2 Study of the Wiener Filter for Noise Reduction

Jacob Benesty[1], Jingdong Chen[2], Yiteng (Arden) Huang[2], and
Simon Doclo[3]

[1] Université du Québec, INRS-EMT
Montréal, QC H5A 1K6, Canada
E-mail: benesty@inrs-emt.uquebec.ca
[2] Bell Laboratories, Lucent Technologies
Murray Hill, NJ 07974, USA
E-mail: {jingdong, arden}@research.bell-labs.com
[3] Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT-SCD)
Leuven 3001, Belgium
E-mail: doclo@esat.kuleuven.ac.be

Abstract. The problem of noise reduction has attracted a considerable amount of research attention over the past several decades. Numerous techniques were developed, and among them is the optimal Wiener filter, which is the most fundamental approach, and has been delineated in different forms and adopted in diversified applications. It is not a secret that the Wiener filter achieves noise reduction with some integrity loss of the speech signal. However, few efforts have been reported to show the inherent relationship between noise reduction and speech distortion. By defining a speech-distortion index and a noise-reduction factor, this chapter studies the quantitative performance behavior of the Wiener filter in the context of noise reduction. We show that for a single-channel Wiener filter, the amount of noise attenuation is in general proportionate to the amount of speech degradation. In other words, the more the noise is reduced, the more the speech is distorted. This may seem discouraging as we always expect an algorithm to have maximal noise attenuation without much speech distortion. Fortunately, we show that the speech distortion can be better managed by properly manipulating the Wiener filter, or by considering some knowledge of the speech signal. The former leads to a sub-optimal Wiener filter where a parameter is introduced to control the tradeoff between speech distortion and noise reduction, and the latter leads to the well-known parametric-model-based noise reduction technique. We also show that speech distortion can even be avoided if we have multiple realizations of the speech signal.

2.1 Introduction

The existence of noise is inevitable in real-world applications of speech processing. In a voice communication system, for example, a desired speech signal, when propagating through an acoustic channel and picked up by a microphone sensor, is corrupted by unwanted noise, which may result in appreciable or even significant degradation in the quality and intelligibility of the recorded speech. Therefore, it is essential for such systems that we can have some effective noise reduction/speech enhancement techniques to extract the desired speech signal from its corrupted observations.

The noise reduction technique has a broad range of applications, from hearing aids, cellular phones, voice-controlling systems, teleconferencing and multiparty teleconferencing, to automatic speech recognition (ASR) systems. The difference between two systems using and not using such techniques can be significant; therefore, the choice can have a great impact on the functioning of the system.

Research on noise reduction/speech enhancement can be traced back to 40 years ago with 2 patents by Schroeder [1], [2] where an analog implementation of the spectral magnitude subtraction method was described. Since then, it has become an area of active research. Over the past several decades, researchers and engineers have approached this challenging problem by exploiting different facets of the properties of the speech and noise signals [3], [4], [5], [6], [7]. A variety of approaches have been developed, including Wiener filter [8], [9], [10], [11], [12], [13], spectral restoration [3], [11], [14], [15], [16], [17], [18], [19], signal subspace method [20], [21], [22], [23], [24], [25], [26], parametric-model-based approach [27], [28], [29], [30], [31], statistical-model-based method [5], [32], [33], [34], [35], [36], [37], and spatio-temporal filtering [38], [39], [40], [41], [42].

Most of these algorithms were developed independently of each other and their performance on noise reduction were evaluated mostly by assessing the improvement of signal-to-noise ratio (SNR) or subjective speech quality when the methods were formulated. It has been noticed that these algorithms, almost with no exception, achieve noise reduction by some integrity loss of the speech signal. Some algorithms are even formulated explicitly based on the tradeoff between noise reduction and speech distortion, such as the subspace method. However, so far, few efforts have been devoted to analyzing such a tradeoff behavior even though it is a very important issue. In this chapter, we attempt to provide an analysis about the compromise between noise reduction and speech distortion. On the one hand, such a study may offer us some insight into the range of the existing algorithms that can be employed in practical noisy environments. On the other hand, a good understanding may help us to find new algorithms that can work more effectively than the existing ones.

Since there are so many algorithms in the literature, it is extremely difficult if not impossible to find a universal analytical tool that can be applied to any algorithm. In this study, we choose the Wiener filter as the basis since it is the most fundamental approach, and many algorithms are closely connected to this technique. For example, the minimum-mean-square-error (MMSE) estimator presented in [15], which belongs to the category of spectral restoration, converges to the Wiener filter at a high SNR. Also it is widely known that the Kalman filter is tightly related to the Wiener filter.

Starting from the optimal Wiener filtering theory, we introduce two new concepts: the speech-distortion index and the noise-reduction factor. We then show that for a single-channel Wiener filter, the amount of noise attenuation is in general proportionate to the amount of speech degradation. In other words, the more the noise is attenuated, the more the speech is distorted. This observation may seem quite discouraging as we always expect an algorithm to have maximal noise attenuation without much speech distortion. Fortunately, we show that the compromise between noise reduction and speech distortion can be better managed by properly manipulating the Wiener filter, or by considering some knowledge of the speech signal. The former leads to a suboptimal Wiener filter where, like in the spectral subtraction, a parameter is introduced to control the tradeoff between speech distortion and noise reduction, and the latter leads to the well-known parametric-model-based noise-reduction technique. We also discuss the possibility to avoid speech distortion by using an array of microphones.

2.2 Estimation of the Clean Speech Samples

We consider a zero-mean clean speech signal $x(n)$ contaminated by a zero-mean noise process $v(n)$ [white or colored but uncorrelated with $x(n)$], so that the noisy speech signal, at the discrete time sample n, is,

$$y(n) = x(n) + v(n). \tag{2.1}$$

Define the error signal between the clean speech sample at time n and its estimate:

$$e_x(n) \overset{\triangle}{=} x(n) - \hat{x}(n) = x(n) - \mathbf{h}^T \mathbf{y}(n), \tag{2.2}$$

where superscript ^{T} denotes transpose of a vector or a matrix,

$$\mathbf{h} = \begin{bmatrix} h_0 & h_1 & \cdots & h_{L-1} \end{bmatrix}^T$$

is an FIR filter of length L, and

$$\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n-1) & \cdots & y(n-L+1) \end{bmatrix}^T$$

is a vector containing the L most recent samples of the observation signal $y(n)$.

We now can write the mean-square error (MSE) criterion:

$$J_x(\mathbf{h}) \overset{\triangle}{=} E\left\{e_x^2(n)\right\}, \tag{2.3}$$

where $E\{\cdot\}$ denotes mathematical expectation. The optimal estimate $\hat{x}_{\mathrm{o}}(n)$ of the clean speech sample $x(n)$ tends to contain less noise than the observation sample $y(n)$, and the optimal filter that forms $\hat{x}_{\mathrm{o}}(n)$ is the Wiener filter which is obtained as follows,

$$\mathbf{h}_{\mathrm{o}} = \arg\min_{\mathbf{h}} J_x(\mathbf{h}). \tag{2.4}$$

Consider the particular filter,

$$\mathbf{u}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T.$$

This means that the observed signal $y(n)$ will pass this filter unaltered (no noise reduction), thus the corresponding MSE is,

$$\begin{aligned} J_x(\mathbf{u}_1) &= E\left\{\left[x(n) - \mathbf{u}_1^T\mathbf{y}(n)\right]^2\right\} = E\left\{\left[x(n) - y(n)\right]^2\right\} \\ &= E\left\{v^2(n)\right\} = \sigma_v^2. \end{aligned} \tag{2.5}$$

In principle, for the optimal filter $\mathbf{h}_{\mathrm{o}}$, we should have,

$$J_x(\mathbf{h}_{\mathrm{o}}) < J_x(\mathbf{u}_1) = \sigma_v^2. \tag{2.6}$$

In other words, the Wiener filter will be able to reduce the level of noise in the noisy speech signal $y(n)$.

From (2.4), we easily find the Wiener-Hopf equation:

$$\mathbf{R}_y\mathbf{h}_{\mathrm{o}} = \mathbf{p}, \tag{2.7}$$

where

$$\mathbf{R}_y = E\left\{\mathbf{y}(n)\mathbf{y}^T(n)\right\} \tag{2.8}$$

is the correlation matrix of the observed signal $y(n)$ and

$$\mathbf{p} = E\left\{\mathbf{y}(n)x(n)\right\} \tag{2.9}$$

is the cross-correlation vector between the noisy and clean speech signals. However, $x(n)$ is unobservable; as a result, an estimation of $\mathbf{p}$ may seem difficult to obtain. But,

$$\begin{aligned} \mathbf{p} &= E\left\{\mathbf{y}(n)x(n)\right\} = E\left\{\mathbf{y}(n)\left[y(n) - v(n)\right]\right\} \\ &= E\left\{\mathbf{y}(n)y(n)\right\} - E\left\{\left[\mathbf{x}(n) + \mathbf{v}(n)\right]v(n)\right\} \\ &= E\left\{\mathbf{y}(n)y(n)\right\} - E\{\mathbf{v}(n)v(n)\} \\ &= \mathbf{r}_y - \mathbf{r}_v. \end{aligned} \tag{2.10}$$

Now $\mathbf{p}$ depends on the correlation vectors $\mathbf{r}_y$ and $\mathbf{r}_v$. The vector $\mathbf{r}_y$ (which is also the first column of $\mathbf{R}_y$) can be easily estimated during speech and noise periods while $\mathbf{r}_v$ can be estimated during noise-only intervals assuming that the statistics of the noise do not change much with time.

Using (2.10) and the fact that $\mathbf{u}_1 = \mathbf{R}_y^{-1}\mathbf{r}_y$, we obtain the optimal filter:

$$\begin{aligned}\mathbf{h}_\mathrm{o} &= \mathbf{u}_1 - \mathbf{R}_y^{-1}\mathbf{r}_v = \left[\mathbf{I} - \mathbf{R}_y^{-1}\mathbf{R}_v\right]\mathbf{u}_1 \\ &= \left[\frac{\mathbf{I}}{\mathrm{SNR}} + \tilde{\mathbf{R}}_v^{-1}\tilde{\mathbf{R}}_x\right]^{-1}\tilde{\mathbf{R}}_v^{-1}\tilde{\mathbf{R}}_x\mathbf{u}_1,\end{aligned} \tag{2.11}$$

where

$$\mathrm{SNR} \triangleq \frac{\sigma_x^2}{\sigma_v^2} \tag{2.12}$$

is the signal-to-noise ratio, $\mathbf{I}$ is the identity matrix, and

$$\begin{aligned}\tilde{\mathbf{R}}_x &\triangleq \frac{\mathbf{R}_x}{\sigma_x^2}, \\ \tilde{\mathbf{R}}_v &\triangleq \frac{\mathbf{R}_v}{\sigma_v^2}.\end{aligned}$$

We have,

$$\lim_{\mathrm{SNR}\to\infty} \mathbf{h}_\mathrm{o} = \mathbf{u}_1, \tag{2.13}$$

$$\lim_{\mathrm{SNR}\to 0} \mathbf{h}_\mathrm{o} = \mathbf{0}. \tag{2.14}$$

The minimum MSE (MMSE) is,

$$J_x(\mathbf{h}_\mathrm{o}) = \sigma_x^2 - \mathbf{p}^T\mathbf{h}_\mathrm{o} = \sigma_v^2 - \mathbf{r}_v^T\mathbf{R}_y^{-1}\mathbf{r}_v = \mathbf{r}_v^T\mathbf{h}_\mathrm{o}. \tag{2.15}$$

We see clearly from the previous expression that $J_x(\mathbf{h}_\mathrm{o}) < J_x(\mathbf{u}_1)$; therefore, noise reduction is possible.

The normalized MMSE is

$$\tilde{J}_x(\mathbf{h}_\mathrm{o}) \triangleq \frac{J_x(\mathbf{h}_\mathrm{o})}{J_x(\mathbf{u}_1)} = \frac{J_x(\mathbf{h}_\mathrm{o})}{\sigma_v^2}, \tag{2.16}$$

and $0 < \tilde{J}_x(\mathbf{h}_\mathrm{o}) < 1$.

2.3 Estimation of the Noise Samples

In this section, we will estimate the noise samples from the observations $y(n)$. Define the error signal between the noise sample at time n and its estimate:

$$e_v(n) = v(n) - \hat{v}(n) = v(n) - \mathbf{g}^T\mathbf{y}(n), \tag{2.17}$$

where

$$\mathbf{g} = \left[g_0 \; g_1 \; \cdots \; g_{L-1}\right]^T$$

is an FIR filter of length L. The MSE criterion associated with (2.17) is,

$$J_v(\mathbf{g}) \stackrel{\triangle}{=} E\left\{e_v^2(n)\right\}. \tag{2.18}$$

The estimation of $v(n)$ in the MSE sense will tend to attenuate the clean speech.

The minimization of (2.18) leads to the Wiener-Hopf equation:

$$\begin{aligned} \mathbf{g}_\mathrm{o} &= \mathbf{R}_y^{-1}\mathbf{r}_v = \mathbf{R}_y^{-1}\mathbf{R}_v\mathbf{u}_1 \\ &= \left[\mathrm{SNR}\cdot\mathbf{I} + \tilde{\mathbf{R}}_x^{-1}\tilde{\mathbf{R}}_v\right]^{-1}\tilde{\mathbf{R}}_x^{-1}\tilde{\mathbf{R}}_v\mathbf{u}_1. \end{aligned}$$

We have,

$$\lim_{\mathrm{SNR}\to\infty}\mathbf{g}_\mathrm{o} = \mathbf{0}, \tag{2.19}$$

$$\lim_{\mathrm{SNR}\to 0}\mathbf{g}_\mathrm{o} = \mathbf{u}_1. \tag{2.20}$$

The MSE for the particular filter $\mathbf{u}_1$ (no clean speech reduction) is,

$$J_v(\mathbf{u}_1) = E\left\{x^2(n)\right\} = \sigma_x^2. \tag{2.21}$$

Therefore, the MMSE and the normalized MMSE are respectively,

$$J_v(\mathbf{g}_\mathrm{o}) = \sigma_v^2 - \mathbf{r}_v^T\mathbf{R}_y^{-1}\mathbf{r}_v = \sigma_v^2 - \mathbf{r}_v^T\mathbf{g}_\mathrm{o}, \tag{2.22}$$

$$\tilde{J}_v(\mathbf{g}_\mathrm{o}) \stackrel{\triangle}{=} \frac{J_v(\mathbf{g}_\mathrm{o})}{J_v(\mathbf{u}_1)} = \frac{J_v(\mathbf{g}_\mathrm{o})}{\sigma_x^2}. \tag{2.23}$$

Since $J_v(\mathbf{g}_\mathrm{o}) < J_v(\mathbf{u}_1)$, the Wiener filter will be able to reduce the level of clean speech of the signal $y(n)$. As a result, $0 < \tilde{J}_v(\mathbf{g}_\mathrm{o}) < 1$.

In the next section, we will see that while the normalized MMSE, $\tilde{J}_x(\mathbf{h}_\mathrm{o})$, of the clean speech estimation plays a key role in noise reduction, the normalized MMSE, $\tilde{J}_v(\mathbf{g}_\mathrm{o})$, of the noise process estimation plays a key role in speech distortion.

2.4 Important Relationships Between Noise Reduction and Speech Distortion

Obviously, there are some important relationships between the estimation of the clean speech and noise samples. We immediately see from (2.15) and (2.22) that the two MMSEs are equal,

$$J_x(\mathbf{h}_\mathrm{o}) = J_v(\mathbf{g}_\mathrm{o}). \tag{2.24}$$

However, the normalized MMSEs are not, in general. Indeed, we have a relation between the two:

$$\begin{aligned} \tilde{J}_v(\mathbf{g}_\mathrm{o}) &= \frac{J_v(\mathbf{g}_\mathrm{o})}{\sigma_x^2} = \frac{J_x(\mathbf{h}_\mathrm{o})}{\sigma_x^2} \\ &= \frac{\sigma_v^2}{\sigma_x^2}\frac{J_x(\mathbf{h}_\mathrm{o})}{\sigma_v^2} = \frac{\tilde{J}_x(\mathbf{h}_\mathrm{o})}{\mathrm{SNR}}. \end{aligned} \tag{2.25}$$

So the only situation where the two normalized MMSEs are equal is when the SNR is equal to 1. For SNR < 1, $\tilde{J}_x(\mathbf{h}_\mathrm{o}) < \tilde{J}_v(\mathbf{g}_\mathrm{o})$ and for SNR > 1, $\tilde{J}_v(\mathbf{g}_\mathrm{o}) < \tilde{J}_x(\mathbf{h}_\mathrm{o})$. Also, $\tilde{J}_x(\mathbf{h}_\mathrm{o}) <$ SNR and $\tilde{J}_v(\mathbf{g}_\mathrm{o}) < 1/$SNR.

From (2.11) and (2.19), we get a relation between the two optimal filters:

$$\mathbf{h}_\mathrm{o} = \mathbf{u}_1 - \mathbf{g}_\mathrm{o}. \tag{2.26}$$

In fact, minimizing $J_x(\mathbf{h})$ or $J_v(\mathbf{u}_1 - \mathbf{h})$ with respect to $\mathbf{h}$ is equivalent. In the same manner, minimizing $J_v(\mathbf{g})$ or $J_x(\mathbf{u}_1 - \mathbf{g})$ with respect to $\mathbf{g}$ is the same thing. At the optimum, we have,

$$\begin{aligned} e_{x,\mathrm{o}}(n) &= x(n) - \mathbf{h}_\mathrm{o}^T \mathbf{y}(n) = x(n) - [\mathbf{u}_1 - \mathbf{g}_\mathrm{o}]^T [\mathbf{x}(n) + \mathbf{v}(n)] \\ &= -v(n) + \mathbf{g}_\mathrm{o}^T \mathbf{y}(n) = -e_{v,\mathrm{o}}(n). \end{aligned} \tag{2.27}$$

We can easily verify the following:

$$\begin{aligned} J_v(\mathbf{h}_\mathrm{o}) &= J_x(\mathbf{g}_\mathrm{o}) \\ &= \sigma_y^2 - 3J_x(\mathbf{h}_\mathrm{o}), \end{aligned} \tag{2.28}$$

which implies that $J_x(\mathbf{h}_\mathrm{o}) < \sigma_y^2/3$. We already know that $J_x(\mathbf{h}_\mathrm{o}) < \sigma_v^2$ and $J_x(\mathbf{h}_\mathrm{o}) < \sigma_x^2$.

The optimal estimation of the clean speech, in the Wiener sense, is in fact what we call noise reduction:

$$\hat{x}_\mathrm{o}(n) = \mathbf{h}_\mathrm{o}^T \mathbf{y}(n), \tag{2.29}$$

or equivalently, if the noise is estimated first:

$$\hat{v}_\mathrm{o}(n) = \mathbf{g}_\mathrm{o}^T \mathbf{y}(n), \tag{2.30}$$

we can use this estimate to reduce the noise from the observed signal:

$$\hat{x}_\mathrm{o}(n) = y(n) - \hat{v}_\mathrm{o}(n). \tag{2.31}$$

The power of the estimated clean speech signal with the optimal Wiener filter is,

$$\begin{aligned} E\left\{\hat{x}_\mathrm{o}^2(n)\right\} &= \mathbf{h}_\mathrm{o}^T \mathbf{R}_y \mathbf{h}_\mathrm{o} = \sigma_x^2 - J_x(\mathbf{h}_\mathrm{o}) \\ &= \mathbf{h}_\mathrm{o}^T \mathbf{R}_x \mathbf{h}_\mathrm{o} + \mathbf{h}_\mathrm{o}^T \mathbf{R}_v \mathbf{h}_\mathrm{o}, \end{aligned} \tag{2.32}$$

which is the sum of two terms. The first one is the power of the attenuated clean speech and the second one is the power of the residual noise (always greater than zero). While noise reduction is feasible with the Wiener filter, expression (2.32) shows that the price to pay for this is also a reduction of the clean speech [by a quantity equal to $J_x(\mathbf{h}_\mathrm{o}) + \mathbf{h}_\mathrm{o}^T \mathbf{R}_v \mathbf{h}_\mathrm{o}$ and this implies distortion], since $\mathbf{h}_\mathrm{o}^T \mathbf{R}_x \mathbf{h}_\mathrm{o} < \sigma_x^2$. In other words, the power of the attenuated clean speech signal is, obviously, always smaller than the power of the clean speech itself; this means that parts of the clean speech are attenuated in the process and as a result, distortion is unavoidable with this approach.

We now define the speech-distortion index due to the optimal filtering operation as,

$$\begin{aligned}\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) &\triangleq \frac{E\left\{\left[x(n) - \mathbf{h}_{\mathrm{o}}^T\mathbf{x}(n)\right]^2\right\}}{\sigma_x^2} \\ &= \frac{\mathbf{g}_{\mathrm{o}}^T\mathbf{R}_x\mathbf{g}_{\mathrm{o}}}{\sigma_x^2} = \frac{1}{\mathrm{SNR}}\left[\tilde{J}_x(\mathbf{h}_{\mathrm{o}}) - \mathbf{h}_{\mathrm{o}}^T\tilde{\mathbf{R}}_v\mathbf{h}_{\mathrm{o}}\right] < \tilde{J}_v(\mathbf{g}_{\mathrm{o}}).\end{aligned} \tag{2.33}$$

Clearly, this index is always between 0 and 1 for the optimal filter. Also,

$$\lim_{\mathrm{SNR}\to 0} \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) = 1, \tag{2.34}$$

$$\lim_{\mathrm{SNR}\to \infty} \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) = 0. \tag{2.35}$$

So when $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ is close to 1, the speech signal is highly distorted and when $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ is near 0, the speech signal is lowly distorted. We deduce that for low SNRs, the Wiener filter can have a disastrous effect on the speech signal.

Similarly, we define the noise-reduction factor due to the Wiener filter as,

$$\begin{aligned}\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) &\triangleq \frac{\sigma_v^2}{E\left\{\left[\mathbf{h}_{\mathrm{o}}^T\mathbf{v}(n)\right]^2\right\}} \\ &= \frac{\sigma_v^2}{\mathbf{h}_{\mathrm{o}}^T\mathbf{R}_v\mathbf{h}_{\mathrm{o}}} = \frac{1}{\mathrm{SNR}\left[\tilde{J}_v(\mathbf{g}_{\mathrm{o}}) - \mathbf{g}_{\mathrm{o}}^T\tilde{\mathbf{R}}_x\mathbf{g}_{\mathrm{o}}\right]} > \frac{1}{\tilde{J}_x(\mathbf{h}_{\mathrm{o}})},\end{aligned} \tag{2.36}$$

and $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) > 1$. The greater is $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$, the more noise reduction we have. Also,

$$\lim_{\mathrm{SNR}\to 0} \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) = \infty, \tag{2.37}$$

$$\lim_{\mathrm{SNR}\to \infty} \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) = 1. \tag{2.38}$$

Using (2.33) and (2.36), we obtain important relations between the speech-distortion index and the noise-reduction factor:

$$\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) = \frac{1}{\mathrm{SNR}}\left[\tilde{J}_x(\mathbf{h}_{\mathrm{o}}) - \frac{1}{\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})}\right], \tag{2.39}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) = \frac{1}{\mathrm{SNR}\left[\tilde{J}_v(\mathbf{g}_{\mathrm{o}}) - \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})\right]}. \tag{2.40}$$

Therefore, for the optimum filter, when the SNR is very large, there is little speech distortion and little noise reduction (which is not really needed in this situation). On the other hand, when the SNR is very small, speech distortion is large as well as noise reduction. Using the fact that $J_x(\mathbf{h}_\mathrm{o}) < \sigma_y^2/3$, we can easily derive from (2.39) and (2.40) that,

$$\begin{cases} \xi_\mathrm{nr}(\mathbf{h}_\mathrm{o}) > \frac{1}{\mathrm{SNR}} & \text{if} \quad \mathrm{SNR} \leq 1/2 \\ \xi_\mathrm{nr}(\mathbf{h}_\mathrm{o}) \geq \frac{3}{\mathrm{SNR}+1} & \text{if } 1/2 < \mathrm{SNR} \leq 2 \\ \xi_\mathrm{nr}(\mathbf{h}_\mathrm{o}) > 1 & \text{if} \quad \mathrm{SNR} > 2 \end{cases}, \tag{2.41}$$

and

$$\begin{cases} \upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{o}) < 1 & \text{if} \quad \mathrm{SNR} \leq 1/2 \\ \upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{o}) < \frac{\mathrm{SNR}+1}{3\mathrm{SNR}} & \text{if } 1/2 < \mathrm{SNR} \leq 2 \\ \upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{o}) < \frac{1}{\mathrm{SNR}} & \text{if} \quad \mathrm{SNR} > 2 \end{cases}. \tag{2.42}$$

Equations (2.41) and (2.42) give the lower bound for the noise-reduction factor and the upper bound for the speech-distortion index respectively. These bounds can be further refined. But before going further, let us first analyze the *a posteriori* SNR, which is defined, after noise reduction with the Wiener filter, as,

$$\begin{aligned} \mathrm{SNR}_\mathrm{o} &\triangleq \frac{\mathbf{h}_\mathrm{o}^T \mathbf{R}_x \mathbf{h}_\mathrm{o}}{\mathbf{h}_\mathrm{o}^T \mathbf{R}_v \mathbf{h}_\mathrm{o}} \\ &= \mathrm{SNR}\, \frac{\mathbf{h}_\mathrm{o}^T \tilde{\mathbf{R}}_x \mathbf{h}_\mathrm{o}}{\mathbf{h}_\mathrm{o}^T \tilde{\mathbf{R}}_v \mathbf{h}_\mathrm{o}} = -1 + \mathrm{SNR}\, \xi_\mathrm{nr}(\mathbf{h}_\mathrm{o}) \left[1 - \tilde{J}_v(\mathbf{g}_\mathrm{o})\right] \\ &= -1 + \frac{1 - \tilde{J}_v(\mathbf{g}_\mathrm{o})}{\tilde{J}_v(\mathbf{g}_\mathrm{o}) - \upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{o})}. \end{aligned} \tag{2.43}$$

It can be easily verified that,

$$\mathrm{SNR}_\mathrm{o} > \frac{\mathrm{SNR}}{\tilde{J}_x(\mathbf{h}_\mathrm{o})} - 2. \tag{2.44}$$

We now give a proposition showing the relationship between the *a priori* SNR and the *a posteriori* SNR.

Proposition: With the Wiener filter, the *a posteriori* SNR and the *a priori* SNR satisfy

$$\mathrm{SNR}_\mathrm{o} = \frac{\mathbf{h}_\mathrm{o}^T \mathbf{R}_x \mathbf{h}_\mathrm{o}}{\mathbf{h}_\mathrm{o}^T \mathbf{R}_v \mathbf{h}_\mathrm{o}} \geq \mathrm{SNR} = \frac{\mathbf{u}_1^T \mathbf{R}_x \mathbf{u}_1}{\mathbf{u}_1^T \mathbf{R}_v \mathbf{u}_1}. \tag{2.45}$$

Proof. From their definitions, we know that all three matrices, $\mathbf{R}_x$, $\mathbf{R}_v$, and $\mathbf{R}_y$ are symmetric, and positive semi-definite. We further assume that $\mathbf{R}_v$ is positive definite so its inverse exists. In addition, based on the independence assumption between the speech signal and noise, we have $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_v$. In case that both $\mathbf{R}_x$ and $\mathbf{R}_v$ are diagonal matrices, or $\mathbf{R}_v$ is a scaled version of $\mathbf{R}_x$ (i.e., $\mathbf{R}_x = \mathrm{SNR} \cdot \mathbf{R}_v$), it can be easily seen that $\mathrm{SNR}_\mathrm{o} = \mathrm{SNR}$. Here, we consider more complicated situations where at least one of the $\mathbf{R}_x$ and $\mathbf{R}_v$ matrices is not diagonal. In this case, according to [45], there exists a linear transform that can simultaneously diagonalize $\mathbf{R}_x$, $\mathbf{R}_v$, and $\mathbf{R}_y$. The process is done as follows.

$$\begin{aligned}
\mathbf{R}_x &= (\mathbf{B}^T)^{-1}\mathbf{\Lambda}\mathbf{B}^{-1}, \\
\mathbf{R}_v &= (\mathbf{B}^T)^{-1}\mathbf{B}^{-1}, \\
\mathbf{R}_y &= (\mathbf{B}^T)^{-1}[\mathbf{I}+\mathbf{\Lambda}]\mathbf{B}^{-1},
\end{aligned} \tag{2.46}$$

where again $\mathbf{I}$ is the identity matrix,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & 0 & \lambda_L \end{bmatrix} \tag{2.47}$$

is the eigenvalue matrix of $\mathbf{R}_v^{-1}\mathbf{R}_x$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0$, $\mathbf{B}$ is the eigenvector matrix of $\mathbf{R}_v^{-1}\mathbf{R}_x$, and

$$\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{B} = \mathbf{B}\mathbf{\Lambda}. \tag{2.48}$$

Note that $\mathbf{B}$ is not necessarily orthogonal since $\mathbf{R}_v^{-1}\mathbf{R}_x$ is not necessarily symmetric. Then from the definition of SNR and SNR_o, we immediately have

$$\mathrm{SNR} = \frac{\mathbf{u}_1^T\mathbf{R}_x\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{R}_v\mathbf{u}_1} = \frac{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Lambda}\mathbf{B}^{-1}\mathbf{u}_1}{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{B}^{-1}\mathbf{u}_1}, \tag{2.49}$$

and

$$\begin{aligned}
\mathrm{SNR}_\mathrm{o} = \frac{\mathbf{h}_\mathrm{o}^T\mathbf{R}_x\mathbf{h}_\mathrm{o}}{\mathbf{h}_\mathrm{o}^T\mathbf{R}_v\mathbf{h}_\mathrm{o}} &= \frac{\mathbf{u}_1^T\mathbf{R}_x^T\mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{R}_x^T\mathbf{R}_y^{-1}\mathbf{R}_v\mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{u}_1} \\
&= \frac{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}\mathbf{B}^{-1}\mathbf{u}_1}{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}\mathbf{B}^{-1}\mathbf{u}_1} \\
&= \frac{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Sigma_1}\mathbf{B}^{-1}\mathbf{u}_1}{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Sigma_2}\mathbf{B}^{-1}\mathbf{u}_1},
\end{aligned} \tag{2.50}$$

where

$$\begin{aligned}
\mathbf{\Sigma_1} &\overset{\triangle}{=} \mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda} \\
&= \begin{bmatrix} \frac{\lambda_1^3}{(1+\lambda_1)^2} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2^3}{(1+\lambda_2)^2} & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & 0 & \frac{\lambda_L^3}{(1+\lambda_L)^2} \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}\mathbf{\Sigma_2} &\stackrel{\triangle}{=} \mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda} \\ &= \begin{bmatrix} \frac{\lambda_1^2}{(1+\lambda_1)^2} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2^2}{(1+\lambda_2)^2} & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & 0 & \frac{\lambda_L^2}{(1+\lambda_L)^2} \end{bmatrix}\end{aligned}$$

are two diagonal matrices. If for the ease of expression we denote $\mathbf{B}^{-1}$ as $\mathbf{A} = \mathbf{B}^{-1} = [a_{ij}]$, then both SNR and SNR_o can be rewritten as

$$\begin{aligned}\text{SNR} &= \frac{\sum_{i=1}^L \lambda_i a_{i1}^2}{\sum_{i=1}^L a_{i1}^2}, \\ \text{SNR}_\text{o} &= \frac{\sum_{i=1}^L \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2}{\sum_{i=1}^L \frac{\lambda_i^2}{(1+\lambda_i)^2} a_{i1}^2}.\end{aligned} \tag{2.51}$$

Since $\sum_{i=1}^L \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2$, $\sum_{i=1}^L \frac{\lambda_i^2}{(1+\lambda_i)^2} a_{i1}^2$, $\sum_{i=1}^L \lambda_i a_{i1}^2$, and $\sum_{i=1}^L a_{i1}^2$ all are nonnegative numbers, as long as we can show that the inequality

$$\sum_{i=1}^L \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^L a_{i1}^2 \geq \sum_{i=1}^L \frac{\lambda_i^2}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^L \lambda_i a_{i1}^2 \tag{2.52}$$

holds, then $\text{SNR}_\text{o} \geq \text{SNR}$. Now we prove this inequality by way of induction.

- Basic Step: If $L = 2$,

$$\begin{aligned}&\sum_{i=1}^2 \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^2 a_{i1}^2 \\ &= \frac{\lambda_1^3}{(1+\lambda_1)^2} a_{11}^4 + \frac{\lambda_2^3}{(1+\lambda_2)^2} a_{21}^4 + \left[\frac{\lambda_1^3}{(1+\lambda_1)^2} + \frac{\lambda_2^3}{(1+\lambda_2)^2}\right] a_{11}^2 a_{21}^2.\end{aligned}$$

  Since $\lambda_i \geq 0$, it is trivial to show that

$$\frac{\lambda_1^3}{(1+\lambda_1)^2} + \frac{\lambda_2^3}{(1+\lambda_2)^2} \geq \frac{\lambda_1^2\lambda_2}{(1+\lambda_1)^2} + \frac{\lambda_1\lambda_2^2}{(1+\lambda_2)^2},$$

  where "=" holds when $\lambda_1 = \lambda_2$. Therefore

$$\sum_{i=1}^2 \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^2 a_{i1}^2$$

$$\geq \frac{\lambda_1^3}{(1+\lambda_1)^2}a_{11}^4 + \frac{\lambda_2^3}{(1+\lambda_2)^2}a_{21}^4 + \left[\frac{\lambda_1^2\lambda_2}{(1+\lambda_1)^2} + \frac{\lambda_1\lambda_2^2}{(1+\lambda_2)^2}\right]a_{11}^2 a_{21}^2$$
$$= \sum_{i=1}^{2} \frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{2} \lambda_i a_{i1}^2,$$

so the property is true for $L = 2$, where "=" holds when any one of a_{11} and a_{21} is equal to 0 (note that a_{11} and a_{21} cannot be zero at the same time since $\mathbf{A}$ is invertible) or when $\lambda_1 = \lambda_2$.

- Inductive Step: Assume that the property is true for $L = n$, i.e.,

$$\sum_{i=1}^{n} \frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{n} a_{i1}^2 \geq \sum_{i=1}^{n} \frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{n} \lambda_i a_{i1}^2.$$

We must prove that it is also true for $L = n + 1$. As a matter of fact,

$$\sum_{i=1}^{n+1} \frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{n+1} a_{i1}^2$$
$$= \left[\sum_{i=1}^{n} \frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2 + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2}a_{n+11}^2\right]\left[\sum_{i=1}^{n} a_{i1}^2 + a_{n+11}^2\right]$$
$$= \left[\sum_{i=1}^{n} \frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2\right]\left[\sum_{i=1}^{n} a_{i1}^2\right] + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2}a_{n+11}^4$$
$$+ \sum_{i=1}^{n}\left[\frac{\lambda_i^3}{(1+\lambda_i)^2} + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2}\right]a_{i1}^2 a_{n+11}^2. \tag{2.53}$$

Using the induction hypothesis, and also the fact that

$$\frac{\lambda_i^3}{(1+\lambda_i)^2} + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2} \geq \frac{\lambda_i^2\lambda_{n+1}}{(1+\lambda_i)^2} + \frac{\lambda_i\lambda_{n+1}^2}{(1+\lambda_{n+1})^2},$$

hence

$$\sum_{i=1}^{n+1} \frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{n+1} a_{i1}^2$$
$$\geq \sum_{i=1}^{n} \frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{n} \lambda_i a_{i1}^2 + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2}a_{n+11}^4$$
$$+ \sum_{i=1}^{n}\left[\frac{\lambda_i^2\lambda_{n+1}}{(1+\lambda_i)^2} + \frac{\lambda_i\lambda_{n+1}^2}{(1+\lambda_{n+1})^2}\right]a_{i1}^2 a_{n+11}^2$$
$$= \sum_{i=1}^{n+1} \frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2 \sum_{i=1}^{n+1} \lambda_i a_{i1}^2, \tag{2.54}$$

where "=" holds when all the λ_i's corresponding to nonzero a_{i1} are equal, where $i = 1, 2, \ldots, n+1$. That completes the proof.

Even though it can improve the SNR, the Wiener filter does not maximize the *a posteriori* SNR. As a matter of fact, (2.43) is well known as the generalized Rayleigh quotient. So the filter that really maximizes the *a posteriori* SNR is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{R}_v^{-1}\mathbf{R}_x$.

Knowing that $\mathrm{SNR}_\mathrm{o} \geq \mathrm{SNR}$, we can now refine the lower bound for $\xi_{\mathrm{nr}}(\mathbf{h}_\mathrm{o})$. As a matter of fact, it follows from (2.43) that

$$\mathrm{SNR}_\mathrm{o} = -1 + \frac{1 - \tilde{J}_v(\mathbf{g}_\mathrm{o})}{\tilde{J}_v(\mathbf{g}_\mathrm{o}) - \upsilon_{\mathrm{sd}}(\mathbf{g}_\mathrm{o})} \geq \mathrm{SNR}.$$

Since $\upsilon_{\mathrm{sd}}(\mathbf{g}_\mathrm{o}) < \tilde{J}_v(\mathbf{g}_\mathrm{o})$, and $0 \leq \upsilon_{\mathrm{sd}}(\mathbf{g}_\mathrm{o}) \leq 1$, it can be easily shown that

$$\xi_{\mathrm{nr}}(\mathbf{h}_\mathrm{o}) \geq \frac{\mathrm{SNR}+2}{\mathrm{SNR}}. \tag{2.55}$$

This lower bound for $\xi_{\mathrm{nr}}(\mathbf{h}_\mathrm{o})$ is tighter than the one given in (2.41). Similarly, we can derive that

$$\upsilon_{\mathrm{sd}}(\mathbf{g}_\mathrm{o}) \leq \frac{1}{2\mathrm{SNR}+1}. \tag{2.56}$$

It can be easily verified that this upper bound for $\upsilon_{\mathrm{sd}}(\mathbf{g}_\mathrm{o})$ is tighter than the one given in (2.42). Figure 2.1 illustrates expressions (2.55) and (2.56).

We now introduce another index for noise reduction:

$$\zeta_{\mathrm{nr}}(\mathbf{h}_\mathrm{o}) \overset{\triangle}{=} 1 - \tilde{J}_x(\mathbf{h}_\mathrm{o}) < 1. \tag{2.57}$$

The closer is $\zeta_{\mathrm{nr}}(\mathbf{h}_\mathrm{o})$ to 1, the more noise reduction we get. This index will be helpful to use in the following sections.

2.5 Particular Case: White Gaussian Noise

In this section, we assume that the additive noise is white, so that,

$$\mathbf{r}_v = \sigma_v^2 \mathbf{u}_1. \tag{2.58}$$

From (2.16) and (2.23), we observe that the two normalized MMSEs are

$$\tilde{J}_x(\mathbf{h}_\mathrm{o}) = h_{\mathrm{o},0}, \tag{2.59}$$

$$\tilde{J}_v(\mathbf{g}_\mathrm{o}) = \frac{1 - g_{\mathrm{o},0}}{\mathrm{SNR}} = \frac{h_{\mathrm{o},0}}{\mathrm{SNR}}, \tag{2.60}$$

where $h_{\mathrm{o},0}$ and $g_{\mathrm{o},0}$ are the first components of vectors $\mathbf{h}_\mathrm{o}$ and $\mathbf{g}_\mathrm{o}$, respectively. Clearly, $0 < h_{\mathrm{o},0} < 1$ and $0 < g_{\mathrm{o},0} < 1$. Hence, the normalized MMSE $\tilde{J}_x(\mathbf{h}_\mathrm{o})$ is completely governed by the first element of the Wiener filter $\mathbf{h}_\mathrm{o}$.

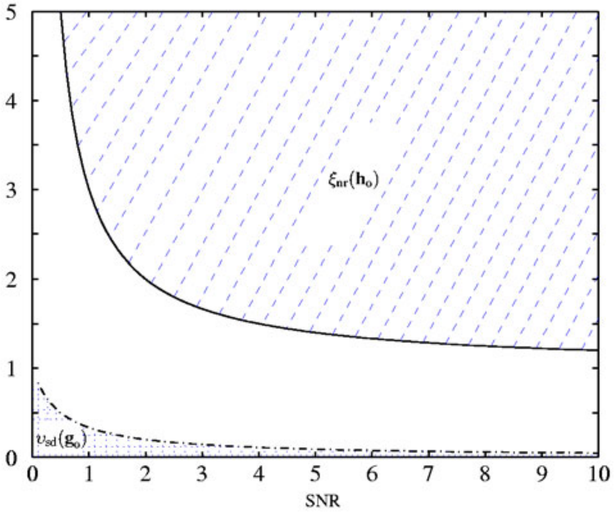


Fig. 2.1. Illustration of the areas where $\xi_{\rm nr}(\mathbf{h}_{\rm o})$ and $\upsilon_{\rm sd}(\mathbf{g}_{\rm o})$ take their values as functions of the SNR. $\xi_{\rm nr}(\mathbf{h}_{\rm o})$ can take any value above the solid line while $\upsilon_{\rm sd}(\mathbf{g}_{\rm o})$ can take any value under the dotted line.

Now, the speech-distortion index and the noise-reduction factor for the optimal filter can be simplified:

$$\upsilon_{\rm sd}(\mathbf{g}_{\rm o}) = \frac{1}{\rm SNR}\left[h_{{\rm o},0} - \mathbf{h}_{\rm o}^T\mathbf{h}_{\rm o}\right] \tag{2.61}$$

$$= \frac{\mathbf{g}_{\rm o}^T\mathbf{h}_{\rm o}}{\rm SNR} = \frac{1}{\rm SNR}\left[g_{{\rm o},0} - \mathbf{g}_{\rm o}^T\mathbf{g}_{\rm o}\right],$$

$$\xi_{\rm nr}(\mathbf{h}_{\rm o}) = \frac{1}{\mathbf{h}_{\rm o}^T\mathbf{h}_{\rm o}}. \tag{2.62}$$

We also deduce from (2.61) that $h_{{\rm o},0} > \mathbf{h}_{\rm o}^T\mathbf{h}_{\rm o}$ and $g_{{\rm o},0} > \mathbf{g}_{\rm o}^T\mathbf{g}_{\rm o}$.

We know from the linear prediction theory that [43],

$$\mathbf{R}_y \begin{bmatrix} 1 \\ -\mathbf{a}_y \end{bmatrix} = \begin{bmatrix} E_y \\ \mathbf{0}_{(L-1)\times 1} \end{bmatrix}, \tag{2.63}$$

where $\mathbf{a}_y$ is the forward linear predictor and E_y is the corresponding error energy. Replacing the previous equation in (2.11), we obtain:

$$\mathbf{h}_{\rm o} = \mathbf{u}_1 - \sigma_v^2\mathbf{R}_y^{-1}\mathbf{u}_1 = \begin{bmatrix} h_{{\rm o},0} \\ \frac{\sigma_v^2}{E_y}\mathbf{a}_y \end{bmatrix}, \tag{2.64}$$

where

$$h_{{\rm o},0} = \tilde{J}_x(\mathbf{h}_{\rm o}) = 1 - \frac{\sigma_v^2}{E_y}. \tag{2.65}$$

Equation (2.64) shows how the Wiener filter is related to the forward predictor of the observed signal $y(n)$. This expression also gives a hint on how to choose the length of the optimal filter $\mathbf{h}_\mathrm{o}$: it should be equal to the length of the predictor $\mathbf{a}_y$ required to have a good prediction of the observed signal $y(n)$. Equation (2.65) contains some very interesting information. Indeed, if the clean speech signal is completely predictable, this means that $E_y \approx \sigma_v^2$ and $\tilde{J}_x(\mathbf{h}_\mathrm{o}) \approx 0$. On the other hand, if $x(n)$ is not predictable, we have $E_y \approx \sigma_y^2$ and $\tilde{J}_x(\mathbf{h}_\mathrm{o}) \approx 1 - \sigma_v^2/\sigma_y^2$. This implies that the Wiener filter is more efficient to reduce the level of noise for predictable signals than for unpredictable ones.

2.6 Better Ways to Manage Noise Reduction and Speech Distortion

For a noise-reduction/speech-enhancement system, we always expect that it can achieve maximal noise reduction without much speech distortion. From the previous section, however, we see that when noise reduction is maximized with the optimal Wiener filter, speech distortion is also maximized. One may ask a legitimate question: are there better ways to control the tradeoff between the conflicting requirements of noise reduction and speech distortion? Examining (2.33), one can see that to control the speech distortion, we have to minimize $E\left\{\left[x(n) - \mathbf{h}_\mathrm{o}^T\mathbf{x}(n)\right]^2\right\}$. This can be achieved by either manipulating $\mathbf{h}_\mathrm{o}$ or exploiting a speech model.

2.6.1 A Suboptimal Filter

Consider the suboptimal filter:

$$\mathbf{h}_\mathrm{s} = \mathbf{u}_1 - \mathbf{g}_\mathrm{s} = \mathbf{u}_1 - \alpha\mathbf{g}_\mathrm{o}, \tag{2.66}$$

where α is a real number. The MSE of the clean speech estimation corresponding to $\mathbf{h}_\mathrm{s}$ is,

$$\begin{aligned} J_x(\mathbf{h}_\mathrm{s}) &= E\left\{\left[x(n) - \mathbf{h}_\mathrm{s}^T\mathbf{y}(n)\right]^2\right\} \\ &= \sigma_v^2 - \alpha(2-\alpha)\mathbf{r}_v^T\mathbf{R}_y^{-1}\mathbf{r}_v, \end{aligned} \tag{2.67}$$

and, obviously, $J_x(\mathbf{h}_\mathrm{s}) \geq J_x(\mathbf{h}_\mathrm{o}),\ \forall\alpha$; we have equality for $\alpha = 1$. In order to have noise reduction, α must be chosen in such a way that $J_x(\mathbf{h}_\mathrm{s}) < J_x(\mathbf{u}_1)$, therefore,

$$0 < \alpha < 2. \tag{2.68}$$

We can check that,

$$J_v(\mathbf{g}_\mathrm{s}) = E\left\{\left[v(n) - \alpha\mathbf{g}_\mathrm{o}^T\mathbf{y}(n)\right]^2\right\} \\ = J_x(\mathbf{h}_\mathrm{s}). \tag{2.69}$$

Let

$$\hat{x}_\mathrm{s}(n) = \mathbf{h}_\mathrm{s}^T\mathbf{y}(n) \tag{2.70}$$

denote the estimation of the clean speech at time n with respect to $\mathbf{h}_\mathrm{s}$. The power of $\hat{x}_\mathrm{s}(n)$ is,

$$\begin{aligned} E\left\{\hat{x}_\mathrm{s}^2(n)\right\} &= \mathbf{h}_\mathrm{s}^T\mathbf{R}_y\mathbf{h}_\mathrm{s} \\ &= \left[\mathbf{u}_1 - \alpha\mathbf{R}_y^{-1}\mathbf{r}_v\right]^T\left[\mathbf{r}_y - \alpha\mathbf{r}_v\right] \\ &= \sigma_x^2 + (1-2\alpha)\sigma_v^2 + \alpha^2\mathbf{r}_v^T\mathbf{R}_y^{-1}\mathbf{r}_v \\ &= \mathbf{h}_\mathrm{s}^T\mathbf{R}_x\mathbf{h}_\mathrm{s} + \mathbf{h}_\mathrm{s}^T\mathbf{R}_v\mathbf{h}_\mathrm{s}. \end{aligned} \tag{2.71}$$

The speech-distortion index corresponding to the filter $\mathbf{h}_\mathrm{s}$ is,

$$\begin{aligned} \upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{s}) &= \frac{E\left\{\left[x(n) - \mathbf{h}_\mathrm{s}^T\mathbf{x}(n)\right]^2\right\}}{\sigma_x^2} \\ &= \alpha^2\mathbf{g}_\mathrm{o}^T\tilde{\mathbf{R}}_x\mathbf{g}_\mathrm{o} = \alpha^2\upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{o}). \end{aligned} \tag{2.72}$$

The previous expression shows that the ratio of the speech-distortion indices corresponding to the two filters $\mathbf{g}_\mathrm{s}$ and $\mathbf{g}_\mathrm{o}$ depends on α only.

In order to have less distortion with the suboptimal filter $\mathbf{h}_\mathrm{s}$ than with the Wiener filter $\mathbf{h}_\mathrm{o}$, we must find α in such a way that,

$$\upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{s}) < \upsilon_\mathrm{sd}(\mathbf{g}_\mathrm{o}), \tag{2.73}$$

hence, the condition on α should be

$$-1 < \alpha < 1. \tag{2.74}$$

Finally, the suboptimal filter $\mathbf{h}_\mathrm{s}$ can reduce the level of noise of the observed signal $y(n)$ but with less distortion than the Wiener filter $\mathbf{h}_\mathrm{o}$ if α is taken such as,

$$0 < \alpha < 1. \tag{2.75}$$

For the extreme cases $\alpha = 0$ and $\alpha = 1$ we obtain respectively $\mathbf{h}_\mathrm{s} = \mathbf{u}_1$, no noise reduction at all but no additional distortion added, and $\mathbf{h}_\mathrm{s} = \mathbf{h}_\mathrm{o}$, maximum noise reduction with maximum speech distortion.

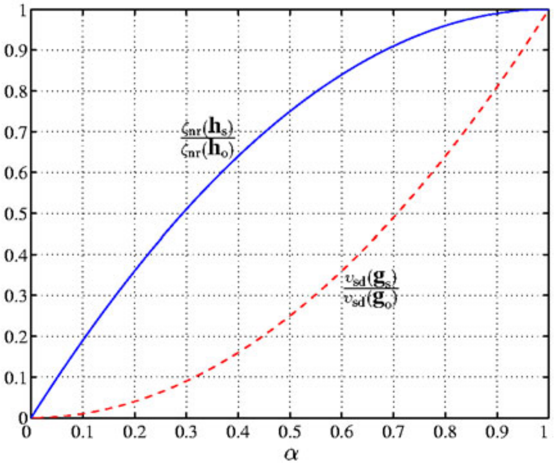


Fig. 2.2. $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ (dashed line) and $\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$ (solid line), both as a function of α.

Since

$$\begin{aligned} J_v(\mathbf{g}_{\mathrm{s}}) &= \mathbf{g}_{\mathrm{s}}^T \mathbf{R}_x \mathbf{g}_{\mathrm{s}} + \mathbf{h}_{\mathrm{s}}^T \mathbf{R}_v \mathbf{h}_{\mathrm{s}} \\ &= \sigma_x^2 \mathbf{g}_{\mathrm{s}}^T \tilde{\mathbf{R}}_x \mathbf{g}_{\mathrm{s}} + \sigma_v^2 \mathbf{h}_{\mathrm{s}}^T \tilde{\mathbf{R}}_v \mathbf{h}_{\mathrm{s}} \\ &= J_x(\mathbf{h}_{\mathrm{s}}), \end{aligned} \tag{2.76}$$

it follows immediately that the speech-distortion index and the noise-reduction factor due to $\mathbf{h}_{\mathrm{s}}$ are,

$$\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}}) = \frac{1}{\mathrm{SNR}} \left[\tilde{J}_x(\mathbf{h}_{\mathrm{s}}) - \frac{1}{\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})} \right], \tag{2.77}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}}) = \frac{1}{\mathrm{SNR} \left[\tilde{J}_v(\mathbf{g}_{\mathrm{s}}) - \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}}) \right]}. \tag{2.78}$$

Unlike $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ which depends on α only, $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$ does not. However, using (2.67) and (2.15), we find that,

$$\frac{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})}{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})} = \frac{1 - \tilde{J}_x(\mathbf{h}_{\mathrm{s}})}{1 - \tilde{J}_x(\mathbf{h}_{\mathrm{o}})} = \alpha(2 - \alpha). \tag{2.79}$$

Figure 2.2 plots $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ and $\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$ as functions of α. For example, for $\alpha = 0.7$, we see that the speech-distortion index with the sub-optimal filter represents 49% of the speech-distortion index with the Wiener filter while the noise-reduction index is 91%.

2.6.2 Noise Reduction Exploiting the Speech Model

Section 2.5 has shown that the Wiener filter is more efficient to attenuate the level of noise for predictable signals than for unpredictable ones. In fact, it is well known that speech can be represented by an autoregressive (AR) process; thus, speech can be seen as the output of an all-pole linear system where the input is a zero-mean white Gaussian process, $w(n)$, with variance σ_w^2. The clean speech signal is then given by,

$$\begin{aligned} x(n) &= \sum_{l=1}^{L} a_{x,l} x(n-l) + w(n) \\ &= \mathbf{a}_x^T \mathbf{x}(n-1) + w(n), \end{aligned} \tag{2.80}$$

where $a_{x,l}$ are the parameters of the AR process. This model is very often combined with the Kalman filter to enhance a noisy speech signal; see, for example, [27], [28], and [29]. The main challenge in this approach is to get an accurate estimate of the AR parameters from the observations.

We can use this model in the Wiener context with some advantages. For that, in this section, we assume that the additive noise, $v(n)$, of the observed signal, $y(n)$, is white. The cross-correlation vector, $\mathbf{p}$, between the noisy and clean speech signals that appears in the Wiener-Hopf equation is now:

$$\begin{aligned} \mathbf{p} = E\{\mathbf{y}(n)x(n)\} &= E\left\{\mathbf{y}(n)\mathbf{x}^T(n-1)\right\}\mathbf{a}_x + \sigma_w^2 \mathbf{u}_1 \\ &= E\left\{\mathbf{y}(n)\left[\mathbf{y}(n-1) - \mathbf{v}(n-1)\right]^T\right\}\mathbf{a}_x + \sigma_w^2 \mathbf{u}_1 \\ &= (\mathbf{R}_{y,1} - \mathbf{R}_{v,1})\mathbf{a}_x + \sigma_w^2 \mathbf{u}_1, \end{aligned} \tag{2.81}$$

where

$$\mathbf{R}_{y,1} \stackrel{\triangle}{=} E\left\{\mathbf{y}(n)\mathbf{y}^T(n-1)\right\},$$

and

$$\begin{aligned} \mathbf{R}_{v,1} &\stackrel{\triangle}{=} E\left\{\mathbf{v}(n)\mathbf{v}^T(n-1)\right\} \\ &= \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \sigma_v^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_v^2 & 0 & \cdots & 0 & 0 \\ & \vdots & & & \vdots & \\ 0 & 0 & 0 & \cdots & \sigma_v^2 & 0 \end{bmatrix}. \end{aligned}$$

We deduce the optimal filter:

$$\mathbf{h}_{\mathrm{o}} = \mathbf{R}_y^{-1}(\mathbf{R}_{y,1} - \mathbf{R}_{v,1})\mathbf{a}_x + \sigma_w^2 \mathbf{R}_y^{-1}\mathbf{u}_1. \tag{2.82}$$

Equation (2.82) shows the relationship between the Wiener filter and the AR parameters of the clean speech signal. When $v(n)$ is a white Gaussian noise signal, (2.82) yields similar results as in (2.64).

2.6.3 Noise Reduction with Multiple Microphones

In more and more applications, multiple microphone signals are available. Therefore, it is interesting to investigate deeply the multichannel case. One of the first papers to do so is a paper written by Doclo and Moonen [42], where the optimal filter is derived as well as a general class of estimators. The authors also show how the generalized singular value decomposition can be used in this spatio-temporal technique. In this section, we take a slightly different approach. We will see, in particular, that we can reduce the level of noise without distorting the speech signal. This result was never observed before.

We suppose that we have a linear array consisting of M microphones whose outputs are denoted as $y_m(n)$, $m = 0, 1, \cdots, M-1$. Without loss of generality, we select microphone 0 as the reference point and to simplify the analysis, we consider the following propagation model:

$$y_m(n) = \beta_m s(n - t - \tau_m) + v_m(n), \ m = 0, 1, \cdots, M-1, \tag{2.83}$$

where β_m is the attenuation factor (with $\beta_0 = 1$), t is the propagation time from the unknown speech source $s(n)$ to microphone 0, $v_m(n)$ is an additive noise signal at the mth microphone, and τ_m is the relative delay between microphones 0 and m, with $\tau_0 = 0$.

In the following, we assume that the relative delays τ_m, $m = 1, \cdots, M-1$, are known or can easily be estimated. So our first step is the design of a simple delay-and-sum beamformer, which spatially aligns the microphone signals to the direction of the speech source. From now on, we will work on the aligned signals:

$$\begin{aligned} z_m(n) &= y_m(n + \tau_m) \\ &= \beta_m s(n - t) + v_m(n + \tau_m), \\ &= x_m(n) + v_m(n + \tau_m), \ m = 0, 1, \cdots, M-1. \end{aligned} \tag{2.84}$$

A straightforward approach for noise reduction is to average the M signals $z_m(n)$,

$$z_{\mathrm{a}}(n) = \frac{1}{M} \sum_{m=0}^{M-1} z_m(n) = \frac{\beta_{\mathrm{a}}}{M} s(n - t) + \frac{1}{M} \sum_{m=0}^{M-1} v_m(n + \tau_m), \tag{2.85}$$

where $\beta_{\mathrm{a}} = \sum_{m=0}^{M-1} \beta_m$. If the noises are added incoherently, the output SNR will, in principle, increase [44]. We can further reduce the noise by passing the signal $z_{\mathrm{a}}(n)$ through a Wiener filter as was shown in the previous sections. This approach has, however, two drawbacks. The first one is that, since for $m \neq i$, $E\{v_m(n + \tau_m) v_i(n + \tau_i)\} \neq 0$ in general, the output SNR will not improve that much; and the second one, as we know already, is speech distortion that the optimal filter introduces.

Let us now define the error signal, for the mth microphone, between the clean speech sample $x_m(n)$ and its estimate as,

$$\begin{aligned} e_{x_m}(n) &\overset{\triangle}{=} x_m(n) - \mathbf{h}_{:m}^T\mathbf{z}(n) \\ &= x_m(n) - \sum_{i=0}^{M-1} \mathbf{h}_{i:m}^T\mathbf{z}_i(n), \end{aligned} \tag{2.86}$$

where $\mathbf{h}_{i:m}$ are filters of length L and,

$$\begin{aligned} \mathbf{h}_{:m} &\overset{\triangle}{=} \left[\mathbf{h}_{0:m}^T \ \mathbf{h}_{1:m}^T \ \cdots \ \mathbf{h}_{M-1:m}^T \right]^T, \\ \mathbf{z}(n) &\overset{\triangle}{=} \left[\mathbf{z}_0^T(n) \ \mathbf{z}_1^T(n) \ \cdots \ \mathbf{z}_{M-1}^T(n) \right]^T. \end{aligned}$$

Since $\mathbf{z}_i(n) = \beta_i\mathbf{s}(n-t) + \mathbf{v}_i(n+\tau_i)$, (2.86) becomes:

$$\begin{aligned} e_{x_m}(n) &= \mathbf{s}^T(n-t)\left[\beta_m\mathbf{u}_1 - \sum_{i=0}^{M-1}\beta_i\mathbf{h}_{i:m}\right] - \sum_{i=0}^{M-1}\mathbf{v}_i^T(n+\tau_i)\mathbf{h}_{i:m} \\ &= \mathbf{s}^T(n-t)\left[\beta_m\mathbf{u}_1 - \mathbf{D}\mathbf{h}_{:m}\right] - \mathbf{v}^T(n)\mathbf{h}_{:m} \\ &= e_{s,m}(n) - e_{v,m}(n), \end{aligned} \tag{2.87}$$

where

$$\begin{aligned} \mathbf{D} &\overset{\triangle}{=} \left[\beta_0\mathbf{I} \ \beta_1\mathbf{I} \ \cdots \ \beta_{M-1}\mathbf{I} \right], \\ \mathbf{v}(n) &\overset{\triangle}{=} \left[\mathbf{v}_0^T(n+\tau_0) \ \mathbf{v}_1^T(n+\tau_1) \ \cdots \ \mathbf{v}_{M-1}^T(n+\tau_{M-1}) \right]^T. \end{aligned}$$

Expression (2.87) is the difference between two error signals; $e_{s,m}(n)$ represents signal distortion and $e_{v,m}(n)$ represents the residual noise. The MSE corresponding to the residual noise with the mth microphone as the reference signal is,

$$\begin{aligned} J_{v,m}(\mathbf{h}_{:m}) &= E\left\{e_{v,m}^2(n)\right\} \\ &= \mathbf{h}_{:m}^T E\left\{\mathbf{v}(n)\mathbf{v}^T(n)\right\}\mathbf{h}_{:m} \\ &= \mathbf{h}_{:m}^T\mathbf{R}_v\mathbf{h}_{:m}. \end{aligned} \tag{2.88}$$

Usually, in the single-channel case, the minimization of the MSE corresponding to the residual noise is done while keeping the signal distortion below a threshold [20]. With no distortion, the optimal filter obtained from this optimization is $\mathbf{u}_1$, hence there is not any noise reduction either. The advantage of multiple microphones is that, actually, we can minimize $J_{v,m}(\mathbf{h}_{:m})$ with the constraint that $\beta_m\mathbf{u}_1 = \mathbf{D}\mathbf{h}_{:m}$ (no speech distortion at all). Therefore, our optimization problem is,

$$\min_{\mathbf{h}_{:m}} J_{v,m}(\mathbf{h}_{:m}) \text{ subject to } \beta_m\mathbf{u}_1 = \mathbf{D}\mathbf{h}_{:m}. \tag{2.89}$$

By using a Lagrange multiplier, we easily find the optimal solution:

$$\mathbf{h}_{\mathrm{o},:m} = \beta_m \mathbf{R}_v^{-1}\mathbf{D}^T \left[\mathbf{D}\mathbf{R}_v^{-1}\mathbf{D}^T\right]^{-1} \mathbf{u}_1, \tag{2.90}$$

where we assumed that the noise signals $v_i(n)$ are not perfectly coherent so that $\mathbf{R}_v$ is not singular.

The MMSE for the mth microphone is,

$$J_{v,m}(\mathbf{h}_{\mathrm{o},:m}) = \beta_m^2 \mathbf{u}_1^T \left[\mathbf{D}\mathbf{R}_v^{-1}\mathbf{D}^T\right]^{-1} \mathbf{u}_1. \tag{2.91}$$

Since we have M microphones, we have M MMSEs as well. The best MMSE from a noise reduction point of view is the smallest one, which is, according to (2.91), the microphone signal with the smallest attenuation factor.

The attenuation factors β_m can be easily determined, if the power of the noise signals are known, by using the formula:

$$\beta_m^2 = \frac{E\{z_m^2(n)\} - E\{v_m^2(n+\tau_m)\}}{E\{z_0^2(n)\} - E\{v_0^2(n)\}}, \ m = 1, 2, \cdots, M-1. \tag{2.92}$$

For the particular case where the noise is spatio-temporally white with a power equal to σ_v^2, the MMSE and the normalized MMSE for the mth microphone are respectively,

$$J_{v,m}(\mathbf{h}_{\mathrm{o},:m}) = \sigma_v^2 \frac{\beta_m^2}{\sum_{i=0}^{M-1} \beta_i^2}, \tag{2.93}$$

$$\tilde{J}_{v,m}(\mathbf{h}_{\mathrm{o},:m}) = \frac{\beta_m^2}{\sum_{i=0}^{M-1} \beta_i^2}. \tag{2.94}$$

We can see that when the number of microphones goes to infinity, the normalized MMSE goes to zero, which means that the noise can be completely removed with no signal distortion at all.

2.7 Simulation Experiments

By defining a noise-reduction factor to quantify the amount of noise being attenuated and a speech-distortion index to valuate the degree to which the speech signal is deformed, we have analytically examined the performance behavior of the Wiener-filter-based noise reduction technique. It is shown that the Wiener filter achieves noise reduction by distorting the speech signal. The more the noise is reduced, the more the speech is distorted. We also proposed several approaches to better manage the tradeoff between noise reduction and speech distortion. To further verify the analysis, and to assess the noise-reduction-and-speech-distortion management schemes, we implemented a time-domain Wiener-filter system. The sampling rate is 8 kHz. The noise signal is estimated in the time-frequency domain using a sequential algorithm presented in [6], [7]. Briefly, this algorithm obtains an estimate of noise using the overlap-add technique on a frame-by-frame basis. The noisy speech signal $y(n)$ is segmented into frames with a frame width of 8 milliseconds and an overlapping factor of 75%. Each frame is then transformed via a DFT into a block of spectral samples. Successive blocks of spectral samples form a two-dimensional time-frequency matrix denoted by $Y_t(j\omega)$, where subscript t is the frame index, denoting the time dimension, and ω is the angular frequency. Then an estimate of the magnitude of the noise spectrum is formulated as

$$\hat{V}_t(\omega) = \begin{cases} \alpha_\mathrm{a}\hat{V}_{t-1}(\omega) + (1-\alpha_\mathrm{a})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| \geq \hat{V}_{t-1}(\omega) \\ \alpha_\mathrm{d}\hat{V}_{t-1}(\omega) + (1-\alpha_\mathrm{d})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| < \hat{V}_{t-1}(\omega) \end{cases}, \tag{2.95}$$

where α_a and α_d are the the "attack" and "decay" coefficients respectively. Meanwhile, to reduce its temporal fluctuation, the magnitude of the noisy speech spectrum is smoothed according to the following recursion:

$$\bar{Y}_t(\omega) = \begin{cases} \beta_\mathrm{a}\bar{Y}_{t-1}(\omega) + (1-\beta_\mathrm{a})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| \geq \bar{Y}_{t-1}(\omega) \\ \beta_\mathrm{d}\bar{Y}_{t-1}(\omega) + (1-\beta_\mathrm{d})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| < \bar{Y}_{t-1}(\omega) \end{cases}, \tag{2.96}$$

where again β_a is the "attack" coefficient and β_d the "decay" coefficient. To further reduce the spectral fluctuation, both $\hat{V}_t(\omega)$ and $\bar{Y}_t(\omega)$ are averaged across the neighboring frequency bins around ω. Finally, an estimate of the noise spectrum is obtained by multiplying $\hat{V}_t(\omega)/\bar{Y}_t(\omega)$ with $Y_t(j\omega)$, and the time-domain noise signal is obtained through IDFT and the overlap-add technique. See [6], [7] for more detailed description of this noise-estimation scheme. Figure 2.3 shows a speech signal corrupted by a car noise (SNR = 10 dB), the waveform and the spectrogram of the car noise that is added to the speech, and the waveform and spectrogram of the noise estimate. It can be seen that during the absence of speech, the estimate is a good approximation of the noise signal. It is also noticed from its spectrogram that the noise estimate consists of some minor speech components during the presence of speech. Our listening test, however, shows that the residual speech remained in the noise estimate is almost inaudible. An apparent advantage of this noise-estimation technique is that it does not require an explicit voice activity detector. In addition, our experimental investigation reveals that such a scheme is able to capture the noise characteristics in both the presence and absence of speech, therefore it does not rely on the assumption that the noise characteristics in the presence of speech stay the same as in the absence of speech.

Based on the implemented system, we evaluate the Wiener filter for noise reduction. The first experiment investigates the influence of the filter length on the noise reduction performance. Instead of using the estimated noise, here we assume that the noise signal is known *a priori*. Therefore this experiment demonstrates the upper limit of the performance of the Wiener filter. We consider two cases. In the first one, both the source signal and the background

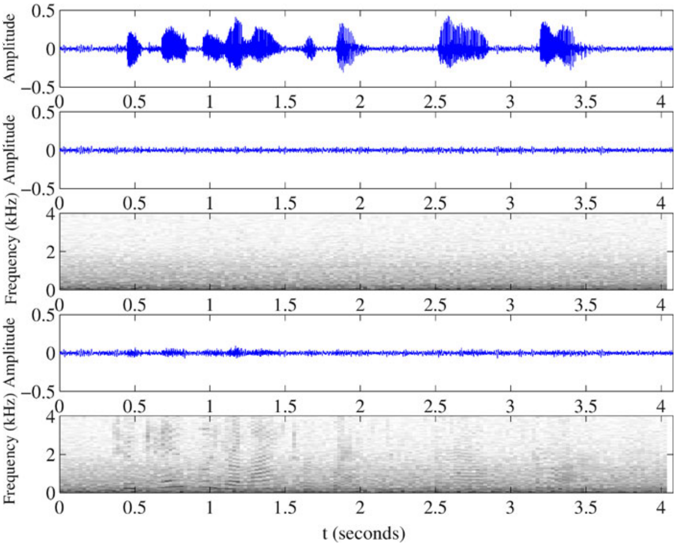
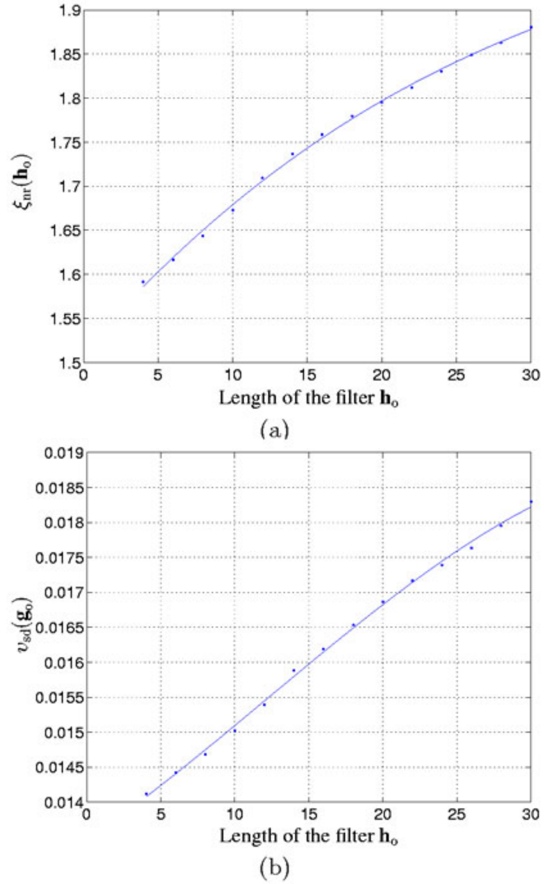


Fig. 2.3. Noise and its estimate. The first trace (from the top) shows the waveform of a speech signal corrupted by a car noise where SNR = 10 dB. The second and third traces plot the waveform and spectrogram of the noise signal. The fourth and fifth traces display the waveform and spectrogram of the noise estimate.

noise are random processes in which the current value of the signal cannot be predicted from its past samples. The source signal is a noise signal recorded from a New York Stock Exchange (NYSE) room. This signal consists of sound from various sources such as speakers, telephone rings, electric fans, etc. The background noise is a computer-generated Gaussian random process. The results for this case is graphically portrayed in Fig. 2.4. It can be seen that both the noise-reduction factor and the speech-distortion index increase linearly with the filter length. Therefore, a longer filter should be applied for more noise reduction. However, the more the noise is attenuated, the more the source signal is deformed, as shown in Fig. 2.4.

In the second case, we test the Wiener filter for noise reduction in the context of speech signal. It is known that a speech signal can be modelled as an AR process, where its current value can be predicted from its past samples. To simplify the situation for the ease of analysis, the source signal used here is an /i:/ sound recorded from a female speaker. Same as in the previous case, the background noise is a computer-generated white Gaussian random process. The results are plotted in Fig. 2.5. Again, the noise-reduction factor, which quantifies the amount of noise being attenuated, increases monotonically with the filter length; but unlike the previous case, the relationship between the noise reduction and the filter length is not linear. Instead, the curve at first grows quickly as the filter length is increased up to 10, and then

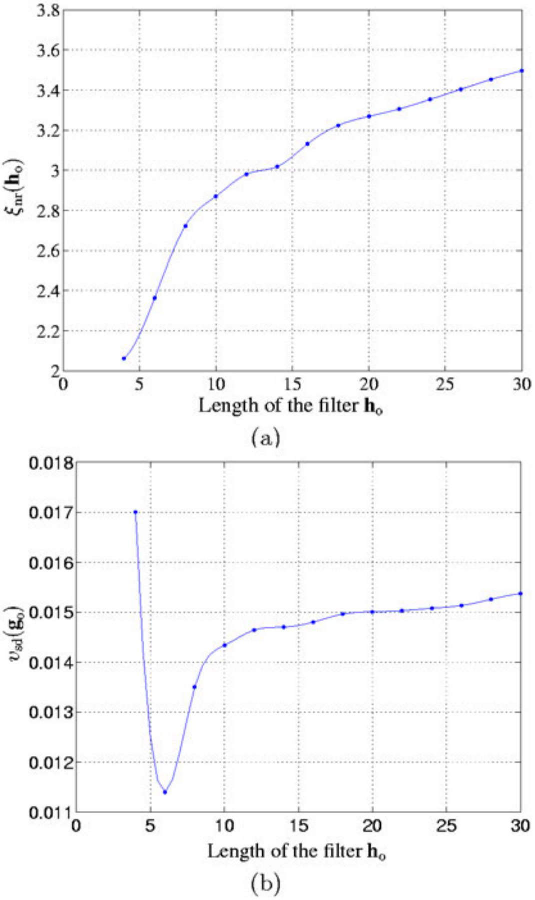


(a)

(b)

Fig. 2.4. Noise-reduction factor and signal-distortion index, both as a function of the filter length: (a) noise reduction; (b) signal distortion. The source is a signal recorded in a NYSE room; the background noise is a computer-generated white Gaussian random process; and SNR = 10 dB.

continues to grow but with a slower rate. Unlike $\xi_{\rm nr}$, the speech-distortion index, i.e., $\upsilon_{\rm sd}$, exhibits a non-monotonic relationship with the filter length. It first decreases to its minimum, and then increases again as the filter length is increased. The reason, as we have explained in Section 2.6.2, is that a speech signal can be modelled as an AR process. Particular to this experiment, the /i:/ sound used here can be well modelled with a 6^{th} order LPC (linear prediction coding) analysis. Therefore, when the filter length is increased to 6, the numerator of (2.33) is minimized, as a result, the speech-distortion index reaches its minimum. Continuing to increase the filter length leads to a higher distortion due to more noise reduction. To further verify this observation, we

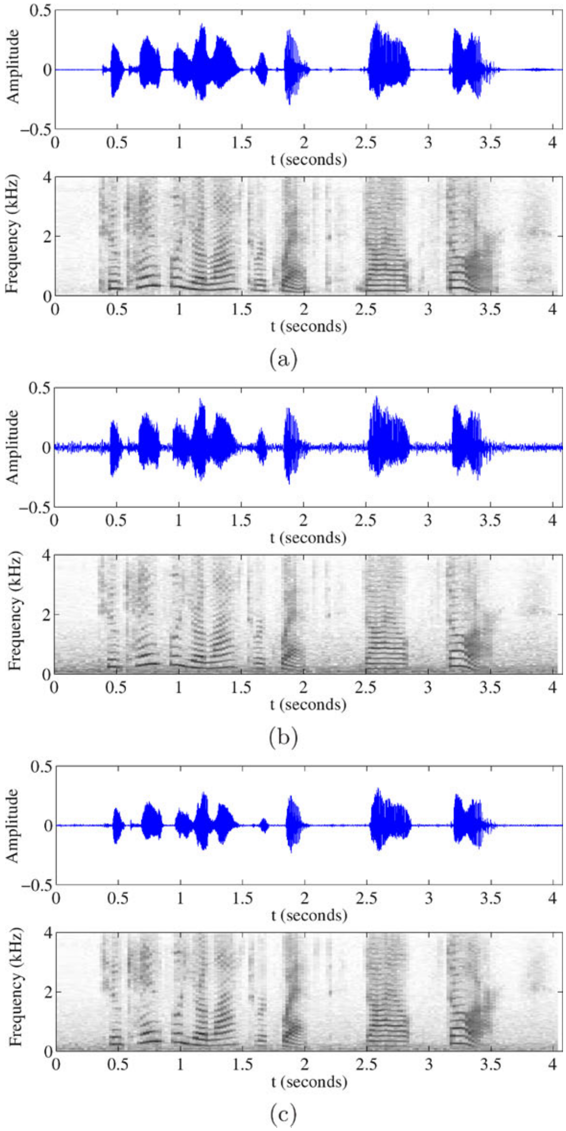


(a)

(b)

Fig. 2.5. Noise-reduction factor and signal-distortion index, both as a function of the filter length: (a) noise reduction; (b) speech distortion. The source signal is an /i:/ sound from a female speaker; the background noise is a computer-generated white Gaussian process; and SNR = 10 dB.

investigated several other vowels, and found that the curve of v_{sd} vs. filter length follows a similar shape, except that the minimum may appear in a slightly different location. Taking into account the sounds other than vowels in speech that may be less predicable, we find that good performance with the Wiener filter (in terms of the compromise between noise reduction and speech distortion) can be achieved when filter length L is chosen around 20. Figure 2.6 plots the output of our Wiener filter system with $L = 20$, where the speech signal is from a female speaker, the background noise is a car noise signal, and SNR = 10 dB.



(a)

(b)

(c)

Fig. 2.6. Noise reduction in a car noise condition where SNR = 10 dB: (a) clean speech and its spectrogram; (b) noisy speech and its spectrogram; (c) noise reduced speech and its spectrogram.

The second experiment tests the noise reduction performance in different SNR conditions. Here the speech signal is recorded from a female speaker as shown in Fig. 2.6. The computer-generated random Gaussian noise is added to the speech signal to control the SNR. The length of the Wiener filter is set to $L = 20$. The results are presented in Fig.2.7, where besides ξ_{nr} and υ_{sd}, we also plotted the Itakura-Saito (IS) distance, a widely used objective quality measure that performs a comparison of spectral envelopes (AR parameters) between the clean and the processed speech [46]. Studies have shown that the IS measure is highly correlated (0.59) with the subjective quality judgements [47]. A recent report reveals that the difference in mean opinion score (MOS) between two processed speech signals would be less than 1.6 if their IS measure is less than 0.5 for various codecs [48]. Many other reported experiments confirmed that two spectra would be perceptually nearly identical if their IS distance is less than 0.1. All these evidences indicates that the IS distance is a reasonably good objective measure of speech quality.

As SNR decreases, the observation signal becomes more noisy. Therefore the Wiener filter is expected to have more noise reduction for low SNRs. This is verified by Fig. 2.7 (a), where significant noise reduction is obtained for low SNR conditions. However, more noise reduction would correspond to more speech distortion. This is confirmed by Fig. 2.7 (b) and (d) where both the speech-distortion index and the IS distance increase as speech becomes more noisy. Comparing the IS distance before [Fig. 2.7 (c)] and after [Fig. 2.7 (d)] noise reduction, one can see that significant gain in the IS distance has been achieved, indicating that the Wiener filter is able to reduce noise and improve speech quality (but not necessarily speech intelligibility).

The last experiment is to verify the performance behavior of the suboptimal filter derived in Section 2.6.1. The experimental conditions are the same as outlined in the previous experiment. The results are presented in Table 2.1, where for the purpose of comparison, besides the speech-distortion index and the noise-reduction factor, we also show three IS distances (between the clean and filtered speeches denoted as ISD^1, between the clean and noise-reduced speeches marked as ISD^2, and between the clean and noisy speeches denoted as ISD^3, respectively). From the results, one can make the following observations:

- The IS distance between the clean and noisy speech signals increases as SNR drops. The reason for this is apparent. When SNR decreases, the speech signal becomes more noisy. As a result, the difference between the spectral envelope (or AR parameters) of the clean speech and that (or those) of the noisy speech tends to be more significant, which leads to a higher IS distance.
- ISD^2 is much smaller than ISD^3. This significant gain in IS distance indicates that the use of noise reduction technique is able to mitigate noise and improve speech quality.
- A better compromise between noise reduction and speech distortion is accomplished by using the suboptimal filter. For example, when SNR = 20 dB, the speech-distortion index for the suboptimal filter with $\alpha = 0.7$ is 0.0006, which is only 54% of that of the Wiener filter; the corresponding

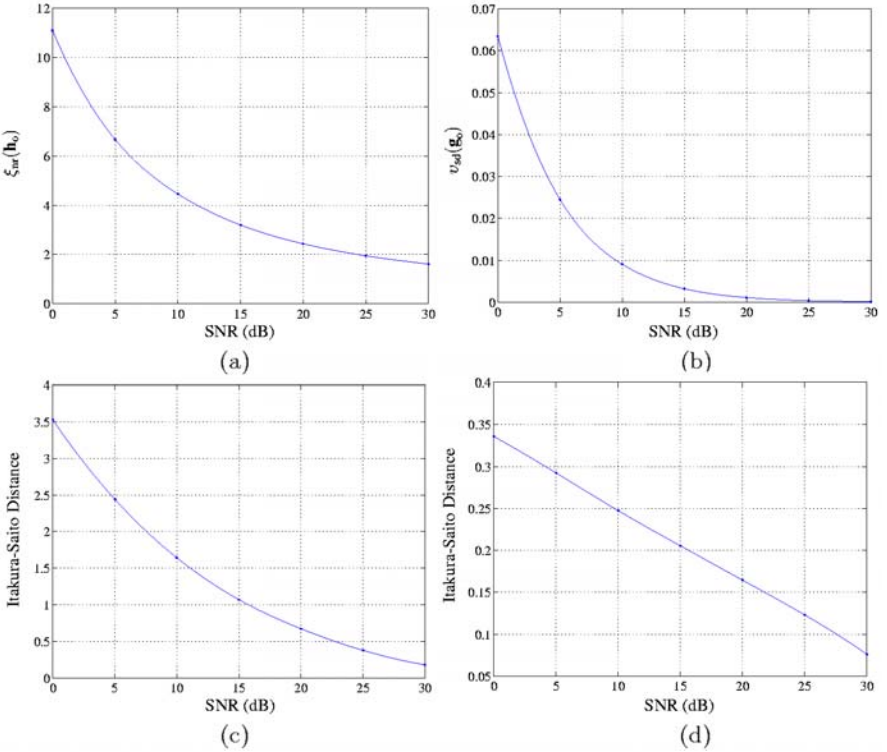


Fig. 2.7. Noise reduction performance as a function of SNR in white Gaussian noise: (a) noise-reduction factor; (b) speech-distortion index; (c) Itakura-Saito distance between the clean and noisy speeches; (d) Itakura-Saito distance between the clean and noise-reduced speeches.

IS distance between the clean and filtered speech is 0.0281, which is only 17% of that of the Wiener filter; but it has achieved a noise reduction of 2.0106, which is 82% of that with the Wiener filter.

- Different from ISD^1, which decreases with α, ISD^2 increases when a smaller α is selected. This is due to the fact that ISD^2 is affected by both speech distortion and the residual noise remained in the noise-reduced speech. As elaborated in Section 2.6.1, as long as α satisfies $0 \leq \alpha \leq 1$, a smaller α would lead to less speech distortion; but a smaller α also means that more residual noise will remain in the noise-reduced speech. While the former may reduce the IS distance, the latter will enlarge the IS distance. As a result, ISD^2 increases when a smaller α is chosen.
- From the analysis shown in Section 2.6.1, we see that both $\frac{\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})}{\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})}$ and $\frac{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})}{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})}$ are independent of SNR but not $\frac{\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})}{\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})}$. From the experimental results, we notice that the ratio between $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})$ and $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ does not

Table 2.1 Noise reduction performance with the suboptimal filter, where ISD[1] is the IS distance between the clean speech [i.e., $x(n)$] and the filtered version of the clean speech [i.e., $\mathbf{h}^T\mathbf{x}(n)$], which purely measures the speech distortion due to the filtering effect; ISD[2] is the IS distance between the clean and noise-reduced speeches; ISD[3] is the IS distance between the clean and noisy speech signals.

SNR		υ_{sd}	ξ_{nr}	ISD[1]	ISD[2]	ISD[3]
20dB	Wiener filter	0.0011	2.4390	0.1691	0.1471	0.6727
	Suboptimal filter ($\alpha = 0.8$)	0.0007	2.1753	0.0423	0.2820	0.6727
	Suboptimal filter ($\alpha = 0.7$)	0.0006	2.0106	0.0281	0.3476	0.6727
15dB	Wiener filter	0.0033	3.1977	0.2133	0.2032	1.0446
	Suboptimal filter ($\alpha = 0.8$)	0.0021	2.7379	0.0488	0.5114	1.0446
	Suboptimal filter ($\alpha = 0.7$)	0.0016	2.4544	0.0352	0.6034	1.0446
10dB	Wiener filter	0.0092	4.4565	0.2622	0.2652	1.5458
	Suboptimal filter ($\alpha = 0.8$)	0.0059	3.5896	0.0582	0.7759	1.5458
	Suboptimal filter ($\alpha = 0.7$)	0.0045	3.0807	0.0441	0.8917	1.5458

vary much when SNR is changed; but $\frac{\xi_{\text{nr}}(\mathbf{h}_{\text{s}})}{\xi_{\text{nr}}(\mathbf{h}_{\text{o}})}$ decreases with SNR. For examples, when $\alpha = 0.7$, the ratio calculated from experiment is 0.82 when SNR = 20 dB, and is 0.77 when SNR = 15 dB. From numerous experiments, we noticed that the speech distortion and noise reduction satisfy $\frac{\xi_{\text{nr}}(\mathbf{h}_{\text{s}})}{\upsilon_{\text{sd}}(\mathbf{g}_{\text{s}})} > \frac{\xi_{\text{nr}}(\mathbf{h}_{\text{o}})}{\upsilon_{\text{sd}}(\mathbf{g}_{\text{o}})}$ if SNR > 5 dB, which indicates that the suboptimal filter can be used to control the tradeoff between noise reduction and speech distortion as long as SNR > 5 dB. The higher is the SNR, the more effective will the suboptimal filter work.

2.8 Conclusions

The problem of speech enhancement has attracted a considerable amount of research attention over the past several decades. Numerous techniques were developed, among them is the optimal Wiener filter, which is the most fundamental approach. It is widely noticed that the Wiener filter achieves noise reduction by deforming the speech signal. However, so far not much has been said on how the Wiener filter really works. This chapter was devoted to analyzing the intrinsic relationship between noise reduction and speech distortion with the Wiener filter. Starting from the speech and noise estimation using the Wiener theory, we introduced a speech-distortion index and a noise-reduction factor. We showed that for the single-channel Wiener filter, the amount of noise attenuation is in general proportionate to the amount of speech degradation, i.e., more noise reduction incurs more speech distortion.

Depending on the nature of the application, some practical noise-reduction systems may require very high-quality speech, but can tolerate a certain amount of noise. While others may want speech as clean as possible even with some degree of speech distortion. Therefore it is necessary that we can have some management schemes to control the contradicting requirements between noise reduction and speech distortion. To do so, we have discussed three approaches. When there is no *a priori* knowledge or no additional information available, a sub-optimal filter with one more free parameter can be used. By setting the free parameter to 0.7, we showed that the sub-optimal filter can achieve 90% of the noise reduction that the Wiener filter can have; but the resulting speech distortion is less than half of that of the Wiener filter. Speech signal can be modeled as an autoregressive (AR) process. If the AR coefficients can be estimated reliably, we showed that these coefficients can be used to construct the Wiener filter for less speech distortion. In scenarios where we can have multiple noisy realizations of the speech signal, then spatio-temporal filtering techniques can be exploited to obtain noise reduction with less or even no speech distortion.

References

1. M. R. Schroeder, U.S. Patent No 3,180,936, filed Dec. 1, 1960, issued Apr. 27, 1965.
2. M. R. Schroeder, U.S. Patent No 3,403,224, filed May 28, 1965, issued Sept. 24, 1968.
3. J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.
4. J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
5. Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, pp. 1526–1554, Oct. 1992.
6. E. J. Diethorn, “Subband noise reduction methods for speech enhancement,” in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., pp. 91–115, Boston, MA: Kluwer, 2004.
7. J. Chen, Y. Huang, and J. Benesty, “Filtering techniques for noise reduction and speech enhancement,” in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds., pp. 129–154, Berlin, Germany: Springer, 2003.
8. B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1985.

9. S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.
10. R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 137–145, Apr. 1980.
11. P. Vary, “Noise suppression by spectral magnitude estimation–mechanism and theoretical limits,” *Signal Processing*, vol. 8, pp. 387–400, July 1985.
12. R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.
13. W. Etter and G. S. Moschytz, “Noise reduction by noise-adaptive spectral magnitude expansion,” *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.
14. D. L. Wang and J. S. Lim, “The unimportance of phase in speech enhancement,” *IEEE Trans. Acoustic. Speech, Signal Processing*, vol. 30, pp. 679–681, Aug. 1982.
15. Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.
16. Y. Ephraim and D. Malah,“Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 443–445, Apr. 1985.
17. N. Virag, “Single channel speech enhancement basd on masking properties of human auditory system,” *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.
18. Y. M. Chang and D. O’Shaughnessy, “Speech enhancement based conceptually on auditory evidence,” *IEEE Trans. Signal Processing*, vol. 39, pp. 1943–1954, Sept. 1991.
19. T. F. Quatieri and R. B. Dunn, “Speech enhancement based on auditory spectral change,” in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 257–260.
20. Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.
21. M. Dendrinos, S. Bakamidis, and G. Garayannis, “Speech enhancement from noise: a regenerative approach,” *Speech Commun.*, vol. 10, pp. 45–57, Feb. 1991.
22. P. S. K. Hansen, *Signal Subspace Methods for Speech Enhancement.* Ph.D. dissertation, Techn. Univ. Denmark, Lyngby, Denmark, 1997.
23. H. Lev-Ari and Y. Ephraim, “Extension of the signal subspace speech enhancement approach to colored noise,” *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 104–106, Apr. 2003.
24. A. Rezayee and S. Gazor, “An adpative KLT approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.
25. U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 159–167, Mar. 2000.
26. Y. Hu and P. C. Loizou, “A generalized subspace approach for enhancing spech corrupted by colored noise,” *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 334–341, July 2003.
27. K. K. Paliwal and A. Basu, “A speech enhancement method based on Kalman filtering,” in *Proc. IEEE ICASSP*, 1987, pp. 177–180.

28. J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. Signal Processing*, vol. 39, pp. 1732–1742, Aug. 1991.
29. M. Gabrea, E. Grivel, and M. Najim, “A single microphone Kalman filter-based noise canceller,” *IEEE Trans. Signal Processing Lett.*, vol. 6, pp. 55–57, Mar. 1999.
30. B. Lee, K. Y. Lee, and S. Ann, “An EM-based approach for parameter enhancement with an application to speech signals,” *Signal Processing*, vol. 46, pp. 1–14, Sept. 1995.
31. S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential Kalman filter-based speech enhancement algorithms,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 373–385, July 1998.
32. Y. Ephraim, D. Malah, and B.-H. Juang, “On the application of hidden Markov models for enhancing noisy speech,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.
33. Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Processing*, vol. 40, pp. 725–735, Apr. 1992.
34. J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, “Particle methods for Bayesian modeling and enhancement of speech signals,” *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 173–185, Mar. 2002.
35. H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 445–455, Sept. 1998.
36. D. Burshtein and S. Gannot, “Speech enhancement using a mixture-maximum model,” *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 341–351, Sept. 2002.
37. J. Vermaak and M. Niranjan, “Markov chain Monte Carlo methods for speech enhancement,” in *Proc. IEEE ICASSP*, vol. 2, 1998, pp. 1013–1016.
38. S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Processing*, vol. 49, pp. 1614–1626, Aug. 2001.
39. S. E. Nordholm, I. Claesson, and N. Grbic, “Performance limits in subband beamforming,” *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 193-203, May 2003.
40. F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 497–507, Sept. 2000.
41. F. Jabloun and B. Champagne, “A multi-microphone signal subspace approach for speech enhancement,” in *Proc. IEEE ICASSP*, pp. 205–208, 2001.
42. S. Doclo and M. Moonen, “GSVD-based optimal filtering for single and multimicrophone speech enhancement,” *IEEE Trans. Signal Processing*, vol. 50, pp. 2230–2244, Sept. 2002.
43. S. Haykin, *Adaptive Filter Theory.* Fourth Edition, Upper Saddle River, NJ: Prentice Hall, 2002.
44. P. M. Clarkson, *Optimal and Adaptive Signal Processing.* Boca Raton, FL: CRC, 1993.
45. K. Fukunaga, *Introduction to Statistial Pattern Recognition.* San Diego, CA: Academic, 1990.
46. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice Hall, 1993.

47. S. Quakenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewook Cliffs, NJ: Prentice Hall, 1988.
48. G. Chen, S. N. Koh, and I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Processing*, vol. 83, pp. 1445–1456, July 2003.