{MVDR}}) = 1, \tag{69}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{MVDR}}) = \frac{\mathrm{oSNR}_{\mathrm{max}}}{\mathrm{iSNR}} \leq \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{W}}), \tag{70}$$

and

$$1 \geq \tilde{J}(\mathbf{h}_{\mathrm{MVDR}}) = \frac{\mathrm{iSNR}}{\mathrm{oSNR}_{\mathrm{max}}} \geq \tilde{J}(\mathbf{h}_{\mathrm{W}}). \tag{71}$$

However, from a global viewpoint, the time-varying scaling factor may put more attenuation in silence periods where the desired speech is absent and less attenuation when speech is present. This weighting process can cause some performance differences between the MVDR and Wiener filters if the performance is evaluated on a long-term basis. We will come back to this point when we discuss the experiments.

C. Tradeoff

In the tradeoff approach, we try to compromise between noise reduction and speech distortion. Instead of minimizing the MSE as we already did to find 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

$$\min_{\mathbf{h}} \quad J_{\mathrm{d}}(\mathbf{h}) \ \text{ subject to } \ J_{\mathrm{r}}(\mathbf{h}) = \beta\sigma_v^2, \tag{72}$$

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

$$\mathbf{h}_{\mathrm{T},\mu} = \sigma_x^2[\sigma_x^2\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T + \mu\mathbf{R}_{\mathrm{in}}]^{-1}\boldsymbol{\gamma}_x = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x}{\mu\sigma_x^{-2} + \boldsymbol{\gamma}_x^T\mathbf{R}_{\mathrm{in}}^{-1}\boldsymbol{\gamma}_x}, \tag{73}$$

where the Lagrange multiplier, μ, satisfies $J_{\mathrm{r}}(\mathbf{h}_{\mathrm{T},\mu}) = \beta\sigma_v^2$. Taking $\mu = 1$, we obtain the Wiener filter while for $\mu = 0$, we get the MVDR filter. With μ, we can make a compromise between noise reduction and speech distortion. Again, we observe here as well that the tradeoff and Wiener filters are equivalent up to a scaling factor. Locally at each time instant k, the scaling factor should not affect the SNR. So, the output SNR of the tradeoff filter is independent of μ and is identical to the output SNR of the Wiener filter, i.e.,

$$\mathrm{oSNR}(\mathbf{h}_{\mathrm{T},\mu}) = \mathrm{oSNR}(\mathbf{h}_{\mathrm{W}}), \quad \forall\mu. \tag{74}$$

VII. THE LCMV FILTER

We can derive an LCMV filter (Frost, 1972; Er and Cantoni, 1983) which can handle more than one linear constraint, by exploiting the structure of the noise signal.

In Sec. IV, we decomposed the vector $\mathbf{x}(k)$ into two orthogonal components to extract the desired signal, $x(k)$. We can also decompose (but not for the same reason) the noise signal vector, $\mathbf{v}(k)$, into two orthogonal vectors:

$$\mathbf{v}(k) = v(k)\boldsymbol{\gamma}_v + \mathbf{v}'(k), \tag{75}$$

where $\boldsymbol{\gamma}_v$ and $\mathbf{v}'(k)$ are defined in a similar way to $\boldsymbol{\gamma}_x$ and $\mathbf{x}'(k)$.

Our problem this time is the following. We wish to perfectly recover our desired signal, $x(k)$, and completely remove the correlated components, $v(k)\boldsymbol{\gamma}_v$. Thus, the two constraints can be put together in a matrix form as

$$\mathbf{C}^T\mathbf{h} = \mathbf{i}, \tag{76}$$

where

$$\mathbf{C} = [\boldsymbol{\gamma}_x \ \boldsymbol{\gamma}_v] \tag{77}$$

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

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

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

$$\mathbf{h}_{\mathrm{LCMV}} = \arg\ \min_{\mathbf{h}} \quad \mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h} \ \text{ subject to } \ \mathbf{C}^T\mathbf{h} = \mathbf{i}. \tag{78}$$

The solution to (78) is given by

$$\mathbf{h}_{\mathrm{LCMV}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{C}[\mathbf{C}^T\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{C}]^{-1}\mathbf{i}. \tag{79}$$

By developing (79), it can easily be shown that the LCMV can be written as a function of the MVDR:

$$\mathbf{h}_{\mathrm{LCMV}} = \frac{1}{1-\rho^2}\mathbf{h}_{\mathrm{MVDR}} - \frac{\rho^2}{1-\rho^2}\mathbf{t}, \tag{80}$$

where

$$\rho^2 = \frac{(\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v)^2}{(\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x)(\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v)}, \tag{81}$$

with $0 \leq \rho^2 \leq 1$, $\mathbf{h}_{\mathrm{MVDR}}$ is defined in (64), and

$$\mathbf{t} = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}{\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}. \quad (82)$$

We observe from (80) that when $\rho^2 = 0$, the LCMV filter becomes the MVDR filter; however, when ρ^2 tends to 1, which happens if and only if $\boldsymbol{\gamma}_x = \boldsymbol{\gamma}_v$, we have no solution since we have conflicting requirements.

Obviously, we always have

$$\text{oSNR}(\mathbf{h}_{\text{LCMV}}) \leq \text{oSNR}(\mathbf{h}_{\text{MVDR}}), \quad (83)$$

$$v_{\text{sd}}(\mathbf{h}_{\text{LCMV}}) = 0, \quad (84)$$

$$\xi_{\text{sr}}(\mathbf{h}_{\text{LCMV}}) = 1, \quad (85)$$

and

$$\xi_{\text{nr}}(\mathbf{h}_{\text{LCMV}}) \leq \xi_{\text{nr}}(\mathbf{h}_{\text{MVDR}}) \leq \xi_{\text{nr}}(\mathbf{h}_{\text{W}}). \quad (86)$$

The LCMV filter is able to remove all the correlated noise but at the price that its overall noise reduction is lower than that of the MVDR filter.

VIII. EXPERIMENTAL RESULTS

We have redefined the error signals, optimization cost functions, and evaluation criteria for the noise reduction problem in the time domain and derived several new optimal noise reduction filters. In this section, we study those filters through experiments.

The clean speech signal used in our experiments was recorded from a female speaker in a quiet office room. It was originally sampled at 16 kHz and then downsampled to 8 kHz. The overall length of the signal is approximately 10 min, but only the first 30 s is used in our experiments. Noisy speech is obtained by adding noise to the clean speech (the noise signal is properly scaled to control the input SNR level). We consider three types of noise: a white Gaussian random process, a babble noise signal recorded in a New York Stock Exchange (NYSE) room, and a competing speech signal recorded from a male speaker. The NYSE noise and the competing speech signal were also digitized with a sampling rate of 16 kHz, but again they were downsampled into 8 kHz. Compared with Gaussian random noise which is stationary and white, the NYSE noise is nonstationary and colored. It consists of sounds from various sources such as electrical fans, telephone rings, and background speech. See Huang *et al.* (2008) for a more detailed description of this babble noise. The male interference speech signal will be used to evaluate the LCMV filter for its performance in reducing correlated noise.

A. Estimation of correlation matrices and vectors

The implementation of most of the noise reduction filters derived in Secs. VI and VII requires the estimation of the correlation matrices $\mathbf{R}_{\mathbf{y}}$, $\mathbf{R}_{\mathbf{x}}$, and $\mathbf{R}_{\mathbf{v}}$, the correlation vector $\boldsymbol{\gamma}_{\text{x}}$, and the signal variance σ_x^2. Computation of $\mathbf{R}_{\mathbf{y}}$ is relatively easy because the noisy signal $y(k)$ is accessible. But in practice, we need a noise estimator or a VAD to compute all the other parameters. The problems regarding noise estimation and VAD have been widely studied in the literature and we have developed a recursive algorithm in our previous research that can achieve reasonably good noise estimation in practical environments (Chen *et al.*, 2006). However, in this paper, we will focus on illustrating the basic ideas while setting aside the noise estimation issues. So, we will not use any noise estimator in the following experiments. Instead, we directly compute the noise statistics from the noise signal. Specifically, at each time instant k, the matrix $\mathbf{R}_{\mathbf{y}}$ (its size is in the range between 4×4 and 80×80) is computed using the most recent 400 samples (50 ms long) of the noisy signal with a short-time average. The matrix $\mathbf{R}_{\mathbf{v}}$ is also computed using a short-time average; but noise is in general stationary (except for the competing speech case where $\mathbf{R}_{\mathbf{y}}$ and $\mathbf{R}_{\mathbf{v}}$ are computed in the same way), so we use 640 samples (80 ms long) to compute $\mathbf{R}_{\mathbf{v}}$. Then the $\hat{\mathbf{R}}_{\mathbf{x}}$ matrix is computed according to $\hat{\mathbf{R}}_{\mathbf{x}} = \hat{\mathbf{R}}_{\mathbf{y}} - \hat{\mathbf{R}}_{\mathbf{v}}$, and $\hat{\boldsymbol{\gamma}}_x$ is calculated using (19).

B. Comparison between the traditional and new performance measures

In this experiment, we compare the traditional performance measures defined based on the filtered speech $x_{\text{f}}(k)$ with the new performance measures defined using the filtered desired signal $x_{\text{fd}}(k)$ and interference $x'_{\text{ri}}(k)$. The Wiener filter given in (51) is used in this experiment, which is the same for both the traditional definition of the error signal shown in (7) and the new decomposition of the error signal given in (26) (since the Wiener filter minimizes the overall MSE, which is not affected by any decomposition form of the error signal). Specifically, at each time instant k, we first compute the correlation matrix $\hat{\mathbf{R}}_{\mathbf{y}}$, the correlation vector $\hat{\boldsymbol{\gamma}}_x$, and the signal variance $\hat{\sigma}_x^2$ as described in the previous subsection. A Wiener filter is then constructed according to (51). Applying this Wiener filter to $y(k)$, $x(k)$, and $v(k)$, we obtain the enhanced signal $\hat{x}(k)$, the filtered signal $x_{\text{f}}(k)$, the filtered desired signal $x_{\text{fd}}(k)$, the interference $x'_{\text{ri}}(k)$, and the residual noise $v_{\text{rn}}(k)$. To illustrate the importance of separating the filtered signal into the filtered desired signal and interference, we computed the ISDs between the clean speech $x(k)$ and the signals $\hat{x}(k)$, $x_{\text{f}}(k)$, $x_{\text{fd}}(k)$, and $x'_{\text{ri}}(k)$ obtained with the Wiener filter. The results as a function of the filter length L for the white Gaussian noise case with iSNR = 10 dB are plotted in Fig. 3(a). It is seen that the ISD between the clean speech $x(k)$ and the filtered desired signal $x_{\text{fd}}(k)$ is approximately 0, indicating that these two signals are almost the same. In comparison, the ISD between $x(k)$ and $x'_{\text{ri}}(k)$ is very large, which shows that these two signals are significantly different in spectrum. Therefore, $x'_{\text{ri}}(k)$ should not be treated as part of the desired signal after filtering, which verifies the necessity of separating the filtered signal into the filtered desired signal and residual interference. The ISD between the clean speech and the filtered signal $x_{\text{f}}(k)$ is larger than that between the clean speech and the filtered desired signal $x_{\text{fd}}(k)$; but it is significantly smaller than the ISD between $x(k)$ and $x'_{\text{ri}}(k)$. This shows that $x_{\text{fd}}(k)$ is the dominant component in $x_{\text{f}}(k)$ while the intensity of $x'_{\text{ri}}(k)$ is

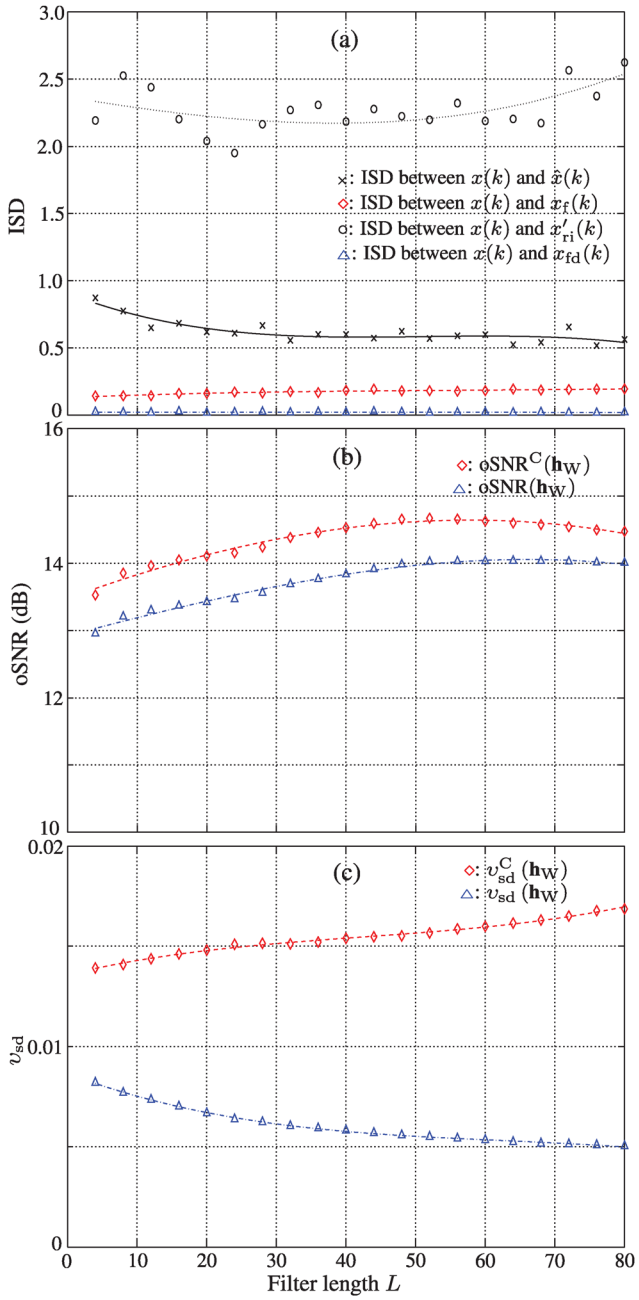

FIG. 3. (Color online) Comparison between the traditional performance measures [based on $x_{\rm f}(k)$] and the new performance measures [based on $x_{\rm fd}(k)$ and $x'_{\rm ri}(k)$] with the Wiener filter. The white Gaussian noise is used and the input SNR is 10 dB.

much lower than that of $x_{\rm fd}(k)$, which is desired in noise reduction. It is noticed that the filter length L does not affect much the ISD between $x(k)$ and $x_{\rm fd}(k)$. But the ISD between $x(k)$ and $x_{\rm f}(k)$ slightly increases with L. This variation is mainly caused by the residual interference. The ISD between the clean speech and its estimate, $\hat{x}(k)$, first decreases as L increases up to 30; but it does not change much if we further increase L. This confirms the results reported in Chen *et al.* (2006) and indicates that for the time-domain Wiener filter with an 8 kHz sampling rate, 30 is a sufficient filter length and, a larger length will not significantly improve the speech quality but would dramatically increase the complexity of the algorithm.

With the new decomposition of the error signal, the output SNR [given in (33)] is defined as the ratio of the intensity of the filtered desired signal over the intensity of the residual interference-plus-noise. Traditionally, however, the whole filtered signal $x_{\rm f}(k)$ is treated as the desired signal after noise reduction, so the output SNR of the Wiener filter is in the following form

$$\text{oSNR}^{\rm C}(\mathbf{h}_{\rm W}) = \frac{\sigma^2_{x_{\rm f}}}{\sigma^2_{v_{\rm rn}}}, \tag{87}$$

where, again, we use the superscript$^{\rm C}$ to indicate the "classical" definition. Both oSNR$(\mathbf{h}_{\rm W})$ and oSNR$^{\rm C}(\mathbf{h}_{\rm W})$ are plotted in Fig. 3(b) as a function of the filter length L. The trends between the two versions of the output SNR and the filter length L are similar. But the new definition should be more accurate as the residual interference is not treated as part of the desired speech signal.

Also plotted in Fig. 3 is the speech distortion index defined in (44). For the purpose of comparison, we also showed the "classical" definition of this index, which is given by

$$v^{\rm C}_{\rm sd}(\mathbf{h}_{\rm W}) = \frac{E\{[x_{\rm f}(k) - x(k)]^2\}}{\sigma^2_x}. \tag{88}$$

It is seen that $v_{\rm sd}(\mathbf{h}_{\rm W})$ decreases with L (rapidly for small L values). But with the classical definition, the speech distortion index increases with L, which is similar to the ISD between $x(k)$ and $x_{\rm f}(k)$ in Fig. 3(a). This difference between the two indices is due to the residual interference.

We also studied the case of the NYSE babble noise. The results are shown in Fig. 4. Again, the appropriateness of the new performance measures is verified.

C. Comparison between the Wiener and MVDR filters

With the new decomposition of the error signal, we have shown that it is now possible to derive an MVDR filter for single-channel noise reduction. The difference between the MVDR and Wiener filters is a scaling factor, which is given in (66). If this scaling factor is time-invariant, the Wiener and MVDR filters have the same performance. However, in speech applications, the desired speech signal is always nonstationary and noise statistics may change with time. As a result, the scaling factor is in general time-varying, which can cause some performance difference between the two filters. This subsection studies the difference between the MVDR and Wiener filters through experiments. Based on the previous experiment, we set the filter length L to 20. White Gaussian noise is used and the correlation matrix $\hat{\mathbf{R}}_{\mathbf{y}}$ is computed using the method described in Sec. VIII A. However, unlike the previous experiment, here we directly compute $\hat{\boldsymbol{\gamma}}_x$ and $\hat{\sigma}^2_x$ from the signal $x(k)$ using a same short-time average as $\hat{\mathbf{R}}_{\mathbf{y}}$. We then estimate the scaling factor between the MVDR and Wiener filters according to (66). The first 3 s of the noisy speech (with iSNR = 10 dB) and the computed scaling factor are plotted in Fig. 5 (in both the linear and dB scales). It is seen that the value of the scaling

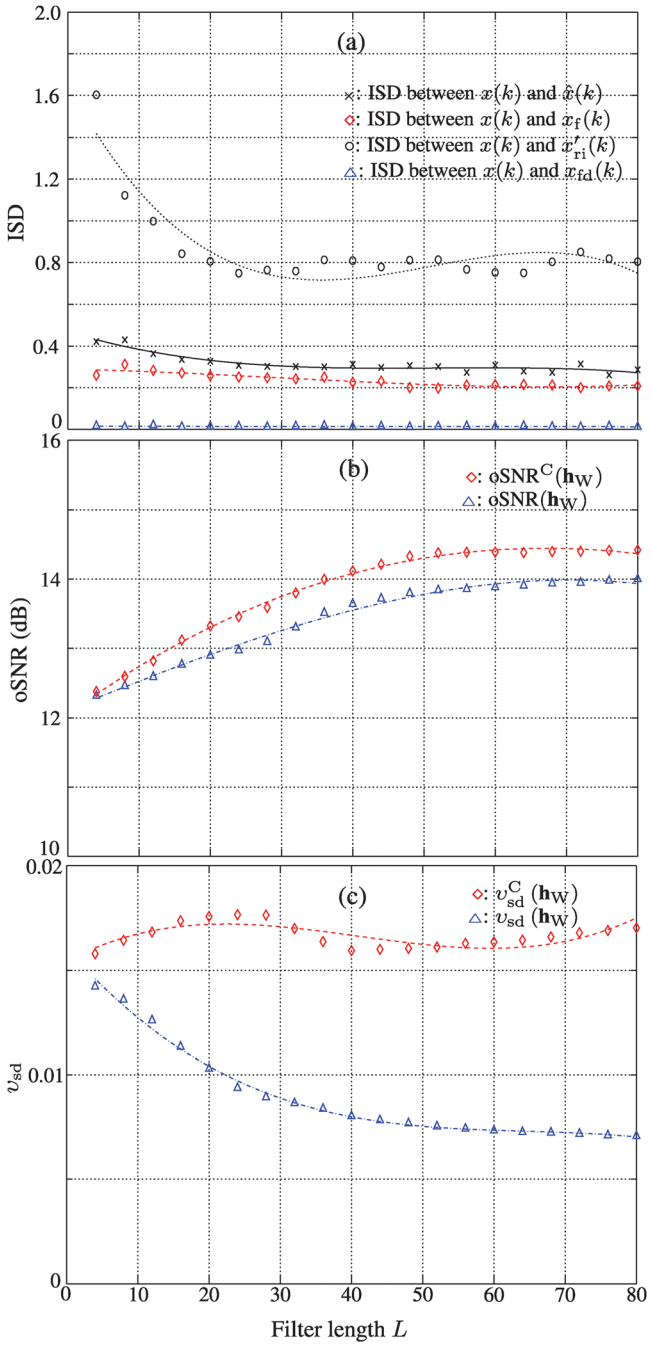

FIG. 4. (Color online) Comparison between the traditional performance measures [based on $x_{\rm f}(k)$] and the new performance measures [based on $x_{\rm fd}(k)$ and $x'_{\rm ri}(k)$] with the Wiener filter. The NYSE noise is used and the input SNR is 10 dB.

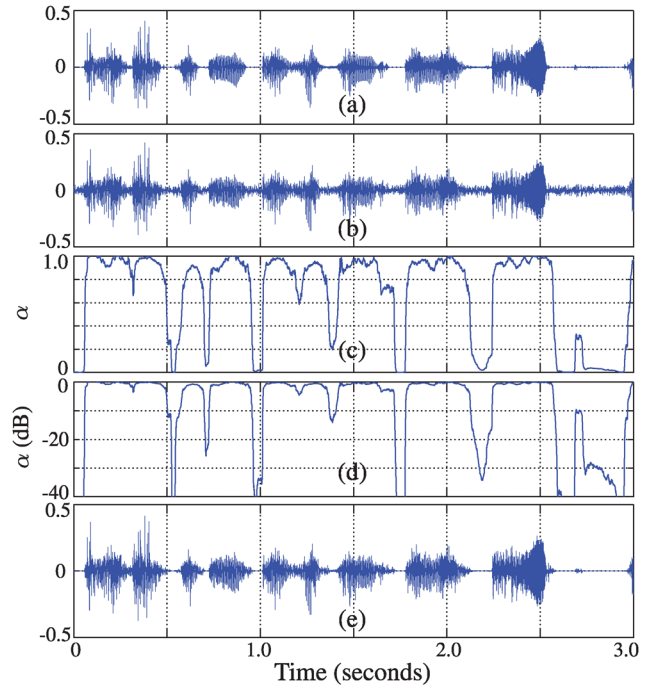

FIG. 5. (Color online) (a) The clean speech waveform, (b) the noisy speech waveform, (c) the scaling factor α between the Wiener and MVDR filters given in (66), (d) the scaling factor in the dB scale, and (e) the noisy speech multiplied by the scaling factor. The white Gaussian noise is used with iSNR = 10 dB and $L = 20$.

factor is large (close to 1) during the presence of speech; but it is very small (close to 0) in silence periods. Figure 5(d) plots the noisy speech multiplied with the scaling factor. It is seen that the noise in silence periods is significantly attenuated while the noise level in the presence of speech remains almost unchanged. This indicates that the Wiener filter is more aggressive in suppressing silence periods while it behaves almost the same as the MVDR filter during the presence of speech.

The performance results for this experiment are sketched in Fig. 6. One can notice that speech distortion either measured by the ISD or by the speech distortion index is zero with the MVDR filter. When the SNR is above 15 dB, the MVDR and Wiener filters have almost the same performance. However, as the SNR decreases, the Wiener filter tends to have more noise reduction, but it has more speech distortion as well. It should be noted that all the performance measures shown in Fig. 6 are computed globally with the use of all the signal samples. If we evaluate the measures on a short-time basis, the two filters would have similar output SNRs during the presence of speech. The reason that the Wiener filter achieves a higher global output SNR is that it suppresses more noise during the absence of speech. This, however, causes some discontinuity in the residual noise level, which is unpleasant to listen to and should be avoided in practice.

Experiments using the NYSE noise were also conducted and the performance difference between the two filters is similar to that in the white Gaussian noise case.

D. The tradeoff filter

The tradeoff filter derived in Sec. VI C introduces a nonnegative parameter μ to control the residual noise level. When $\mu = 0$, the tradeoff filter degenerates to the MVDR filter. This experiment is to investigate the effect of μ on the output SNR and speech distortion. Similar to the MVDR and Wiener filters, we need to know the matrix $\hat{\mathbf{R}}_{\mathbf{y}}$, the vector $\hat{\boldsymbol{\gamma}}_x$, and the signal variance $\hat{\sigma}_x^2$. Again, these parameters are computed using the method described in Sec. VIII A. The results for both the white Gaussian and NYSE noise cases are shown in Fig. 7. It is seen that both the output SNR and speech distortion index increase as μ increases. Theoretically at each time instant k, increasing μ should not affect the output SNR. But the value of the parameter μ controls how

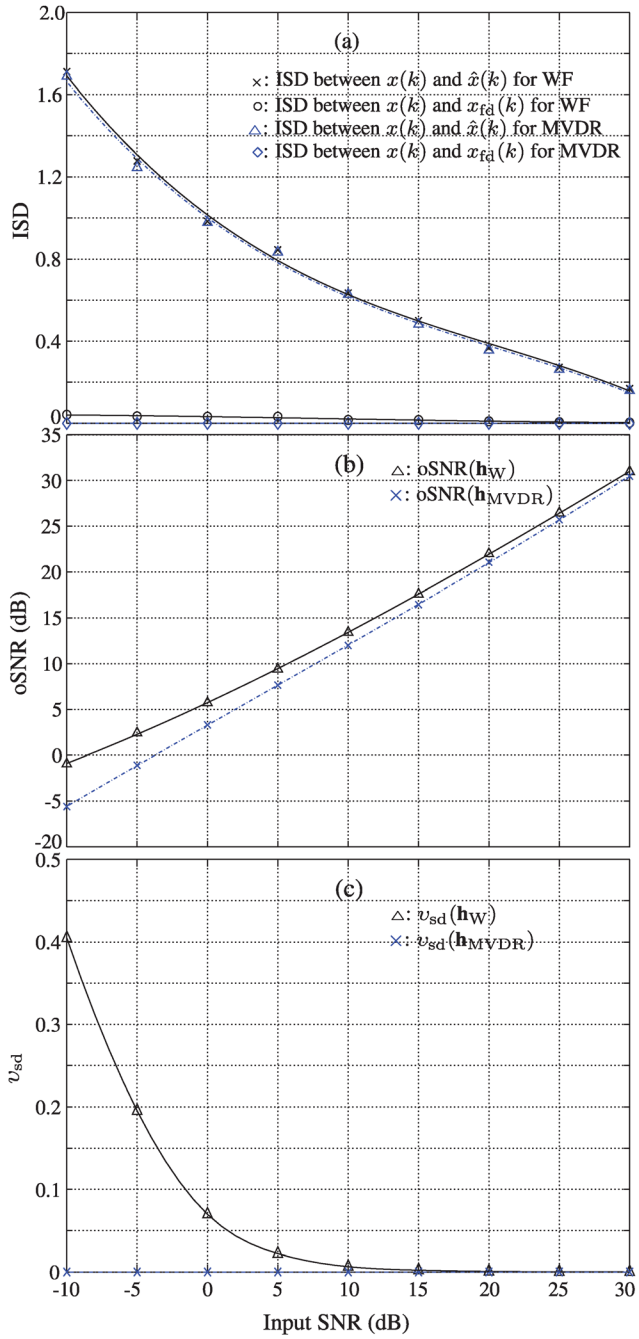

FIG. 6. (Color online) Comparison between the Wiener and MVDR filters in the white Gaussian noise case with $L = 20$.

aggressive the filter suppresses silence periods where the desired speech is absent. A larger value of μ indicates that the filter is more aggressive in suppressing silence periods. So, when we evaluate the output SNR globally, we have more SNR gain for a larger value of μ.

It is noticed that in the NYSE noise case, the output SNR is lower and the speech distortion index is larger. This is due to the fact that this noise is nonstationary and, hence, more difficult to deal with than with the white Gaussian noise.

E. The LCMV filter

The LCMV filter is derived based on the constraints that the desired speech ought to be perfectly recovered while the

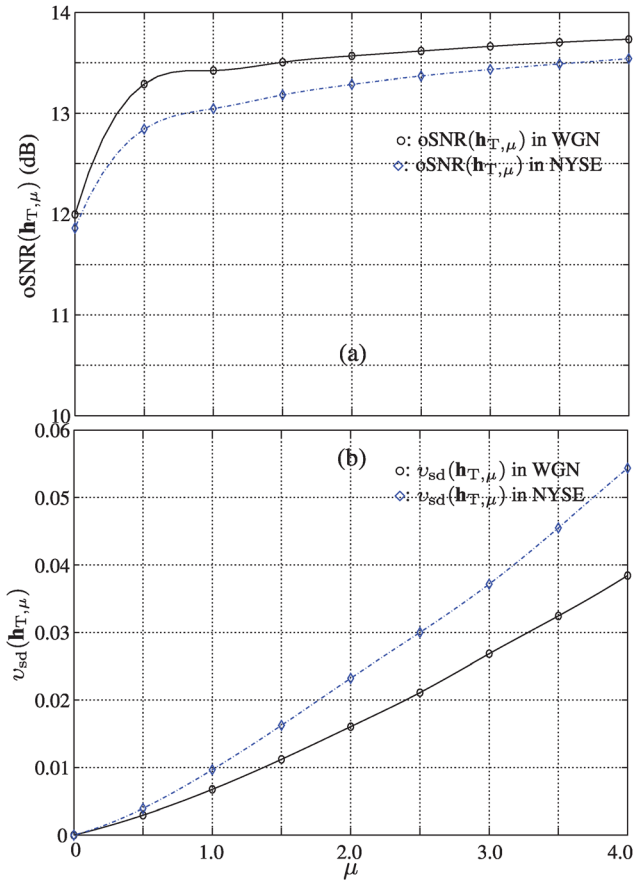

FIG. 7. (Color online) Performance of the tradeoff filter as a function of μ. The input SNR is 10 dB and $L = 20$.

correlated components in noise (if any) should be removed. In other words, the LCMV filter needs to meet two constraints simultaneously, i.e., $\boldsymbol{\gamma}_x^T\mathbf{h} = 1$ and $\boldsymbol{\gamma}_v^T\mathbf{h} = 0$. Consequently, the filter length of the LCMV filter should be much longer than that of the MVDR filter since the latter only needs to satisfy $\boldsymbol{\gamma}_x^T\mathbf{h} = 1$ while minimizing the interference-plus-noise. The performance of the LCMV filter depends not only on the filter length, but also on the degree of speech and noise self correlation. In this experiment, we investigate the LCMV filter in three different noise backgrounds: white Gaussian noise, NYSE noise, and speech from a competing talker. In the first case, the noise samples are completely uncorrelated. Therefore, we have $\boldsymbol{\gamma}_v = \mathbf{i}_0$. Forcing $\boldsymbol{\gamma}_v^T\mathbf{h} = 0$ in this case means h_0 should be 0, which implies that the current speech sample $x(k)$ is completely predicted from the previous $L-1$ samples. In the NYSE noise case, there will be some but weak correlation among neighboring samples, while in the competing speech case, the correlation between signal samples can be very strong. The results of this experiment are shown in Fig. 8. The input SNR is 10 dB, and again the noisy correlation matrix and the speech and noise correlation vectors are directly computed from the noisy, clean, and noise signals. It is seen that the speech distortion index in the three conditions is very small (of the order of 10^{-14}), indicating that the desired speech signal is estimated without speech distortion. The output SNR in three conditions increases with the filter length L. When L is reasonably large (>50), slightly more SNR gain is achieved in the NYSE noise case than in the

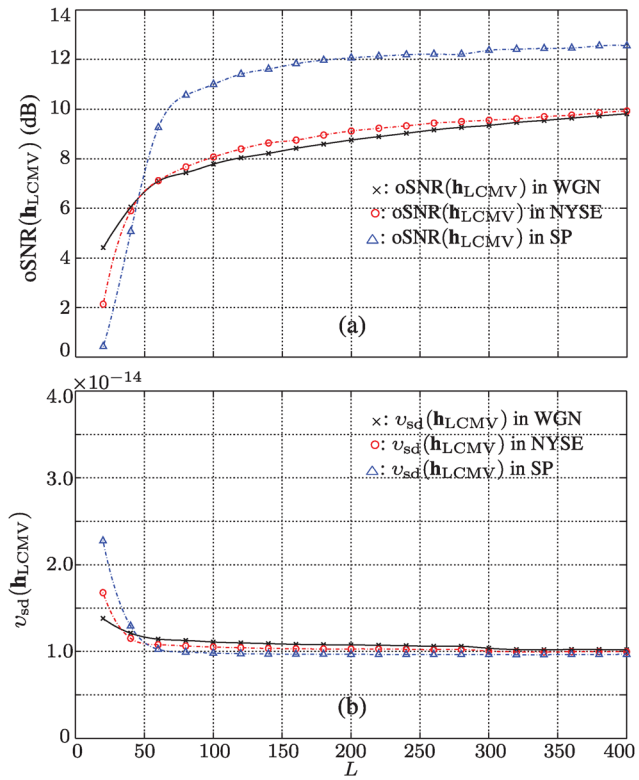

FIG. 8. (Color online) Performance of the LCMV filter as a function of L in difference noise conditions with an input SNR of 10 dB.

white Gaussian noise case, while the output SNR for the competing speech situation is significantly higher than those in the NYSE and Gaussian noise cases. This coincides with the theoretical analysis that the LCMV is designed to remove correlated components in noise. The stronger the noise self correlation, the higher the SNR improvement.

It is also noticed that for the white Gaussian and NYSE noise cases, the output SNR is lower than the input SNR. This indicates that the LCMV filter boosts the uncorrelated components of noise while removing its correlated components. This problem is more serious when the filter length is short. One way to circumvent this issue is to put a penalty term in the cost function so as the filter does not amplify the uncorrelated noise components, i.e.,

$$\mathbf{h}_{\mathrm{LCMV}} = \arg\min_{\mathbf{h}}(\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h} + \delta\mathbf{h}^T\mathbf{h}) \quad \text{subject to} \quad \mathbf{C}^T\mathbf{h} = \mathbf{i}, \tag{89}$$

where δ is a positive constant that controls how strongly we impose the penalty. The solution to (89) is given by

$$\mathbf{h}_{\mathrm{LCMV}} = (\mathbf{R}_{\mathbf{y}} + \delta\mathbf{I})^{-1}\mathbf{C}[\mathbf{C}^T(\mathbf{R}_{\mathbf{y}} + \delta\mathbf{I})^{-1}\mathbf{C}]^{-1}\mathbf{i}. \tag{90}$$

Comparing (90) with (79), one can see that adding a penalty term is identical to putting a regularization parameter when computing the inverse of the noisy correlation matrix. By choosing a proper value of this regularization, the LCMV filter can be more robust to the uncorrelated noise component. But we should note that, unlike the MVDR filter, the LCMV is not designed to reduce the self uncorrelated noise. So, no matter how we control the regularization factor, we should not expect much SNR improvement if the noise is white.

IX. CONCLUSIONS

This paper studied the noise reduction problem in the time domain. We presented a new way to decompose the clean speech vector into two orthogonal components: one is correlated and the other is uncorrelated with the current clean speech sample. While the correlated component helps estimate the clean speech, the uncorrelated component interferes with the estimation, just as the additive noise. With this new decomposition, we discussed how to redefine the error signal and form different cost functions and addressed the issue of how to design different optimal noise reduction filters by optimizing these new cost functions. We showed that with the redefined error signal, the maximum SNR filter is equivalent to the widely known Wiener filter. We demonstrated that it is possible to derive an MVDR filter that can reduce noise without adding any speech distortion in the single-channel case. This new MVDR filter is different from the Wiener filter only by a scaling factor, where by adjusting this scaling factor, the Wiener filter tends to be more aggressive in suppressing noise during silence periods, which can cause significant discontinuity in the residual noise level that is unpleasant to listen to. We also showed that an LCMV filter can be developed to remove correlated components in noise without adding distortion to the desired speech signal. Furthermore, several performance measures have been defined based on the new orthogonal decomposition of the clean speech vector, which are more appropriate than the traditional ones for the evaluation of the noise reduction filters.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (**2009**). *Noise Reduction in Speech Processing* (Springer-Verlag, Berlin), pp. 1–229.

Benesty, J., Chen, J., Huang, Y., and Doclo, S. (**2005**). "Study of the wiener filter for noise reduction," in *Speech Enhancement*, edited by J. Benesty, S. Makino, and J. Chen (Springer-Verlag, Berlin), Chap. 2, pp. 9–41.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-27**, 113–120.

Capon, J. (**1969**). "High resolution frequency-wavenumber spectrum analysis," Proc. IEEE **57**, 1408–1418.

Chen, G., Koh, S. N., and Soon, I. Y. (**2003**). "Enhanced itakura measure incorporating masking properties of human auditory system," Signal Process. **83**, 1445–1456.

Chen, J., Benesty, J., and Huang, Y. (**2009**). "Study of the noise-reduction problem in the karhunen-loève expansion domain," IEEE Trans. Audio, Speech, Language Process. **17**, 787–802.

Chen, J., Benesty, J., Huang, Y., and Doclo, S. (**2006**). "New insights into the noise reduction wiener filter," IEEE Trans. Audio, Speech, Language Process. **14**, 1218–1234.

Cohen, I., Benesty, J., and Gannot, S., eds. (**2010**). *Speech Processing in Modern Communication–Challenges and Perspectives* (Springer, Berlin), pp. 1–342.

Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-32**, 1109–1121.

Er, M., and Cantoni, A. (**1983**). "Derivative constraints for broad-band element space antenna array processors," IEEE Trans. Acoust., Speech, Signal Process. **31**, 1378–1393.

Frost, O. (**1972**). "An algorithm for linearly constrained adaptive array processing" Proc. IEEE **60**, 926–935.

Huang, Y., Benesty, J., and Chen, J. (**2008**). "Analysis and comparison of multichannel noise reduction methods in a common framework," IEEE Trans. Audio, Speech, Language Process. **16**, 957–968.

Itakura, F., and Saito, S. (**1970**). "A statistical method for estimation of speech spectral density and formant frequencies," Electron. Commun. Japan **53A**, 36–43.
Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), pp. 1–585.
Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech, Audio Process. **9**, 504–512.
Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (**1988**). *Objective Measures of Speech Quality* (Prentice-Hall, Englewood Cliffs, NJ), pp. 1–355.
Vary, P. (**1985**). "Noise suppression by spectral magnitude estimation–mechanism and theoretical limits," Signal Process. **8**, 387–400.
Vary, P., and Martin, R. (**2006**). *Digital Speech Transmission: Enhancement, Coding and Error Concealment* (John Wiley and Sons, Chichester, England), pp. 1–625.