ON A DUAL-GAIN APPROACH TO NOISE REDUCTION IN THE STFT DOMAIN

Chao Pan[1]*, Jingdong Chen*[1]*, and Jacob Benesty*[2]

[1]: Northwestern Polytechnical University
127 Youyi West Road
Xi'an, Shaanxi 710072, China

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

ABSTRACT

Noise reduction is typically achieved by applying a gain filter to the complex spectrum of the noisy speech signal in the short-time Fourier transform (STFT) domain. However, such an approach does not take into account the noncircularity property of the complex speech spectrum. Recently, a widely linear (WL) filtering framework was developed, which can fully take advantage of the second-order statistics of the noncircular speech spectrum for noise reduction. But the optimal WL filters are more complicated to estimate as the estimation involves the use of interframe information. In this paper, we investigate a dual-gain approach, which achieves noise reduction by applying one gain to filter the real part and another gain to filter the imaginary part of the complex noisy spectrum. We show that this approach can be viewed as a particular case of the WL framework. Compared to the classical method with a single gain, this new approach is shown to be able to achieve better noise reduction performance. Another advantage is that the optimal filters with this approach can be implemented using only the current frame of spectra without the need of the interframe information.

Index Terms— Noise reduction, speech enhancement, STFT domain, maximum SNR gains, Wiener gains, tradeoff gains.

1. SIGNAL MODEL AND PROBLEM FORMULATION

The noise reduction problem considered in this study is one of recovering the desired signal (or clean speech) $x(t)$, t being the time index, of zero mean from the noisy observation (microphone signal) [1]:

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

where $v(t)$ is the unwanted additive noise, which is assumed to be a zero-mean random process, white or colored, but uncorrelated with $x(t)$. All signals are considered to be real and broadband.

Using the STFT, (1) can be rewritten in the time-frequency domain as

$$Y(k,n) = X(k,n) + V(k,n), \tag{2}$$

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at frequency bin $k \in \{0, 1, \ldots, K-1\}$ and time frame n. Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the variance of $Y(k,n)$ is

$$\phi_Y(k,n) \triangleq E\left[|Y(k,n)|^2\right] = \phi_X(k,n) + \phi_V(k,n), \tag{3}$$

where $E[\cdot]$ denotes mathematical expectation, and $\phi_X(k,n)$ and $\phi_V(k,n)$ are defined in a similar way to $\phi_Y(k,n)$. The core issue of noise reduction with the signal model given in (2) is to estimate $X(k,n)$ from $Y(k,n)$.

2. CLASSICAL APPROACH WITH A SINGLE GAIN

In the classical approach, the desired signal spectrum, $X(k,n)$, is estimated from the noisy spectrum, $Y(k,n)$, as follows:

$$\widehat{X}(k,n) = H(k,n)Y(k,n) = X_{\rm fd}(k,n) + V_{\rm rn}(k,n), \tag{4}$$

where $\widehat{X}(k,n)$ is supposed to be the estimate of $X(k,n)$, $H(k,n)$ is a gain that needs to be determined, $X_{\rm fd}(k,n) \triangleq H(k,n)X(k,n)$ and $V_{\rm rn}(k,n) \triangleq H(k,n)V(k,n)$ are, respectively, the filtered desired signal and residual noise.

The error signal between the estimated and desired signals at frequency bin k and time frame n is then

$$\mathcal{E}(k,n) \triangleq \widehat{X}(k,n) - X(k,n) = \mathcal{E}_{\rm d}(k,n) + \mathcal{E}_{\rm r}(k,n), \tag{5}$$

where

$$\mathcal{E}_{\rm d}(k,n) \triangleq [H(k,n) - 1]\, X(k,n), \tag{6a}$$

$$\mathcal{E}_{\rm r}(k,n) \triangleq H(k,n)V(k,n), \tag{6b}$$

are, respectively, the speech distortion and residual noise.

The mean-square error (MSE) criterion is

$$J(k,n) \triangleq E\left[|\mathcal{E}(k,n)|^2\right] = J_{\rm d}(k,n) + J_{\rm r}(k,n), \tag{7}$$

where

$$J_{\rm d}(k,n) \triangleq E\left[|\mathcal{E}_{\rm d}(k,n)|^2\right] = |1 - H(k,n)|^2\, \phi_X(k,n), \tag{8a}$$

$$J_{\rm r}(k,n) \triangleq E\left[|\mathcal{E}_{\rm r}(k,n)|^2\right] = |H(k,n)|^2\, \phi_V(k,n). \tag{8b}$$

Given the above MSE criteria, the objective of noise reduction is to find an optimal gain that would either minimize $J(k,n)$ or minimize $J_{\rm d}(k,n)$ or $J_{\rm r}(k,n)$ subject to some constraint. For example, we find the classical Wiener gain by minimizing $J(k,n)$ [2]:

$$H_{\rm W}(k,n) = \frac{\phi_X(k,n)}{\phi_X(k,n) + \phi_V(k,n)}. \tag{9}$$

One can see that the Wiener gain is always real, positive, and $0 \leq H_{\rm W}(k,n) \leq 1$ for any k and n. The estimated desired signal with the Wiener gain is

$$\widehat{X}_{\rm W}(k,n) = \frac{\phi_X(k,n)}{\phi_X(k,n) + \phi_V(k,n)} Y(k,n). \tag{10}$$

Clearly, $\widehat{X}_{\rm W}(k,n)$ has the same phase as $Y(k,n)$. In other words, the Wiener approach estimates the desired signal spectrum by only filtering the magnitude spectrum of the noisy signal.

It is well known that the Wiener gain adds distortion to the desired speech signal, and the amount of distortion is, in general, proportional to the amount of noise reduction [3]. To control the compromise between the degree of speech distortion and the amount of noise reduction, a more general form of Wiener, called the tradeoff gain was introduced (see [4] and references therein), which is shown as follows:

$$H_{{\rm T},\mu}(k,n) = \frac{\phi_X(k,n)}{\phi_X(k,n) + \mu\phi_V(k,n)}, \tag{11}$$

978-1-4799-0972-8/13/$31.00 ©2013IEEE

where $\mu \geq 0$. If $\mu = 1$, (11) degenerates to the Wiener gain as in (9). If $\mu = 0$, we have $H_{\mathrm{T},0}(k,n) = 1$, which does not introduce any speech distortion, but does not produce any noise reduction either. Generally, for $\mu > 1$, we can achieve more noise reduction with the tradeoff gain as compared to the Wiener gain filter, but the resulting speech distortion is larger than that of the Wiener gain. For $\mu < 1$, we achieve less noise reduction, but the resulting speech distortion is smaller than that of the Wiener gain.

3. WIDELY LINEAR APPROACH

The classical optimal gains discussed previously are derived from the linear estimation theory based on the assumption that the complex STFT coefficients of nonstationary speech signals at each frequency bin are second-order circular. (A random complex variable is second-order circular or noncircular depending on whether its pseudo-variance is equal to zero or not [5], [6]. However, as shown in [7], the STFT coefficients of nonstationary speech signals are noncircular. To fully use the second-order statistics of the complex speech spectra in noise reduction, a widely linear (WL) filtering approach was developed [7], [8], where the desired signal, $X(k,n)$, is estimated according to

$$\begin{aligned}\widehat{X}(k,n) &= H_1(k,n)Y(k,n) + H_2(k,n)Y^*(k,n)\\ &= \mathbf{h}^H(k,n)\mathbf{y}(k,n),\end{aligned} \tag{12}$$

where the superscripts * and ^{H} are, respectively, the complex-conjugate and transpose-conjugate operators, and

$$\mathbf{h}(k,n) \triangleq [H_1^*(k,n)\ \ H_2^*(k,n)]^T,$$
$$\mathbf{y}(k,n) \triangleq [Y(k,n)\ \ Y^*(k,n)]^T.$$

With this estimate, the MSE can be written as

$$J(k,n) \triangleq E\left[|\mathcal{E}(k,n)|^2\right]. \tag{13}$$

where

$$\mathcal{E}(k,n) \triangleq \widehat{X}(k,n) - X(k,n) = \mathbf{h}^H(k,n)\mathbf{y}(k,n) - X(k,n). \tag{14}$$

Taking the gradient of $J(k,n)$ with respect to $\mathbf{h}(k,n)$ and equating the result to zero, we find the WL Wiener filter:

$$\begin{aligned}\mathbf{h}_{\mathrm{WLW}}(k,n) &= \mathbf{\Phi}_{\mathbf{y}}^{-1}(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_1 \\ &= \frac{\phi_X(k,n)}{\phi_Y(k,n)}\mathbf{\Gamma}_{\mathbf{y}}^{-1}(k,n)\mathbf{\Gamma}_{\mathbf{x}}(k,n)\mathbf{i}_1,\end{aligned} \tag{15}$$

where

$$\mathbf{\Phi}_{\mathbf{y}}(k,n) \triangleq E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right], \tag{16}$$

$$\mathbf{\Gamma}_{\mathbf{y}}(k,n) \triangleq \begin{bmatrix} 1 & \gamma_Y(k,n) \\ \gamma_Y^*(k,n) & 1 \end{bmatrix}, \tag{17}$$

$$\gamma_Y(k,n) \triangleq \frac{E\left[Y^2(k,n)\right]}{E\left[|Y(k,n)|^2\right]}, \tag{18}$$

$$\mathbf{i}_1 \triangleq [1\ \ 0]^T, \tag{19}$$

and $\mathbf{\Phi}_{\mathbf{x}}(k,n)$, $\gamma_X(k,n)$, and $\mathbf{\Gamma}_{\mathbf{x}}(k,n)$ are defined in a similar way to $\mathbf{\Phi}_{\mathbf{y}}(k,n)$, $\gamma_Y(k,n)$, and $\mathbf{\Gamma}_{\mathbf{y}}(k,n)$, respectively.

The WL Wiener filter achieves better noise reduction performance as compared to the classical Wiener gain given in (9) [7]. However, its implementation requires the knowledge of the circularity quotients $\gamma_Y(k,n)$ and $\gamma_X(k,n)$. In practice, it is challenging to obtain an accurate estimate of these two quotients since it involves the use of the interframe information. One cannot approximate the expectation operation in (18) with its instantaneous value since this would make the $\mathbf{\Gamma}_{\mathbf{y}}(k,n)$ matrix rank deficient.

4. DUAL-GAIN APPROACH

Another way to achieve noise reduction in the STFT domain is to process the real and imaginary parts of the noisy speech spectra separately [9]. In this section, we introduce a dual-gain approach. Let us rewrite (2) as

$$Y(k,n) = Y_{\mathrm{R}}(k,n) + jY_{\mathrm{I}}(k,n), \tag{20}$$

where $j = \sqrt{-1}$ is the imaginary unit and

$$Y_{\mathrm{R}}(k,n) = X_{\mathrm{R}}(k,n) + V_{\mathrm{R}}(k,n), \tag{21}$$
$$Y_{\mathrm{I}}(k,n) = X_{\mathrm{I}}(k,n) + V_{\mathrm{I}}(k,n), \tag{22}$$

are the real and imaginary parts of $Y(k,n)$, respectively.

Now, we estimate the real and imaginary parts of the desired signal separately with two real gains, i.e.,

$$\widehat{X}_{\square}(k,n) = H_{\square}(k,n)Y_{\square}(k,n), \tag{23}$$

where the subscript $\square \in \{\mathrm{R},\mathrm{I}\}$, $\widehat{X}_{\square}(k,n)$ is supposed to be the estimate of the real or imaginary part of the desired signal, and $H_{\square}(k,n)$ is a real-valued gain. In this approach, we implicitly assume that the real and imaginary parts of the signals are uncorrelated and therefore can be processed separately. The estimate of the desired signal is then

$$\begin{aligned}\widehat{X}(k,n) &= \widehat{X}_{\mathrm{R}}(k,n) + j\widehat{X}_{\mathrm{I}}(k,n)\\ &= H_{\mathrm{R}}(k,n)Y_{\mathrm{R}}(k,n) + jH_{\mathrm{I}}(k,n)Y_{\mathrm{I}}(k,n)\\ &= X_{\mathrm{fd}}(k,n) + V_{\mathrm{rn}}(k,n),\end{aligned} \tag{24}$$

where

$$X_{\mathrm{fd}}(k,n) \triangleq H_{\mathrm{R}}(k,n)X_{\mathrm{R}}(k,n) + jH_{\mathrm{I}}(k,n)X_{\mathrm{I}}(k,n) \tag{25}$$

is the filtered desired signal and

$$V_{\mathrm{rn}}(k,n) \triangleq H_{\mathrm{R}}(k,n)V_{\mathrm{R}}(k,n) + jH_{\mathrm{I}}(k,n)V_{\mathrm{I}}(k,n) \tag{26}$$

is the residual noise. The error signal between the estimated and desired signals at frequency bin k and time frame n can now be written as

$$\mathcal{E}(k,n) \triangleq \widehat{X}(k,n) - X(k,n) = \mathcal{E}_{\mathrm{d}}(k,n) + \mathcal{E}_{\mathrm{r}}(k,n), \tag{27}$$

where

$$\begin{aligned}\mathcal{E}_{\mathrm{d}}(k,n) &\triangleq [H_{\mathrm{R}}(k,n) - 1]\, X_{\mathrm{R}}(k,n)\\ &\quad + j\,[H_{\mathrm{I}}(k,n) - 1]\, X_{\mathrm{I}}(k,n),\end{aligned} \tag{28a}$$
$$\mathcal{E}_{\mathrm{r}}(k,n) \triangleq H_{\mathrm{R}}(k,n)V_{\mathrm{R}}(k,n) + jH_{\mathrm{I}}(k,n)V_{\mathrm{I}}(k,n), \tag{28b}$$

are the speech distortion and residual noise, respectively. The MSE is then

$$J(k,n) \triangleq E\left[|\mathcal{E}(k,n)|^2\right] = J_{\mathrm{d}}(k,n) + J_{\mathrm{r}}(k,n), \tag{29}$$

where

$$\begin{aligned}J_{\mathrm{d}}(k,n) &\triangleq E\left[|\mathcal{E}_{\mathrm{d}}(k,n)|^2\right] \\ &= [1 - H_{\mathrm{R}}(k,n)]^2\,\phi_{X_{\mathrm{R}}}(k,n)\\ &\quad + [1 - H_{\mathrm{I}}(k,n)]^2\,\phi_{X_{\mathrm{I}}}(k,n),\end{aligned} \tag{30a}$$
$$\begin{aligned}J_{\mathrm{r}}(k,n) &\triangleq E\left[|\mathcal{E}_{\mathrm{r}}(k,n)|^2\right] \\ &= H_{\mathrm{R}}^2(k,n)\phi_{V_{\mathrm{R}}}(k,n) + H_{\mathrm{I}}^2(k,n)\phi_{V_{\mathrm{I}}}(k,n).\end{aligned} \tag{30b}$$

Having defined the MSE criteria $J(k,n)$, $J_{\mathrm{d}}(k,n)$, and $J_{\mathrm{r}}(k,n)$, we can now start to derive different optimal dual-gain filters.

4.1 Wiener Gains

By minimizing the MSE criterion [eq. (29)], we easily find the Wiener gains:

$$H_{\mathrm{R,W}}(k,n) = \frac{\phi_{X_{\mathrm{R}}}(k,n)}{\phi_{Y_{\mathrm{R}}}(k,n)} = \frac{\phi_{X_{\mathrm{R}}}(k,n)}{\phi_{X_{\mathrm{R}}}(k,n)+\phi_{V_{\mathrm{R}}}(k,n)}, \quad (31a)$$

$$H_{\mathrm{I,W}}(k,n) = \frac{\phi_{X_{\mathrm{I}}}(k,n)}{\phi_{Y_{\mathrm{I}}}(k,n)} = \frac{\phi_{X_{\mathrm{I}}}(k,n)}{\phi_{X_{\mathrm{I}}}(k,n)+\phi_{V_{\mathrm{R}}}(k,n)}. \quad (31b)$$

The minimum MSE of the dual-gain Wiener filter can be found by substituting (31) into (29):

$$J_{\mathrm{DGW}}(k,n) = \frac{\phi_{X_{\mathrm{R}}}(k,n)\phi_{V_{\mathrm{R}}}(k,n)}{\phi_{X_{\mathrm{R}}}(k,n)+\phi_{V_{\mathrm{R}}}(k,n)} + \frac{\phi_{X_{\mathrm{I}}}(k,n)\phi_{V_{\mathrm{I}}}(k,n)}{\phi_{X_{\mathrm{I}}}(k,n)+\phi_{V_{\mathrm{I}}}(k,n)}. \quad (32)$$

The MSE for the classical Wiener gain is obtained by substituting (9) into (7):

$$\begin{aligned} J_{\mathrm{W}}(k,n) &= \frac{\phi_X\phi_V}{\phi_X+\phi_V} \quad (33)\\ &= \frac{[\phi_{X_{\mathrm{R}}}(k,n)+\phi_{X_{\mathrm{I}}}(k,n)]\,[\phi_{V_{\mathrm{R}}}(k,n)+\phi_{V_{\mathrm{I}}}(k,n)]}{\phi_{X_{\mathrm{R}}}(k,n)+\phi_{X_{\mathrm{I}}}(k,n)+\phi_{V_{\mathrm{R}}}(k,n)+\phi_{V_{\mathrm{I}}}(k,n)}. \end{aligned}$$

It is easy to check that

$$J_{\mathrm{DGW}}(k,n) \leq J_{\mathrm{W}}(k,n), \quad (34)$$

where the equality holds if and only if $\phi_{X_{\mathrm{R}}}(k,n)/\phi_{V_{\mathrm{R}}}(k,n) = \phi_{X_{\mathrm{I}}}(k,n)/\phi_{V_{\mathrm{I}}}(k,n)$. So, the minimum MSE of the dual-gain Wiener filter is generally smaller than that of the classical Wiener gain. Note that both $J_{\mathrm{DGW}}(k,n)$ and $J_{\mathrm{W}}(k,n)$ consist of two parts, i.e., speech distortion part and residual noise part. Therefore, the inequality in (34) indicates that the total amount of speech distortion and residual noise of the dual-gain Wiener filter is generally smaller than that of the classical Wiener gain.

4.2 Tradeoff Gains

The tradeoff gains are obtained by minimizing the speech distortion with the constraint that the residual noise level is equal to a value smaller than the level of the original noise. This is equivalent to solving the following optimization problem:

$$\min_{H_{\mathrm{R}}(k,n),H_{\mathrm{I}}(k,n)} \; J_{\mathrm{d}}(k,n) \;\text{ subject to }\; J_{\mathrm{r}}(k,n) \leq \beta\cdot\phi_V(k,n), \quad (35)$$

where $0 < \beta < 1$ in order to have some noise reduction at the frequency bin k. Using a Lagrange multiplier, $\mu \geq 0$, to adjoin the constraint to the cost function, we can solve the above optimization problem and obtain the tradeoff gains:

$$H_{\mathrm{R,T},\mu}(k,n) = \frac{\phi_{X_{\mathrm{R}}}(k,n)}{\phi_{X_{\mathrm{R}}}(k,n)+\mu\phi_{V_{\mathrm{R}}}(k,n)}, \quad (36a)$$

$$H_{\mathrm{I,T},\mu}(k,n) = \frac{\phi_{X_{\mathrm{I}}}(k,n)}{\phi_{X_{\mathrm{I}}}(k,n)+\mu\phi_{V_{\mathrm{R}}}(k,n)}. \quad (36b)$$

The particular cases of $\mu = 1$ and $\mu = 0$ correspond to the Wiener and identity gains, respectively. More generally, we can write the tradeoff gains into the following forms:

$$H_{\mathrm{R,T},\mu_{\mathrm{R}}}(k,n) = \frac{\phi_{X_{\mathrm{R}}}(k,n)}{\phi_{X_{\mathrm{R}}}(k,n)+\mu_{\mathrm{R}}\phi_{V_{\mathrm{R}}}(k,n)}, \quad (37a)$$

$$H_{\mathrm{I,T},\mu_{\mathrm{I}}}(k,n) = \frac{\phi_{X_{\mathrm{I}}}(k,n)}{\phi_{X_{\mathrm{I}}}(k,n)+\mu_{\mathrm{I}}\phi_{V_{\mathrm{R}}}(k,n)}. \quad (37b)$$

4.3 Maximum SNR Gains

We define the subband output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual noise [4], i.e.,

$$\begin{aligned} \mathrm{oSNR}(k,n) &\triangleq \frac{\phi_{X_{\mathrm{fd}}}(k,n)}{\phi_{V_{\mathrm{rn}}}(k,n)}\\ &= \frac{H_{\mathrm{R}}^2(k,n)\phi_{X_{\mathrm{R}}}(k,n)+H_{\mathrm{I}}^2(k,n)\phi_{X_{\mathrm{I}}}(k,n)}{H_{\mathrm{R}}^2(k,n)\phi_{V_{\mathrm{R}}}(k,n)+H_{\mathrm{I}}^2(k,n)\phi_{V_{\mathrm{I}}}(k,n)}. \quad (38) \end{aligned}$$

In the maximum SNR technique, we find the gains that maximize the subband output SNR. It is clear that we need to find the maximum eigenvector of the matrix $\mathbf{D}_V^{-1}(k,n)\mathbf{D}_X(k,n)$, where $\mathbf{D}_V(k,n) = \mathrm{diag}\,[\phi_{V_{\mathrm{R}}}(k,n), \phi_{V_{\mathrm{I}}}(k,n)]$ and $\mathbf{D}_X(k,n) = \mathrm{diag}\,[\phi_{X_{\mathrm{R}}}(k,n), \phi_{X_{\mathrm{I}}}(k,n)]$ are two diagonal matrices. Since $\mathbf{D}_V^{-1}(k,n)\mathbf{D}_X(k,n)$ is also a diagonal matrix, we deduce that the maximum gains are

$$\begin{aligned} &\begin{cases} H_{\mathrm{R,max}}(k,n) = 1, \\ H_{\mathrm{I,max}}(k,n) = 0, \end{cases} \quad \text{if } \tfrac{\phi_{X_{\mathrm{R}}}(k,n)}{\phi_{V_{\mathrm{R}}}(k,n)} \geq \tfrac{\phi_{X_{\mathrm{I}}}(k,n)}{\phi_{V_{\mathrm{I}}}(k,n)}\,, \\ &\begin{cases} H_{\mathrm{R,max}}(k,n) = 0, \\ H_{\mathrm{I,max}}(k,n) = 1, \end{cases} \quad \text{if } \tfrac{\phi_{X_{\mathrm{R}}}(k,n)}{\phi_{V_{\mathrm{R}}}(k,n)} < \tfrac{\phi_{X_{\mathrm{I}}}(k,n)}{\phi_{V_{\mathrm{I}}}(k,n)}\,. \quad (39) \end{aligned}$$

It is interesting to see that the result of the maximum SNR filter is similar to the widely known binary masking technique [11]. It selects either the real part or the imaginary part at the frequency bin k and time frame n depending whose SNR is larger.

5. CONNECTION BETWEEN THE DUAL-GAIN AND WIDELY LINEAR APPROACHES

In this section, we show that the dual-gain Wiener filter is equivalent to the WL Wiener filter if $X_{\mathrm{R}}(k,n)$, $V_{\mathrm{R}}(k,n)$, and $Y_{\mathrm{R}}(k,n)$ are uncorrelated, respectively, with $X_{\mathrm{I}}(k,n)$, $V_{\mathrm{I}}(k,n)$, and $Y_{\mathrm{I}}(k,n)$.

Proof: If $X_{\mathrm{R}}(k,n)$, $V_{\mathrm{R}}(k,n)$, and $Y_{\mathrm{R}}(k,n)$ are uncorrelated, respectively, with $X_{\mathrm{I}}(k,n)$, $V_{\mathrm{I}}(k,n)$, and $Y_{\mathrm{I}}(k,n)$, we have $E\,[X_{\mathrm{R}}(k,n)X_{\mathrm{I}}(k,n)] = 0$, $E\,[V_{\mathrm{R}}(k,n)V_{\mathrm{I}}(k,n)] = 0$, and $E\,[Y_{\mathrm{R}}(k,n)Y_{\mathrm{I}}(k,n)] = 0$. Then, the covariance matrices, $\mathbf{\Phi}_{\mathbf{y}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{x}}(k,n)$, can be written into the following forms:

$$\mathbf{\Phi}_{\mathbf{y}}(k,n) = \begin{bmatrix} \phi_{Y_{\mathrm{R}}}(k,n)+\phi_{Y_{\mathrm{I}}}(k,n) & \phi_{Y_{\mathrm{R}}}(k,n)-\phi_{Y_{\mathrm{I}}}(k,n) \\ \phi_{Y_{\mathrm{R}}}(k,n)-\phi_{Y_{\mathrm{I}}}(k,n) & \phi_{Y_{\mathrm{R}}}(k,n)+\phi_{Y_{\mathrm{I}}}(k,n) \end{bmatrix},$$

$$\mathbf{\Phi}_{\mathbf{x}}(k,n) = \begin{bmatrix} \phi_{X_{\mathrm{R}}}(k,n)+\phi_{X_{\mathrm{I}}}(k,n) & \phi_{X_{\mathrm{R}}}(k,n)-\phi_{X_{\mathrm{I}}}(k,n) \\ \phi_{X_{\mathrm{R}}}(k,n)-\phi_{X_{\mathrm{I}}}(k,n) & \phi_{X_{\mathrm{R}}}(k,n)+\phi_{X_{\mathrm{I}}}(k,n) \end{bmatrix}.$$

Substituting the above two equations into (15), we can readily get

$$\mathbf{h}_{\mathrm{WLW}}(k,n) = \frac{1}{2}\begin{bmatrix} H_{\mathrm{R,W}}(k,n)+H_{\mathrm{I,W}}(k,n) \\ H_{\mathrm{R,W}}(k,n)-H_{\mathrm{I,W}}(k,n) \end{bmatrix}, \quad (41)$$

where $H_{\mathrm{R,W}}(k,n)$ and $H_{\mathrm{I,W}}(k,n)$ are the dual-gain Wiener filters given in (31).

With the WL Wiener filter, the estimate of the desired signal from the noisy observation, $\mathbf{y}(k,n) = [Y(k,n)\;\; Y^*(k,n)]^T$, is given by

$$\widehat{X}_{\mathrm{WLW}}(k,n) = \mathbf{h}_{\mathrm{WLW}}^H(k,n)\mathbf{y}(k,n). \quad (42)$$

Substituting (41) into (42) yields

$$\begin{aligned} \widehat{X}_{\mathrm{WLW}}(k,n) &= H_{\mathrm{R,W}}(k,n)Y_{\mathrm{R}}(k,n) + jH_{\mathrm{I,W}}(k,n)Y_{\mathrm{I}}(k,n)\\ &= \widehat{X}_{\mathrm{DGW}}(k,n), \quad (43) \end{aligned}$$

where $\widehat{X}_{\mathrm{DGW}}(k,n)$ is the estimate of desired signal obtained from the dual-gain Wiener filter. That completes the proof.

However, in practice, some correlation may exist between $X_{\rm R}(k,n)$ and $X_{\rm I}(k,n)$, and $Y_{\rm R}(k,n)$ and $Y_{\rm I}(k,n)$ due to the facts that speech signals are nonstationary and the STFT length is limited. In this case, the WL Wiener filter may achieve better noise reduction performance than the dual-gain Wiener filter.

6. EXPERIMENTS

In the section, we evaluate the performance of the optimal dual-gain Wiener filter using experiments and compare it with the Wiener filters derived from the classical single-gain as well the WL approaches. The clean speech used is recorded in a quiet room with a sampling rate of 8 kHz and the overall length of the signal is 30 seconds. The noisy signal is generated by adding some pre-recorded noise signal to the clean speech where the noise signal is properly scaled to control the input SNR to 10 dB. The overlap add technique is used in the implementation. The frame size is set to 64 and the overlap between the successive frames is 75%. To minimize frequency aliasing, a Kaiser window is applied both before the STFT and after the inverse STFT.

We use the output SNR and speech distortion index as the performance measures, which are defined, respectively, as [4]:

$$\text{oSNR} \triangleq \frac{E\left[x_{\rm fd}^2(t)\right]}{E\left[v_{\rm rn}^2(t)\right]}, \text{ and } \upsilon_{\rm sd} \triangleq \frac{E\left\{\left[x_{\rm fd}(t)-x(t)\right]^2\right\}}{E\left[x^2(t)\right]}, \quad (44)$$

where $x_{\rm fd}(t)$ and $v_{\rm rn}(t)$ are the filtered desired signal and residual noise reconstructed from $X_{\rm fd}(k,n)$ and $V_{\rm rn}(k,n)$, respectively. Note that for the WL Wiener filter, there is also a residual interference term, which is treated same as the residual noise.

To implement different Wiener filters, we need to know the variance parameters and circularity quotients of the noisy and clean speech signals in the STFT domain. Since the noisy signal is accessible, all the parameters associated with this signal can be easily computed. However, in order to estimate the parameters associated with the clean and noise signals, we would need a noise estimator, which generally relies on a voice activity detector (VAD). However, due to space limitation, we will put aside the VAD issues in this paper and directly compute the circularity quotients from the corresponding signals with a short-time average using the most recently 20 frames. The variance parameters of the noisy and noise signals are computed by approximating the mathematical expectation with a recursive average. Specifically, the variance $\phi_Y(k,n)$ is estimated according to

$$\widehat{\phi}_Y(k,n) = \lambda_Y \widehat{\phi}_Y(k,n-1) + (1-\lambda_Y)\left|Y(k,n)\right|^2, \quad (45)$$

where $\lambda_Y \in (0,1)$ is a forgetting factor, and $\phi_{Y_{\rm R}}(k,n)$ and $\phi_{Y_{\rm I}}(k,n)$ are computed in a same way. The parameter $\phi_V(k,n)$, $\phi_{V_{\rm R}}(k,n)$ and $\phi_{V_{\rm I}}(k,n)$ are computed similarly but with a forgetting factor of λ_V. Then, the variance of the clean speech is computed by subtracting the variance of the noise signal from that of the noisy signal (the result is forced to 0 if it is negative). With this way of estimation of the variance parameters, the values of the forgetting factors λ_Y and λ_V play an important role on the noise-reduction performance. If they are too small, the estimation variance of the signal statistics would be large, which will be translated into speech distortion. If they are too large, the estimated statistics would not follow the nonstationary property of the speech and noise signals. It is difficult, of course, to determine the optimal values of λ_Y and λ_V in an analytically manner. So, we use experiments to study their impact on noise reduction performance. We investigated three noise conditions: white Gaussian noise, a car noise recorded in a Volvo sedan running at 55 MPH and a babble noise recorded in a New York stock exchange (NYSE) noise. Due to space limit, we only report the results in the NYSE noise as show in Fig. 1, where we set λ_V to 0.96 and vary λ_Y from 0.3 to 0.96.

As seen, the three Wiener filters can achieve reasonably good performance when λ_Y is in the range between 0.8 and 0.9. Among the three filters, the dual Wiener gains yield a performance better than that of the classical Wiener filter but worse than the WL wiener filter, which validates the previous theoretical analysis.

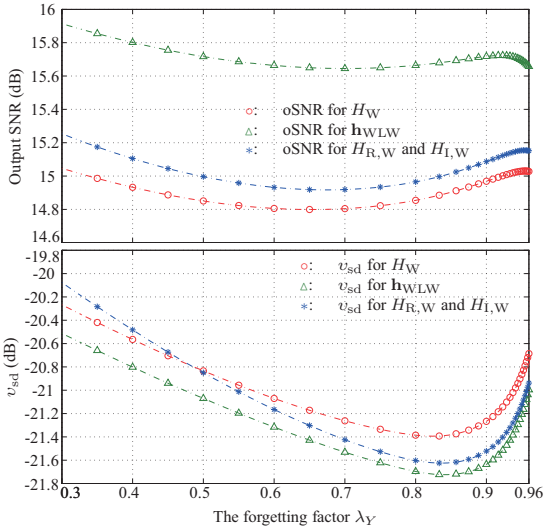

Figure 1: Performance of the classical, WL, and dual-gain Wiener filters in the NYSE babble noise where the input SNR is 10 dB and $\lambda_V = 0.96$.

7. CONCLUSIONS

In this paper, we investigated a dual-gain approach to noise reduction in the STFT domain. Unlike the classical approach that achieves noise reduction by applying a single gain to the noisy speech spectrum, this dual-gain approach applies one gain to filter the real part and another gain to filter the imaginary part of the complex noisy spectrum. With this formulation, we derived the Wiener, tradeoff, and maximum SNR filters. We showed that the dual Wiener gains can be viewed as a particular case of the WL Wiener filter. Experiments demonstrated that the dual-gain approach has the potential to achieve a larger output SNR and a smaller speech distortion than the classical single-gain method.

8. REFERENCES

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

[2] P. Vary, "Noise suppression by spectral magnitude estimation–mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, July 1985.

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

[4] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer Briefs in Electrical and Computer Engineering, 2011.

[5] B. Picinbono, "On circularity," *IEEE Trans. Signal Process.*, vol. 42, pp. 3473–3482, Dec. 1994.

[6] D. P. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. Wiley, 2009.

[7] J. Benesty, J. Chen, and Y. Huang, "A widely linear distortionless filter for single-channel noise reduction," *IEEE Signal Process. Lett.*, vol. 17, pp. 469–472, May 2010.

[8] J. Benesty, J. Chen, and Y. Huang, "On widely linear Wiener and tradeoff filters for noise reduction," *Speech Communication*, vol. 52, pp. 427–439, 2010.

[9] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 845–856, Sept. 2005.

[10] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Sig. Lett.*, vol. 15, pp. 213–126, 2008

[11] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Piscateway, NJ: IEEE Press, 2006.