Chapter 3
Single-Channel Noise Reduction in the Time Domain

One of the most important schemes in the fundamental topic of speech enhancement is single-channel noise reduction in the time domain since most communication devices have only one microphone and the time-domain processing seems intuitive and natural. This approach has been very well studied in the literature (see [1] for example). In this chapter, we revisit this method from the perspective proposed in Chapter 2.

3.1 Signal Model

The noise reduction problem considered in this chapter is one of recovering the desired signal (or clean speech) $x(t)$, t being the discrete-time index, of zero mean from the noisy observation (microphone signal) [1], [2]:

$$y(t) = x(t) + v(t), \tag{3.1}$$

where the zero-mean random process $v(t)$ is the unwanted additive noise, which is assumed to be uncorrelated with $x(t)$. In this context, all signals are real.

The signal model given in (3.1) can be put into a vector form by considering the L most recent successive time samples, i.e.,

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{v}(t), \tag{3.2}$$

where

$$\mathbf{y}(t) = \big[\, y(t)\ y(t-1)\ \cdots\ y(t-L+1) \,\big]^T \tag{3.3}$$

is a vector of length L, superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}(t)$ and $\mathbf{v}(t)$ are defined in a similar way to $\mathbf{y}(t)$ from (3.3). Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the correlation matrix (of size $L \times L$) of the noisy signal can be written as

© The Author(s) 2015

J. Benesty and J. Chen, *A Conceptual Framework for Noise Reduction*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-12955-6_3

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}} &= E\left[\mathbf{y}(t)\mathbf{y}^T(t)\right] \\ &= \mathbf{\Phi}_{\mathbf{x}} + \mathbf{\Phi}_{\mathbf{v}},\end{aligned} \tag{3.4}$$

where

$$\mathbf{\Phi}_{\mathbf{x}} = E\left[\mathbf{x}(t)\mathbf{x}^T(t)\right], \tag{3.5}$$

$$\mathbf{\Phi}_{\mathbf{v}} = E\left[\mathbf{v}(t)\mathbf{v}^T(t)\right], \tag{3.6}$$

are the correlation matrices of $\mathbf{x}(t)$ and $\mathbf{v}(t)$, respectively. The objective of noise reduction in the time domain and with a single microphone is then to find a "good" estimate of the sample $x(t)$ given the vector $\mathbf{y}(t)$, in the sense that the additive noise is significantly reduced while the desired signal is not much distorted. This is what will be studied in this chapter.

Since $x(t)$ is the signal of interest, it is important to write the vector $\mathbf{y}(t)$ as an explicit function of $x(t)$. For that, we need first to decompose $\mathbf{x}(t)$ into two orthogonal components: one proportional to the desired signal, $x(t)$, and the other one corresponding to the interference. Indeed, it is easy to see that this decomposition is

$$\mathbf{x}(t) = x(t)\boldsymbol{\rho}_{\mathbf{x}x} + \mathbf{x}_{\mathrm{i}}(t), \tag{3.7}$$

where

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}x} &= \left[1\ \rho_x(1)\ \cdots\ \rho_x(L-1)\right]^T \\ &= \frac{E\left[\mathbf{x}(t)x(t)\right]}{E\left[x^2(t)\right]}\end{aligned} \tag{3.8}$$

is the normalized [with respect to $x(t)$] correlation vector (of length L) between $\mathbf{x}(t)$ and $x(t)$,

$$\rho_x(l) = \frac{E\left[x(t-l)x(t)\right]}{E\left[x^2(t)\right]},\ l = 0, 1, \ldots, L-1 \tag{3.9}$$

is the correlation coefficient between $x(t-l)$ and $x(t)$,

$$\mathbf{x}_{\mathrm{i}}(t) = \mathbf{x}(t) - x(t)\boldsymbol{\rho}_{\mathbf{x}x} \tag{3.10}$$

is the interference signal vector, and

$$E\left[\mathbf{x}_{\mathrm{i}}(t)x(t)\right] = \mathbf{0}_{L\times 1}, \tag{3.11}$$

where $\mathbf{0}_{L\times 1}$ is a vector of length L containing only zeroes.

Substituting (3.7) into (3.2), the signal model for noise reduction in the time domain can be expressed as

$$\mathbf{y}(t) = x(t)\boldsymbol{\rho}_{\mathbf{x}x} + \mathbf{x}_{\mathrm{i}}(t) + \mathbf{v}(t). \tag{3.12}$$

This formulation will be extensively used in the following sections.

3.2 Linear Filtering

In this chapter, we try to estimate the desired signal sample, $x(t)$, by applying a finite-impulse-response (FIR) filter to the observation signal vector, $\mathbf{y}(t)$, i.e.,

$$\begin{aligned}\widehat{x}(t) &= \sum_{l=0}^{L-1} h_l y(t-l) \\ &= \mathbf{h}^T \mathbf{y}(t),\end{aligned} \tag{3.13}$$

where $\widehat{x}(t)$ is the estimate of $x(t)$ and

$$\mathbf{h} = \begin{bmatrix} h_0 \; h_1 \; \cdots \; h_{L-1} \end{bmatrix}^T \tag{3.14}$$

is a real-valued filter of length L. This procedure is called single-channel noise reduction in the time domain with a linear filter.

Using (3.12), we can express (3.13) as

$$\begin{aligned}\widehat{x}(t) &= \mathbf{h}^T \left[x(t)\boldsymbol{\rho}_{\mathbf{x}x} + \mathbf{x}_{\mathrm{i}}(t) + \mathbf{v}(t) \right] \\ &= x_{\mathrm{fd}}(t) + x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t),\end{aligned} \tag{3.15}$$

where

$$x_{\mathrm{fd}}(t) = x(t)\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x} \tag{3.16}$$

is the filtered desired signal,

$$x_{\mathrm{ri}}(t) = \mathbf{h}^T \mathbf{x}_{\mathrm{i}}(t) \tag{3.17}$$

is the residual interference, and

$$v_{\mathrm{rn}}(t) = \mathbf{h}^T \mathbf{v}(t) \tag{3.18}$$

is the residual noise.

Since the estimate of the desired signal at time t is the sum of three terms that are mutually uncorrelated, the variance of $\widehat{x}(t)$ is

$$\begin{aligned}\phi_{\widehat{x}} &= \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{y}} \mathbf{h} \\ &= \phi_{x_{\mathrm{fd}}} + \phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}},\end{aligned} \tag{3.19}$$

where

$$\phi_{x_{\mathrm{fd}}} = \phi_x \left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2 \tag{3.20}$$
$$= \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{d}}} \mathbf{h},$$
$$\phi_{x_{\mathrm{ri}}} = \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}} \mathbf{h} \tag{3.21}$$
$$= \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{x}} \mathbf{h} - \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{d}}} \mathbf{h},$$
$$\phi_{v_{\mathrm{rn}}} = \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{v}} \mathbf{h}, \tag{3.22}$$

$\phi_x = E\left[x^2(t)\right]$ is the variance of the desired signal, $\boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{d}}} = \phi_x \boldsymbol{\rho}_{\mathbf{x}x} \boldsymbol{\rho}_{\mathbf{x}x}^T$ is the correlation matrix (whose rank is equal to 1) of $\mathbf{x}_{\mathrm{d}}(t) = x(t)\boldsymbol{\rho}_{\mathbf{x}x}$, and $\boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}} = E\left[\mathbf{x}_{\mathrm{i}}(t)\mathbf{x}_{\mathrm{i}}^T(t)\right]$ is the correlation matrix of $\mathbf{x}_{\mathrm{i}}(t)$. The variance of $\widehat{x}(t)$ is useful in the definitions of the performance measures.

3.3 Performance Measures

In this section, we extend the performance measures given in Chapter 2 for the conceptual framework to the single-channel noise reduction problem in the time domain.

The input SNR, derived from (3.1), is defined as

$$\mathrm{iSNR} = \frac{\phi_x}{\phi_v}, \tag{3.23}$$

where $\phi_v = E\left[v^2(t)\right]$ is the variance of the additive noise.

The output SNR[1] helps quantify the level of noise remaining at the filter output signal. The output SNR is obtained from (3.19):

$$\mathrm{oSNR}\left(\mathbf{h}\right) = \frac{\phi_{x_{\mathrm{fd}}}}{\phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}}} \tag{3.24}$$
$$= \frac{\phi_x \left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2}{\mathbf{h}^T \boldsymbol{\Phi}_{\mathrm{in}} \mathbf{h}},$$

where

$$\boldsymbol{\Phi}_{\mathrm{in}} = \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}} + \boldsymbol{\Phi}_{\mathbf{v}} \tag{3.25}$$

is the interference-plus-noise correlation matrix. Basically, (3.24) is the variance of the first signal (filtered desired) from the right-hand side of (3.19) over the variance of the two other signals (filtered interference-plus-noise). The objective of the noise reduction filter is to make the output SNR greater than the input SNR. Consequently, the quality of the noisy signal may be enhanced.

[1] In this work, we consider the uncorrelated interference as part of the noise in the definitions of the performance measures.

For the particular filter:

$$\mathbf{h} = \mathbf{i}_{\text{id}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \tag{3.26}$$

of length L, which corresponds to the first column of the identity matrix $\mathbf{I}_L$ of size $L \times L$, we have

$$\text{oSNR}\left(\mathbf{i}_{\text{id}}\right) = \text{iSNR}. \tag{3.27}$$

With the identity filter, $\mathbf{i}_{\text{id}}$, the SNR cannot be improved.

For any two vectors $\mathbf{h}$ and $\boldsymbol{\rho}_{\mathbf{x}x}$ and a positive definite matrix $\boldsymbol{\Phi}_{\text{in}}$, we have

$$\left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2 \leq \left(\mathbf{h}^T \boldsymbol{\Phi}_{\text{in}} \mathbf{h}\right) \left(\boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}\right), \tag{3.28}$$

with equality if and only if $\mathbf{h} = \varsigma \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}$, where $\varsigma (\neq 0)$ is an arbitrary real number. Using the inequality (3.28) in (3.24), we deduce an upper bound for the output SNR:

$$\text{oSNR}\left(\mathbf{h}\right) \leq \phi_x \boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}, \ \forall \mathbf{h} \tag{3.29}$$

and, clearly,

$$\text{oSNR}\left(\mathbf{i}_{\text{id}}\right) \leq \phi_x \boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}, \tag{3.30}$$

which implies that

$$\phi_v \boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \geq 1. \tag{3.31}$$

The maximum output SNR is then

$$\text{oSNR}_{\text{max}} = \phi_x \boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \tag{3.32}$$

and

$$\text{oSNR}_{\text{max}} \geq \text{iSNR}. \tag{3.33}$$

We also observe that this maximum output SNR is achieved with the maximum SNR filter:

$$\mathbf{h}_{\text{max}} = \varsigma \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}. \tag{3.34}$$

We define the maximum gain in SNR as

$$\begin{aligned} \mathcal{G}_{\text{max}} &= \frac{\text{oSNR}_{\text{max}}}{\text{iSNR}} \\ &= \phi_v \boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \geq 1. \end{aligned} \tag{3.35}$$

We define the partial speech intelligibility index as

$$\begin{aligned} \upsilon_{\mathrm{i}}(\mathbf{h}) &= \frac{\phi_x - \phi_{x_{\mathrm{fd}}}}{\phi_x} \\ &= 1 - \left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2. \end{aligned} \tag{3.36}$$

The larger is $\upsilon_{\mathrm{i}}(\mathbf{h})$, the less intelligible is the estimated desired signal, $\widehat{x}(t)$.

The speech quality index is defined as the ratio of the residual noise over the variance of the additive noise, i.e.,

$$\begin{aligned} \upsilon_{\mathrm{q}}(\mathbf{h}) &= \frac{\phi_{v_{\mathrm{rn}}}}{\phi_v} \\ &= \frac{\mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{v}} \mathbf{h}}{\phi_v}. \end{aligned} \tag{3.37}$$

For a fixed value of the input SNR, the quality of the signal improves as $\upsilon_{\mathrm{q}}(\mathbf{h})$ decreases.

From the two previous expressions, we deduce the global speech intelligibility index:

$$\upsilon'_{\mathrm{i}}(\mathbf{h}) = (1 - \varpi)\,\upsilon_{\mathrm{i}}(\mathbf{h}) + \varpi \upsilon_{\mathrm{q}}(\mathbf{h}). \tag{3.38}$$

The variance of the estimated desired signal can be rewritten as a function of the two indices $\upsilon_{\mathrm{i}}(\mathbf{h})$ and $\upsilon_{\mathrm{q}}(\mathbf{h})$, i.e.,

$$\phi_{\widehat{x}} = [1 - \upsilon_{\mathrm{i}}(\mathbf{h})]\,\phi_x + \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}} \mathbf{h} + \upsilon_{\mathrm{q}}(\mathbf{h})\,\phi_v. \tag{3.39}$$

3.4 MSE-Based Criterion

For any MSE-type criterion, an error signal is needed. We define the error signal between the estimated and desired signals as

$$\begin{aligned} e(t) &= \widehat{x}(t) - x(t) \\ &= x_{\mathrm{fd}}(t) + x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t) - x(t), \end{aligned} \tag{3.40}$$

which can be written as the sum of two other uncorrelated error signals:

$$e(t) = e_{\mathrm{i}}(t) + e_{\mathrm{q}}(t), \tag{3.41}$$

where

$$E\left[e_{\mathrm{i}}(t) e_{\mathrm{q}}(t)\right] = 0, \tag{3.42}$$

$$\begin{aligned} e_{\mathrm{i}}(t) &= x_{\mathrm{fd}}(t) - x(t) \\ &= \left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x} - 1\right) x(t) \end{aligned} \tag{3.43}$$

is the speech distortion due to the FIR filter, which affects the partial intelligibility, and

$$\begin{aligned} e_{\mathrm{q}}(t) &= x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t) \\ &= \mathbf{h}^T \mathbf{x}_{\mathrm{i}}(t) + \mathbf{h}^T \mathbf{v}(t) \end{aligned} \tag{3.44}$$

represents the residual interference-plus-noise, which affects the quality as well as the other part of intelligibility.

The classical MSE criterion is then

$$\begin{aligned} J(\mathbf{h}) &= E\left[e^2(t)\right] \\ &= \phi_x + \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{y}} \mathbf{h} - 2\mathbf{h}^T E\left[\mathbf{x}(t)x(t)\right] \\ &= \phi_x + \mathbf{h}^T \boldsymbol{\Phi}_{\mathbf{y}} \mathbf{h} - 2\phi_x \mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x} \\ &= J_{\mathrm{i}}(\mathbf{h}) + J_{\mathrm{q}}(\mathbf{h}), \end{aligned} \tag{3.45}$$

where

$$\begin{aligned} J_{\mathrm{i}}(\mathbf{h}) &= E\left[e_{\mathrm{i}}^2(t)\right] \\ &= \phi_x \left(\mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x} - 1\right)^2 \end{aligned} \tag{3.46}$$

and

$$\begin{aligned} J_{\mathrm{q}}(\mathbf{h}) &= E\left[e_{\mathrm{q}}^2(t)\right] \\ &= \mathbf{h}^T \boldsymbol{\Phi}_{\mathrm{in}} \mathbf{h}. \end{aligned} \tag{3.47}$$

The two particular filters $\mathbf{h} = \mathbf{i}_{\mathrm{id}}$ and $\mathbf{h} = \mathbf{0}_{L\times 1}$ described in the previous section are of interest to us. With the first one (identity filter), we achieve the worst quality and the best partial intelligibility, while with the second one (zero filter), we have the best quality and the worst intelligibility. For these two particular filters, the MSEs are

$$J(\mathbf{i}_{\mathrm{id}}) = J_{\mathrm{q}}(\mathbf{i}_{\mathrm{id}}) = \phi_v, \tag{3.48}$$

$$J(\mathbf{0}_{L\times 1}) = J_{\mathrm{i}}(\mathbf{0}_{L\times 1}) = \phi_x. \tag{3.49}$$

As a result,

$$\mathrm{iSNR} = \frac{J(\mathbf{0}_{L\times 1})}{J(\mathbf{i}_{\mathrm{id}})}. \tag{3.50}$$

We define the NMSE with respect to $J(\mathbf{i}_{\mathrm{id}})$ as

$$J_{\mathrm{n},1}(\mathbf{h}) = \frac{J(\mathbf{h})}{J(\mathbf{i}_{\mathrm{id}})} \tag{3.51}$$
$$= \mathrm{iSNR} \times \left(1 - \mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2 + \frac{\mathbf{h}^T \boldsymbol{\Phi}_{\mathrm{in}} \mathbf{h}}{\phi_v}.$$

We define the NMSE with respect to $J(\mathbf{0}_{L\times 1})$ as

$$J_{\mathrm{n},2}(\mathbf{h}) = \frac{J(\mathbf{h})}{J(\mathbf{0}_{L\times 1})} \tag{3.52}$$
$$= \left(1 - \mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2 + \frac{\mathbf{h}^T \boldsymbol{\Phi}_{\mathrm{in}} \mathbf{h}}{\phi_x}$$

and, obviously,

$$J_{\mathrm{n},1}(\mathbf{h}) = \mathrm{iSNR} \times J_{\mathrm{n},2}(\mathbf{h}). \tag{3.53}$$

Expressions (3.51) and (3.52) show how the NMSEs and the different MSEs are implicitly related to the performance measures.

We are only interested in filters for which

$$J_{\mathrm{i}}(\mathbf{i}_{\mathrm{id}}) \leq J_{\mathrm{i}}(\mathbf{h}) < J_{\mathrm{i}}(\mathbf{0}_{L\times 1}), \tag{3.54}$$
$$J_{\mathrm{q}}(\mathbf{0}_{L\times 1}) < J_{\mathrm{q}}(\mathbf{h}) < J_{\mathrm{q}}(\mathbf{i}_{\mathrm{id}}). \tag{3.55}$$

From the two previous expressions, we deduce that

$$0 \leq \left(1 - \mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2 < 1, \tag{3.56}$$
$$0 < \frac{\mathbf{h}^T \boldsymbol{\Phi}_{\mathrm{in}} \mathbf{h}}{\phi_v} < 1. \tag{3.57}$$

For this reason, we propose to use the more general MSE-based criterion:

$$J_{\mu}(\mathbf{h}) = \mu \frac{J_{\mathrm{i}}(\mathbf{h})}{\phi_x} + \frac{J_{\mathrm{q}}(\mathbf{h})}{\phi_v} \tag{3.58}$$
$$= \mu \left(1 - \mathbf{h}^T \boldsymbol{\rho}_{\mathbf{x}x}\right)^2 + \frac{\mathbf{h}^T \boldsymbol{\Phi}_{\mathrm{in}} \mathbf{h}}{\phi_v},$$

where μ is a positive real number allowing to compromise between $\upsilon_{\mathrm{i}}(\mathbf{h})$ and $\upsilon_{\mathrm{q}}(\mathbf{h})$.

3.5 Optimal Filters

Taking the gradient of (3.58) with respect to $\mathbf{h}$ and equating the result to zero, we get the optimal filter:

$$\mathbf{h}_{\mathrm{o},\mu} = \mu \left(\mu \boldsymbol{\rho}_{\mathbf{x}x} \boldsymbol{\rho}_{\mathbf{x}x}^T + \frac{\boldsymbol{\Phi}_{\mathrm{in}}}{\phi_v} \right)^{-1} \boldsymbol{\rho}_{\mathbf{x}x}. \tag{3.59}$$

Using the decomposition:

$$\boldsymbol{\Phi}_{\mathbf{y}} = \phi_x \boldsymbol{\rho}_{\mathbf{x}x} \boldsymbol{\rho}_{\mathbf{x}x}^T + \boldsymbol{\Phi}_{\mathrm{in}}, \tag{3.60}$$

we can rewrite the optimal filter as

$$\mathbf{h}_{\mathrm{o},\mu} = \mu \left[(\mu - \mathrm{iSNR}) \, \boldsymbol{\rho}_{\mathbf{x}x} \boldsymbol{\rho}_{\mathbf{x}x}^T + \frac{\boldsymbol{\Phi}_{\mathbf{y}}}{\phi_v} \right]^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \tag{3.61}$$

and the vector $\boldsymbol{\rho}_{\mathbf{x}x}$ can be expressed as a function of the statistics of $y(t)$ and $v(t)$, i.e.,

$$\begin{aligned} \boldsymbol{\rho}_{\mathbf{x}x} &= \frac{E\left[\mathbf{y}(t) y(t)\right] - E\left[\mathbf{v}(t) v(t)\right]}{\phi_y - \phi_v} \\ &= \frac{\phi_y \boldsymbol{\rho}_{\mathbf{y}y} - \phi_v \boldsymbol{\rho}_{\mathbf{v}v}}{\phi_y - \phi_v}, \end{aligned} \tag{3.62}$$

so that $\mathbf{h}_{\mathrm{o},\mu}$ can be estimated from the statistics of $y(t)$ and $v(t)$ only.

Using the Woodbury's identity in (3.59), it can easily be shown that the optimal filter can be reformulated as

$$\begin{aligned} \mathbf{h}_{\mathrm{o},\mu} &= \frac{\mu \dfrac{\phi_x}{\mathrm{iSNR}}}{1 + \mu \dfrac{\mathrm{oSNR}_{\max}}{\mathrm{iSNR}}} \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \\ &= \frac{\mu \phi_v}{1 + \mu \mathcal{G}_{\max}} \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}. \end{aligned} \tag{3.63}$$

Comparing $\mathbf{h}_{\mathrm{o},\mu}$ with $\mathbf{h}_{\max}$ [eq. (3.34)], we see that the two filters are equivalent up to a scaling factor. As a result, $\mathbf{h}_{\mathrm{o},\mu}$ also maximizes the output SNR, i.e.,

$$\mathrm{oSNR}\left(\mathbf{h}_{\mathrm{o},\mu}\right) = \mathrm{oSNR}_{\max}, \ \forall \mu > 0. \tag{3.64}$$

From (3.63), we deduce the partial speech intelligibility index:

$$v_{\mathrm{i}}\left(\mathbf{h}_{\mathrm{o},\mu}\right) = 1 - \left(\frac{\mu \mathcal{G}_{\max}}{1 + \mu \mathcal{G}_{\max}} \right)^2 \tag{3.65}$$

and the speech quality index:

$$v_{\mathrm{q}}\left(\mathbf{h}_{\mathrm{o},\mu}\right) = \frac{\mu^2 \phi_v \boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\Phi}_{\mathbf{v}} \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}}{\left(1 + \mu \mathcal{G}_{\max}\right)^2}. \tag{3.66}$$

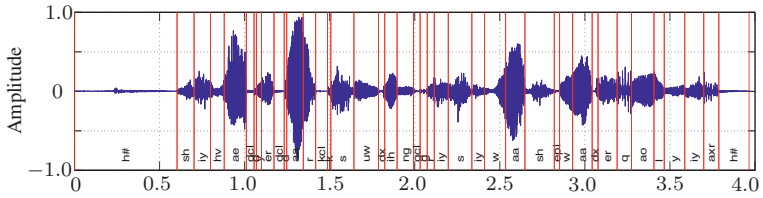

Fig. 3.1 A speech signal from the speaker FAKS0 of the TIMIT database.

Taking $\mu = \text{iSNR}$ in (3.63), we find the well-known Wiener filter [1]:

$$\begin{aligned} \mathbf{h}_{\text{W}} &= \frac{\phi_x}{1+\text{oSNR}_{\max}} \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \\ &= \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\Phi}_{\mathbf{x}} \mathbf{i}_{\text{id}} \\ &= \left(\mathbf{I}_L - \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\Phi}_{\mathbf{v}}\right) \mathbf{i}_{\text{id}} \end{aligned} \tag{3.67}$$

and taking $\mu = \infty$ in (3.63), we find the MVDR filter [1]:

$$\begin{aligned} \mathbf{h}_{\text{MVDR}} &= \frac{\phi_x}{\text{oSNR}_{\max}} \boldsymbol{\Phi}_{\text{in}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x} \\ &= \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}}{\boldsymbol{\rho}_{\mathbf{x}x}^T \boldsymbol{\Phi}_{\mathbf{y}}^{-1} \boldsymbol{\rho}_{\mathbf{x}x}} \\ &= \frac{1+\text{oSNR}_{\max}}{\text{oSNR}_{\max}} \mathbf{h}_{\text{W}}. \end{aligned} \tag{3.68}$$

A value of μ in (3.63) greater (resp. smaller) than the input SNR will result in a filter that will favor partial intelligibility (resp. quality) over quality (resp. partial intelligibility) as compared to the Wiener filter.

3.6 Simulations

In this section, we illustrate the performance of the optimal filters derived above through simulations. The clean speech used is from the TIMIT database [3], [4]. This database was originally designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition (ASR) systems; but it has now been used in various applications including noise reduction [5]. The database consists of a total of 6300 sentences spoken by 630 speakers with 10 sentences by each speaker. All speech signals were recorded with a 16-kHz sampling rate and a 16-bit quantization. Each signal is accompanied by manually segmented phonetic (based on 61 phonemes) transcripts as illustrated in Fig. 3.1. In the simulations of this chapter, we take all the ten sentences from the speaker

FAKS0 and downsample the signals from 16 kHz to 8 kHz. We then use these downsampled signals as the clean speech. The corresponding noisy signals are obtained by adding noise to the clean speech, where the noise signal is properly scaled to control the input SNR level. We consider two types of noise: white Gaussian and a babble signal recorded in a New York Stock Exchange (NYSE) room. In comparison with the Gaussian random noise, which is stationary and white, the NYSE noise is nonstationary and colored. This babble noise consists of sounds from various sources such as electrical fans, telephone rings, and background speech.

The implementation of the noise reduction filters derived in Section 3.5 requires the estimation of the correlation matrices $\mathbf{\Phi_y}$ and $\mathbf{\Phi_v}$, and the correlation vector $\boldsymbol{\rho}_{\mathbf{x}x}$. Here, we directly compute the $\mathbf{\Phi_y}$ matrix from $y(t)$ using a short-time average, i.e., at every time instant t, an estimate of $\mathbf{\Phi_y}$ is computed as

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(t) = \frac{1}{P}\sum_{p=0}^{P-1} \mathbf{y}(t-p)\mathbf{y}^T(t-p), \tag{3.69}$$

where P is the total number of samples used in the short-time average. In our simulations, we choose $P = 320$, i.e., using the most recent 40 ms samples. In a similar way, we compute the $\mathbf{\Phi_v}$ matrix and the $\boldsymbol{\rho}_{\mathbf{x}x}$ vector at time instant t. Substituting the estimated correlation matrices and vector into (3.67) and (3.68), we obtain the Wiener and MVDR filters, respectively.

We use the partial speech intelligibility index, υ_{i}, the speech quality index, υ_{q}, and the output SNR as the performance measures to evaluate the implemented Wiener and MVDR filters. Figure 3.2 plots the performance of the Wiener filter as a function of the filter length, L, in the white Gaussian noise. As it can be seen, the partial speech intelligibility index decreases monotonically with L. So, the larger the filter length, the more intelligible is the enhanced speech with the Wiener filter. In comparison, the quality index first decreases and then increases with L, which means that the quality of the enhanced signal with the Wiener filter is not a monotonic function of L. The quality first increases and then decreases as the filter length increases. The output SNR is seen to increase with L for the studied range of filter length; but it first increases quickly and then starts to saturate when L is large. In real applications, the choice of the value of L has to take into consideration both the noise reduction performance and complexity. If this value is too small, the performance improvement may not be significant for the listener to appreciate, while if it is too large, the complexity can be very high and, meanwhile, the estimation of the correlation matrices and vector may become less reliable, resulting degradation in noise reduction performance.

The performance of the Wiener filter as a function of the filter length, L, in the NYSE noise is plotted in Fig. 3.3. Comparing Figs. 3.2 and 3.3, one can see that there is some difference between the performance of the Wiener filter

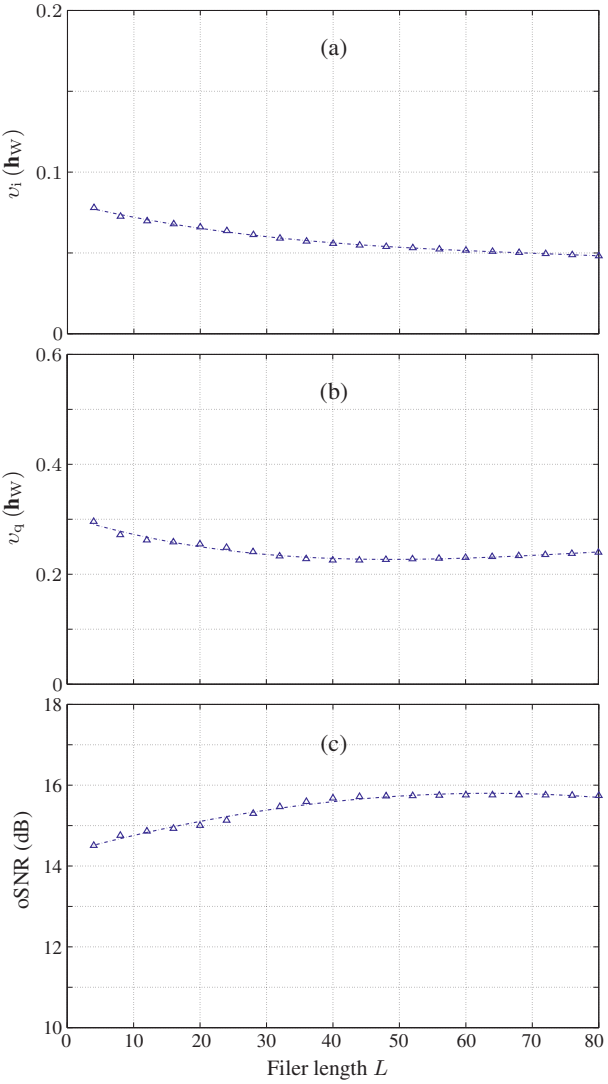

Fig. 3.2 Performance of the Wiener filter as a function of the filter length, L, in the white Gaussian noise: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The input SNR is 10 dB.

in the NYSE noise and that in the white Gaussian noise; but the performance trend as a function of the filter length in the two noise conditions is similar.

Now, let us fix the filter length, L, to 40 and investigate the performance behavior of the Wiener and MVDR filters in different SNR conditions. Figure 3.4 plots the results in the white Gaussian noise. It is seen that the partial

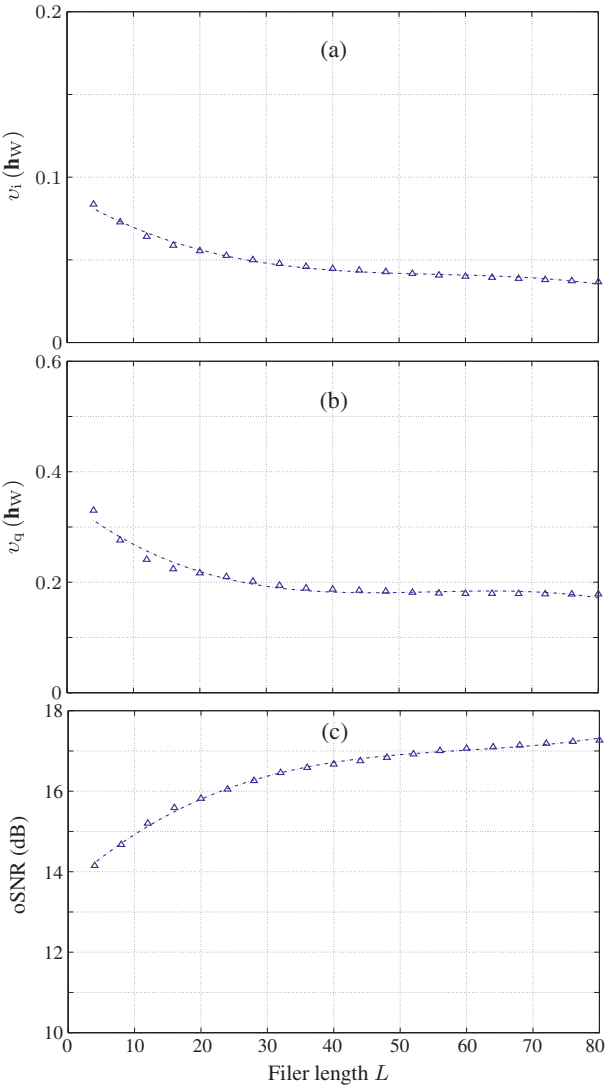

Fig. 3.3 Performance of the Wiener filter as a function of the filter length, L, in the NYSE noise: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The input SNR is 10 dB.

speech intelligibility index, $v_{\rm i}$, of the MVDR filter is always 0 regardless of the SNR level. In comparison, this index is not zero for the Wiener filter and it decreases as the input SNR increases. The SNR improvement (i.e., the difference between the input and output SNRs) decreases as the input SNR increases. It can be seen that the speech quality index for both the Wiener

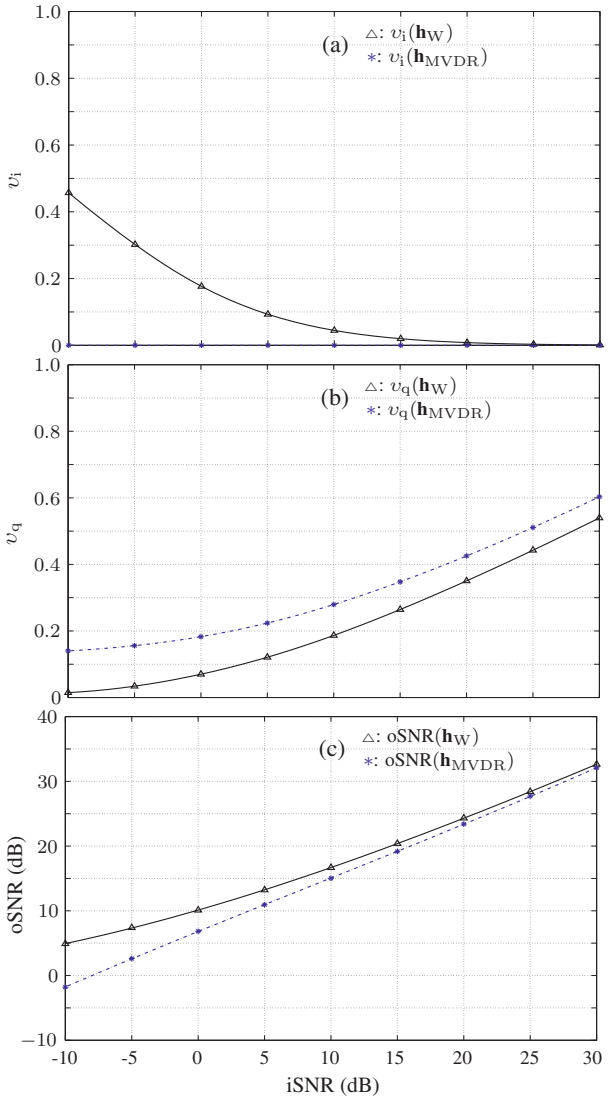

Fig. 3.4 Performance of the Wiener and MVDR filters in the white Gaussian noise at different input SNRs: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The filter length $L = 40$.

and MVDR filters increases with the input SNR. It should be pointed out that the speech quality index, from its definition, measures the amount of noise reduction. The value of this index depends on many factors including the nature of the noise, the SNR condition, the noise reduction filter that is used, etc. In a given noise and SNR condition, this index measures partially

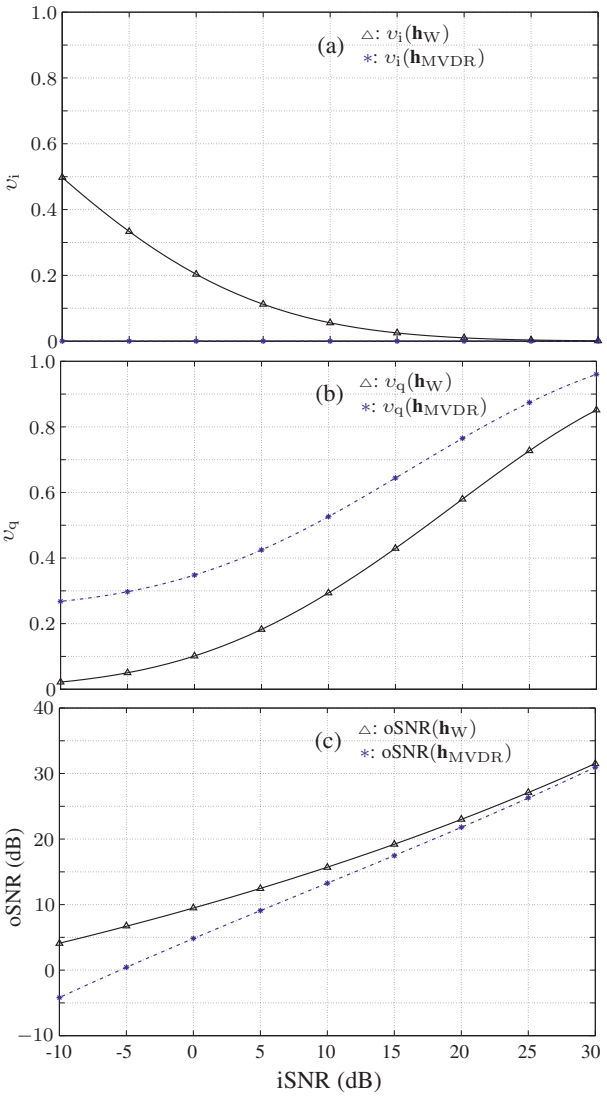

Fig. 3.5 Performance of the Wiener and MVDR filters in the NYSE noise at different input SNRs: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The filter length $L = 40$.

the speech quality after noise reduction: the smaller is this index, the better is the speech quality. In a particular noise environment and for a particular noise reduction filter, we see that the value of this index increases with the input SNR. In this case, this index measures the quality improvement. So, the smaller is this index, the larger is the quality improvement. To summarize,

if the input SNR is high, the speech quality index gets closer to 1, since the improvement can be very small in this case. Consequently, the speech quality index makes sense only when combined with the input SNR.

Figure 3.5 plots the performance of the Wiener and MVDR filters in the NYSE noise. Comparing Figs. 3.5 and 3.4, one can see that the performance trend of the two filters in the NYSE noise is similar to that in the white Gaussian noise though the partial speech intelligibility index, the speech quality index, and the output SNR of each filter differ slightly in values in the two different noise cases with the same input SNR.

Note that one can also make a compromise in performance between the Wiener and the MVDR filters by adjusting the parameter μ in the tradeoff filter in (3.61). Simulations of this filter are left to the reader's investigation.

References

1. J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters–A Theoretical Study*. Berlin, Germany: SpringerBriefs in Electrical and Computer Engineering, 2011.
2. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
3. "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT)," from the NIST TIMIT Speech Disc.
4. K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1641–1648, Nov. 1989.
5. 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.