ON SINGLE-CHANNEL NOISE REDUCTION IN THE TIME DOMAIN

Jingdong Chen[1], *Jacob Benesty*[2], *Yiteng (Arden) Huang*[1], *and Tomas Gaensler*[3]

[1]: WeVoice, Inc.
1065 Route 22 West, Suite 2E
Bridgewater, NJ 08807, USA

[2]: INRS-EMT, University of Quebec
800 de la Gauchetiere Ouest, Suite 6900
Montreal, QC H5A 1K6, Canada

[3]: mh acoustics LLC
25A Summit Avenue
Summit, NJ 07901, USA

ABSTRACT

In this paper, we revisit the noise-reduction problem in the time domain and present a way to decompose the filtered speech into two uncorrelated (orthogonal) components: the desired speech and the interference. Based on this new decomposition, we discuss how to form different optimization cost functions and address the issue of how to design different noise-reduction filters by optimizing these new cost functions. Particularly, we cover the design of the maximum signal-to-noise-ratio (SNR), the Wiener, the minimum variance distortionless response (MVDR), and the tradeoff filters. It is interesting that with this new decomposition, we can now design the MVDR filter that can achieve noise reduction without adding speech distortion in the single-channel case, which has never been seen before. We also demonstrate that the maximum SNR, Wiener, and tradeoff filters are identical to the MVDR filter up to a scaling factor. From a theoretical point of view, this scaling factor is not significant and should not affect the output SNR at any processing time. But from a practical viewpoint, the scaling factor can be time-varying due to the nonstationarity of the speech and possibly the noise and can cause discontinuity in the residual noise level, which is unpleasant to listen to. As a result, it is essential to have the scaling factor right from one processing sample (or frame) to another in order to avoid large distortions and for this reason, it is recommended to use the MVDR filter in speech enhancement applications.

Index Terms— Single-channel noise reduction, Wiener filter, maximum SNR filter, minimum variance distortionless response (MVDR) filter, tradeoff filter.

1. PROBLEM FORMULATION

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) [1], [2], [3]

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

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.

The signal model given in (1) can be put into a vector form:

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

where

$$\mathbf{y}(k) \triangleq \left[\begin{array}{cccc} y(k) & y(k-1) & \cdots & y(k-L+1) \end{array} \right]^T \tag{3}$$

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 an equivalent 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

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

where $E[\cdot]$ denotes mathematical expectation, and $\mathbf{R}_\mathbf{x} \triangleq E\left[\mathbf{x}(k)\mathbf{x}^T(k)\right]$ and $\mathbf{R}_\mathbf{v} \triangleq E\left[\mathbf{v}(k)\mathbf{v}^T(k)\right]$ 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 our discussion on the estimation of $x(k)$ only. In other words, we consider to estimate the desired signal on a sample-by-sample basis. It should be noted, though, that any approach developed here should be easily extended to the estimation of $\mathbf{x}(k)$. Specifically, an estimate of the desired signal sample $x(k)$ is obtained by applying a finite-impulse-response (FIR) filter to the observation signal vector $\mathbf{y}(k)$ [4], i.e.,

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

where

$$\mathbf{h} \triangleq \left[\begin{array}{cccc} h_0 & h_1 & \cdots & h_{L-1} \end{array} \right]^T \tag{6}$$

is an FIR filter of length L, $x_\mathrm{f}(k) \triangleq \mathbf{h}^T\mathbf{x}(k)$ is the filtered speech, and $v_\mathrm{rn}(k) \triangleq \mathbf{h}^T\mathbf{v}(k)$ is the residual noise. With this filtering model, the noise-reduction problem then becomes one of finding an "optimal" filter that can significantly reduce the additive noise while keeping the speech distortion due to the filter as small as possible. In order to find such an "optimal" filter, we need to define an error signal and a cost function.

Traditionally, the error signal for the estimator given in (5) is defined as

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

where

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

denotes 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 "C" to denote the classical methods. The mean-square error (MSE) is then

$$J\left(\mathbf{h}\right) \triangleq E\left[e^2(k)\right] = J_\mathrm{d}^\mathrm{C}\left(\mathbf{h}\right) + J_\mathrm{r}^\mathrm{C}\left(\mathbf{h}\right), \tag{10}$$

where

$$J_\mathrm{d}^\mathrm{C}\left(\mathbf{h}\right) \triangleq E\left[e_\mathrm{d}^2(k)\right] \tag{11}$$

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

and

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

Given the above 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.

It is seen that the filtered speech is treated as the desired signal in the definitions of the error signal and MSE in the classical methods. However, this definition of the desired speech incurs many problems for both the design and evaluation of noise-reduction filters. For example, the filter that maximizes the output SNR should intuitively be a good filter. However, we found that such a filter introduces too much speech distortion that renders it useless in practice.

In this paper, we present a new way to decompose the filtered speech into two uncorrelated (orthogonal) components: the desired speech and the interference. Specifically, since our desired signal at time k is only the sample $x(k)$, we can decompose the whole vector $\mathbf{x}(k)$ into the following form:

$$\mathbf{x}(k) = x(k)\boldsymbol{\gamma}_x + \mathbf{x}'(k) = \mathbf{x}_{\mathrm{d}}(k) + \mathbf{x}'(k), \tag{13}$$

where $\mathbf{x}_{\mathrm{d}}(k) \triangleq x(k)\boldsymbol{\gamma}_x$, $\mathbf{x}'(k) \triangleq \mathbf{x}(k) - x(k)\boldsymbol{\gamma}_x$,

$$\begin{aligned}\boldsymbol{\gamma}_x &= \begin{bmatrix} \gamma_{x,0} & \gamma_{x,1} & \cdots & \gamma_{x,L-1} \end{bmatrix}^T \\ &= \begin{bmatrix} 1 & \gamma_{x,1} & \cdots & \gamma_{x,L-1} \end{bmatrix}^T \\ &= \frac{E\left[x(k)\mathbf{x}(k)\right]}{E\left[x^2(k)\right]}\end{aligned} \tag{14}$$

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

$$\gamma_{x,l} \triangleq \frac{E\left[x(k)x(k-l)\right]}{E\left[x^2(k)\right]} \tag{15}$$

is the correlation coefficient between $x(k)$ and $x(k-l)$ with $-1 \leq \gamma_{x,l} \leq 1$. It is easy to see that $\mathbf{x}_{\mathrm{d}}(k)$ is correlated with $x(k)$, while $\mathbf{x}'(k)$ is orthogonal to $x(k)$, i.e.,

$$E\left[x(k)\mathbf{x}'(k)\right] = \mathbf{0}. \tag{16}$$

Substituting (13) into (5), we get

$$\begin{aligned}\hat{x}(k) &= \mathbf{h}^T\left[x(k)\boldsymbol{\gamma}_x + \mathbf{x}'(k) + \mathbf{v}(k)\right] \\ &= x_{\mathrm{fd}}(k) + x'_{\mathrm{ri}}(k) + v_{\mathrm{rn}}(k),\end{aligned} \tag{17}$$

where $x_{\mathrm{fd}}(k) \triangleq x(k)\mathbf{h}^T\boldsymbol{\gamma}_x$ is the filtered desired signal, $x'_{\mathrm{ri}}(k) \triangleq \mathbf{h}^T\mathbf{x}'(k)$ is the interference, and $v_{\mathrm{rn}}(k) \triangleq \mathbf{h}^T\mathbf{v}(k)$, as in the classical methods, represents the residual noise. 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

$$\sigma_{\hat{x}}^2 = \sigma_{x_{\mathrm{fd}}}^2 + \sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2, \tag{18}$$

where

$$\sigma_{x_{\mathrm{fd}}}^2 = \sigma_x^2\left(\mathbf{h}^T\boldsymbol{\gamma}_x\right)^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\mathbf{h}, \tag{19}$$

$$\sigma_{x'_{\mathrm{ri}}}^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{x}'}\mathbf{h} = \mathbf{h}^T\mathbf{R}_{\mathbf{x}}\mathbf{h} - \sigma_x^2\left(\mathbf{h}^T\boldsymbol{\gamma}_x\right)^2, \tag{20}$$

$$\sigma_{v_{\mathrm{rn}}}^2 = \mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h}, \tag{21}$$

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

Comparing (17) with (5), one can see the difference between the traditional and new definitions of the desired signal after filtering. Specifically, the whole filtered speech $x_{\mathrm{f}}(k) = \mathbf{h}^T\mathbf{x}(k)$ is treated as the desired speech in the traditional methods, while in our new formulation, $x_{\mathrm{fd}}(k) = x(k)\mathbf{h}^T\boldsymbol{\gamma}_x$ is considered as the desired speech. Notice that in the new formulation, there appears an interference term after filtering. This term should be treated as noise and should be minimized.

Now, the error signal between the estimated and desired signals can be defined as

$$e(k) \triangleq \hat{x}(k) - x(k) = e_{\mathrm{d}}(k) + e_{\mathrm{r}}(k), \tag{22}$$

where

$$e_{\mathrm{d}}(k) \triangleq x_{\mathrm{fd}}(k) - x(k) \tag{23}$$

is the signal distortion due to the FIR filter and

$$e_{\mathrm{r}}(k) \triangleq x'_{\mathrm{ri}}(k) + v_{\mathrm{rn}}(k) \tag{24}$$

represents the residual interference plus noise.

The MSE is then

$$J(\mathbf{h}) = E\left[e^2(k)\right] = J_{\mathrm{d}}(\mathbf{h}) + J_{\mathrm{r}}(\mathbf{h}), \tag{25}$$

where

$$J_{\mathrm{d}}(\mathbf{h}) = E\left[e_{\mathrm{d}}^2(k)\right] = \sigma_x^2\left(\mathbf{h}^T\boldsymbol{\gamma}_x - 1\right)^2 \tag{26}$$

and

$$J_{\mathrm{r}}(\mathbf{h}) = E\left[e_{\mathrm{r}}^2(k)\right] = \sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2. \tag{27}$$

It is clear that the objective of noise reduction 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. Comparing (22) with (7) and (25) with (10), one can clearly see the difference between the new decompositions of the error signal and MSE and their traditional decompositions.

2. OPTIMAL FILTERS

2.1. Maximum SNR filter

As far as noise reduction is concerned, it is desirable to design a filter that can maximize the SNR of the output signal. With the signal model given in (1), the input SNR is

$$\mathrm{iSNR} \triangleq \frac{\sigma_x^2}{\sigma_v^2}, \tag{28}$$

where $\sigma_v^2 \triangleq E\left[v^2(k)\right]$ is the variance of the noise.

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, i.e.,

$$\mathrm{oSNR}(\mathbf{h}) = \frac{\sigma_{x_{\mathrm{fd}}}^2}{\sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2} = \frac{\sigma_x^2\left(\mathbf{h}^T\boldsymbol{\gamma}_x\right)^2}{\mathbf{h}^T\mathbf{R}_{\mathrm{in}}\mathbf{h}} = \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathrm{in}}\mathbf{h}}, \tag{29}$$

where

$$\mathbf{R}_{\mathrm{in}} = \mathbf{R}_{\mathbf{x}'} + \mathbf{R}_{\mathbf{v}} \tag{30}$$

is the interference plus noise covariance matrix. The term in the most right-hand side of (29) is known as the generalized Rayleigh quotient. So the filter that maximizes the output SNR is

$$\mathbf{h}_{\max} = \mathbf{q}_{\max}\left(\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}\right), \tag{31}$$

where $\mathbf{q}_{\max}\left(\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}\right)$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}$, i.e., $\lambda_{\max}\left(\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}\right)$. With this filter, the output SNR is

$$\text{oSNR}\left(\mathbf{h}_{\max}\right) = \text{oSNR}_{\max} = \lambda_{\max}\left(\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}\right). \tag{32}$$

Since the rank of the matrix $\mathbf{R}_{\mathbf{x}_{\text{d}}}$ is equal to 1, we also have

$$\text{oSNR}_{\max} = \text{tr}\left[\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}\right] = \sigma_x^2 \boldsymbol{\gamma}_x^T \mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x, \tag{33}$$

where tr[·] denotes the trace of a square matrix. The quantity $\text{oSNR}_{\max}$ corresponds to the maximum SNR that can be achieved through filtering. As a result, we have

$$\text{oSNR}\left(\mathbf{h}\right) \leq \text{oSNR}_{\max}, \ \forall \mathbf{h} \tag{34}$$

and

$$\text{oSNR}\left(\mathbf{h}_{\max}\right) = \text{oSNR}_{\max} \geq \text{iSNR}. \tag{35}$$

2.2. Wiener

The Wiener filter is easily derived by taking the gradient of the MSE, i.e., $J\left(\mathbf{h}\right)$ defined in (25), with respect to $\mathbf{h}$ and equating the result to zero:

$$\mathbf{h}_{\text{W}} = \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}\mathbf{i}_0 = \left[\mathbf{I} - \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\right]\mathbf{i}_0, \tag{36}$$

where $\mathbf{I}$ is the identity matrix of size $L \times L$ and $\mathbf{i}_0$ corresponds to the first column of $\mathbf{I}$. Since

$$\mathbf{R}_{\mathbf{x}}\mathbf{i}_0 = \sigma_x^2 \boldsymbol{\gamma}_x, \tag{37}$$

we can rewrite (36) as

$$\mathbf{h}_{\text{W}} = \sigma_x^2 \mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x. \tag{38}$$

From Section 1, it is easy to verify that

$$\mathbf{R}_{\mathbf{y}} = \sigma_x^2 \boldsymbol{\gamma}_x \boldsymbol{\gamma}_x^T + \mathbf{R}_{\text{in}}. \tag{39}$$

Determining the inverse of $\mathbf{R}_{\mathbf{y}}$ from (39) with the Woodbury's identity

$$\mathbf{R}_{\mathbf{y}}^{-1} = \mathbf{R}_{\text{in}}^{-1} - \frac{\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T\mathbf{R}_{\text{in}}^{-1}}{\sigma_x^{-2} + \boldsymbol{\gamma}_x^T\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x} \tag{40}$$

and substituting the result into (38), we get another interesting formulation of the Wiener filter:

$$\mathbf{h}_{\text{W}} = \frac{\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x}{\sigma_x^{-2} + \boldsymbol{\gamma}_x^T\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x}, \tag{41}$$

that we can rewrite as

$$\begin{aligned} \mathbf{h}_{\text{W}} &= \frac{\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{y}} - \mathbf{I}}{1 - L + \text{tr}\left[\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{y}}\right]}\mathbf{i}_0 \\ &= \frac{\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}}{1 + \text{oSNR}_{\max}}\mathbf{i}_0. \end{aligned} \tag{42}$$

Using (41), we deduce that the output SNR is

$$\text{oSNR}(\mathbf{h}_{\text{W}}) = \text{oSNR}_{\max} = \text{tr}\left[\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{y}}\right] - L.$$

So, the Wiener filter maximizes the output SNR. The two filters $\mathbf{h}_{\text{W}}$ and $\mathbf{h}_{\max}$ are equivalent (different only by a scaling factor).

2.3. Minimum Variance Distortionless Response

The celebrated minimum variance distortionless response (MVDR) filter proposed by Capon [5] is usually derived in a context where we have at least two sensors (microphones) available. Interestingly, with the new formulation, we can also derive the MVDR (with one sensor only) by minimizing the MSE of the residual interference plus noise, $J_{\text{r}}\left(\mathbf{h}\right)$, with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\text{in}}\mathbf{h} \quad \text{subject to} \quad \mathbf{h}^T\boldsymbol{\gamma}_x = 1. \tag{43}$$

The solution to the above optimization problem is

$$\mathbf{h}_{\text{MVDR}} = \frac{\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x}, \tag{44}$$

which can also be written as

$$\mathbf{h}_{\text{MVDR}} = \frac{\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{y}} - \mathbf{I}}{\text{tr}\left[\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{y}}\right] - L}\mathbf{i}_0 = \frac{\mathbf{R}_{\text{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\text{d}}}}{\text{oSNR}_{\max}}\mathbf{i}_0. \tag{45}$$

Obviously, we can rewrite the MVDR as

$$\mathbf{h}_{\text{MVDR}} = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x}. \tag{46}$$

The Wiener and MVDR filters are simply related as follows

$$\mathbf{h}_{\text{W}} = \alpha_{\mathbf{h}_{\text{W}}}\mathbf{h}_{\text{MVDR}}, \tag{47}$$

where

$$\alpha_{\mathbf{h}_{\text{W}}} = \mathbf{h}_{\text{W}}^T\boldsymbol{\gamma}_x = \frac{\text{oSNR}_{\max}}{1 + \text{oSNR}_{\max}}. \tag{48}$$

So, the two filters $\mathbf{h}_{\text{W}}$ and $\mathbf{h}_{\text{MVDR}}$ are equivalent up to a scaling factor. It is easy to see that $0 \leq \alpha_{\mathbf{h}_{\text{W}}} \leq 1$. On a short-time basis, a scaling factor does not affect the SNR. So, we have

$$\text{oSNR}\left(\mathbf{h}_{\text{MVDR}}\right) = \text{oSNR}\left(\mathbf{h}_{\text{W}}\right). \tag{49}$$

However, the scaling factor is in general time-varying due to the nonstationarity of the speech and possibly the noise. It acts as a weighting process that puts more attenuation in silence periods where the desired speech is absent and less attenuation when speech is present. As a result, The Wiener filter may have a higher output SNR if we evaluate the SNR on a long-time basis. But this weighting process can also cause discontinuity in the residual noise level, which is unpleasant to listen to and should be avoided in practice.

2.4. Tradeoff

In the tradeoff approach, we try to compromise between the amount of noise reduction and the degree of speech distortion. Instead of minimizing the MSE as we already did to find the Wiener filter, we could minimize $J_{\text{d}}\left(\mathbf{h}\right)$ with the constraint that the residual interference plus noise is less than the original noise level. Mathematically, this is equivalent to

$$\min_{\mathbf{h}} J_{\text{d}}\left(\mathbf{h}\right) \quad \text{subject to} \quad J_{\text{r}}\left(\mathbf{h}\right) = \beta\sigma_v^2, \tag{50}$$

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:

$$\mathbf{h}_{\text{T},\mu} = \sigma_x^2\left[\sigma_x^2\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T + \mu\mathbf{R}_{\text{in}}\right]^{-1}\boldsymbol{\gamma}_x = \frac{\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x}{\mu\sigma_x^{-2} + \boldsymbol{\gamma}_x^T\mathbf{R}_{\text{in}}^{-1}\boldsymbol{\gamma}_x}, \tag{51}$$

where the Lagrange multiplier, μ, satisfies $J_{\text{r}}\left(\mathbf{h}_{\text{T},\mu}\right) = \beta\sigma_v^2$. Taking $\mu = 1$, we obtain the Wiener filter while for $\mu = 0$, we get the MVDR. By playing on the value of μ, we can make a compromise between noise reduction and speech distortion. Again, we observe here as well that the tradeoff and MVDR filters are equivalent up to a scaling factor, i.e.,

$$\mathbf{h}_{\mathrm{T},\mu} = \alpha_{\mathbf{h}_{\mathrm{T},\mu}} \mathbf{h}_{\mathrm{MVDR}}, \tag{52}$$

where

$$\alpha_{\mathbf{h}_{\mathrm{T},\mu}} = \frac{\alpha_{\mathbf{h}_{\mathrm{W}}}}{\alpha_{\mathbf{h}_{\mathrm{W}}} + \mu\left(1 - \alpha_{\mathbf{h}_{\mathrm{W}}}\right)} = \frac{\mathrm{oSNR}_{\max}}{\mu + \mathrm{oSNR}_{\max}}. \tag{53}$$

Again, we have $0 \leq \alpha_{\mathbf{h}_{\mathrm{T},\mu}} \leq 1$. 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 that of the MVDR filter, i.e.,

$$\mathrm{oSNR}\left(\mathbf{h}_{\mathrm{T},\mu}\right) = \mathrm{oSNR}\left(\mathbf{h}_{\mathrm{MVDR}}\right), \ \forall \mu. \tag{54}$$

3. EXPERIMENTAL RESULTS

In this section, we use experiments to illustrate the relationship between the Wiener, MVDR, and tradeoff filters. Note that the maximum SNR filter is a unit vector, which has to be properly scaled according to the signal level at every time instant k, which will not be discussed here.

The clean speech signal used in the experiments was recorded from a female speaker in a quiet office room. It was sampled at 8 kHz. The overall length of the signal is 30 seconds. The noisy speech is obtained by adding a car noise signal (recorded in a car running at 50 miles/hour on a high way) to the clean speech and the noise signal is properly scaled to control the input SNR to 10 dB. The first 5 seconds of the clean and noisy signals are shown in Fig. 1 (a) and (b).

Implementation of the noise-reduction filters derived in Section 2 requires estimation of the correlation matrices $\mathbf{R_y}$, $\mathbf{R_x}$, and $\mathbf{R_v}$, the correlation vector $\boldsymbol{\gamma}_x$, and the signal variance σ_x^2. Computation of $\mathbf{R_y}$ is relatively easy because the noisy signal $y(k)$ is accessible. But in practice, we need a noise estimator or voice activity detector (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 [4]. 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 our experiments. Instead, we directly compute the noise statistics from the noise signal. Specifically, we set the filter length L to 20. At each time instance k, the matrix $\mathbf{R_y}$ is computed using the most recent 480 samples (60-ms long) of the noisy signal with a short-time average. The matrix $\mathbf{R_v}$ is also computed using a short-time average; but noise is more stationary so we use 960 samples (120-ms long). Then all the other parameters are computed in the following way: $\hat{\mathbf{R}}_\mathbf{x} = \hat{\mathbf{R}}_\mathbf{y} - \hat{\mathbf{R}}_\mathbf{v}$, $\hat{\sigma}_x^2$ is taken as the first element of $\hat{\mathbf{R}}_\mathbf{x}$, and $\hat{\boldsymbol{\gamma}}_x$ is equal to the first column of $\hat{\mathbf{R}}_\mathbf{x}$ normalized by $\hat{\sigma}_x^2$.

With the computed covariance matrices and correlation vectors, we then constructed the Wiener, MVDR, and tradeoff (with $\mu = 0.01$) filters according to (41), (44), and (51) respectively. Figure 1(c) plots the estimated scaling factors $\alpha_{\mathbf{h}_{\mathrm{W}}}$ and $\alpha_{\mathbf{h}_{\mathrm{T},\mu}}$ in the dB scale. It is seen that the value of the two scaling factors is approximately 1 (0 dB) during the presence of speech while it is rather small in the absence of speech. 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. Similarly, the tradeoff filter can have more noise attenuation as compared to the MVDR filter, but only in the silence periods. The aggressiveness of the tradeoff filter in suppressing silence periods depends on the value of μ. The larger the value of μ, the more aggressive is the tradeoff filter in suppressing the noise in silence periods. Figure 1(d) and (e) plot the outputs of the MVDR and Wiener filters. We found that the enhanced signal by the MVDR filter is more pleasant to listen to than the output of the Wiener filter because the residual noise level with the Wiener filter changes significantly from time to time while it remains almost the same with the MVDR filter. Therefore, it is recommended to use the MVDR filter in practice.

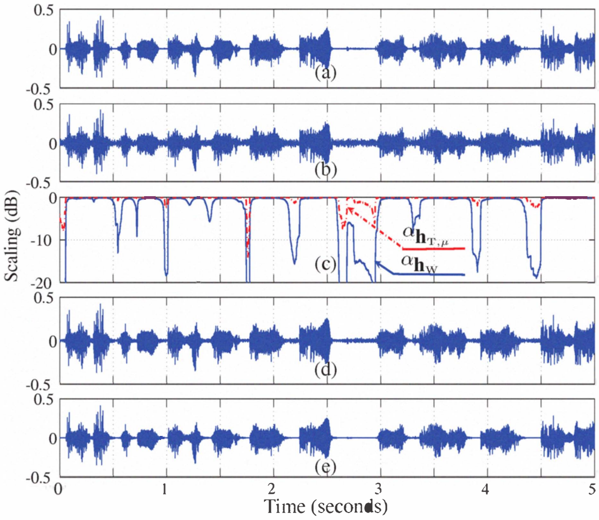

Fig. 1. (a) The clean speech waveform, (b) the noisy speech waveform, (c) the scaling factor of the Wiener and tradeoff ($\mu = 0.01$) filters compared to the MVDR filter, (d) the enhanced signal by the MVDR filter, and (e) the enhanced signal by the Wiener filter. Car noise with $\mathrm{iSNR} = 10$ dB and $L = 20$.

4. CONCLUSIONS

In this paper, we reexamined the noise-reduction problem in the time domain and presented a new way to decompose the filtered speech into two uncorrelated (orthogonal) components: the desired speech and the interference. Based on this new decomposition, we discussed how to form different optimization cost functions. By optimizing these cost functions, we showed how to derive the maximum SNR, the Wiener, the MVDR, and the tradeoff filters. Through both theoretical analysis and experimental results, we demonstrated that the maximum SNR, Wiener, and tradeoff filters are identical to the MVDR filter up to a scaling factor. From a theoretical point of view, this scaling factor is not significant and should not affect the output SNR at any processing time. But from a practical viewpoint, the scaling factor can cause significant discontinuity in the residual noise level, which is unpleasant to listen to. As a result, it is essential to have the scaling factor right from one processing sample (or frame) to another in order to avoid large distortions and for this reason, it is recommended to use the MVDR filter in speech enhancement applications.

5. REFERENCES

[1] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Chichester, England: John Wiley & Sons Ltd, 2006.

[3] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[4] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1218–1234, July 2006.

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