SPRINGER BRIEFS IN
ELECTRICAL AND COMPUTER ENGINEERING

Jacob Benesty
Jingdong Chen

A Conceptual Framework for Noise Reduction

Springer

SpringerBriefs in Electrical and Computer Engineering

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50–125 pages, the series covers a range of content from professional to academic. Typical topics might include: timely report of state-of-the art analytical techniques, a bridge between new research results, as published in journal articles, and a contextual literature review, a snapshot of a hot or emerging topic, an in-depth case study or clinical example and a presentation of core concepts that students must understand in order to make independent contributions.

More information about this series at http://www.springer.com/series/10059

Jacob Benesty · Jingdong Chen

A Conceptual Framework for Noise Reduction

Springer

Jacob Benesty
INRS-EMT, University of Quebec
Montreal, QC
Canada

Jingdong Chen
Northwestern Polytechnical University
Xi'an, Shaanxi
China

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Electrical and Computer Engineering
ISBN 978-3-319-12954-9 ISBN 978-3-319-12955-6 (eBook)
DOI 10.1007/978-3-319-12955-6

Library of Congress Control Number: 2015936620

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Abstract

The noise reduction (or speech enhancement) problem has been studied for at least five decades but its understanding and the development of reliable solutions are more than ever very welcome. Therefore, by having a solid grasp of this problem, it will certainly become easier to design a well-targeted solution for a well-defined application. In this work, we propose a conceptual framework that can be applied to the many different aspects of noise reduction. As a consequence, the monaural or binaural noise reduction problem, in the time domain or in the frequency domain, with a single microphone or with multiple microphones, is presented in a unified way. Moreover, the derivation of optimal linear filters is simplified as well as the performance measures for their evaluation.

Contents

Chapter 1
Introduction

In this chapter, we very briefly introduce the problem of noise reduction. For more details, the reader is invited to consult the rich literature on this topic.

1.1 Noise Reduction

The problem of noise reduction (or speech enhancement) is an important part of speech processing [1] and all the ideas developed in this topic can be easily applied to the general problem of signal enhancement. It is well known that any speech communication system suffers from the ubiquitous presence of additive noise [2], [3]. Typical examples of such products are cellular phones and hearing aids. In these systems, the noise degrades the perceptual quality of the speech and will impair the speech intelligibility when the signal-to-noise ratio (SNR) comes down to a certain level. Therefore, the objective of noise reduction is to suppress such additive noise for purposes of speech enhancement.

The first noise reduction algorithm was proposed by Schroeder more than 50 years ago [4], [5]; it is basically the spectral magnitude subtraction method. Since then a great deal of progress has been made. Not only we can better exploit the temporal/spectral information of the signals but also the spatial information by using multiple microphones. Thanks to the spatial information, we can better compromise between noise reduction and speech distortion, which is a fundamental limitation of single-channel noise reduction algorithms. Even though today we have a better understanding of this problem and some interesting solutions available, it is far from being solved.

Monaural or binaural noise reduction can be performed in the time domain or in the frequency domain, with one single microphone or with multiple microphones. Very often, in the literature, most approaches seem to be very different and their evaluation does not seem to be consistent. In this work, we present a framework that has the potential to simplify the study of the

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_1

general speech enhancement problem. Because of their great flexibility, we limit this investigation to linear filters. Within the proposed framework, the most important performance measures are derived as well as general forms of the optimal filters.

1.2 Organization of the Work

This book consists of six chapters including this introduction. In Chapter 2, we present a conceptual framework for studying the general problem of noise reduction. Furthermore, we introduce two important performance measures: the speech intelligibility index and the speech quality index. In Chapters 3, 4, 5, and 6, we show how this concept is applied to the single-channel noise reduction in time domain, to the single-channel noise reduction in the short-time Fourier transform (STFT) domain, to the binaural noise reduction in the time domain, and to the multichannel noise reduction in the STFT domain, respectively.

References

1. J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2007.
2. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
3. P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.
4. M. R. Schroeder, U.S. Patent No. 3,180,936, filed Dec. 1, 1960, issued Apr. 27, 1965.
5. M. R. Schroeder, U.S. Patent No. 3,403,224, filed May 28, 1965, issued Sept. 24, 1968.

Chapter 2
Conceptual Framework

In this chapter, a conceptual framework for noise reduction is proposed. This new formulation gives a better insight into this fundamental problem. Within this framework, we define all important performance measures and criteria that will be of great help in the derivation of the most well-known estimators. Some key discussions concern also the definitions of speech intelligibility and speech quality that will be used in the rest of this work.

2.1 Signal Model

We consider the conventional signal model [1], [2], [3]:

$$y = x + v, \tag{2.1}$$

where y is the noisy observation, x is the desired (speech) signal, and v is the unwanted additive noise. These signals can be in the time, frequency, or any other domain. Therefore, in this chapter, we are interested in the general case of complex random variables (CRVs). Furthermore, we assume that x and v are uncorrelated, stationary, and zero mean. In this context, the variance of y is

$$\begin{aligned} \phi_y &= E\left(|y|^2\right) \\ &= \phi_x + \phi_v, \end{aligned} \tag{2.2}$$

where $E(\cdot)$ denotes mathematical expectation, and

$$\phi_x = E\left(|x|^2\right), \tag{2.3}$$

$$\phi_v = E\left(|v|^2\right), \tag{2.4}$$

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_2

are the variances of x and v, respectively.

2.2 Principle of the Conceptual Framework

The objective of noise reduction/speech enhancement in any domain is to find a "good" estimate[1], $\widehat{x}$, of the desired signal, x, given y and y^*, where the superscript * denotes complex conjugation, with an appropriate function $f(\cdot)$, i.e.,

$$\widehat{x} = f\left(y, y^*\right). \tag{2.5}$$

In order to be able to define consistent performance measures for any function $f\left(y, y^*\right)$, we need to decompose this latter into two orthogonal components; one component that is proportional to the desired signal, x, and will, therefore, correspond to a linear function of x, and the other component that is uncorrelated with the desired signal and will, therefore, correspond to the residual interference-plus-noise. As a result, we can express (2.5) as

$$\begin{aligned} \widehat{x} &= x_{\mathrm{ld}} + x_{\mathrm{ri}} + v_{\mathrm{rn}} \\ &= x_{\mathrm{ld}} + u \\ &= \rho^* x + u, \end{aligned} \tag{2.6}$$

where

$$x_{\mathrm{ld}} = \rho^* x \tag{2.7}$$

is a linear version of the desired signal,

$$\begin{aligned} \rho &= \frac{E\left(x\widehat{x}^*\right)}{\phi_x} \\ &= \frac{\phi_{x\widehat{x}}}{\phi_x} \end{aligned} \tag{2.8}$$

is the normalized (with respect to x) correlation between x and $\widehat{x}$,

$$\phi_{x\widehat{x}} = E\left(x\widehat{x}^*\right) \tag{2.9}$$

is the correlation between between x and $\widehat{x}$,

$$\begin{aligned} u &= x_{\mathrm{ri}} + v_{\mathrm{rn}} \\ &= \widehat{x} - \rho^* x \end{aligned} \tag{2.10}$$

[1] By "good" estimate, we mean that the additive noise is significantly reduced while the desired signal is lowly (or not) distorted.

is the residual interference-plus-noise, x_{ri} is a speech component (called here interference) that is uncorrelated with x_{ld} (and x), v_{rn} is the residual noise, and

$$\phi_{x_{\mathrm{ri}}v_{\mathrm{rn}}} = E\left(x_{\mathrm{ri}}v_{\mathrm{rn}}^*\right) = 0, \tag{2.11}$$

$$\phi_{xu} = E\left(xu^*\right) = 0. \tag{2.12}$$

Since the three components on the right-hand side of (2.6) are uncorrelated, the variance of $\widehat{x}$ is

$$\begin{aligned}\phi_{\widehat{x}} &= \phi_{x_{\mathrm{ld}}} + \phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}} \\ &= |\rho|^2 \phi_x + \phi_u,\end{aligned} \tag{2.13}$$

where

$$\phi_{x_{\mathrm{ld}}} = |\rho|^2 \phi_x, \tag{2.14}$$

$$\phi_{x_{\mathrm{ri}}} = E\left(|x_{\mathrm{ri}}|^2\right), \tag{2.15}$$

$$\phi_{v_{\mathrm{rn}}} = E\left(|v_{\mathrm{rn}}|^2\right), \tag{2.16}$$

$$\begin{aligned}\phi_u &= E\left(|u|^2\right) \\ &= \phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}},\end{aligned} \tag{2.17}$$

are the variances of x_{ld}, x_{ri}, v_{rn}, and u, respectively.

In the rest, it is assumed that $f\left(y, y^*\right)$ does not amplify the estimated desired signal, i.e.,

$$\phi_{x_{\mathrm{ld}}} \leq \phi_x, \tag{2.18}$$

which is equivalent to saying that

$$|\rho|^2 \leq 1. \tag{2.19}$$

We see from (2.6) that we should try to derive $f\left(y, y^*\right)$ in such a way that $\rho^* = 1$ and $u = 0$ (and, hence, $\widehat{x} = x$); but this is, in general, almost impossible in practice. In most situations, the best we can do is to approach $\widehat{x}$ to x by paying a price. We conclude that when $\phi_u \to 0$ then $|\rho|^2 \to 0$; indeed, we have

$$\phi_u = E\left(|\widehat{x}|^2\right) - E\left(x\widehat{x}^*\right) \tag{2.20}$$

and since $x \neq \widehat{x}$, this implies that $|\rho|^2 \to 0$ when $\phi_u \to 0$. In other words, complete removal of the noise may lead to the cancellation of the desired signal (full distortion). This explains the classical compromise between noise reduction and speech distortion.

We also observe from (2.6) that two different distortions affect the estimated desired signal. The first distortion is due to the scaling factor[2], ρ^* (and, possibly, to the residual interference, x_{ri}), and the second one is due to the additive residual noise, v_{rn}. We will refer to these two distortions as distortion 1 and distortion 2, respectively. It is reasonable to say that distortion 1 affects the intelligibility of the estimated signal since when ρ^* is small, not much energy of $\widehat{x}$ is left in $\phi_{\widehat{x}}$, and when ρ^* is close to 1, almost the whole desired signal is in $\widehat{x}$. Distortion 2 affects both the quality and intelligibility of the estimated signal since the smaller is the variance of v_{rn}, the more pleasant it is to hear to $\widehat{x}$ and the better is its intelligibility. To summarize, distortion 1 is related to speech intelligibility while distortion 2 is related to both speech quality and intelligibility.

2.3 Performance Measures

In this section, we derive the most useful performance measures for noise reduction with the conceptual framework, where any function $f(y, y^*)$ can be used.

The signal-to-noise ratio (SNR) is the most important performance measure in the problem of speech enhancement since it gives a precise information on the level of the noise before and after processing. We have the input SNR (before processing) and the output SNR (after processing).

The input SNR is derived from (2.1). It is defined as

$$\begin{aligned} \mathrm{iSNR} &= \frac{\phi_x}{\phi_v} \\ &= \frac{|\gamma_{xy}|^2}{1-|\gamma_{xy}|^2}, \end{aligned} \tag{2.21}$$

where

$$\begin{aligned} |\gamma_{xy}|^2 &= \frac{|\phi_{xy}|^2}{\phi_x \phi_y} \\ &= \frac{|E(xy^*)|^2}{\phi_x \phi_y} \end{aligned} \tag{2.22}$$

is the magnitude squared correlation coefficient (MSCC) between x and y. It is clear that $0 \leq |\gamma_{xy}|^2 \leq 1$.

[2] The scaling factor distorts the desired signal. In the frequency domain, the value of the scaling factor is different from one bin to the other; as a consequence, when the estimated desired signal is reconstructed into the time domain, it will be up to a filter and the desired signal may be badly affected. The processing in the time domain has a similar effect because of the nonstationarity of the speech signal.

To quantify the level of the interference-plus-noise remaining after the noise reduction processing via the function $f(y, y^*)$, we define the output SNR as the ratio of the variance of the linear version of the desired signal over the variance of the residual interference-plus-noise [see eq. (2.6)], i.e.,

$$\begin{aligned} \text{oSNR} &= \frac{\phi_{x_{\text{ld}}}}{\phi_u} \\ &= \frac{|\rho|^2 \phi_x}{\phi_u}. \end{aligned} \tag{2.23}$$

Clearly, the function $f(y, y^*)$ must be found in such a way that oSNR $\geq$ iSNR, which will be assumed in the rest of this section. In this scenario, we should have

$$\frac{\phi_u}{\phi_v} \leq |\rho|^2 \leq 1, \tag{2.24}$$

which implies that the variance of the residual interference-plus-noise is smaller than the variance of the additive noise.

The output SNR can be rewritten as

$$\text{oSNR} = \frac{|\gamma_{x\widehat{x}}|^2}{1 - |\gamma_{x\widehat{x}}|^2}, \tag{2.25}$$

where $|\gamma_{x\widehat{x}}|^2$ is the MSCC between x and $\widehat{x}$. When $\widehat{x} = y$, the input and output SNRs are equal. The output SNR is always upper bounded.

The gain in SNR is defined as

$$\mathcal{G} = \frac{\text{oSNR}}{\text{iSNR}}. \tag{2.26}$$

Using (2.21) and (2.23), (2.26) becomes

$$\begin{aligned} \mathcal{G} &= \frac{|\rho|^2 \phi_v}{\phi_u} \\ &= \frac{|\gamma_{x\widehat{x}}|^2}{|\gamma_{xy}|^2} \times \frac{1 - |\gamma_{xy}|^2}{1 - |\gamma_{x\widehat{x}}|^2}. \end{aligned} \tag{2.27}$$

The function $f(y, y^*)$ must be derived in such a way that $|\gamma_{x\widehat{x}}|^2 \geq |\gamma_{xy}|^2$, i.e., $\widehat{x}$ is more correlated with x than y is correlated with x. The gain depends on the variances of the additive noise and residual interference-plus-noise, and the normalized correlation between x and $\widehat{x}$.

Let us now open a short parenthesis on a widely used definition in the literature of the SNR after processing, often called SNR improvement[3]. It is defined as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{imp}} &= \frac{\phi_x}{E\left(|x-\widehat{x}|^2\right)} \\ &= \frac{\phi_x}{|1-\rho^*|^2\phi_x+\phi_u}.\end{aligned} \tag{2.28}$$

The SNR improvement is related to the output SNR as follows:

$$\mathrm{SNR}_{\mathrm{imp}} = \frac{\mathrm{oSNR}}{|1-\rho^*|^2\,\mathrm{oSNR}+|\rho|^2}. \tag{2.29}$$

In some situations, $\mathrm{SNR}_{\mathrm{imp}}$ can be close to oSNR. However, in general, these measures can be very much different. Moreover, only oSNR is the true definition of the output SNR and should be the one to be compared to the input SNR.

To evaluate how $f(y, y^*)$ affects intelligibility, we define the partial speech intelligibility index (from distortion 1) as the (normalized) difference between the variance of the original speech signal and the variance of the processed one, i.e.,

$$\begin{aligned}\upsilon_{\mathrm{i}} &= \frac{\phi_x-\phi_{x_{\mathrm{ld}}}}{\phi_x} \\ &= 1-|\rho|^2.\end{aligned} \tag{2.30}$$

The larger is υ_{i}, the less intelligible is the estimated desired signal, $\widehat{x}$.

The speech quality index (from distortion 2) is obtained by comparing the variance of the additive noise from the observation signal to the variance of the additive residual noise after processing with $f(y, y^*)$. We have[4]

$$\upsilon_{\mathrm{q}} = \frac{\phi_{v_{\mathrm{rn}}}}{\phi_v}. \tag{2.31}$$

For a fixed value of the input SNR, the quality of the signal degrades as υ_{q} increases.

It can be checked that

---

[3] In our previous work, we defined the inverse of the SNR improvement, i.e., $\upsilon_{\mathrm{sd}} = \phi_x^{-1}E\left(|x-\widehat{x}|^2\right)$, as the speech distortion index. This is, indeed, a good measure of distortion.

[4] In our previous work, we defined the inverse of the speech quality index, i.e., $\xi_{\mathrm{nr}} = \phi_v/\phi_{v_{\mathrm{rn}}}$, as the noise reduction factor. That definition also makes sense as it compared the original level of noise to the residual noise.

$$\frac{\phi_x - \phi_{\widehat{x}}}{\phi_x} = \upsilon_{\mathrm{i}} - \frac{\upsilon_{\mathrm{q}}}{\mathrm{iSNR}} - \frac{\phi_{x_{\mathrm{ri}}}}{\phi_x} \tag{2.32}$$

or

$$\phi_{\widehat{x}} = (1 - \upsilon_{\mathrm{i}})\, \phi_x + \phi_{x_{\mathrm{ri}}} + \upsilon_{\mathrm{q}} \phi_v. \tag{2.33}$$

Since υ_{q} also affects intelligibility, we can define the global speech intelligibility index (from distortions 1 and 2) as

$$\upsilon'_{\mathrm{i}} = (1 - \varpi)\, \upsilon_{\mathrm{i}} + \varpi \upsilon_{\mathrm{q}}, \tag{2.34}$$

where ϖ $(0 < \varpi < 1)$ is a weighting factor that allows to emphasize more on one of the two distortions if desired.

Ideally, we would like to have a large gain in SNR with υ_{i} and υ_{q} as small as possible. However, υ_{i} and υ_{q} are related by the function $f(y, y^*)$ and depending on how this latter is optimized, we will have to compromise between distortion 1 and distortion 2. When υ_{q} is small (i.e., good quality of the estimated desired signal), we observe that $1 - \upsilon_{\mathrm{i}}$ should also get small; as a result, the partial intelligibility decreases. In other words, quality can always be improved but at the expense, at some point, of intelligibility degradation.

2.4 Mean-Squared-Error (MSE) Based Criterion

The mean-squared-error (MSE) is very convenient to use as a criterion in many practical problems when the underlying parameters of the function $f(y, y^*)$ need to be optimized.

We define the error signal between the estimated and desired signals as

$$\begin{aligned} e &= \widehat{x} - x \\ &= x_{\mathrm{ld}} + u - x, \end{aligned} \tag{2.35}$$

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

$$e = e_{\mathrm{i}} + e_{\mathrm{q}}, \tag{2.36}$$

where

$$e_{\mathrm{i}} = (\rho^* - 1)\, x \tag{2.37}$$

is the speech distortion, which affects the partial intelligibility, and

$$e_{\mathrm{q}} = u \tag{2.38}$$

is the residual interference-plus-noise, which affects the quality (and the other part of intelligibility). It is easy to verify that

$$E\left(e_{\mathrm{i}}e_{\mathrm{q}}^{*}\right)=0. \tag{2.39}$$

The classical MSE criterion is then

$$\begin{aligned} J\left[f\left(y,y^{*}\right)\right] &= E\left(\left|e\right|^{2}\right) \\ &= \phi_{x}-\phi_{x\widehat{x}}-\phi_{x\widehat{x}}^{*}+\phi_{\widehat{x}} \\ &= \left|1-\rho^{*}\right|^{2}\phi_{x}+\phi_{u} \\ &= J_{\mathrm{i}}\left[f\left(y,y^{*}\right)\right]+J_{\mathrm{q}}\left[f\left(y,y^{*}\right)\right], \end{aligned} \tag{2.40}$$

where

$$\begin{aligned} J_{\mathrm{i}}\left[f\left(y,y^{*}\right)\right] &= E\left(\left|e_{\mathrm{i}}\right|^{2}\right) \\ &= \left|1-\rho^{*}\right|^{2}\phi_{x} \end{aligned} \tag{2.41}$$

and

$$\begin{aligned} J_{\mathrm{q}}\left[f\left(y,y^{*}\right)\right] &= E\left(\left|e_{\mathrm{q}}\right|^{2}\right) \\ &= \phi_{u}. \end{aligned} \tag{2.42}$$

Two particular functions are of great interest: $f_1\left(y,y^{*}\right)=y$ and $f_0\left(y,y^{*}\right)=0$. With the first one, the partial intelligibility of the noisy signal is not affected but there is no improvement of quality either. With the second one, the estimated signal is totally unintelligible (since the desired signal is completely cancelled) but the quality is maximum (since no residual noise is left). For both functions, however, it can be verified that the output SNR is equal to the input SNR. For these two particular functions, the MSEs are

$$J\left[f_1\left(y,y^{*}\right)\right]=J_{\mathrm{q}}\left[f_1\left(y,y^{*}\right)\right]=\phi_{v}, \tag{2.43}$$

$$J\left[f_0\left(y,y^{*}\right)\right]=J_{\mathrm{i}}\left[f_0\left(y,y^{*}\right)\right]=\phi_{x}. \tag{2.44}$$

As a result,

$$\mathrm{iSNR}=\frac{J\left[f_0\left(y,y^{*}\right)\right]}{J\left[f_1\left(y,y^{*}\right)\right]}. \tag{2.45}$$

We define the normalized MSE (NMSE) with respect to $J\left[f_1\left(y,y^{*}\right)\right]$ as

$$\begin{aligned} J_{\mathrm{n},1}\left[f\left(y,y^{*}\right)\right] &= \frac{J\left[f\left(y,y^{*}\right)\right]}{J\left[f_1\left(y,y^{*}\right)\right]} \\ &= \mathrm{iSNR}\times\left|1-\rho^{*}\right|^{2}+\frac{\phi_{u}}{\phi_{v}}. \end{aligned} \tag{2.46}$$

We define the NMSE with respect to $J\left[f_0\left(y,y^*\right)\right]$ as

$$J_{\mathrm{n},2}\left[f\left(y,y^*\right)\right]=\frac{J\left[f\left(y,y^*\right)\right]}{J\left[f_0\left(y,y^*\right)\right]}=\left|1-\rho^*\right|^2+\frac{\phi_u}{\phi_x} \tag{2.47}$$

and, obviously,

$$J_{\mathrm{n},1}\left[f\left(y,y^*\right)\right]=\mathrm{iSNR}\times J_{\mathrm{n},2}\left[f\left(y,y^*\right)\right]. \tag{2.48}$$

We are only interested in functions for which

$$J_{\mathrm{i}}\left[f_1\left(y,y^*\right)\right]\leq J_{\mathrm{i}}\left[f\left(y,y^*\right)\right]<J_{\mathrm{i}}\left[f_0\left(y,y^*\right)\right], \tag{2.49}$$

$$J_{\mathrm{q}}\left[f_0\left(y,y^*\right)\right]<J_{\mathrm{q}}\left[f\left(y,y^*\right)\right]<J_{\mathrm{q}}\left[f_1\left(y,y^*\right)\right]. \tag{2.50}$$

From the two previous expressions, we deduce that

$$0\leq\left|1-\rho^*\right|^2<1, \tag{2.51}$$

$$0<\frac{\phi_u}{\phi_v}<1. \tag{2.52}$$

By minimizing the MSE criterion, $J\left[f\left(y,y^*\right)\right]$, we obtain the classical Wiener estimate [4], [5], [6]. Let us denote by $\widehat{x}_{\mathrm{W}}$ this optimal estimate. Using the orthogonality principle, i.e., $E\left[\widehat{x}_{\mathrm{W}}^*\left(x-\widehat{x}_{\mathrm{W}}\right)\right]=0$, we find that

$$\phi_{x\widehat{x}_{\mathrm{W}}}=\phi_{\widehat{x}_{\mathrm{W}}}. \tag{2.53}$$

As a result, the minimum MSE (MMSE) is

$$J_{\min}\left[f\left(y,y^*\right)\right]=\phi_x-\phi_{\widehat{x}_{\mathrm{W}}}. \tag{2.54}$$

We deduce that $\phi_{\widehat{x}_{\mathrm{W}}}\leq\phi_x$ [i.e., the function $f\left(y,y^*\right)$ does not amplify the estimated desired signal],

$$\rho=\frac{\phi_{\widehat{x}_{\mathrm{W}}}}{\phi_x}\leq 1 \tag{2.55}$$

is always real and positive,

$$\left|\gamma_{x\widehat{x}_{\mathrm{W}}}\right|^2=\rho, \tag{2.56}$$

$$J_{\min}\left[f\left(y,y^*\right)\right]=\phi_x\left[1-\left|\gamma_{x\widehat{x}_{\mathrm{W}}}\right|^2\right], \tag{2.57}$$

$$\mathrm{oSNR}=\frac{\rho}{1-\rho}, \tag{2.58}$$

$$\phi_u = \rho\,(1-\rho)\,\phi_x \leq \phi_v, \tag{2.59}$$

and

$$\frac{J_{\min}\left[f\left(y,y^*\right)\right]}{\phi_x} = \upsilon_{\mathrm{i}} - \frac{\upsilon_{\mathrm{q}}}{\mathrm{iSNR}} - \frac{\phi_{x_{\mathrm{ri}}}}{\phi_x}. \tag{2.60}$$

In order to better compromise between distortion 1 and distortion 2, we propose to use the more powerful MSE-based criterion:

$$\begin{aligned} J_\mu\left[f\left(y,y^*\right)\right] &= \mu\frac{J_{\mathrm{i}}\left[f\left(y,y^*\right)\right]}{\phi_x} + \frac{J_{\mathrm{q}}\left[f\left(y,y^*\right)\right]}{\phi_v} \\ &= \mu\left|1-\rho^*\right|^2 + \frac{\phi_u}{\phi_v}, \end{aligned} \tag{2.61}$$

where μ is a positive real number allowing to compromise between υ_{i} and υ_{q}.

For $\mu = \mathrm{iSNR}$, it is clear that minimizing $J_\mu\left[f\left(y,y^*\right)\right]$ is equivalent to minimizing the MSE criterion, $J\left[f\left(y,y^*\right)\right]$.

For $\mu = \infty$, minimizing $J_\mu\left[f\left(y,y^*\right)\right]$ is equivalent to minimizing $J\left[f\left(y,y^*\right)\right]$ with the constraint that $\rho^* = 1$. In other words, we don't affect much the partial intelligibility while we maximize quality (and, hence, the other portion of intelligibility). This approach is equivalent to the well-known minimum variance distortionless response (MVDR) technique [7], [8]. Comparing Wiener with MVDR, we understand that the former will affect intelligibility but quality will be better than the latter, which does not affect much the desired signal. The smallest output SNR should be obtained with the MVDR.

Taking $\mu \leq \mathrm{iSNR}$ (resp. $\mu \geq \mathrm{iSNR}$), will result to a noise reduction method that will decrease the partial intelligibility (resp. quality and the other portion of intelligibility) and increase the quality and the other portion of intelligibility (resp. partial intelligibility). The output SNR should improve as μ decreases but up to a certain point.

2.5 Summary

After giving a broad definition of the signal model, we presented a conceptual framework for noise reduction. Within this context, we defined the most important performance measures, namely, the input and output SNRs, and the speech intelligibility and quality indices. We then proposed a general MSE-based criterion from which all known estimators can be deduced. In the rest of this work, we will show how to apply these different concepts to all classical noise reduction schemes.

References

1. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
2. P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Chichester, England: John Wiley & Sons Ltd, 2006.
3. P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.
4. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: John Wiley & Sons, 1949.
5. J. Benesty, J. Chen, Y. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., Berlin, Germany: Springer-Verlag, 2005, Chapter 2, pp. 9–41.
6. J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1218–1234, July 2006.
7. J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.
8. R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, pp. 661–675, Aug. 1971.

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

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:

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

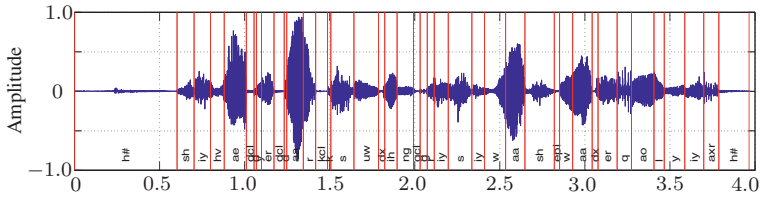

Fig. 3.1 A speech signal from the speaker FAKS0 of the TIMIT database.

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

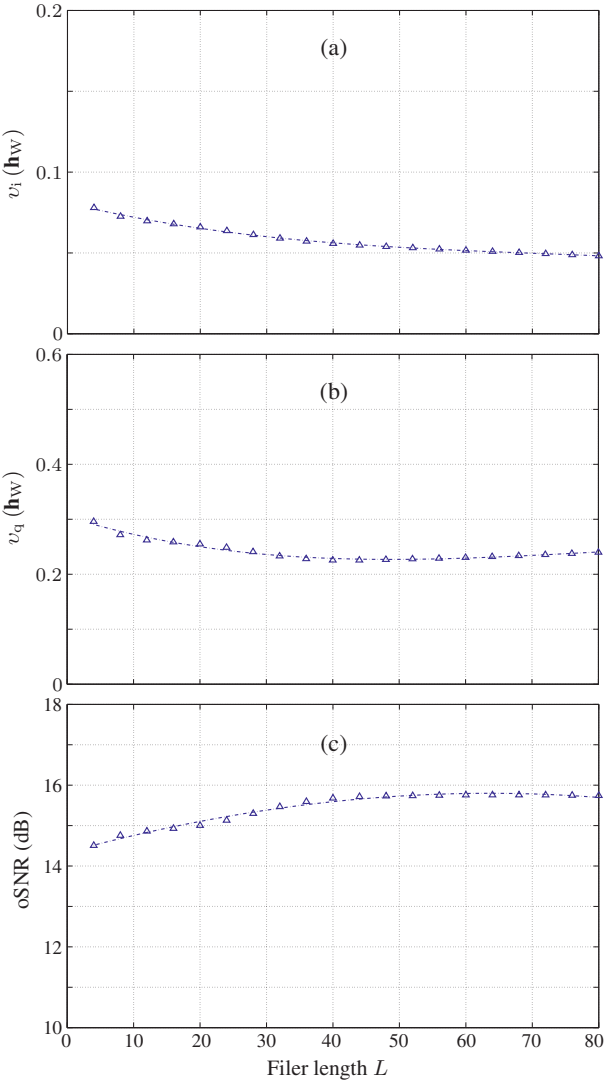

Fig. 3.2 Performance of the Wiener filter as a function of the filter length, L, in the white Gaussian noise: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The input SNR is 10 dB.

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

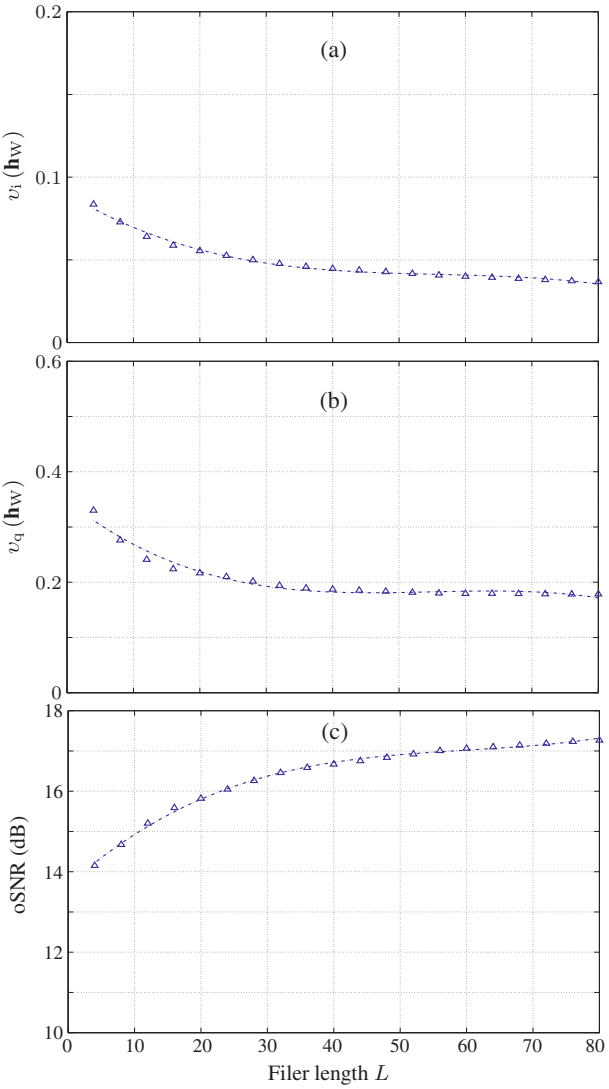

Fig. 3.3 Performance of the Wiener filter as a function of the filter length, L, in the NYSE noise: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The input SNR is 10 dB.

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

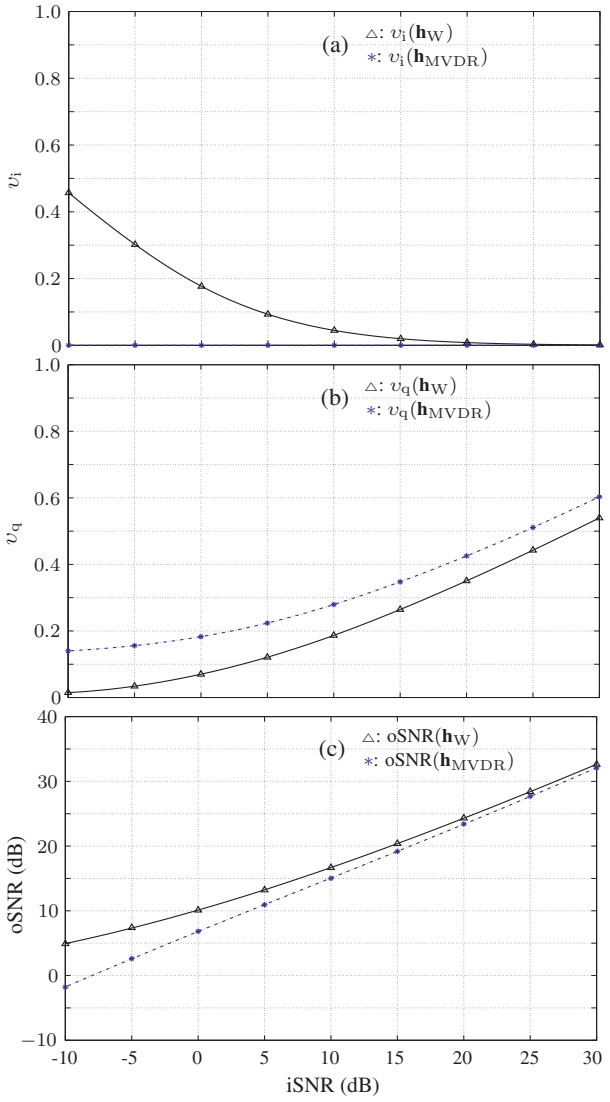

Fig. 3.4 Performance of the Wiener and MVDR filters in the white Gaussian noise at different input SNRs: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The filter length $L = 40$.

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

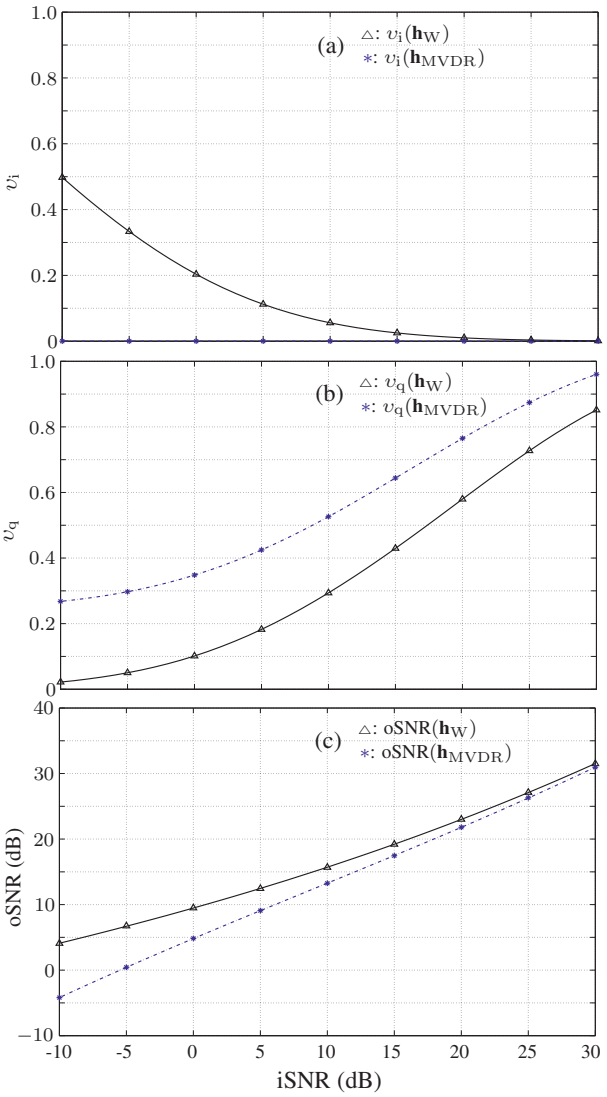

Fig. 3.5 Performance of the Wiener and MVDR filters in the NYSE noise at different input SNRs: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The filter length $L = 40$.

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.

Chapter 4
Single-Channel Noise Reduction in the STFT Domain with Interframe Correlation

In the previous chapter, we studied single-channel noise reduction in the time domain. In this chapter, we study the same problem but in the more convenient short-time Fourier transform (STFT) domain. Contrary to most conventional approaches, we do not assume here that successive STFT frames are uncorrelated. As a consequence, the interframe correlation is now taken into account and a filter is used in each subband instead of just a gain to enhance the noisy signal.

4.1 Signal Model

Using the short-time Fourier transform (STFT), (3.1) can be rewritten in the time-frequency domain as [1]

$$Y(k,n) = X(k,n) + V(k,n), \tag{4.1}$$

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

$$\begin{aligned}\phi_Y(k,n) &= E\left[|Y(k,n)|^2\right] \\ &= \phi_X(k,n) + \phi_V(k,n),\end{aligned} \tag{4.2}$$

and

$$\phi_X(k,n) = E\left[|X(k,n)|^2\right], \tag{4.3}$$

$$\phi_V(k,n) = E\left[|V(k,n)|^2\right], \tag{4.4}$$

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_4

are the variances of $X(k,n)$ and $V(k,n)$, respectively.

By considering the L most recent time frames of the signals, we can express (4.1) as

$$\begin{aligned}\mathbf{y}(k,n) &= \left[\, Y(k,n)\; Y(k,n-1)\; \cdots\; Y(k,n-L+1) \,\right]^T \\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n),\end{aligned} \tag{4.5}$$

where $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are also vectors of length L defined similarly to $\mathbf{y}(k,n)$.

At the time frame n, our desired signal is $X(k,n)$ [and not the whole vector $\mathbf{x}(k,n)$]. However, the vector $\mathbf{x}(k,n)$ in (4.5) contains both the desired signal, $X(k,n)$, and the components $X(k,n-l),\ l \neq 0$, which are not the desired signals at time frame n but signals that are correlated with $X(k,n)$. Therefore, the elements $X(k,n-l),\ l \neq 0$, contain both a part of the desired signal and a component that we consider as an interference. This suggests that we should decompose $X(k,n-l)$ into two orthogonal components corresponding to the part of the desired signal and interference, i.e.,

$$X(k,n-l) = \rho^*_{X,l}(k,n)X(k,n) + X_{\mathrm{i},l}(k,n), \tag{4.6}$$

where

$$X_{\mathrm{i},l}(k,n) = X(k,n-l) - \rho^*_{X,l}(k,n)X(k,n), \tag{4.7}$$

$$E\left[X(k,n)X^*_{\mathrm{i},l}(k,n)\right] = 0, \tag{4.8}$$

and

$$\rho_{X,l}(k,n) = \frac{E\left[X(k,n)X^*(k,n-l)\right]}{E\left[\left|X(k,n)\right|^2\right]} \tag{4.9}$$

is the interframe correlation coefficient of the signal $X(k,n)$. Hence, we can write the vector $\mathbf{x}(k,n)$ as

$$\begin{aligned}\mathbf{x}(k,n) &= X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) + \mathbf{x}_{\mathrm{i}}(k,n) \\ &= \mathbf{x}_{\mathrm{d}}(k,n) + \mathbf{x}_{\mathrm{i}}(k,n),\end{aligned} \tag{4.10}$$

where

$$\mathbf{x}_{\mathrm{d}}(k,n) = X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) \tag{4.11}$$

is the desired signal vector,

$$\mathbf{x}_{\mathrm{i}}(k,n) = \left[\, X_{\mathrm{i},0}(k,n)\; X_{\mathrm{i},1}(k,n)\; \cdots\; X_{\mathrm{i},L-1}(k,n) \,\right]^T$$

is the interference signal vector, and

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{x}X}(k,n) &= \left[\,\rho^*_{X,0}(k,n)\;\rho^*_{X,1}(k,n)\;\cdots\;\rho^*_{X,L-1}(k,n)\,\right]^T \\ &= \left[\,1\;\rho^*_{X,1}(k,n)\;\cdots\;\rho^*_{X,L-1}(k,n)\,\right]^T \\ &= \frac{E\left[\mathbf{x}(k,n)X^*(k,n)\right]}{E\left[\left|X(k,n)\right|^2\right]}\end{aligned} \tag{4.12}$$

is the (normalized) interframe correlation vector between $\mathbf{x}(k,n)$ and $X(k,n)$.

Substituting (4.10) into (4.5), the signal model for noise reduction in the STFT domain can be expressed as

$$\mathbf{y}(k,n) = X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) + \mathbf{x}_{\mathrm{i}}(k,n) + \mathbf{v}(k,n). \tag{4.13}$$

We will see how this important expression will be used in the following sections.

4.2 Linear Filtering

Since the interframe correlation is taken into account, we estimate $X(k,n),\ k = 0,1,\ldots,K-1$, by passing $Y(k,n),\ k = 0,1,\ldots,K-1$, from consecutive time frames through an FIR filter of length L, i.e.,

$$\begin{aligned}\widehat{X}(k,n) &= \sum_{l=0}^{L-1} H_l^*(k,n)Y(k,n-l) \\ &= \mathbf{h}^H(k,n)\mathbf{y}(k,n),\ k = 0,1,\ldots,K-1,\end{aligned} \tag{4.14}$$

where L is the number of consecutive time frames, the superscript ^{H} is the conjugate-transpose operator, and

$$\mathbf{h}(k,n) = \left[\,H_0(k,n)\;H_1(k,n)\;\cdots\;H_{L-1}(k,n)\,\right]^T$$

is a complex-valued filter of length L. The case $L = 1$ corresponds to the conventional STFT-domain approach where the consecutive time frames are assumed to be uncorrelated.

Substituting (4.13) into (4.14), we get

$$\begin{aligned}\widehat{X}(k,n) &= \mathbf{h}^H(k,n)\left[X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) + \mathbf{x}_{\mathrm{i}}(k,n) + \mathbf{v}(k,n)\right] \\ &= X_{\mathrm{fd}}(k,n) + X_{\mathrm{ri}}(k,n) + V_{\mathrm{rn}}(k,n),\end{aligned} \tag{4.15}$$

where

$$X_{\mathrm{fd}}(k,n) = X(k,n)\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) \tag{4.16}$$

is the filtered desired signal,

$$X_{\mathrm{ri}}(k,n) = \mathbf{h}^H(k,n)\mathbf{x}_{\mathrm{i}}(k,n) \tag{4.17}$$

is the residual interference, and

$$V_{\mathrm{rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{v}(k,n) \tag{4.18}$$

is the residual noise. We observe that the estimate of the desired signal is the sum of three terms that are mutually uncorrelated. The first one is clearly the filtered desired signal while the two others are the filtered undesired signals (interference-plus-noise). Therefore, the variance of $\widehat{X}(k,n)$ is

$$\begin{aligned}\phi_{\widehat{X}}(k,n) &= \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{y}}(k,n)\mathbf{h}(k,n) \\ &= \phi_{X_{\mathrm{fd}}}(k,n) + \phi_{X_{\mathrm{ri}}}(k,n) + \phi_{V_{\mathrm{rn}}}(k,n),\end{aligned} \tag{4.19}$$

where

$$\mathbf{\Phi}_{\mathbf{y}}(k,n) = E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] \tag{4.20}$$

is the correlation matrix of $\mathbf{y}(k,n)$,

$$\begin{aligned}\phi_{X_{\mathrm{fd}}}(k,n) &= \phi_X(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2 \\ &= \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}_{\mathrm{d}}}(k,n)\mathbf{h}(k,n),\end{aligned} \tag{4.21}$$

$$\begin{aligned}\phi_{X_{\mathrm{ri}}}(k,n) &= \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}_{\mathrm{i}}}(k,n)\mathbf{h}(k,n) \\ &= \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{h}(k,n) - \phi_X(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2,\end{aligned} \tag{4.22}$$

$$\phi_{V_{\mathrm{rn}}}(k,n) = \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n), \tag{4.23}$$

$$\mathbf{\Phi}_{\mathbf{x}_{\mathrm{d}}}(k,n) = \phi_X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n) \tag{4.24}$$

is the correlation matrix (whose rank is equal to 1) of $\mathbf{x}_{\mathrm{d}}(k,n)$, and

$$\mathbf{\Phi}_{\mathbf{a}}(k,n) = E\left[\mathbf{a}(k,n)\mathbf{a}^H(k,n)\right] \tag{4.25}$$

is the correlation matrix of $\mathbf{a}(k,n) \in \{\mathbf{x}(k,n), \mathbf{x}_{\mathrm{i}}(k,n), \mathbf{v}(k,n)\}$. In the rest, it is assumed that the rank of $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ is equal to L so that its inverse exists.

4.3 Performance Measures

In this section, the performance measures, which are tailored for noise reduction in the STFT domain with interframe correlation, are defined. We need to distinguish between subband and fullband measures.

We define the subband and fullband input SNRs at time frame n as [1]

$$\text{iSNR}(k,n) = \frac{\phi_X(k,n)}{\phi_V(k,n)},\ k = 0,1,\ldots,K-1, \tag{4.26}$$

$$\text{iSNR}(n) = \frac{\sum_{k=0}^{K-1}\phi_X(k,n)}{\sum_{k=0}^{K-1}\phi_V(k,n)}. \tag{4.27}$$

It is easy to show that

$$\text{iSNR}(n) \leq \max_k\ \text{iSNR}(k,n). \tag{4.28}$$

In words, the fullband input SNR can never exceed the maximum subband input SNR.

To quantify the level of noise remaining at the output of the FIR filter, we define the subband output SNR as

$$\begin{aligned}\text{oSNR}\left[\mathbf{h}(k,n)\right] &= \frac{\phi_{X_{\text{fd}}}(k,n)}{\phi_{X_{\text{ri}}}(k,n)+\phi_{V_{\text{rn}}}(k,n)} \\ &= \frac{\phi_X(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2}{\mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\text{in}}(k,n)\mathbf{h}(k,n)},\ k = 0,1,\ldots,K-1,\end{aligned} \tag{4.29}$$

where

$$\boldsymbol{\Phi}_{\text{in}}(k,n) = \boldsymbol{\Phi}_{\mathbf{x}_{\text{i}}}(k,n) + \boldsymbol{\Phi}_{\mathbf{v}}(k,n) \tag{4.30}$$

is the interference-plus-noise correlation matrix. For the particular filter $\mathbf{h}(k,n) = \mathbf{i}_{\text{id}}$ (identity filter), where $\mathbf{i}_{\text{id}}$ is the first column of the identity matrix $\mathbf{I}_L$ (of size $L \times L$), we have

$$\text{oSNR}\left[\mathbf{i}_{\text{id}}(k,n)\right] = \text{iSNR}(k,n). \tag{4.31}$$

And for the particular case $L = 1$, we also have

$$\text{oSNR}\left[H_0(k,n)\right] = \text{iSNR}(k,n). \tag{4.32}$$

Hence, in the two previous scenarios, the subband SNR cannot be improved.

Now, let us define the quantity:

$$\begin{aligned}\text{oSNR}_{\max}(k,n) &= \text{tr}\left[\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\Phi}_{\mathbf{x}_{\text{d}}}(k,n)\right] \\ &= \phi_X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n)\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n),\end{aligned} \tag{4.33}$$

where tr[·] denotes the trace of a square matrix. This quantity corresponds to the maximum eigenvalue, $\lambda_{\max}(k,n)$, of the matrix $\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\Phi}_{\mathbf{x}_{\text{d}}}(k,n)$. It also corresponds to the maximum subband output SNR since the filter, $\mathbf{h}_{\max}(k,n)$, that maximizes $\text{oSNR}\left[\mathbf{h}(k,n)\right]$ [eq. (4.29)] is the maximum eigenvector of $\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\Phi}_{\mathbf{x}_{\text{d}}}(k,n)$ for which its corresponding eigenvalue is

$\lambda_{\max}(k,n)$. As a result, we have

$$\text{oSNR}\left[\mathbf{h}(k,n)\right] \leq \text{oSNR}_{\max}(k,n) = \lambda_{\max}(k,n),\ \forall \mathbf{h}(k,n) \tag{4.34}$$

and

$$\text{oSNR}_{\max}(k,n) = \text{oSNR}\left[\mathbf{h}_{\max}(k,n)\right] \geq \text{oSNR}\left[\mathbf{i}_{\text{id}}(k,n)\right] = \text{iSNR}(k,n). \tag{4.35}$$

The maximum SNR filter is then

$$\mathbf{h}_{\max}(k,n) = \varsigma(k,n)\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n), \tag{4.36}$$

where $\varsigma(k,n) \neq 0$ is an arbitrary complex number. We will show in the next section that the optimal filters are equivalent to $\mathbf{h}_{\max}(k,n)$ up to $\varsigma(k,n)$.

We define the maximum subband gain in SNR as

$$\begin{aligned}\mathcal{G}_{\max}(k,n) &= \frac{\text{oSNR}_{\max}(k,n)}{\text{iSNR}(k,n)} \\ &= \phi_V(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n)\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) \geq 1.\end{aligned} \tag{4.37}$$

We define the fullband output SNR at time frame n as

$$\text{oSNR}\left[\mathbf{h}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_X(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2}{\sum_{k=0}^{K-1}\mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\text{in}}(k,n)\mathbf{h}(k,n)} \tag{4.38}$$

and it can be verified that

$$\text{oSNR}\left[\mathbf{h}(:,n)\right] \leq \max_k\ \text{oSNR}\left[\mathbf{h}(k,n)\right]. \tag{4.39}$$

The fullband output SNR with the maximum SNR filter is

$$\text{oSNR}\left[\mathbf{h}_{\max}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\frac{\left|\varsigma(k,n)\right|^2\lambda_{\max}^2(k,n)}{\phi_X(k,n)}}{\sum_{k=0}^{K-1}\frac{\left|\varsigma(k,n)\right|^2\lambda_{\max}(k,n)}{\phi_X(k,n)}}. \tag{4.40}$$

We see that the performance (in terms of SNR improvement) of the maximum SNR filter is quite dependent on the values of $\varsigma(k,n)$. We can express (4.40) as

$$\text{oSNR}\left[\mathbf{h}_{\max}(:,n)\right] = \frac{\boldsymbol{\varsigma}^H(n)\mathbf{D}_1(n)\boldsymbol{\varsigma}(n)}{\boldsymbol{\varsigma}^H(n)\mathbf{D}_2(n)\boldsymbol{\varsigma}(n)}, \tag{4.41}$$

where

$$\boldsymbol{\varsigma}(n) = \left[\varsigma(0,n)\ \varsigma(1,n)\ \cdots\ \varsigma(K-1,n)\right]^T \tag{4.42}$$

is a vector of length K containing all the scaling factors and

$$\mathbf{D}_1(n) = \text{diag}\left[\frac{\lambda^2_{\max}(0,n)}{\phi_X(0,n)}, \frac{\lambda^2_{\max}(1,n)}{\phi_X(1,n)}, \ldots, \frac{\lambda^2_{\max}(K-1,n)}{\phi_X(K-1,n)}\right], \tag{4.43}$$

$$\mathbf{D}_2(n) = \text{diag}\left[\frac{\lambda_{\max}(0,n)}{\phi_X(0,n)}, \frac{\lambda_{\max}(1,n)}{\phi_X(1,n)}, \ldots, \frac{\lambda_{\max}(K-1,n)}{\phi_X(K-1,n)}\right], \tag{4.44}$$

are two diagonal matrices. Now, if we maximize (4.41) with respect to $\boldsymbol{\varsigma}(n)$, we find that the solution, $\boldsymbol{\varsigma}_{\max}(n)$, is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{D}_2^{-1}(n)\mathbf{D}_1(n)$. Since this matrix is diagonal, its maximum eigenvalue is its largest diagonal element, i.e., $\max_k\ \lambda_{\max}(k,n)$. We deduce that

$$\text{oSNR}\left[\mathbf{h}(:,n)\right] \leq \max_k\ \lambda_{\max}(k,n),\ \forall \mathbf{h}(k,n). \tag{4.45}$$

This result is very interesting on its own since it shows that the fullband output SNR of any filter can never exceed its maximum subband output SNR.

The partial speech intelligibility index quantifies the amount of the desired signal that is cancelled by the filter. The subband and fullband partial speech intelligibility indices are then

$$\begin{aligned} \upsilon_{\text{i}}\left[\mathbf{h}(k,n)\right] &= \frac{\phi_X(k,n) - \phi_{X_{\text{fd}}}(k,n)}{\phi_X(k,n)} \\ &= 1 - \left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2,\ k = 0,1,\ldots,K-1 \end{aligned} \tag{4.46}$$

and

$$\begin{aligned} \upsilon_{\text{i}}\left[\mathbf{h}(:,n)\right] &= \frac{\sum_{k=0}^{K-1}\left[\phi_X(k,n) - \phi_{X_{\text{fd}}}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_X(k,n)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_X(k,n)\left[1 - \left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2\right]}{\sum_{k=0}^{K-1}\phi_X(k,n)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_X(k,n)\upsilon_{\text{i}}\left[\mathbf{h}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_X(k,n)}. \end{aligned} \tag{4.47}$$

The partial speech intelligibility indices are expected to be upper bounded by 1 for optimal filters. Lower values of the partial speech intelligibility indices imply a higher intelligible signal.

The quality of the signal is measured with the speech quality index. Therefore, we define the subband and fullband speech quality indices as

$$\upsilon_{\mathrm{q}}\left[\mathbf{h}(k,n)\right] = \frac{\phi_{V_{\mathrm{rn}}}(k,n)}{\phi_V(k,n)} = \frac{\mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n)}{\phi_V(k,n)},\ k=0,1,\ldots,K-1 \tag{4.48}$$

and

$$\upsilon_{\mathrm{q}}\left[\mathbf{h}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_{V_{\mathrm{rn}}}(k,n)}{\sum_{k=0}^{K-1}\phi_V(k,n)} = \frac{\sum_{k=0}^{K-1}\mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n)}{\sum_{k=0}^{K-1}\phi_V(k,n)} = \frac{\sum_{k=0}^{K-1}\phi_V(k,n)\upsilon_{\mathrm{q}}\left[\mathbf{h}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_V(k,n)}. \tag{4.49}$$

The speech quality indices are also expected to be upper bounded by 1 for optimal filters. Low values of the speech quality indices imply a good signal quality.

The global speech intelligibility index quantifies the amount of the desired signal that is affected by the filter. The subband and fullband global speech intelligibility indices are derived from the previous definitions:

$$\upsilon_{\mathrm{i}}'\left[\mathbf{h}(k,n)\right] = (1-\varpi)\,\upsilon_{\mathrm{i}}\left[\mathbf{h}(k,n)\right] + \varpi\upsilon_{\mathrm{q}}\left[\mathbf{h}(k,n)\right],\ k=0,1,\ldots,K-1 \tag{4.50}$$

and

$$\upsilon_{\mathrm{i}}'\left[\mathbf{h}(:,n)\right] = (1-\varpi)\,\upsilon_{\mathrm{i}}\left[\mathbf{h}(:,n)\right] + \varpi\upsilon_{\mathrm{q}}\left[\mathbf{h}(:,n)\right]. \tag{4.51}$$

The variance of the estimated desired signal can be rewritten as a function of the subband speech intelligibility and quality indices, i.e.,

$$\phi_{\widehat{X}}(k,n) = \left\{1-\upsilon_{\mathrm{i}}\left[\mathbf{h}(k,n)\right]\right\}\phi_X(k,n) + \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}_{\mathrm{i}}}(k,n)\mathbf{h}(k,n) + \upsilon_{\mathrm{q}}\left[\mathbf{h}(k,n)\right]\phi_V(k,n), \tag{4.52}$$

which is interesting to compare to the variance of the observation signal, i.e.,

$$\phi_Y(k,n) = \phi_X(k,n) + \phi_V(k,n). \tag{4.53}$$

We see how any optimal filter will try to compromise between speech intelligibility and speech quality.

4.4 MSE-Based Criterion

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

$$\begin{aligned}\mathcal{E}(k,n) &= \widehat{X}(k,n) - X(k,n) \\ &= \mathbf{h}^H(k,n)\mathbf{y}(k,n) - X(k,n).\end{aligned} \tag{4.54}$$

We can rewrite (4.54) as

$$\mathcal{E}(k,n) = \mathcal{E}_{\mathrm{i}}(k,n) + \mathcal{E}_{\mathrm{q}}(k,n), \tag{4.55}$$

where

$$\begin{aligned}\mathcal{E}_{\mathrm{i}}(k,n) &= X_{\mathrm{fd}}(k,n) - X(k,n) \\ &= \left[\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) - 1\right] X(k,n)\end{aligned} \tag{4.56}$$

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

$$\begin{aligned}\mathcal{E}_{\mathrm{q}}(k,n) &= X_{\mathrm{ri}}(k,n) + V_{\mathrm{rn}}(k,n) \\ &= \mathbf{h}^H(k,n)\mathbf{x}_{\mathrm{i}}(k,n) + \mathbf{h}^H(k,n)\mathbf{v}(k,n)\end{aligned} \tag{4.57}$$

represents the residual interference-plus-noise, which affects the quality and the other portion of intelligibility. It is obvious that

$$E\left[\mathcal{E}_{\mathrm{i}}(k,n)\mathcal{E}_{\mathrm{q}}^*(k,n)\right] = 0. \tag{4.58}$$

Having defined the error signal, we can now write the subband MSE criterion:

$$\begin{aligned}J\left[\mathbf{h}(k,n)\right] &= E\left[\left|\mathcal{E}(k,n)\right|^2\right] \\ &= J_{\mathrm{i}}\left[\mathbf{h}(k,n)\right] + J_{\mathrm{q}}\left[\mathbf{h}(k,n)\right],\end{aligned} \tag{4.59}$$

where

$$\begin{aligned}J_{\mathrm{i}}\left[\mathbf{h}(k,n)\right] &= E\left[\left|\mathcal{E}_{\mathrm{i}}(k,n)\right|^2\right] \\ &= E\left[\left|X_{\mathrm{fd}}(k,n) - X(k,n)\right|^2\right] \\ &= \phi_X(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) - 1\right|^2\end{aligned} \tag{4.60}$$

and

$$\begin{aligned} J_{\mathrm{q}}\left[\mathbf{h}(k,n)\right] &= E\left[\left|\mathcal{E}_{\mathrm{q}}(k,n)\right|^2\right] \\ &= E\left[\left|X_{\mathrm{ri}}(k,n)\right|^2\right] + E\left[\left|V_{\mathrm{rn}}(k,n)\right|^2\right] \\ &= \phi_{X_{\mathrm{ri}}}(k,n) + \phi_{V_{\mathrm{rn}}}(k,n). \end{aligned} \tag{4.61}$$

For the two particular filters $\mathbf{h}(k,n) = \mathbf{i}_{\mathrm{id}}$ and $\mathbf{h}(k,n) = \mathbf{0}_{L\times 1}$, we get

$$J\left[\mathbf{i}_{\mathrm{id}}(k,n)\right] = J_{\mathrm{q}}\left[\mathbf{i}_{\mathrm{id}}(k,n)\right] = \phi_V(k,n), \tag{4.62}$$

$$J\left[\mathbf{0}_{L\times 1}(k,n)\right] = J_{\mathrm{i}}\left[\mathbf{0}_{L\times 1}(k,n)\right] = \phi_X(k,n). \tag{4.63}$$

We then find that the subband NMSE with respect to $J\left[\mathbf{i}_{\mathrm{id}}(k,n)\right]$ is

$$\begin{aligned} J_{\mathrm{n},1}\left[\mathbf{h}(k,n)\right] &= \frac{J\left[\mathbf{h}(k,n)\right]}{J\left[\mathbf{i}_{\mathrm{id}}(k,n)\right]} \\ &= \mathrm{iSNR}(k,n) \times \left|1 - \mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2 \\ &\quad + \frac{\mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathrm{in}}(k,n)\mathbf{h}(k,n)}{\phi_V(k,n)} \end{aligned} \tag{4.64}$$

and the subband NMSE with respect to $J\left[\mathbf{0}_{L\times 1}(k,n)\right]$ is

$$\begin{aligned} J_{\mathrm{n},2}\left[\mathbf{h}(k,n)\right] &= \frac{J\left[\mathbf{h}(k,n)\right]}{J\left[\mathbf{0}_{L\times 1}(k,n)\right]} \\ &= \left|1 - \mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2 + \frac{\mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathrm{in}}(k,n)\mathbf{h}(k,n)}{\phi_X(k,n)}. \end{aligned} \tag{4.65}$$

We have

$$J_{\mathrm{n},1}\left[\mathbf{h}(k,n)\right] = \mathrm{iSNR}(k,n) \times J_{\mathrm{n},2}\left[\mathbf{h}(k,n)\right]. \tag{4.66}$$

Expressions (4.64) and (4.65) show how the subband NMSEs and the different subband MSEs are implicitly related to the subband performance measures.

We are only interested in complex filters for which

$$J_{\mathrm{i}}\left[\mathbf{i}_{\mathrm{id}}(k,n)\right] \leq J_{\mathrm{i}}\left[\mathbf{h}(k,n)\right] < J_{\mathrm{i}}\left[\mathbf{0}_{L\times 1}(k,n)\right], \tag{4.67}$$

$$J_{\mathrm{q}}\left[\mathbf{0}_{L\times 1}(k,n)\right] < J_{\mathrm{q}}\left[\mathbf{h}(k,n)\right] < J_{\mathrm{q}}\left[\mathbf{i}_{\mathrm{id}}(k,n)\right]. \tag{4.68}$$

From the two previous expressions, we deduce that

$$0 \leq \left|1 - \mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2 < 1, \tag{4.69}$$

$$0 < \frac{\mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathrm{in}}(k,n)\mathbf{h}(k,n)}{\phi_V(k,n)} < 1. \tag{4.70}$$

As we have shown in previous chapters, to better compromise between speech intelligibility and speech quality, we propose to use the more general

subband MSE-based criterion:

$$\begin{aligned} J_{\mu}\left[\mathbf{h}(k,n)\right] &= \mu(k,n)\frac{J_{\mathrm{i}}\left[\mathbf{h}(k,n)\right]}{\phi_X(k,n)} + \frac{J_{\mathrm{q}}\left[\mathbf{h}(k,n)\right]}{\phi_V(k,n)} \\ &= \mu(k,n)\left|1-\mathbf{h}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2 \\ &\quad + \frac{\mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathrm{in}}(k,n)\mathbf{h}(k,n)}{\phi_V(k,n)}, \end{aligned} \tag{4.71}$$

where $\mu(k,n)$ is a positive real number allowing this compromise.

4.5 Optimal Filters

By minimizing $J_{\mu}\left[\mathbf{h}(k,n)\right]$ [eq. (4.71)] with respect to $\mathbf{h}(k,n)$, we find the complex optimal filter:

$$\mathbf{h}_{\mathrm{o},\mu}(k,n) = \mu(k,n)\left[\mu(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n) + \frac{\boldsymbol{\Phi}_{\mathrm{in}}(k,n)}{\phi_V(k,n)}\right]^{-1}\boldsymbol{\rho}_{\mathbf{x}X}(k,n). \tag{4.72}$$

From the decomposition:

$$\boldsymbol{\Phi}_{\mathbf{y}}(k,n) = \phi_X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n) + \boldsymbol{\Phi}_{\mathrm{in}}(k,n), \tag{4.73}$$

we can rewrite the optimal filter as

$$\begin{aligned} &\mathbf{h}_{\mathrm{o},\mu}(k,n) = \mu(k,n)\times \\ &\left\{\left[\mu(k,n) - \mathrm{iSNR}(k,n)\right]\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n) + \frac{\boldsymbol{\Phi}_{\mathbf{y}}(k,n)}{\phi_V(k,n)}\right\}^{-1}\boldsymbol{\rho}_{\mathbf{x}X}(k,n) \end{aligned} \tag{4.74}$$

and the vector $\boldsymbol{\rho}_{\mathbf{x}X}(k,n)$ can be expressed as a function of the statistics of $Y(k,n)$ and $V(k,n)$, i.e.,

$$\begin{aligned} \boldsymbol{\rho}_{\mathbf{x}X}(k,n) &= \frac{E\left[\mathbf{y}(k,n)Y^*(k,n)\right] - E\left[\mathbf{v}(k,n)V^*(k,n)\right]}{\phi_Y(k,n) - \phi_V(k,n)} \\ &= \frac{\phi_Y(k,n)\boldsymbol{\rho}_{\mathbf{y}Y}(k,n) - \phi_V(k,n)\boldsymbol{\rho}_{\mathbf{v}V}(k,n)}{\phi_Y(k,n) - \phi_V(k,n)}, \end{aligned} \tag{4.75}$$

so that $\mathbf{h}_{\mathrm{o},\mu}(k,n)$ can be estimated from the statistics of $Y(k,n)$ and $V(k,n)$ only.

Now, by using the Woodbury's identity in (4.72), it can easily be shown that the optimal filter can be reformulated as

$$\mathbf{h}_{\mathrm{o},\mu}(k,n) = \frac{\mu(k,n)\dfrac{\phi_X(k,n)}{\mathrm{iSNR}(k,n)}}{1+\mu(k,n)\dfrac{\mathrm{oSNR}_{\max}(k,n)}{\mathrm{iSNR}(k,n)}}\boldsymbol{\Phi}_{\mathrm{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n) \tag{4.76}$$
$$= \frac{\mu(k,n)\phi_V(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\boldsymbol{\Phi}_{\mathrm{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n).$$

Comparing $\mathbf{h}_{\mathrm{o},\mu}(k,n)$ with $\mathbf{h}_{\max}(k,n)$ [eq. (4.36)], we observe that the two filters are equivalent up to a scaling factor. As a result, $\mathbf{h}_{\mathrm{o},\mu}(k,n)$ also maximizes the subband output SNR, i.e.,

$$\mathrm{oSNR}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right] = \mathrm{oSNR}_{\max}(k,n),\ \forall \mu(k,n) > 0 \tag{4.77}$$

and

$$\mathrm{oSNR}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right] \geq \mathrm{iSNR}(k,n),\ \forall \mu(k,n) \geq 0. \tag{4.78}$$

From (4.76), we deduce that the subband partial speech intelligibility index and the subband speech quality index are, respectively,

$$\upsilon_{\mathrm{i}}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right] = 1 - \left[\frac{\mu(k,n)\mathcal{G}_{\max}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\right]^2 \tag{4.79}$$
$$= 1 - \left|\mathbf{h}_{\mathrm{o},\mu}^H(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)\right|^2$$

and

$$\upsilon_{\mathrm{q}}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right] = \frac{\mathbf{h}_{\mathrm{o},\mu}^H(k,n)\boldsymbol{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}_{\mathrm{o},\mu}(k,n)}{\phi_V(k,n)}. \tag{4.80}$$

Obviously, $\forall \mu(k,n) \geq 0$, we have

$$0 \leq \upsilon_{\mathrm{i}}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right] \leq 1, \tag{4.81}$$
$$0 \leq \upsilon_{\mathrm{q}}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right] \leq 1. \tag{4.82}$$

We deduce that the fullband indices are

$$\upsilon_{\mathrm{i}}\left[\mathbf{h}_{\mathrm{o},\mu}(:,n)\right] = 1 - \frac{\sum_{k=0}^{K-1}\phi_X(k,n)\upsilon_{\mathrm{i}}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_X(k,n)}, \tag{4.83}$$

$$\upsilon_{\mathrm{q}}\left[\mathbf{h}_{\mathrm{o},\mu}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_V(k,n)\upsilon_{\mathrm{q}}\left[\mathbf{h}_{\mathrm{o},\mu}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_V(k,n)}, \tag{4.84}$$

and, $\forall \mu(k,n) \geq 0$, we also have

$$0 \leq \upsilon_{\mathrm{i}}\left[\mathbf{h}_{\mathrm{o},\mu}(:,n)\right] \leq 1, \tag{4.85}$$
$$0 \leq \upsilon_{\mathrm{q}}\left[\mathbf{h}_{\mathrm{o},\mu}(:,n)\right] \leq 1. \tag{4.86}$$

It is easy to check that the fullband output SNR is

$$\text{oSNR}\left[\mathbf{h}_{\text{o},\mu}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_X(k,n)\left[\dfrac{\mu(k,n)\mathcal{G}_{\max}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\right]^2}{\sum_{k=0}^{K-1}\phi_V(k,n)\dfrac{\mu^2(k,n)\mathcal{G}_{\max}(k,n)}{\left[1+\mu\mathcal{G}_{\max}(k,n)\right]^2}}. \quad (4.87)$$

Taking $\mu(k,n)=\infty$ in (4.76), we find the MVDR filter [2], [3]:

$$\mathbf{h}_{\text{MVDR}}(k,n) = \frac{\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)}{\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n)\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)}. \quad (4.88)$$

We deduce that

$$\upsilon_{\text{i}}\left[\mathbf{h}_{\text{MVDR}}(k,n)\right] = 0, \quad (4.89)$$
$$\upsilon_{\text{i}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right] = 0. \quad (4.90)$$

Taking $\mu(k,n)=\text{iSNR}(k,n)$ in (4.76), we find the Wiener filter [2]:

$$\begin{aligned}\mathbf{h}_{\text{W}}(k,n) &= \frac{\phi_X(k,n)\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)}{1+\phi_X(k,n)\boldsymbol{\rho}_{\mathbf{x}X}^H(k,n)\boldsymbol{\Phi}_{\text{in}}^{-1}(k,n)\boldsymbol{\rho}_{\mathbf{x}X}(k,n)} \\ &= \boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,n)\boldsymbol{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{\text{id}} \\ &= \left[\mathbf{I}_L - \boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,n)\boldsymbol{\Phi}_{\mathbf{v}}(k,n)\right]\mathbf{i}_{\text{id}}.\end{aligned} \quad (4.91)$$

It can be verified that

$$\upsilon_{\text{i}}\left[\mathbf{h}_{\text{W}}(:,n)\right] > \upsilon_{\text{i}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right], \quad (4.92)$$
$$\upsilon_{\text{q}}\left[\mathbf{h}_{\text{W}}(:,n)\right] < \upsilon_{\text{q}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right]. \quad (4.93)$$

Therefore, we can expect a better signal quality with Wiener than MVDR and a more intelligible signal with MVDR than Wiener.

It can also be verified that for $\mu(k,n)\geq\text{iSNR}(k,n)$, we have

$$\upsilon_{\text{i}}\left[\mathbf{h}_{\text{W}}(:,n)\right] \geq \upsilon_{\text{i}}\left[\mathbf{h}_{\text{o},\mu}(:,n)\right] \geq \upsilon_{\text{i}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right], \quad (4.94)$$
$$\upsilon_{\text{q}}\left[\mathbf{h}_{\text{W}}(:,n)\right] \leq \upsilon_{\text{q}}\left[\mathbf{h}_{\text{o},\mu}(:,n)\right] \leq \upsilon_{\text{q}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right], \quad (4.95)$$

and for $\mu(k,n)\leq\text{iSNR}(k,n)$, we have

$$\upsilon_{\text{i}}\left[\mathbf{h}_{\text{o},\mu}(:,n)\right] \geq \upsilon_{\text{i}}\left[\mathbf{h}_{\text{W}}(:,n)\right] > \upsilon_{\text{i}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right], \quad (4.96)$$
$$\upsilon_{\text{q}}\left[\mathbf{h}_{\text{o},\mu}(:,n)\right] \leq \upsilon_{\text{q}}\left[\mathbf{h}_{\text{W}}(:,n)\right] < \upsilon_{\text{q}}\left[\mathbf{h}_{\text{MVDR}}(:,n)\right]. \quad (4.97)$$

4.6 Particular Case

In this section, we briefly study the particular case of $L = 1$.

When the interframe correlation is not taken into account, i.e., when $L = 1$, we get back to the conventional STFT-domain approach [1]. In this case, some of the main variables simplify to

$$\begin{aligned} \boldsymbol{\rho}_{\mathbf{x}X}(k,n) &= 1, \\ \boldsymbol{\Phi}_{\mathrm{in}}(k,n) &= \phi_V(k,n), \\ \mathcal{G}_{\max}(k,n) &= 1. \end{aligned}$$

As a result, the complex optimal filter becomes a real positive gain:

$$H_{\mathrm{o},\mu}(k,n) = \frac{\mu(k,n)}{1+\mu(k,n)}. \tag{4.98}$$

For $\mu(k,n) = \mathrm{iSNR}(k,n)$, we get the conventional noncausal Wiener gain [1]:

$$H_{\mathrm{W}}(k,n) = \frac{\mathrm{iSNR}(k,n)}{1+\mathrm{iSNR}(k,n)}, \tag{4.99}$$

while for $\mu(k,n) = \infty$, we obtain the unity (distortionless) gain:

$$H_{\mathrm{DL}}(k,n) = 1, \tag{4.100}$$

for which the estimated desired signal, $\widehat{X}_{\mathrm{DL}}(k,n)$, is equal to the observation signal, $Y(k,n)$.

4.7 Simulations

In this section, we briefly study the performance of the STFT-domain noise reduction filters derived above through simulations and illustrate the benefit of using multiple STFT frames in improving the performance of single-channel noise reduction. The simulation setup is the same as in Section 3.6. Again, the clean speech is taken from the speaker FAKS0 in the TIMIT database. We continue to study the narrowband case, so the original signals are downsampled from 16 kHz to 8 kHz and these downsampled signals are used as the clean speech in all the simulations. The noisy signals are obtained by adding noise to the clean speech, where the noise signal is properly scaled to control the input SNR level. Similar to the study in Chapter 3, two types of noise are considered here: white Gaussian and NYSE babble.

To implement the STFT-domain noise reduction filters, the noisy speech signal is partitioned into overlapping frames. Our targeted application is full-

duplex voice communications, which allows only a small delay by a noise reduction processor, generally in the magnitude of 10 ms, so we set the frame size to 8 ms in our simulations with a 75% overlapping with neighboring frames (note that if a longer delay is permissible with the applications, one can use a larger frame size, which may slightly improve the performance of noise reduction). A Kaiser window is then applied to each frame (to reduce the aliasing effect due to circular convolution) and the windowed signal is subsequently transformed into the STFT domain using a 64-point fast Fourier transform (FFT). A noise reduction filter is then constructed and applied to the noisy STFT coefficients in every subband. After noise reduction, the inverse FFT (IFFT) with the overlap-add method is used for signal reconstruction in the time domain. A same Kaiser window is applied to the output of the IFFT before the overlap-add process, again, to reduce the aliasing effect caused by circular convolution.

Obviously, the most critical step in the above implementation process is the computation of the noise reduction filters in the STFT subbands. It is seen from (4.72) or (4.74) that we need to know the signal statistics $\mathbf{\Phi}_{\text{in}}(k,n)$ and $\boldsymbol{\rho}_{\mathbf{x}X}(k,n)$ or $\mathbf{\Phi}_{\mathbf{y}}(k,n)$ and $\boldsymbol{\rho}_{\mathbf{x}X}(k,n)$ in order to compute the optimal noise reduction filters. In our simulations, these statistics are estimated as follows. We first estimate the $\mathbf{\Phi}_{\mathbf{y}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ matrices using the following recursions [note that we assume that the noise signal, $V(k,n)$, is accessible so that the process of noise estimation is avoided]:

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) = \alpha_y \widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n-1) + (1-\alpha_y)\,\mathbf{y}(k,n)\mathbf{y}^H(k,n), \tag{4.101}$$

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n) = \alpha_v \widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n-1) + (1-\alpha_v)\,\mathbf{v}(k,n)\mathbf{v}^H(k,n), \tag{4.102}$$

where $\alpha_y \in (0,1)$ and $\alpha_v \in (0,1)$ are the forgetting factors that control the influence of the previous data samples on the current correlation matrix estimate (the initial estimates of these two matrices are obtained from the first 100 signal frames with a short-time average). After the estimates of the $\mathbf{\Phi}_{\mathbf{y}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ matrices are available at time frame n, the estimate of $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ is computed as $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) - \widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$. And then, the estimate of the interframe correlation vector $\boldsymbol{\rho}_{\mathbf{x}X}(k,n)$ is taken as the first column of $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,n)$ normalized by its first element.

With the previous way of statistics estimation, the noise reduction performance of the STFT-domain optimal filters certainly depends on the filter length, L, and the forgetting factors, α_y and α_v. To evaluate this dependency, we can examine either the subband performance measures as defined in Section 4.3 or the fullband measures as defined in Section 3.3. The difference is that the subband measures in Section 4.3 can be used to derive and assess every subband noise reduction filter while the measures in Section 3.3 are convenient to analyze the overall performance. In what follows, we will use the three fullband performance measures defined in Section 3.3, i.e., the partial speech intelligibility index, the speech quality index, and the

output SNR. We first reconstruct the time-domain signals $x_{\rm fd}(t)$, $x_{\rm ri}(t)$, and $v_{\rm rn}(t)$, from their respective STFT-domain counterparts $X_{\rm fd}(k,n)$, $X_{\rm ri}(k,n)$, and $V_{\rm rn}(k,n)$. The fullband partial speech intelligibility index, speech quality index, and output SNR are then computed according to their definitions in Section 3.3 by replacing the mathematical expectation with a long-time average.

Now, suppose that the noise is stationary so that we can set α_v to a large value that is close to 1. The noise reduction performance of the STFT-domain filters is then a function of the filter length, L, and the forgetting factors α_y. Figure 4.1 plots the performance of the STFT-domain MVDR filter as a function of the forgetting factor α_y for different values of the filter length, L, in the white Gaussian noise with $\alpha_v = 0.98$. Note that the MVDR filter degenerates to the unity gain when $L = 1$, which does not change the noisy signal; in this case, the output SNR is equal to the input SNR, the partial speech intelligibility index is zero, and the speech quality index is maximal.

One important observation one can make from Fig. 4.1 is that the value of α_y plays an important role on the noise reduction performance. This role is even more critical as the filter length, L, increases. The reason is that the size of the correlation matrix that needs to be inverted grows as the filter length increases and, as a result, a larger value of α_y needs to be used to make the correlation matrix estimate numerically well defined. Consequently, the optimal forgetting factor that produces the best noise reduction performance increases with L. However, regardless of the value of L, the value of α_y cannot be too large. If it is too large, the estimated statistics cannot capture the time-varying property of the nonstationary speech signals, leading to performance degradation. In general, the optimal value of α_y depends on both the stationarity of the signal of interest and the noise as well as the filter length, L. It should be tuned based on the application scenario for the best noise reduction performance.

Another important observation we can make from Fig. 4.1 is that using multiple STFT frames can greatly help improve noise reduction performance. In comparison with the single-frame case where it is a unity gain, the MVDR filter using two consecutive frames can improve the SNR by more than 3 dB if the forgetting factor α_y is properly chosen as seen from Fig. 4.1. When the filter length is 8, more than 5-dB SNR improvement is achieved with a proper value of α_y. Comparing the case with $L = 16$ and that with $L = 8$, one can see that there is no performance improvement. This performance saturation is primarily due to the fact that there exists not much correlation between distant frames.

In practice, noise can also be nonstationary and, therefore, choosing a proper value of α_v is also important. One easy way is to set $\alpha_y = \alpha_v$. Figure 4.2 plots the performance of the MVDR filter also in the white Gaussian noise, but this time with $\alpha_y = \alpha_v$. Again, one can clearly see the dependency of the noise reduction performance of the MVDR filter on the forgetting factors and the usefulness of using multiple STFT frames to help improve noise

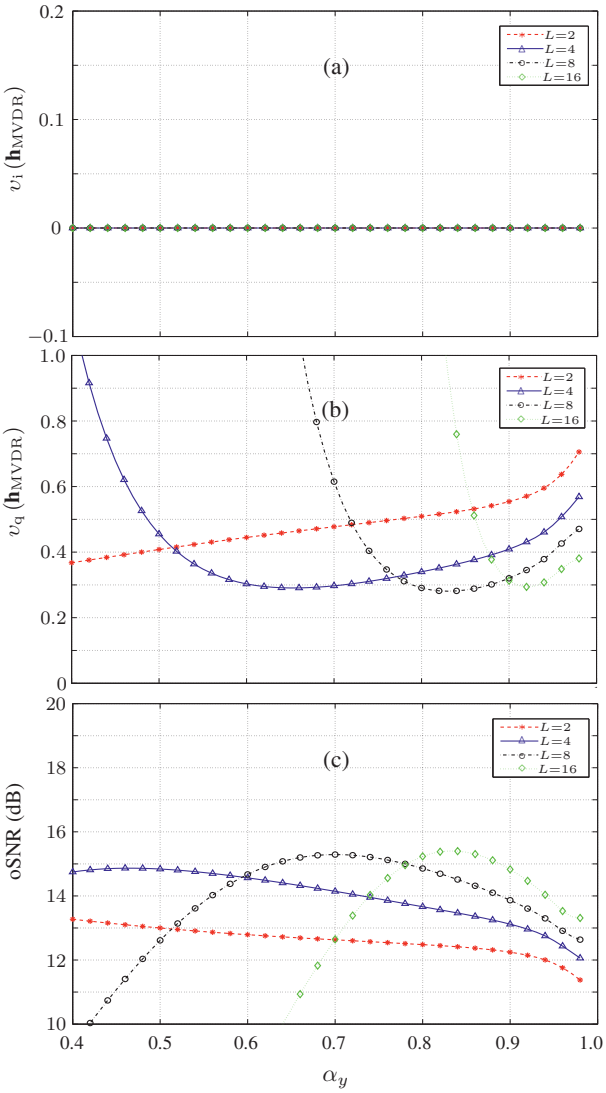

Fig. 4.1 Fullband performance of the MVDR filter for different values of the filter length, L, as a function of the forgetting factor α_y in the white Gaussian noise. The forgetting factor is $\alpha_v = 0.98$, the window size is $K = 64$ (8 ms) with a 75% overlap, and the fullband input SNR is 10 dB.

reduction performance. Comparing Figs. 4.1 and 4.2, we can see that varying the forgetting factor α_v can help optimize the performance even when the noise is stationary. In practice, the two forgetting factors α_y and α_v can be

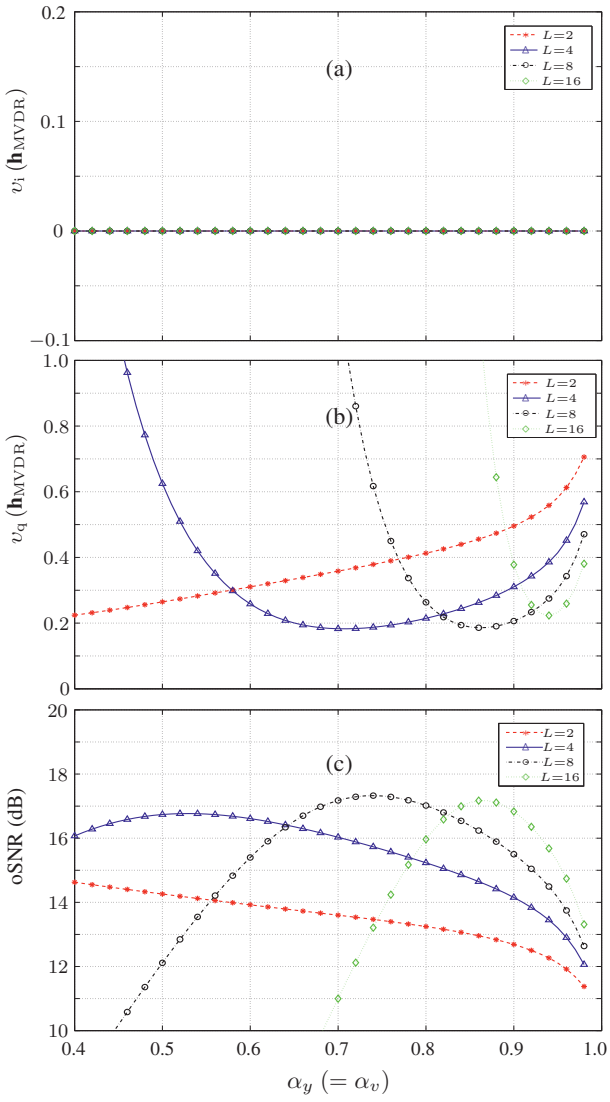

Fig. 4.2 Fullband performance of the MVDR filter for different values of the filter length, L, as a function of the forgetting factor $\alpha_y\ (=\alpha_v)$ in the white Gaussian noise. The window size is $K = 64$ (8 ms) with a 75% overlap and the fullband input SNR is 10 dB.

optimized separately in an iterative way, which will be left to the reader's investigation.

Figure 4.3 plots the performance of the MVDR filter in the NYSE noise. The conditions of this simulation are the same as those in the previous one except that, this time, we take the NYSE noise. Comparing Figs. 4.3 and

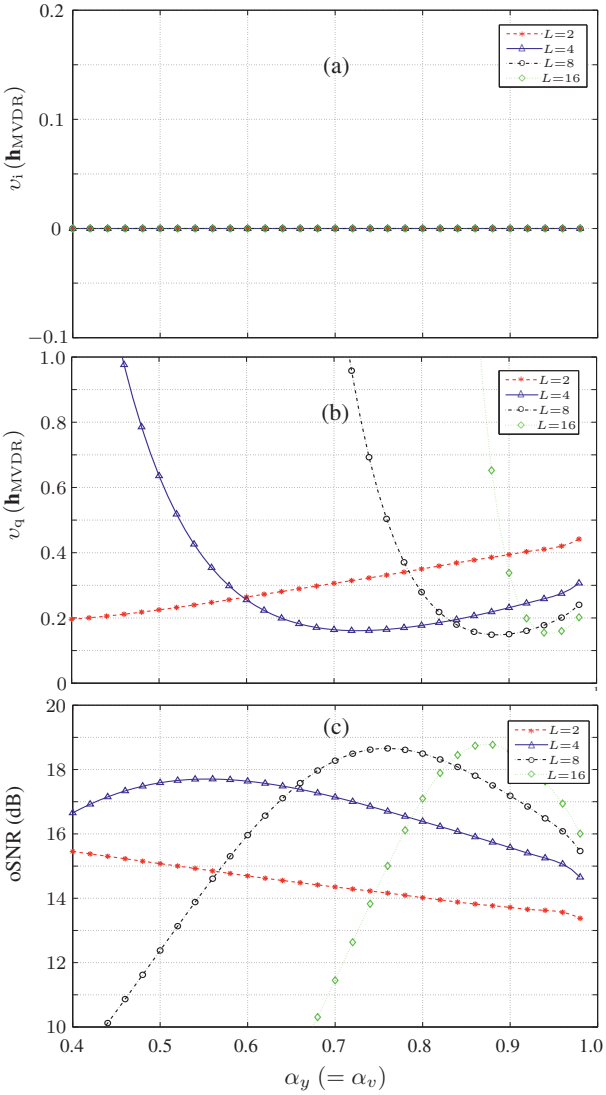

Fig. 4.3 Fullband performance of the MVDR filters for different values of the filter length, L, as a function of the forgetting factor α_y $(=\alpha_v)$ in the NYSE noise. The window size is $K = 64$ (8 ms) with a 75% overlap and the fullband input SNR is 10 dB.

4.2, one can see that the performance trend of the MVDR filter 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 differ slightly in values in the two different noise cases with the same input SNR.

The fullband partial speech intelligibility index is always 0 with the MVDR filter. If some degradation of this index is allowed, one may use the general optimal filter given in either (4.72), (4.74), or (4.76) by setting a proper value of $\mu(k,n)$ or the Wiener filter given in (4.91) to improve the SNR; the results are not presented here.

References

1. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
2. J. Benesty and Y. Huang, *A Perspective on Single-Channel Frequency-Domain Speech Enhancement*. Morgan & Claypool Publishers, 2011.
3. J. Benesty and Y. Huang,"A single-channel noise reduction MVDR filter," in *Proc. IEEE ICASSP*, 2011, pp. 273–276.

Chapter 5
Binaural Noise Reduction in the Time Domain

Binaural noise reduction is an important problem in applications where there is a need to produce two "clean" outputs from noisy observations picked up by multiple microphones. But the mitigation of the noise should be made in such a way that no audible distortion is added to the two outputs (this is the same as in the single-channel case) and meanwhile the spatial information of the desired sound source should be preserved so that, after noise reduction, the remote listener will still be able to localize the sound source thanks to his/her binaural hearing mechanism. In this chapter, we approach this problem with the widely linear theory in the time domain, where both the temporal and spatial information is exploited.

5.1 Signal Model

In this study, we consider the signal model in which $2M$ microphones[1] capture a source signal convolved with acoustic impulse responses in some noise field. The signal received at the ith microphone is then expressed as

$$\begin{aligned} y_{\mathrm{r},i}(t) &= g_i(t) * s(t) + v_{\mathrm{r},i}(t) \\ &= x_{\mathrm{r},i}(t) + v_{\mathrm{r},i}(t),\ i = 1, 2, \ldots, 2M, \end{aligned} \tag{5.1}$$

where $g_i(t)$ is the acoustic impulse response from the unknown speech source, $s(t)$, location to the ith microphone, $*$ stands for linear convolution, and $v_{\mathrm{r},i}(t)$ is the additive noise at microphone i. We assume that the impulse responses are time invariant. We also assume that the signals $x_{\mathrm{r},i}(t) = g_i(t) * s(t)$ and $v_{\mathrm{r},i}(t)$ are uncorrelated, zero mean, real, and broadband.

In this chapter, we consider the problem of recovering the signals $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},M+1}(t)$ given the observations $y_{\mathrm{r},i}(t),\ i = 1, 2, \ldots, 2M$. This means

[1] The generalization to an odd number of microphones is straightforward.

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_5

that the desired signals in our problem are the speech signals received at the first and $(M+1)$th microphones[2]. It is clear then that we have two objectives. The first one is to attenuate the contribution of the noise terms $v_{\mathrm{r},1}(t)$ and $v_{\mathrm{r},M+1}(t)$ as much as possible. The second objective is to preserve $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},M+1}(t)$ with their spatial information, so that with the enhanced signals, along with our binaural hearing process, we will still be able to localize the source $s(t)$. This is the well-known problem of binaural noise reduction.

Since we have binaural signals, it is more convenient to work in the complex domain in order that the original (binaural) problem is transformed to the conventional (monaural) noise reduction processing with a microphone array [1], [2]. Indeed, from the $2M$ real microphone signals given in (5.1), we can form M complex microphone signals as

$$\begin{aligned} y_m(t) &= y_{\mathrm{r},m}(t) + jy_{\mathrm{r},M+m}(t) \\ &= x_m(t) + v_m(t),\ m = 1, 2, \ldots, M, \end{aligned} \tag{5.2}$$

where $j = \sqrt{-1}$,

$$x_m(t) = x_{\mathrm{r},m}(t) + jx_{\mathrm{r},M+m}(t),\ m = 1, 2, \ldots, M \tag{5.3}$$

is the complex convolved speech signal, and

$$v_m(t) = v_{\mathrm{r},m}(t) + jv_{\mathrm{r},M+m}(t),\ m = 1, 2, \ldots, M \tag{5.4}$$

is the complex additive noise. Now, our problem may be stated as follows: given the M complex microphone signals, $y_m(t),\ m = 1, 2, \ldots, M$, which are a mixture of the uncorrelated complex signals $x_m(t)$ and $v_m(t)$, our goal is to recover $x_1(t) = x_{\mathrm{r},1}(t) + jx_{\mathrm{r},M+1}(t)$ (i.e., our desired signal) the best way we can, including the phase, which is important for the localization of the source signal.

The signal model given in (5.2) can be put into a vector form if we accumulate L successive time samples:

$$\mathbf{y}_m(t) = \mathbf{x}_m(t) + \mathbf{v}_m(t),\ m = 1, 2, \ldots, M, \tag{5.5}$$

where

$$\mathbf{y}_m(t) = \big[\, y_m(t)\ y_m(t-1)\ \cdots\ y_m(t-L+1) \,\big]^T \tag{5.6}$$

is a vector of length L, and $\mathbf{x}_m(t)$ and $\mathbf{v}_m(t)$ are defined in a similar way to $\mathbf{y}_m(t)$. Concatenating the M vectors in (5.5) together, we get the vector of length ML:

[2] We can try to recover the desired signals from any other pair of microphones, if we like.

$$\begin{aligned}\underline{\mathbf{y}}(t) &= \left[\mathbf{y}_1^T(t)\ \mathbf{y}_2^T(t)\cdots\ \mathbf{y}_M^T(t)\right]^T \\ &= \underline{\mathbf{x}}(t) + \underline{\mathbf{v}}(t),\end{aligned} \tag{5.7}$$

where $\underline{\mathbf{x}}(t)$ and $\underline{\mathbf{v}}(t)$ are defined in a similar way to $\underline{\mathbf{y}}(t)$.

Since $x_m(t)$ and $v_m(t)$ are uncorrelated by assumption, the correlation matrix (of size $ML \times ML$) of the noisy signal is

$$\begin{aligned}\mathbf{\Phi}_{\underline{\mathbf{y}}} &= E\left[\underline{\mathbf{y}}(t)\underline{\mathbf{y}}^H(t)\right] \\ &= \mathbf{\Phi}_{\underline{\mathbf{x}}} + \mathbf{\Phi}_{\underline{\mathbf{v}}},\end{aligned} \tag{5.8}$$

where

$$\mathbf{\Phi}_{\underline{\mathbf{x}}} = E\left[\underline{\mathbf{x}}(t)\underline{\mathbf{x}}^H(t)\right], \tag{5.9}$$

$$\mathbf{\Phi}_{\underline{\mathbf{v}}} = E\left[\underline{\mathbf{v}}(t)\underline{\mathbf{v}}^H(t)\right], \tag{5.10}$$

are the correlation matrices of $\underline{\mathbf{x}}(t)$ and $\underline{\mathbf{v}}(t)$, respectively.

5.2 Widely Linear Filtering

As it can be noticed from the model given in (5.2), we deal with complex random variables. A very important statistical characteristic of a complex random variable (CRV) is the so-called circularity property or lack of it (noncircularity) [3], [4]. A zero-mean CRV, z, is circular if and only if the only nonnull moments and cumulants are the moments and cumulants constructed with the same power in z and z^* [5], [6]. In particular, z is said to be a second-order circular CRV (CCRV) if its so-called pseudo-variance [3] is equal to zero, i.e., $E\left(z^2\right) = 0$, while its variance is nonnull, i.e., $E\left(|z|^2\right) \neq 0$. This means that the second-order behavior of a CCRV is well described by its variance. If the pseudo-variance $E\left(z^2\right)$ is not equal to 0, the CRV z is then noncircular. A good measure of the second-order circularity is the circularity quotient [3] defined as the ratio between the pseudo-variance and the variance, i.e.,

$$\gamma_z = \frac{E\left(z^2\right)}{E\left(|z|^2\right)}. \tag{5.11}$$

It is easy to show that $0 \leq |\gamma_z| \leq 1$. If $\gamma_z = 0$, z is a second-order CCRV; otherwise, z is noncircular, and a larger value of $|\gamma_z|$ indicates that the CRV z is more noncircular.

Now, let us examine whether the complex desired signal, $x_1(t) = x_{\mathrm{r},1}(t) + jx_{\mathrm{r},M+1}(t)$, is second-order circular. We have

$$\begin{aligned}\gamma_{x_1} &= \frac{E\left[x_1^2(t)\right]}{E\left[|x_1(t)|^2\right]} \\ &= \frac{E\left[x_{\mathrm{r},1}^2(t)\right] - E\left[x_{\mathrm{r},M+1}^2(t)\right] + 2jE\left[x_{\mathrm{r},1}(t)x_{\mathrm{r},M+1}(t)\right]}{\phi_{x_1}},\end{aligned} \tag{5.12}$$

where $\phi_{x_1} = E\left[|x_1(t)|^2\right]$ is the variance of $x_1(t)$. One can check from (5.12) that the CRV $x_1(t)$ is second-order circular (i.e., $\gamma_{x_1} = 0$) if and only if

$$E\left[x_{\mathrm{r},1}^2(t)\right] = E\left[x_{\mathrm{r},M+1}^2(t)\right] \quad \text{and} \quad E\left[x_{\mathrm{r},1}(t)x_{\mathrm{r},M+1}(t)\right] = 0. \tag{5.13}$$

Since the signals $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},M+1}(t)$ come from the same source, they are in general correlated. As a result, the second condition in (5.13) should not be true. Therefore, we can safely state that the complex desired signal, $x_1(t)$, is noncircular, and so is the complex microphone signal, $y_1(t)$. If we assume that the noise terms at the two microphones are uncorrelated and have the same power then $\gamma_{v_1} = 0$ [i.e., $v(t)$ is a second-order CCRV].

Since we deal with noncircular CRVs as demonstrated above, the classical linear estimation technique [7], which is developed for processing real signals or CCRVs, cannot be applied. Instead, an estimate of $x_1(t)$ should be obtained using the widely linear (WL) estimation theory as [4], [8]

$$\begin{aligned}\widehat{x}_1(t) &= \underline{\mathbf{h}}^H\underline{\mathbf{y}}(t) + \underline{\mathbf{h}}'^H\underline{\mathbf{y}}^*(t) \\ &= \widetilde{\mathbf{h}}^H\widetilde{\mathbf{y}}(t),\end{aligned} \tag{5.14}$$

where $\underline{\mathbf{h}}$ and $\underline{\mathbf{h}}'$ are two complex FIR filters of length ML and

$$\widetilde{\mathbf{h}} = \begin{bmatrix}\underline{\mathbf{h}} \\ \underline{\mathbf{h}}'\end{bmatrix}, \tag{5.15}$$

$$\widetilde{\mathbf{y}}(t) = \begin{bmatrix}\underline{\mathbf{y}}(t) \\ \underline{\mathbf{y}}^*(t)\end{bmatrix}, \tag{5.16}$$

are the augmented WL filter and observation vector, respectively, both of length $2ML$. We can rewrite (5.14) as

$$\begin{aligned}\widehat{x}_1(t) &= \widetilde{\mathbf{h}}^H\left[\widetilde{\mathbf{x}}(t) + \widetilde{\mathbf{v}}(t)\right] \\ &= x_{\mathrm{f}}(t) + v_{\mathrm{rn}}(t),\end{aligned} \tag{5.17}$$

where $\widetilde{\mathbf{x}}(t)$ and $\widetilde{\mathbf{v}}(t)$ are defined in a similar way to $\widetilde{\mathbf{y}}(t)$,

$$x_{\mathrm{f}}(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{x}}(t) \tag{5.18}$$

is a filtered version of the desired signal, and

$$v_{\mathrm{rn}}(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{v}}(t) \tag{5.19}$$

is the residual noise. From (5.17), we see that $\widehat{x}_1(t)$ depends on the vector $\widetilde{\mathbf{x}}(t)$. However, our desired signal at time t is only $x_1(t)$ [and not the whole vector $\widetilde{\mathbf{x}}(t)$]; so we should decompose the vector $\widetilde{\mathbf{x}}(t)$ into two orthogonal vectors: one corresponding to the desired signal at time t and the other corresponding to the interference. Therefore, we have

$$\begin{aligned}\widetilde{\mathbf{x}}(t) &= x_1(t)\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} + \widetilde{\mathbf{x}}_{\mathrm{i}}(t) \\ &= \widetilde{\mathbf{x}}_{\mathrm{d}}(t) + \widetilde{\mathbf{x}}_{\mathrm{i}}(t),\end{aligned} \tag{5.20}$$

where

$$\widetilde{\mathbf{x}}_{\mathrm{d}}(t) = x_1(t)\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \tag{5.21}$$

is the desired signal vector,

$$\widetilde{\mathbf{x}}_{\mathrm{i}}(t) = \widetilde{\mathbf{x}}(t) - \widetilde{\mathbf{x}}_{\mathrm{d}}(t) \tag{5.22}$$

is the interference signal vector,

$$\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} = \frac{E\left[\widetilde{\mathbf{x}}(t)x_1^*(t)\right]}{E\left[\left|x_1(t)\right|^2\right]} \tag{5.23}$$

is the normalized [with respect to $x_1(t)$] correlation vector between $\widetilde{\mathbf{x}}(t)$ and $x_1(t)$, and

$$E\left[\widetilde{\mathbf{x}}_{\mathrm{i}}(t)x_1^*(t)\right] = \mathbf{0}_{2ML\times 1}. \tag{5.24}$$

Substituting (5.20) into (5.17), we obtain

$$\begin{aligned}\widehat{x}_1(t) &= \widetilde{\mathbf{h}}^H\left[x_1(t)\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} + \widetilde{\mathbf{x}}_{\mathrm{i}}(t) + \widetilde{\mathbf{v}}(t)\right] \\ &= x_{\mathrm{fd}}(t) + x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t),\end{aligned} \tag{5.25}$$

where

$$x_{\mathrm{fd}}(t) = x_1(t)\widetilde{\mathbf{h}}^H\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \tag{5.26}$$

is the filtered desired signal and

$$x_{\mathrm{ri}}(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{x}}_{\mathrm{i}}(t) \tag{5.27}$$

is the residual interference. We observe that the estimate of the desired signal at time t is the sum of three terms that are mutually uncorrelated. Therefore, the variance of $\widehat{x}_1(t)$ is

$$\phi_{\widehat{x}_1} = \phi_{x_{\mathrm{fd}}} + \phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}}, \tag{5.28}$$

where

$$\phi_{x_{\mathrm{fd}}} = \phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \right|^2 \tag{5.29}$$
$$= \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{d}}} \widetilde{\mathbf{h}},$$
$$\phi_{x_{\mathrm{ri}}} = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} \widetilde{\mathbf{h}} \tag{5.30}$$
$$= \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}} \widetilde{\mathbf{h}} - \phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \right|^2,$$
$$\phi_{v_{\mathrm{rn}}} = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} \widetilde{\mathbf{h}}, \tag{5.31}$$

$\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{d}}} = \phi_{x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H$ is the correlation matrix (whose rank is equal to 1) of $\widetilde{\mathbf{x}}_{\mathrm{d}}(t)$, and $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} = E\left[\widetilde{\mathbf{x}}_{\mathrm{i}}(t)\widetilde{\mathbf{x}}_{\mathrm{i}}^H(t)\right]$, $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}} = E\left[\widetilde{\mathbf{x}}(t)\widetilde{\mathbf{x}}^H(t)\right]$, and $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} = E\left[\widetilde{\mathbf{v}}(t)\widetilde{\mathbf{v}}^H(t)\right]$ are the correlation matrices of $\widetilde{\mathbf{x}}_{\mathrm{i}}(t)$, $\widetilde{\mathbf{x}}(t)$, and $\widetilde{\mathbf{v}}(t)$, respectively.

It is clear from (5.25) that the objective of our noise reduction problem is to find optimal filters that can minimize the effect of $x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t)$ while preserving the desired signal, $x_1(t)$. But before deriving such filters, we first give some very useful performance measures for the evaluation of the time-domain binaural noise reduction problem with the WL model.

5.3 Performance Measures

Since the complex microphone signal $y_1(t)$ is our reference signal, all measures are defined with respect to this signal.

The input SNR is defined as

$$\mathrm{iSNR} = \frac{\phi_{x_1}}{\phi_{v_1}}, \tag{5.32}$$

where $\phi_{v_1} = E\left[\left|v_1(t)\right|^2\right]$ is the variance of the complex additive noise at the first complex microphone.

To quantify the level of noise remaining at the output of the complex WL filter, we define the output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual interference-plus-noise, i.e.,

$$\mathrm{oSNR}\left(\widetilde{\mathbf{h}}\right) = \frac{\phi_{x_{\mathrm{fd}}}}{\phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}}} \tag{5.33}$$
$$= \frac{\phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \right|^2}{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}}$$
$$= \frac{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{d}}} \widetilde{\mathbf{h}}}{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}},$$

where

$$\mathbf{\Phi}_{\text{in}} = \mathbf{\Phi}_{\widetilde{\mathbf{x}}_{\text{i}}} + \mathbf{\Phi}_{\widetilde{\mathbf{v}}} \tag{5.34}$$

is the interference-plus-noise correlation matrix. The objective of the WL noise reduction filter is to make the output SNR greater than the input SNR so that the quality of the noisy signal may be enhanced.

For the particular filter $\widetilde{\mathbf{h}} = \widetilde{\mathbf{i}}_{\text{id}}$, where the identity filter, $\widetilde{\mathbf{i}}_{\text{id}}$, is the first column of the identity matrix, $\mathbf{I}_{2L}$ of size $2L \times 2L$, we have

$$\text{oSNR}\left(\widetilde{\mathbf{i}}_{\text{id}}\right) = \text{iSNR}. \tag{5.35}$$

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

Now, let us introduce the quantity $\text{oSNR}_{\max}$, which is defined as the maximum output SNR that can be achieved through filtering so that

$$\text{oSNR}\left(\widetilde{\mathbf{h}}\right) \leq \text{oSNR}_{\max}, \ \forall \widetilde{\mathbf{h}}. \tag{5.36}$$

It can be checked from (5.33) that this quantity is equal to the maximum eigenvalue of the matrix $\mathbf{\Phi}_{\text{in}}^{-1}\mathbf{\Phi}_{\widetilde{\mathbf{x}}_{\text{d}}}$, i.e.,

$$\text{oSNR}_{\max} = \lambda_{\max}. \tag{5.37}$$

The filter that can achieve $\text{oSNR}_{\max}$ is called the maximum SNR filter and is denoted by $\widetilde{\mathbf{h}}_{\max}$. It is easy to see from (5.37) that $\widetilde{\mathbf{h}}_{\max}$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{\Phi}_{\text{in}}^{-1}\mathbf{\Phi}_{\widetilde{\mathbf{x}}_{\text{d}}}$, i.e.,

$$\widetilde{\mathbf{h}}_{\max} = \varsigma \mathbf{\Phi}_{\text{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}, \tag{5.38}$$

where $\varsigma \neq 0$ is an arbitrary complex number. Clearly, we have

$$\text{oSNR}_{\max} = \text{oSNR}\left(\widetilde{\mathbf{h}}_{\max}\right) \geq \text{oSNR}\left(\widetilde{\mathbf{i}}_{\text{id}}\right) = \text{iSNR}. \tag{5.39}$$

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

$$\begin{aligned} \text{oSNR}_{\max} &= \text{tr}\left(\mathbf{\Phi}_{\text{in}}^{-1}\mathbf{\Phi}_{\widetilde{\mathbf{x}}_{\text{d}}}\right) \\ &= \phi_{x_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^{H}\mathbf{\Phi}_{\text{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}. \end{aligned} \tag{5.40}$$

We define the array gain as

$$\mathcal{G}\left(\widetilde{\mathbf{h}}\right) = \frac{\text{oSNR}\left(\widetilde{\mathbf{h}}\right)}{\text{iSNR}}. \tag{5.41}$$

We easily deduce that the maximum array gain is

$$\mathcal{G}_{\max} = \phi_{v_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^{H}\mathbf{\Phi}_{\text{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \geq 1. \tag{5.42}$$

The partial speech intelligibility index measures the amount of the desired signal, $x_1(t)$, that is cancelled by the WL filter. It is defined as

$$\begin{aligned} \upsilon_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right) &= \frac{\phi_{x_1} - \phi_{x_{\mathrm{fd}}}}{\phi_{x_1}} \\ &= 1 - \left|\widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}\right|^2. \end{aligned} \tag{5.43}$$

A high value of $\upsilon_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right)$ implies a high distortion of the estimated desired signal.

The speech quality index measures the amount of the residual noise left after the WL filtering process. We define it as

$$\begin{aligned} \upsilon_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right) &= \frac{\phi_{v_{\mathrm{rn}}}}{\phi_{v_1}} \\ &= \frac{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} \widetilde{\mathbf{h}}}{\phi_{v_1}}. \end{aligned} \tag{5.44}$$

A high value of $\upsilon_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right)$ implies a low quality of the estimated desired signal or a high input SNR.

We deduce that the global speech intelligibility index is

$$\upsilon'_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right) = (1 - \varpi)\,\upsilon_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right) + \varpi \upsilon_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right). \tag{5.45}$$

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

$$\phi_{\widehat{x}_1} = \left[1 - \upsilon_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right)\right] \phi_{x_1} + \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} \widetilde{\mathbf{h}} + \upsilon_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right) \phi_{v_1}. \tag{5.46}$$

5.4 MSE-Based Criterion

The error signal between the estimated and desired signals is defined as

$$\begin{aligned} e(t) &= \widehat{x}_1(t) - x_1(t) \\ &= x_{\mathrm{fd}}(t) + x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t) - x_1(t). \end{aligned} \tag{5.47}$$

We can express (5.47) as

$$e(t) = e_{\mathrm{i}}(t) + e_{\mathrm{q}}(t), \tag{5.48}$$

where

$$\begin{aligned} e_{\mathrm{i}}(t) &= x_{\mathrm{fd}}(t) - x_1(t) \\ &= \left(\widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} - 1\right) x_1(t) \end{aligned} \tag{5.49}$$

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

$$\begin{aligned} e_{\mathrm{q}}(t) &= x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t) \\ &= \widetilde{\mathbf{h}}^H \widetilde{\mathbf{x}}_{\mathrm{i}}(t) + \widetilde{\mathbf{h}}^H \widetilde{\mathbf{v}}(t) \end{aligned} \tag{5.50}$$

represents the residual interference-plus-noise, which affects the quality and the other part of intelligibility. The two error signals $e_{\mathrm{i}}(t)$ and $e_{\mathrm{q}}(t)$ are clearly uncorrelated.

The classical MSE criterion is then

$$\begin{aligned} J\left(\widetilde{\mathbf{h}}\right) &= E\left[\left|e(t)\right|^2\right] \\ &= \phi_{x_1} + \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}} \widetilde{\mathbf{h}} - \phi_{x_1} \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} - \phi_{x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^H \widetilde{\mathbf{h}} \\ &= J_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right) + J_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right), \end{aligned} \tag{5.51}$$

where $\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}$ is the correlation matrix of $\widetilde{\mathbf{y}}(t)$,

$$\begin{aligned} J_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right) &= E\left[\left|e_{\mathrm{i}}(t)\right|^2\right] \\ &= \phi_{x_1} \left|\widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} - 1\right|^2, \end{aligned} \tag{5.52}$$

and

$$\begin{aligned} J_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right) &= E\left[\left|e_{\mathrm{q}}(t)\right|^2\right] \\ &= \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}. \end{aligned} \tag{5.53}$$

The two particular filters $\widetilde{\mathbf{h}} = \widetilde{\mathbf{i}}_{\mathrm{id}}$ and $\widetilde{\mathbf{h}} = \mathbf{0}_{2ML\times 1}$ 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

$$J\left(\widetilde{\mathbf{i}}_{\mathrm{id}}\right) = J_{\mathrm{q}}\left(\widetilde{\mathbf{i}}_{\mathrm{id}}\right) = \phi_{v_1}, \tag{5.54}$$

$$J\left(\mathbf{0}_{2ML\times 1}\right) = J_{\mathrm{i}}\left(\mathbf{0}_{2ML\times 1}\right) = \phi_{x_1}. \tag{5.55}$$

As a result,

$$\mathrm{iSNR} = \frac{J\left(\mathbf{0}_{2ML\times 1}\right)}{J\left(\widetilde{\mathbf{i}}_{\mathrm{id}}\right)}. \tag{5.56}$$

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

$$J_{\mathrm{n},1}\left(\widetilde{\mathbf{h}}\right) = \frac{J\left(\widetilde{\mathbf{h}}\right)}{J\left(\widetilde{\mathbf{i}}_{\mathrm{id}}\right)} = \mathrm{iSNR} \times \left|1 - \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\right|^2 + \frac{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}}{\phi_{v_1}}. \tag{5.57}$$

We define the NMSE with respect to $J\left(\mathbf{0}_{2ML\times 1}\right)$ as

$$J_{\mathrm{n},2}\left(\widetilde{\mathbf{h}}\right) = \frac{J\left(\widetilde{\mathbf{h}}\right)}{J\left(\mathbf{0}_{2ML\times 1}\right)} = \left|1 - \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\right|^2 + \frac{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}}{\phi_{x_1}} \tag{5.58}$$

and

$$J_{\mathrm{n},1}\left(\widetilde{\mathbf{h}}\right) = \mathrm{iSNR} \times J_{\mathrm{n},2}\left(\widetilde{\mathbf{h}}\right). \tag{5.59}$$

Expressions (5.57) and (5.58) show how the NMSEs and the different MSEs are implicitly related to the performance measures.

We are interested in WL filters for which

$$J_{\mathrm{i}}\left(\widetilde{\mathbf{i}}_{\mathrm{id}}\right) \leq J_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right) < J_{\mathrm{i}}\left(\mathbf{0}_{2ML\times 1}\right), \tag{5.60}$$

$$J_{\mathrm{q}}\left(\mathbf{0}_{2ML\times 1}\right) < J_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right) < J_{\mathrm{q}}\left(\widetilde{\mathbf{i}}_{\mathrm{id}}\right). \tag{5.61}$$

From the two previous expressions, we deduce that

$$0 \leq \left|1 - \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\right|^2 < 1, \tag{5.62}$$

$$0 < \frac{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}}{\phi_{v_1}} < 1. \tag{5.63}$$

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

$$J_{\mu}\left(\widetilde{\mathbf{h}}\right) = \mu \frac{J_{\mathrm{i}}\left(\widetilde{\mathbf{h}}\right)}{\phi_{x_1}} + \frac{J_{\mathrm{q}}\left(\widetilde{\mathbf{h}}\right)}{\phi_{v_1}} = \mu \left|1 - \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\right|^2 + \frac{\widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}}{\phi_{v_1}}, \tag{5.64}$$

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

5.5 Optimal Filters

Taking the gradient of (5.64) with respect to $\widetilde{\mathbf{h}}$ and equating the result to zero, we get the optimal filter:

$$\widetilde{\mathbf{h}}_{\mathrm{o},\mu} = \mu\left(\mu\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^H + \frac{\boldsymbol{\Phi}_{\mathrm{in}}}{\phi_{v_1}}\right)^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}. \tag{5.65}$$

Using the decomposition:

$$\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}} = \phi_{x_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^H + \boldsymbol{\Phi}_{\mathrm{in}}, \tag{5.66}$$

we can express the optimal filter as

$$\widetilde{\mathbf{h}}_{\mathrm{o},\mu} = \mu\left[(\mu - \mathrm{iSNR})\,\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^H + \frac{\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}}{\phi_{v_1}}\right]^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \tag{5.67}$$

and the vector $\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}$ can be expressed as a function of the statistics of $\widetilde{\mathbf{y}}(t)$ and $\widetilde{\mathbf{v}}(t)$, i.e.,

$$\begin{aligned}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} &= \frac{E\left[\widetilde{\mathbf{y}}(t)y_1(t)\right] - E\left[\widetilde{\mathbf{v}}(t)v_1(t)\right]}{\phi_{y_1} - \phi_{v_1}} \\ &= \frac{\phi_{y_1}\boldsymbol{\rho}_{\widetilde{\mathbf{y}}y_1} - \phi_{v_1}\boldsymbol{\rho}_{\widetilde{\mathbf{v}}v_1}}{\phi_{y_1} - \phi_{v_1}},\end{aligned} \tag{5.68}$$

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

Using the Woodbury's identity in (5.65), we reformulate the optimal filter as

$$\begin{aligned}\widetilde{\mathbf{h}}_{\mathrm{o},\mu} &= \frac{\mu\dfrac{\phi_{x_1}}{\mathrm{iSNR}}}{1 + \mu\dfrac{\mathrm{oSNR}_{\max}}{\mathrm{iSNR}}}\boldsymbol{\Phi}_{\mathrm{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \\ &= \frac{\mu\phi_{v_1}}{1 + \mu\mathcal{G}_{\max}}\boldsymbol{\Phi}_{\mathrm{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}.\end{aligned} \tag{5.69}$$

Comparing $\widetilde{\mathbf{h}}_{\mathrm{o},\mu}$ with $\widetilde{\mathbf{h}}_{\max}$ [eq. (5.38)], we see that the two filters are equivalent up to a scaling factor. As a result, $\widetilde{\mathbf{h}}_{\mathrm{o},\mu}$ also maximizes the output SNR, i.e.,

$$\text{oSNR}\left(\widetilde{\mathbf{h}}_{\text{o},\mu}\right) = \text{oSNR}_{\max},\ \forall \mu > 0. \tag{5.70}$$

From (5.69), we deduce the partial speech intelligibility index:

$$\upsilon_{\text{i}}\left(\widetilde{\mathbf{h}}_{\text{o},\mu}\right) = 1 - \left(\frac{\mu\mathcal{G}_{\max}}{1+\mu\mathcal{G}_{\max}}\right)^2 \tag{5.71}$$

and the speech quality index:

$$\upsilon_{\text{q}}\left(\widetilde{\mathbf{h}}_{\text{o},\mu}\right) = \frac{\mu^2\phi_{v_1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}^T\boldsymbol{\Phi}_{\text{in}}^{-1}\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}\boldsymbol{\Phi}_{\text{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}}{\left(1+\mu\mathcal{G}_{\max}\right)^2}. \tag{5.72}$$

We see clearly that, $\forall \mu \geq 0$, we have

$$0 \leq \upsilon_{\text{i}}\left(\widetilde{\mathbf{h}}_{\text{o},\mu}\right) \leq 1, \tag{5.73}$$

$$0 \leq \upsilon_{\text{q}}\left(\widetilde{\mathbf{h}}_{\text{o},\mu}\right) \leq 1. \tag{5.74}$$

Taking $\mu = \text{iSNR}$ in (5.69), we find the Wiener filter:

$$\begin{aligned}\widetilde{\mathbf{h}}_{\text{W}} &= \frac{\phi_{x_1}}{1+\text{oSNR}_{\max}}\boldsymbol{\Phi}_{\text{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \\ &= \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1}\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}\widetilde{\mathbf{i}}_{\text{id}} \\ &= \left(\mathbf{I}_{2L} - \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1}\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}\right)\widetilde{\mathbf{i}}_{\text{id}}\end{aligned} \tag{5.75}$$

and taking $\mu = \infty$ in (5.69), we find the MVDR filter:

$$\begin{aligned}\widetilde{\mathbf{h}}_{\text{MVDR}} &= \frac{\phi_{x_1}}{\text{oSNR}_{\max}}\boldsymbol{\Phi}_{\text{in}}^{-1}\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \\ &= \frac{1+\text{oSNR}_{\max}}{\text{oSNR}_{\max}}\widetilde{\mathbf{h}}_{\text{W}}.\end{aligned} \tag{5.76}$$

A value of μ in (5.69) greater (resp. smaller) than the input SNR will result in a filter that will favor intelligibility (resp. quality) over quality (resp. intelligibility) as compared to the Wiener filter.

5.6 Simulations

In this section, we illustrate the performance of the optimal binaural noise reduction filters deduced above. Simulations are conducted using impulse responses measured in a reverberant room. The room is 6 m long and 5 m wide. For ease of exposition, positions in the room are designated by (x,y) coordinates with reference to one corner of the room, $0 \leq x \leq 6$ and $0 \leq y \leq 5$.

An equispaced linear array with six omnidirectional microphones is configured where the first and last microphones are, respectively, at (3.4, 0.5) and (3.9, 0.5), and the spacing between two neighboring microphones is 0.1 m. A loudspeaker is placed at (1.3, 3.0) to simulate a speech source. To make the experiments repeatable, the acoustic channel impulse responses were measured from the source position to all the six microphones. During experiments, the microphones' outputs are generated by convolving the source signal with the corresponding measured impulse responses and noise is then added to the convolved results to control the input SNR level. The source signal is taken from the speaker FAKS0 in the TIMIT database. We continue to study the narrowband case, so the original signals from the TIMIT database are downsampled from 16 kHz to 8 kHz before convolving them with the measured impulse responses.

With the microphone array of six sensors, we divide the performance study of binaural noise reduction into three cases: with two microphones, with four microphones, and with six microphones. In the two-microphone case, the first and fourth microphones are used while in the four-microphone case, the first, second, fourth, and fifth microphones are used. The reason to select the microphones in this way is to enable fair comparison of performances among the three cases.

To implement the optimal binaural noise reduction filters given in the previous section, we need to know either the noisy correlation matrix $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$ or the interference-plus-noise correlation matrix $\mathbf{\Phi}_{\text{in}}$ and the correlation vector $\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}$. In our simulations, we compute the $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$ matrix from the noisy signals using a short-time average. Specifically, at each time instant t, an estimate of $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$ is computed using the most recent 640 samples (80-ms long) of the noisy signals, $y_m(k), m = 1, 2, 3$. To obtain an estimate of the correlation vector $\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}$, we assume that the noise signals are accessible and then compute this vector according to the relationship given in (5.68) where all the statistics ϕ_{y_1}, ϕ_{v_1}, $\boldsymbol{\rho}_{\widetilde{\mathbf{y}}y_1}$, and $\boldsymbol{\rho}_{\widetilde{\mathbf{v}}v_1}$ are computed directly from the respective signals, again, with a short-time average using the most recent 640 samples. Substituting these statistics estimates into (5.67), (5.75), and (5.76), we implement, respectively, the general optimal filter, the Wiener filter, and the MVDR filter.

Figure 5.1 plots the partial speech intelligibility index, the speech quality index, and the output SNR of the binaural WL Wiener filter, all as a function of the filter length, L. The noise in this simulation is white Gaussian and the input SNR is 10 dB. It is seen that the performance of the WL Wiener filter with four microphones ($M = 2$) is better than that with two microphones ($M = 1$); not only the output SNR is larger, the values of both the partial speech intelligibility index and the speech quality index are also smaller. Similarly, the performance with 6 microphones ($M = 3$) is better than that with four microphones. This clearly shows the benefit of using more microphones to help improve binaural noise reduction.

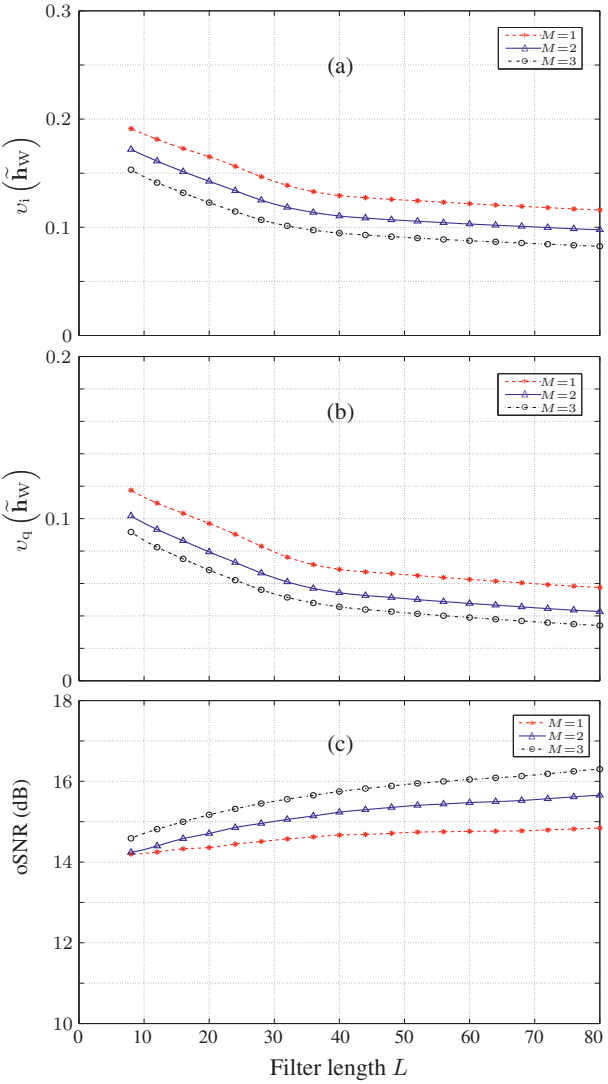

Fig. 5.1 Performance of the WL Wiener filter as a function of the filter length, L, in the white Gaussian noise with 2 ($M = 1$), 4 ($M = 2$), and 6 ($M = 3$) microphones: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The input SNR is 10 dB.

Regardless of how many microphones are used, the value of the filter length, L, plays an important role on the noise reduction performance as seen in Fig. 5.1. Not only the output SNR increases with L, the values of the partial speech intelligibility index and the speech quality index also decrease with L. Therefore, sufficiently large filter lengths should be used for good

performance. However, as the value of L increases, so is the dimension of the correlation matrix to be inverted. This would not only increase the computational complexity, but may also cause numerical instability for matrix inversion. Therefore, a proper value of L should be a compromise between the noise reduction performance and the algorithm complexity and stability.

References

1. J. Benesty, J. Chen, and Y. Huang, "Binaural noise reduction in the time domain with a stereo setup," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, pp. 2260–2272, Nov. 2011.
2. J. Chen and J. Benesty, "A time-domain widely linear MVDR filter for binaural noise reduction," in *Proc. IEEE WASPAA*, 2011, pp. 105–108.
3. E. Ollila, "On the circularity of a complex random variable," *IEEE Signal Process. Lett.*, vol. 15, pp. 841–844, 2008.
4. D. P. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. Wiley, 2009.
5. P. O. Amblard, M. Gaeta, and J. L. Lacoume, "Statistics for complex variables and signals–Part I: variables," *Signal Process.*, vol. 53, pp. 1–13, 1996.
6. P. O. Amblard, M. Gaeta, and J. L. Lacoume, "Statistics for complex variables and signals–Part II: signals," *Signal Process.*, vol. 53, pp. 15–25, 1996.
7. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
8. B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Process.*, vol. 43, pp. 2030–2033, Aug. 1995.

Chapter 6
Multichannel Noise Reduction in the STFT Domain

In Chapters 3 and 4, we exploited the temporal (and spectral) information from a single microphone signal to derive different techniques for noise reduction in the time and STFT domains. In this chapter, we exploit the spatial information available from signals picked up by a determined number of microphones at different positions in the acoustics space in order to mitigate the noise effect. The processing is performed in the STFT domain.

6.1 Signal Model

We consider the conventional signal model in which a microphone array with M sensors captures a convolved source signal in some noise field. The received signals are expressed as [1], [2]

$$\begin{aligned} y_m(t) &= g_m(t) * s(t) + v_m(t) \\ &= x_m(t) + v_m(t),\ m = 1, 2, \ldots, M, \end{aligned} \tag{6.1}$$

where $g_m(t)$ is the acoustic impulse response from the unknown speech source, $s(t)$, location to the mth microphone, $*$ stands for linear convolution, and $v_m(t)$ is the additive noise at microphone m. We assume that the impulse responses are time invariant. We also assume that the signals $x_m(t) = g_m(t) * s(t)$ and $v_m(t)$ are uncorrelated, zero mean, real, and broadband. By definition, the convolved speech signals, $x_m(t),\ m = 1, 2, \ldots, M$, are coherent across the array. The noise terms, $v_m(t),\ m = 1, 2, \ldots, M$, are typically only partially coherent across the array.

In this work, our desired signal is designated by the clean (but convolved) speech signal received at microphone 1, namely $x_1(t)$. Obviously, any signal $x_m(t)$ could be taken as the reference. Our problem then may be stated as follows [3]: given M mixtures of two uncorrelated signals $x_m(t)$ and $v_m(t)$,

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_6

our aim is to preserve $x_1(t)$ while minimizing the contribution of the noise terms, $v_m(t),\ m = 1, 2, \ldots, M$, at the array output.

Expression (6.1) can be rewritten in the STFT domain as

$$\begin{aligned} Y_m(k,n) &= G_m(k)S(k,n) + V_m(k,n) \\ &= X_m(k,n) + V_m(k,n),\ m = 1, 2, \ldots, M, \end{aligned} \tag{6.2}$$

where $Y_m(k,n)$, $G_m(k)$, $S(k,n)$, $X_m(k,n) = G_m(k)S(k,n)$, and $V_m(k,n)$ are the STFT-domain representations of $y_m(t)$, $g_m(t)$, $s(t)$, $x_m(t)$, and $v_m(t)$, respectively. The zero-mean complex random variable $X_1(k,n)$ is our desired signal in the time-frequency domain.

It is more convenient to write the M STFT-domain microphone signals in a vector notation:

$$\begin{aligned} \overleftarrow{\mathbf{y}}(k,n) &= S(k,n)\overleftarrow{\mathbf{g}}(k) + \overleftarrow{\mathbf{v}}(k,n) \\ &= \overleftarrow{\mathbf{x}}(k,n) + \overleftarrow{\mathbf{v}}(k,n) \\ &= X_1(k,n)\overleftarrow{\mathbf{d}}(k) + \overleftarrow{\mathbf{v}}(k,n), \end{aligned} \tag{6.3}$$

where

$$\begin{aligned} \overleftarrow{\mathbf{y}}(k,n) &= \left[Y_1(k,n)\ Y_2(k,n) \cdots\ Y_M(k,n) \right]^T, \\ \overleftarrow{\mathbf{x}}(k,n) &= \left[X_1(k,n)\ X_2(k,n) \cdots\ X_M(k,n) \right]^T \\ &= S(k,n)\overleftarrow{\mathbf{g}}(k), \\ \overleftarrow{\mathbf{g}}(k) &= \left[G_1(k)\ G_2(k) \cdots\ G_M(k) \right]^T, \\ \overleftarrow{\mathbf{v}}(k,n) &= \left[V_1(k,n)\ V_2(k,n) \cdots\ V_M(k,n) \right]^T, \end{aligned}$$

and

$$\begin{aligned} \overleftarrow{\mathbf{d}}(k) &= \left[1\ \frac{G_2(k)}{G_1(k)} \cdots\ \frac{G_M(k)}{G_1(k)} \right]^T \\ &= \frac{\overleftarrow{\mathbf{g}}(k)}{G_1(k)}. \end{aligned} \tag{6.4}$$

Let us note that we assume that $G_1(k) \neq 0$. Expression (6.3) depends explicitly on the desired signal, $X_1(k,n)$; as a result, (6.3) is the STFT-domain signal model for noise reduction. The vector $\overleftarrow{\mathbf{d}}(k)$ is obviously the STFT-domain steering vector for noise reduction [3] since the acoustic impulse responses ratios from the broadband source to the aperture convey information about the position of the source.

There is another interesting way to write (6.3). First, it is easy to see that

$$X_m(k,n) = \rho^*_{X_1X_m}(k,n)X_1(k,n),\ m = 1, 2, \ldots, M, \tag{6.5}$$

where

$$\rho_{X_1X_m}(k,n) = \frac{E\left[X_1(k,n)X_m^*(k,n)\right]}{E\left[\left|X_1(k,n)\right|^2\right]} = \frac{G_m^*(k)}{G_1^*(k)},\ m = 1,2,\ldots,M \tag{6.6}$$

is the partially normalized [with respect to $X_1(k,n)$] correlation coefficient between $X_1(k,n)$ and $X_m(k,n)$. Using (6.5), we can rewrite (6.3) as

$$\begin{aligned}\overleftarrow{\mathbf{y}}(k,n) &= X_1(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) + \overleftarrow{\mathbf{v}}(k,n)\\ &= \overleftarrow{\mathbf{x}}(k,n) + \overleftarrow{\mathbf{v}}(k,n),\end{aligned} \tag{6.7}$$

where

$$\overleftarrow{\mathbf{x}}(k,n) = X_1(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) \tag{6.8}$$

is the speech signal vector and

$$\begin{aligned}\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) &= \left[1\ \rho_{X_1X_2}^*(k,n)\ \cdots\ \rho_{X_1X_M}^*(k,n)\right]^T\\ &= \frac{E\left[\overleftarrow{\mathbf{x}}(k,n)X_1^*(k,n)\right]}{E\left[\left|X_1(k,n)\right|^2\right]}\\ &= \overleftarrow{\mathbf{d}}(k)\end{aligned} \tag{6.9}$$

is the partially normalized [with respect to $X_1(k,n)$] correlation vector (of length M) between $\overleftarrow{\mathbf{x}}(k,n)$ and $X_1(k,n)$.

We see that $\overleftarrow{\mathbf{y}}(k,n)$ is the sum of two uncorrelated components. Therefore, the correlation matrix of $\overleftarrow{\mathbf{y}}(k,n)$ is

$$\begin{aligned}\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n) &= E\left[\overleftarrow{\mathbf{y}}(k,n)\overleftarrow{\mathbf{y}}^H(k,n)\right]\\ &= \phi_{X_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}^H(k,n) + \boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\\ &= \boldsymbol{\Phi}_{\overleftarrow{\mathbf{x}}}(k,n) + \boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n),\end{aligned} \tag{6.10}$$

where $\phi_{X_1}(k,n) = E\left[\left|X_1(k,n)\right|^2\right]$ is the variance of $X_1(k,n)$,

$$\boldsymbol{\Phi}_{\overleftarrow{\mathbf{x}}}(k,n) = \phi_{X_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}^H(k,n) \tag{6.11}$$

is the correlation matrix (whose rank is equal to 1) of $\overleftarrow{\mathbf{x}}(k,n)$, and

$$\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n) = E\left[\overleftarrow{\mathbf{v}}(k,n)\overleftarrow{\mathbf{v}}^H(k,n)\right] \tag{6.12}$$

is the correlation matrices of $\overleftarrow{\mathbf{v}}(k,n)$.

6.2 Linear Filtering

In the STFT domain, the conventional multichannel noise reduction is performed by applying a complex weight to the output of each sensor, at frequency bin k, and summing across the aperture [3], [4]:

$$\widehat{X}_1(k,n) = \overleftarrow{\mathbf{h}}^H(k,n)\overleftarrow{\mathbf{y}}(k,n), \tag{6.13}$$

where $\widehat{X}_1(k,n)$ is the estimate of $X_1(k,n)$ and $\overleftarrow{\mathbf{h}}(k,n)$ is a complex-valued filter of length M containing all the complex gains applied to the microphone outputs at frequency bin k.

We can express (6.13) as a function of the steering vector, i.e.,

$$\begin{aligned}\widehat{X}_1(k,n) &= \overleftarrow{\mathbf{h}}^H(k,n)\left[X_1(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) + \overleftarrow{\mathbf{v}}(k,n)\right] \\ &= X_{\mathrm{fd}}(k,n) + V_{\mathrm{rn}}(k,n),\end{aligned} \tag{6.14}$$

where

$$X_{\mathrm{fd}}(k,n) = X_1(k,n)\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) \tag{6.15}$$

is the filtered desired signal and

$$V_{\mathrm{rn}}(k,n) = \overleftarrow{\mathbf{h}}^H(k,n)\overleftarrow{\mathbf{v}}(k,n) \tag{6.16}$$

is the residual noise.

The two terms on the right-hand side of (6.14) are uncorrelated. Hence, the variance of $\widehat{X}_1(k,n)$ is also the sum of two variances:

$$\begin{aligned}\phi_{\widehat{X}_1}(k,n) &= \overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)\overleftarrow{\mathbf{h}}(k,n) \\ &= \phi_{X_{\mathrm{fd}}}(k,n) + \phi_{V_{\mathrm{rn}}}(k,n),\end{aligned} \tag{6.17}$$

where

$$\phi_{X_{\mathrm{fd}}}(k,n) = \phi_{X_1}(k,n)\left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2, \tag{6.18}$$

$$\phi_{V_{\mathrm{rn}}}(k,n) = \overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n). \tag{6.19}$$

The different variances in (6.17) are important in the definitions of the performance measures.

6.3 Performance Measures

Since microphone 1 is our reference, all measures are defined with respect to the signal from this microphone.

The subband and fullband input SNRs at time frame n are defined as

$$\text{iSNR}(k,n) = \frac{\phi_{X_1}(k,n)}{\phi_{V_1}(k,n)},\ k = 0, 1, \ldots, K-1, \tag{6.20}$$

$$\text{iSNR}(n) = \frac{\sum_{k=0}^{K-1} \phi_{X_1}(k,n)}{\sum_{k=0}^{K-1} \phi_{V_1}(k,n)}, \tag{6.21}$$

where $\phi_{V_1}(k,n) = E\left[|V_1(k,n)|^2\right]$ is the variance of $V_1(k,n)$. It is easy to show that

$$\text{iSNR}(n) \leq \max_k \text{ iSNR}(k,n). \tag{6.22}$$

In words, the fullband input SNR can never be greater than the maximum subband input SNR.

The subband output SNR is obtained from (6.17):

$$\begin{aligned} \text{oSNR}\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= \frac{\phi_{X_{\text{fd}}}(k,n)}{\phi_{V_{\text{rn}}}(k,n)} \\ &= \frac{\phi_{X_1}(k,n)\left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2}{\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)},\ k = 0, 1, \ldots, K-1. \end{aligned} \tag{6.23}$$

For the particular filter $\overleftarrow{\mathbf{h}}(k,n) = \overleftarrow{\mathbf{i}}_{\text{id}}$, where $\overleftarrow{\mathbf{i}}_{\text{id}}$ is the first column of the identity matrix, $\mathbf{I}_M$ (of size $M \times M$), we have

$$\text{oSNR}\left[\overleftarrow{\mathbf{i}}_{\text{id}}(k,n)\right] = \text{iSNR}(k,n). \tag{6.24}$$

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

For any two vectors $\overleftarrow{\mathbf{h}}(k,n)$ and $\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)$ and a positive definite matrix $\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)$, we have

$$\begin{aligned} \left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2 \leq &\left[\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)\right] \times \\ &\left[\boldsymbol{\rho}^H_{\overleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right], \end{aligned} \tag{6.25}$$

with equality if and only if $\overleftarrow{\mathbf{h}}(k,n) \propto \boldsymbol{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)$. Using the previous inequality in (6.23), we deduce an upper bound for the subband output SNR:

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}(k,n)\right] \leq \phi_{X_1}(k,n) \times \boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n),\ \forall \overleftarrow{\mathbf{h}}(k,n) \tag{6.26}$$

and, clearly,

$$\text{oSNR}\left[\overleftarrow{\mathbf{i}}_{\text{id}}(k,n)\right] \leq \phi_{X_1}(k,n)\boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n), \tag{6.27}$$

which implies that

$$\boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n) \geq \frac{1}{\phi_{V_1}(k,n)}. \tag{6.28}$$

The role of the filter is to produce a signal whose subband SNR is higher than that of the subband input SNR. This is measured by the subband array gain:

$$\mathcal{G}\left[\overleftarrow{\mathbf{h}}(k,n)\right] = \frac{\text{oSNR}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\text{iSNR}(k,n)},\ k = 0,1,\ldots,K-1. \tag{6.29}$$

From (6.26), we deduce that the maximum subband array gain is

$$\mathcal{G}_{\max}(k,n) = \phi_{V_1}(k,n)\boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n) \geq 1. \tag{6.30}$$

The maximum SNR filter, $\overleftarrow{\mathbf{h}}_{\max}(k,n)$, is obtained by maximizing the subband output SNR as given above. In (6.23), we recognize the generalized Rayleigh quotient. It is well known that this quotient is maximized with the maximum eigenvector of the matrix $\phi_{X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)$. Let us denote by $\lambda_{\max}(k,n)$ the maximum eigenvalue corresponding to this maximum eigenvector. Since the rank of the mentioned matrix is equal to 1, we have

$$\begin{aligned}\lambda_{\max}(k,n) &= \text{tr}\left[\phi_{X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)\right] \\ &= \phi_{X_1}(k,n)\boldsymbol{\rho}^H_{\underleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\Phi}^{-1}_{\underleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\underleftarrow{\mathbf{x}}X_1}(k,n).\end{aligned} \tag{6.31}$$

As a result,

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}_{\max}(k,n)\right] = \lambda_{\max}(k,n), \tag{6.32}$$

which corresponds to the maximum possible subband output SNR and

$$\mathcal{G}\left[\overleftarrow{\mathbf{h}}_{\max}(k,n)\right] = \mathcal{G}_{\max}(k,n). \tag{6.33}$$

Let us denote by $\mathcal{G}^{(m)}_{\max}(k,n)$ the maximum subband array gain of a microphone array with m sensors. By virtue of the inclusion principle [5] for the

matrix $\phi_{X_1}(k,n)\mathbf{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}^H_{\overleftarrow{\mathbf{x}}X_1}(k,n)$, we have

$$\mathcal{G}^{(M)}_{\max}(k,n) \geq \mathcal{G}^{(M-1)}_{\max}(k,n) \geq \cdots \geq \mathcal{G}^{(2)}_{\max}(k,n) \geq \mathcal{G}^{(1)}_{\max}(k,n) = 1. \quad (6.34)$$

This shows that by increasing the number of microphones, we necessarily increase the gain. If there is one microphone only, the subband SNR cannot be improved as expected [1] (if the interframe correlation is not taken into account, which is the case here).

Obviously, we also have

$$\overleftarrow{\mathbf{h}}_{\max}(k,n) = \varsigma(k,n)\mathbf{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n), \quad (6.35)$$

where $\varsigma(k,n)$ is an arbitrary scaling factor different from zero. While this factor has no effect on the subband output SNR, it has on the fullband output SNR.

We define the fullband output SNR at time frame n as

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)\left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2}{\sum_{k=0}^{K-1}\overleftarrow{\mathbf{h}}^H(k,n)\mathbf{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)} \quad (6.36)$$

and it can be verified that

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}(:,n)\right] \leq \max_k \text{ oSNR}\left[\overleftarrow{\mathbf{h}}(k,n)\right], \ \forall\overleftarrow{\mathbf{h}}(k,n). \quad (6.37)$$

The previous inequality tells us that the fullband output SNR can never exceed the maximum subband output SNR for any filter $\overleftarrow{\mathbf{h}}(k,n)$.

The fullband output SNR with the maximum SNR filter is

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}_{\max}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\frac{|\varsigma(k,n)|^2\lambda^2_{\max}(k,n)}{\phi_{X_1}(k,n)}}{\sum_{k=0}^{K-1}\frac{|\varsigma(k,n)|^2\lambda_{\max}(k,n)}{\phi_{X_1}(k,n)}}. \quad (6.38)$$

We see that the performance (in terms of fullband SNR improvement) of the maximum SNR filter is quite dependent on the values of $\varsigma(k,n)$.

We also define the fullband array gain as

$$\mathcal{G}\left[\overleftarrow{\mathbf{h}}(:,n)\right] = \frac{\text{oSNR}\left[\overleftarrow{\mathbf{h}}(:,n)\right]}{\text{iSNR}(n)}. \quad (6.39)$$

We define the subband and fullband partial speech intelligibility indices as

$$\begin{aligned}\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= \frac{\phi_{X_1}(k,n)-\phi_{X_{\mathrm{fd}}}(k,n)}{\phi_{X_1}(k,n)} \\ &= 1-\left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2,\ k=0,1,\ldots,K-1\end{aligned} \tag{6.40}$$

and

$$\begin{aligned}\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(:,n)\right] &= \frac{\sum_{k=0}^{K-1}\left[\phi_{X_1}(k,n)-\phi_{X_{\mathrm{fd}}}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)\left[1-\left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2\right]}{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)}.\end{aligned} \tag{6.41}$$

The higher is the value of the partial speech intelligibility index, the less intelligible is the estimated desired signal.

We define the subband and fullband speech quality indices as

$$\begin{aligned}\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= \frac{\phi_{V_{\mathrm{rn}}}(k,n)}{\phi_{V_1}(k,n)} \\ &= \frac{\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)}{\phi_{V_1}(k,n)},\ k=0,1,\ldots,K-1\end{aligned} \tag{6.42}$$

and

$$\begin{aligned}\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(:,n)\right] &= \frac{\sum_{k=0}^{K-1}\phi_{V_{\mathrm{rn}}}(k,n)}{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)} \\ &= \frac{\sum_{k=0}^{K-1}\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)}{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)}.\end{aligned} \tag{6.43}$$

The quality of the estimated desired signal decreases as the value of the speech quality index increases.

We deduce from the previous definitions that the subband and fullband global speech intelligibility indices are

$$\upsilon_{\mathrm{i}}'\left[\overleftarrow{\mathbf{h}}(k,n)\right] = (1-\varpi)\,\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] + \varpi\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right],\ k=0,1,\ldots,K-1 \tag{6.44}$$

and

$$\upsilon_{\mathrm{i}}'\left[\overleftarrow{\mathbf{h}}(:,n)\right] = (1-\varpi)\,\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(:,n)\right] + \varpi\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(:,n)\right]. \tag{6.45}$$

The variance of the estimated desired signal can be rewritten as a function of the subband speech intelligibility and quality indices, i.e.,

$$\phi_{\widehat{X}_1}(k,n) = \left\{1-\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]\right\}\phi_{X_1}(k,n) + \upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]\phi_{V_1}(k,n), \tag{6.46}$$

which is interesting to compare to the variance of the observation signal at the reference microphone, i.e.,

$$\phi_{Y_1}(k,n) = \phi_{X_1}(k,n) + \phi_{V_1}(k,n). \tag{6.47}$$

We see how any optimal filter will try to compromise between speech intelligibility and speech quality.

We easily derive the fundamental relations:

$$\frac{\mathrm{oSNR}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\mathrm{iSNR}(k,n)} = \frac{1-\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]},\ k=0,1,\ldots,K-1, \tag{6.48}$$

$$\frac{\mathrm{oSNR}\left[\overleftarrow{\mathbf{h}}(:,n)\right]}{\mathrm{iSNR}(n)} = \frac{1-\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(:,n)\right]}{\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(:,n)\right]}. \tag{6.49}$$

6.4 MSE-Based Criterion

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

$$\begin{aligned}\mathcal{E}(k,n) &= \widehat{X}_1(k,n) - X_1(k,n) \\ &= \overleftarrow{\mathbf{h}}^H(k,n)\overleftarrow{\mathbf{y}}(k,n) - X_1(k,n),\end{aligned} \tag{6.50}$$

which can be rewritten as the sum of two other uncorrelated error signals, i.e.,

$$\mathcal{E}(k,n) = \mathcal{E}_{\mathrm{i}}(k,n) + \mathcal{E}_{\mathrm{q}}(k,n), \tag{6.51}$$

where

$$\begin{aligned}\mathcal{E}_{\mathrm{i}}(k,n) &= X_{\mathrm{fd}}(k,n) - X_1(k,n) \\ &= \left[\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) - 1\right]X_1(k,n)\end{aligned} \tag{6.52}$$

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

$$\begin{aligned}\mathcal{E}_{\mathrm{q}}(k,n) &= V_{\mathrm{rn}}(k,n) \\ &= \overleftarrow{\mathbf{h}}^{H}(k,n)\overleftarrow{\mathbf{v}}(k,n)\end{aligned} \tag{6.53}$$

represents the residual noise, which affects the quality as well as the other portion of intelligibility.

From the error signal defined in (6.50), we can now form the subband MSE criterion:

$$\begin{aligned}J\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= E\left[\left|\mathcal{E}(k,n)\right|^2\right] \\ &= J_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] + J_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right],\end{aligned} \tag{6.54}$$

where

$$\begin{aligned}J_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= E\left[\left|\mathcal{E}_{\mathrm{i}}(k,n)\right|^2\right] \\ &= E\left[\left|X_{\mathrm{fd}}(k,n) - X_1(k,n)\right|^2\right] \\ &= \phi_{X_1}(k,n)\left|\overleftarrow{\mathbf{h}}^{H}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) - 1\right|^2\end{aligned} \tag{6.55}$$

and

$$\begin{aligned}J_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= E\left[\left|\mathcal{E}_{\mathrm{q}}(k,n)\right|^2\right] \\ &= E\left[\left|V_{\mathrm{rn}}(k,n)\right|^2\right] \\ &= \phi_{V_{\mathrm{rn}}}(k,n) \\ &= \phi_{V_1}(k,n)\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right].\end{aligned} \tag{6.56}$$

For the two particular filters $\overleftarrow{\mathbf{h}}(k,n) = \overleftarrow{\mathbf{i}}_{\mathrm{id}}$ and $\overleftarrow{\mathbf{h}}(k,n) = \mathbf{0}_{M\times 1}$, we get

$$J\left[\overleftarrow{\mathbf{i}}_{\mathrm{id}}(k,n)\right] = J_{\mathrm{q}}\left[\overleftarrow{\mathbf{i}}_{\mathrm{id}}(k,n)\right] = \phi_{V_1}(k,n), \tag{6.57}$$

$$J\left[\mathbf{0}_{M\times 1}(k,n)\right] = J_{\mathrm{i}}\left[\mathbf{0}_{M\times 1}(k,n)\right] = \phi_{X_1}(k,n). \tag{6.58}$$

We then find that the subband NMSE with respect to $J\left[\overleftarrow{\mathbf{i}}_{\mathrm{id}}(k,n)\right]$ is

$$\begin{aligned}J_{\mathrm{n},1}\left[\overleftarrow{\mathbf{h}}(k,n)\right] &= \frac{J\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{J\left[\overleftarrow{\mathbf{i}}_{\mathrm{id}}(k,n)\right]} \\ &= \mathrm{iSNR}(k,n) \times \left|1 - \overleftarrow{\mathbf{h}}^{H}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2 + \upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]\end{aligned} \tag{6.59}$$

and the subband NMSE with respect to $J\left[\mathbf{0}_{M\times 1}(k,n)\right]$ is

$$J_{\mathrm{n},2}\left[\overleftarrow{\mathbf{h}}(k,n)\right] = \frac{J\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{J\left[\mathbf{0}_{M\times 1}(k,n)\right]} \tag{6.60}$$
$$= \left|1-\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2 + \frac{1-\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\mathrm{oSNR}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}.$$

We have

$$J_{\mathrm{n},1}\left[\overleftarrow{\mathbf{h}}(k,n)\right] = \mathrm{iSNR}(k,n)\times J_{\mathrm{n},2}\left[\overleftarrow{\mathbf{h}}(k,n)\right]. \tag{6.61}$$

Expressions (6.59) and (6.60) show how the subband NMSEs and the different subband MSEs are related to the subband performance measures.

We are interested in complex filters for which a reasonable compromise can be made between speech quality and speech intelligibility, i.e.,

$$J_{\mathrm{i}}\left[\overleftarrow{\mathbf{i}}_{\mathrm{id}}(k,n)\right] \leq J_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] < J_{\mathrm{i}}\left[\mathbf{0}_{M\times 1}(k,n)\right], \tag{6.62}$$
$$J_{\mathrm{q}}\left[\mathbf{0}_{M\times 1}(k,n)\right] < J_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] < J_{\mathrm{q}}\left[\overleftarrow{\mathbf{i}}_{\mathrm{id}}(k,n)\right]. \tag{6.63}$$

From the two previous expressions, we deduce that

$$0 \leq \left|1-\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2 < 1, \tag{6.64}$$
$$0 < \upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right] < 1. \tag{6.65}$$

In order to derive the filters we are looking for, we propose to use the general subband MSE-based criterion:

$$J_{\mu}\left[\overleftarrow{\mathbf{h}}(k,n)\right] = \mu(k,n)\frac{J_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\phi_{X_1}(k,n)} + \frac{J_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right]}{\phi_{V_1}(k,n)} \tag{6.66}$$
$$= \mu(k,n)\left|1-\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2 + \upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}(k,n)\right],$$

where $\mu(k,n)$ is a positive real number. This parameter allows us to design a large class of flexible filters that can compromise between speech intelligibility and speech quality.

6.5 Optimal Filters

The minimization of (6.66) with respect to $\overleftarrow{\mathbf{h}}(k,n)$ leads to the optimal filter:

$$\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(k,n) = \tag{6.67}$$
$$\mu(k,n)\left[\mu(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}^H_{\overleftarrow{\mathbf{x}}X_1}(k,n) + \frac{\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)}{\phi_{V_1}(k,n)}\right]^{-1}\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n).$$

From the decomposition of $\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)$ given in (6.10), we can express the optimal filter as

$$\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(k,n) = \mu(k,n)\times \tag{6.68}$$
$$\left\{[\mu(k,n) - \mathrm{iSNR}(k,n)]\,\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\boldsymbol{\rho}^H_{\overleftarrow{\mathbf{x}}X_1}(k,n) + \frac{\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)}{\phi_{V_1}(k,n)}\right\}^{-1}\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)$$

and the vector $\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)$ can be expressed as a function of the statistics of $\overleftarrow{\mathbf{y}}(k,n)$ and $\overleftarrow{\mathbf{v}}(k,n)$, i.e.,

$$\begin{aligned}\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) &= \frac{E\left[\overleftarrow{\mathbf{y}}(k,n)Y_1^*(k,n)\right] - E\left[\overleftarrow{\mathbf{v}}(k,n)V_1^*(k,n)\right]}{\phi_{Y_1}(k,n) - \phi_{V_1}(k,n)} \\ &= \frac{\phi_{Y_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{y}}Y_1}(k,n) - \phi_{V_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{v}}V_1}(k,n)}{\phi_{Y_1}(k,n) - \phi_{V_1}(k,n)},\end{aligned} \tag{6.69}$$

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

Now, by using the Woodbury's identity in (6.67), it can easily be shown that the optimal filter can be reformulated as

$$\begin{aligned}\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(k,n) &= \frac{\mu(k,n)\dfrac{\phi_{X_1}(k,n)}{\mathrm{iSNR}(k,n)}}{1+\mu(k,n)\dfrac{\lambda_{\max}(k,n)}{\mathrm{iSNR}(k,n)}}\boldsymbol{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) \\ &= \frac{\mu(k,n)\phi_{V_1}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\boldsymbol{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n) \\ &= \frac{\mu(k,n)\mathrm{iSNR}^{-1}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\boldsymbol{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{x}}}(k,n)\overleftarrow{\mathbf{i}}_{\mathrm{id}} \\ &= \frac{\mu(k,n)\mathrm{iSNR}^{-1}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\left[\boldsymbol{\Phi}^{-1}_{\overleftarrow{\mathbf{v}}}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n) - \mathbf{I}_M\right]\overleftarrow{\mathbf{i}}_{\mathrm{id}}.\end{aligned} \tag{6.70}$$

Comparing $\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(k,n)$ with $\overleftarrow{\mathbf{h}}_{\max}(k,n)$ [eq. (6.35)], we observe that the two filters are equivalent up to a scaling factor. As a result, $\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(k,n)$ also maximizes the subband output SNR, i.e.,

$$\mathrm{oSNR}\left[\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(k,n)\right] = \lambda_{\max}(k,n),\ \forall \mu(k,n) > 0 \tag{6.71}$$

and

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(k,n)\right] \geq \text{iSNR}(k,n),\ \forall \mu(k,n) \geq 0. \tag{6.72}$$

From (6.70), we deduce that the subband partial speech intelligibility index and the subband speech quality index are, respectively,

$$\begin{aligned} \upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(k,n)\right] &= 1 - \left[\frac{\mu(k,n)\mathcal{G}_{\max}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\right]^2 \\ &= 1 - \left|\overleftarrow{\mathbf{h}}_{\text{o},\mu}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2 \end{aligned} \tag{6.73}$$

and

$$\begin{aligned} \upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(k,n)\right] &= \frac{\mu^2(k,n)\mathcal{G}_{\max}(k,n)}{\left[1+\mu\mathcal{G}_{\max}(k,n)\right]^2} \\ &= \frac{\left|\overleftarrow{\mathbf{h}}_{\text{o},\mu}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2}{\mathcal{G}_{\max}(k,n)}. \end{aligned} \tag{6.74}$$

Clearly, $\forall \mu(k,n) \geq 0$, we have

$$0 \leq \upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(k,n)\right] \leq 1, \tag{6.75}$$

$$0 \leq \upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(k,n)\right] \leq 1. \tag{6.76}$$

We deduce that the fullband indices are

$$\upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] = 1 - \frac{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)\left|\overleftarrow{\mathbf{h}}_{\text{o},\mu}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2}{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)}, \tag{6.77}$$

$$\upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)\frac{\left|\overleftarrow{\mathbf{h}}_{\text{o},\mu}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2}{\mathcal{G}_{\max}(k,n)}}{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)}, \tag{6.78}$$

and, $\forall \mu(k,n) \geq 0$, we also have

$$0 \leq \upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] \leq 1, \tag{6.79}$$

$$0 \leq \upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] \leq 1. \tag{6.80}$$

It is easy to check that the fullband output SNR is

$$\text{oSNR}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_{X_1}(k,n)\left[\frac{\mu(k,n)\mathcal{G}_{\max}(k,n)}{1+\mu(k,n)\mathcal{G}_{\max}(k,n)}\right]^2}{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)\frac{\mu^2(k,n)\mathcal{G}_{\max}(k,n)}{[1+\mu\mathcal{G}_{\max}(k,n)]^2}}. \quad (6.81)$$

Taking $\mu(k,n)=\infty$ in (6.70), we find the MVDR filter [1]:

$$\begin{aligned}\overleftarrow{\mathbf{h}}_{\text{MVDR}}(k,n) &= \frac{\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)}{\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)} \\ &= \frac{\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)-\mathbf{I}_M}{\text{tr}\left[\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)\right]-M}\overleftarrow{\mathbf{i}}_{\text{id}}.\end{aligned} \quad (6.82)$$

We deduce that

$$\upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(k,n)\right] = 0, \quad (6.83)$$

$$\upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(:,n)\right] = 0, \quad (6.84)$$

$$\upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(k,n)\right] = \mathcal{G}_{\max}^{-1}(k,n), \quad (6.85)$$

$$\upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)\mathcal{G}_{\max}^{-1}(k,n)}{\sum_{k=0}^{K-1}\phi_{V_1}(k,n)}. \quad (6.86)$$

Taking $\mu(k,n)=\text{iSNR}(k,n)$ in (6.70), we find the Wiener filter [1]:

$$\begin{aligned}\overleftarrow{\mathbf{h}}_{\text{W}}(k,n) &= \frac{\phi_{X_1}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)}{1+\phi_{X_1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)} \\ &= \frac{\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)-\mathbf{I}_M}{1+\text{tr}\left[\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}^{-1}(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)\right]-M}\overleftarrow{\mathbf{i}}_{\text{id}}.\end{aligned} \quad (6.87)$$

It can be verified that

$$\upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{W}}(:,n)\right] > \upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(:,n)\right], \quad (6.88)$$

$$\upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{W}}(:,n)\right] < \upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(:,n)\right]. \quad (6.89)$$

Therefore, we can expect a better signal quality with Wiener than MVDR and a more intelligible signal with MVDR than Wiener.

It can also be verified that for $\mu(k,n)\geq\text{iSNR}(k,n)$, we have

$$\upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{W}}(:,n)\right] \geq \upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] \geq \upsilon_{\text{i}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(:,n)\right], \quad (6.90)$$

$$\upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{W}}(:,n)\right] \leq \upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{o},\mu}(:,n)\right] \leq \upsilon_{\text{q}}\left[\overleftarrow{\mathbf{h}}_{\text{MVDR}}(:,n)\right], \quad (6.91)$$

and for $\mu(k,n) \leq \mathrm{iSNR}(k,n)$, we have

$$\upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(:,n)\right] \geq \upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}_{\mathrm{W}}(:,n)\right] > \upsilon_{\mathrm{i}}\left[\overleftarrow{\mathbf{h}}_{\mathrm{MVDR}}(:,n)\right], \tag{6.92}$$

$$\upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}_{\mathrm{o},\mu}(:,n)\right] \leq \upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}_{\mathrm{W}}(:,n)\right] < \upsilon_{\mathrm{q}}\left[\overleftarrow{\mathbf{h}}_{\mathrm{MVDR}}(:,n)\right]. \tag{6.93}$$

6.6 Simulations

In this section, we illustrate the performance of the optimal multichannel noise reduction filters deduced above. The simulation setup is similar to the one in Section 5.6. An equispaced linear array with six omnidirectional microphones is configured and placed in a reverberant room of size 6 m long and 5 m wide. For ease of exposition, positions in the room are designated by (x, y) coordinates with reference to one corner of the room, $0 \leq x \leq 6$ and $0 \leq y \leq 5$. The positions of the six microphones are, respectively, at (3.4, 0.5), (3.5, 0.5), (3.6, 0.5), (3.7, 0.5), (3.8, 0.5), and (3.9, 0.5). A loudspeaker is placed at (1.3, 3.0) to simulate a speech source. To make the experiments repeatable, the acoustic channel impulse responses from the source position to all the six microphones were measured. During experiments, the microphones' outputs are generated by convolving the source signal with the corresponding measured impulse responses and noise is then added to the convolved results to control the input SNR level. The source signal is taken from the speaker FAKS0 in the TIMIT database. We continue to focus on the narrowband case, so the original signals from the TIMIT database are downsampled from 16 kHz to 8 kHz before convolving with the measured impulse responses. We consider two types of noise: white Gaussian (in this case the noise signals at different sensors are uncorrelated) and a point source (in this case the noise signals at different sensors are coherent). To simulate the noise from a point source, a loudspeaker is placed at (5.3, 3.0) to play back the pre-recorded NYSE noise (see Section 3.6). Again, to make the simulations repeatable, the acoustic impulse responses from this loudspeaker to all the six microphones are measured. The point-source noise at each sensor is generated by convolving the NYSE noise with the corresponding measured acoustic impulse response. This convolution result is scaled according to the input SNR level and then added into the multichannel speech signals.

To implement the STFT-domain noise reduction filters, the noisy speech signals received at the array are partitioned into overlapping frames with a frame size of 8 ms and an overlapping factor of 75%. A Kaiser window is then applied to each frame (to reduce the aliasing effect due to circular convolution) and the windowed signals of all the channels are subsequently transformed into the STFT domain using a 64-point fast Fourier transform (FFT). A multichannel noise reduction filter is then constructed and applied to the multichannel noisy STFT coefficients in every subband to obtain an

estimate of the clean speech at the first microphone. After noise reduction, the inverse FFT (IFFT) with the overlap-add method is used for signal reconstruction in the time domain. A same Kaiser window is applied to the output of the IFFT before the overlap-add process, again, to reduce the aliasing effect caused by circular convolution.

With the above implementation process, the most critical step is the computation of the multichannel noise reduction filters in the STFT subbands. It is seen from (6.67) or (6.68) that we need to know the statistics $\mathbf{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)$, $\mathbf{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)$, and $\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)$ in order to compute the optimal noise reduction filters. In our simulation, these statistics are estimated as follows. We first estimate the $\mathbf{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)$ and $\mathbf{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)$ matrices using the following recursions:

$$\widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{v}}}(k,n) = \alpha_v \widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{v}}}(k,n-1) + (1-\alpha_v)\,\overleftarrow{\mathbf{v}}(k,n)\overleftarrow{\mathbf{v}}^H(k,n), \tag{6.94}$$

$$\widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{y}}}(k,n) = \alpha_y \widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{y}}}(k,n-1) + (1-\alpha_y)\,\overleftarrow{\mathbf{y}}(k,n)\overleftarrow{\mathbf{y}}^H(k,n), \tag{6.95}$$

where $\alpha_v \in (0,1)$ and $\alpha_y \in (0,1)$ are the forgetting factors that control the influence of the previous data samples on the current correlation matrix estimate (the initial estimates of these two matrices are obtained from the first 50 signal frames with a short-time average).

After the estimates of the $\mathbf{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)$ and $\mathbf{\Phi}_{\overleftarrow{\mathbf{y}}}(k,n)$ matrices are available at time frame n, the estimate of the $\mathbf{\Phi}_{\overleftarrow{\mathbf{x}}}(k,n)$ matrix is computed as $\widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{x}}}(k,n) = \widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{y}}}(k,n) - \widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{y}}}(k,n)$. To ensure that this $\widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{x}}}(k,n)$ matrix is positive semi-definite, all the negative eigenvalues of this matrix are forced to be 0. And then, the estimate of the correlation vector $\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)$ is taken as the first column of $\widehat{\mathbf{\Phi}}_{\overleftarrow{\mathbf{x}}}(k,n)$ normalized by its first element.

With the previous way of statistics estimation, the noise reduction performance of the STFT-domain multichannel optimal filters depends on the forgetting factors, α_y and α_v, and the number of microphones, i.e., M. As how the two forgetting factors affect the noise reduction performance and how the optimal values can be found, the reader can follow the study in Section 4.7. In the following simulation, we evaluate the dependency of the noise reduction performance on the number of microphones. For that purpose, We can examine either the subband or the fullband performance measures defined in Section 6.3. For ease of visualization, we will use, in the following simulations, the long-term average of the three fullband performance measures defined in Section 6.3, i.e., the long-term average partial speech intelligibility index, the long-term average speech quality index, and the long-term output SNR to evaluate performance. They are computed, respectively, as

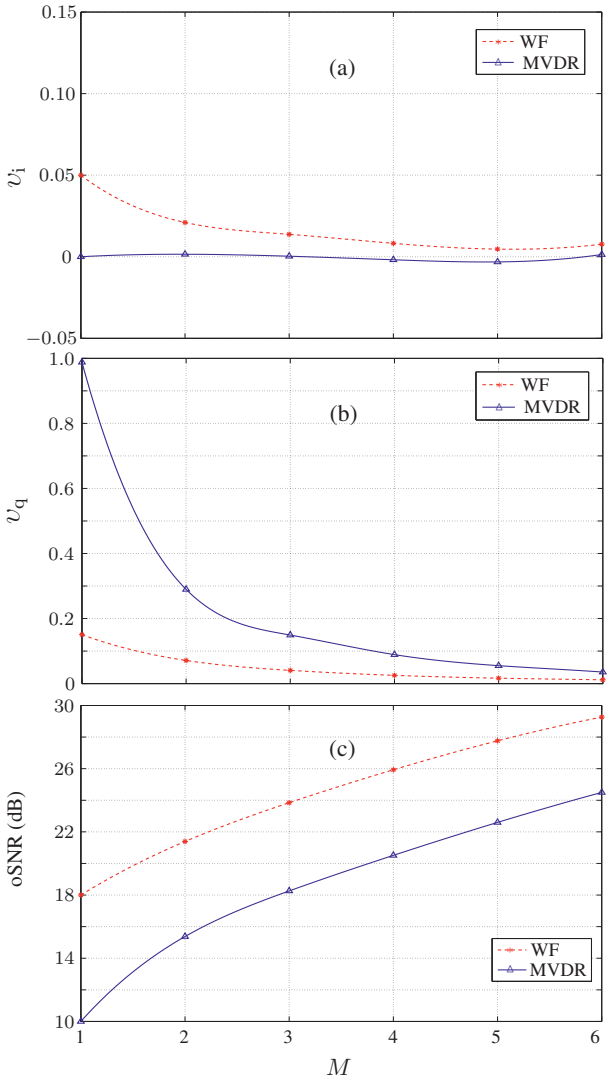

Fig. 6.1 Performance of the multichannel Wiener and MVDR filters as a function of the number of microphones, M, in the white Gaussian noise: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The window size is $K = 64$ (8 ms) with a 75% overlap, the fullband input SNR is 10 dB, and the forgetting factors are $\alpha_y = \alpha_v = 0.8$.

$$v_{\mathrm{i}}\left(\overleftarrow{\mathbf{h}}\right) = \frac{\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left[\phi_{X_1}(k,n) - \phi_{X_{\mathrm{fd}}}(k,n)\right]}{\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\phi_{X_1}(k,n)}, \tag{6.96}$$

$$v_{\mathrm{q}}\left(\overleftarrow{\mathbf{h}}\right) = \frac{\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)}{\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\phi_{V_1}(k,n)}, \tag{6.97}$$

$$\mathrm{oSNR}\left(\overleftarrow{\mathbf{h}}\right) = \frac{\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\phi_{X_1}(k,n)\left|\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\rho}_{\overleftarrow{\mathbf{x}}X_1}(k,n)\right|^2}{\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\overleftarrow{\mathbf{h}}^H(k,n)\boldsymbol{\Phi}_{\overleftarrow{\mathbf{v}}}(k,n)\overleftarrow{\mathbf{h}}(k,n)}, \tag{6.98}$$

where N is the total number of frames. Note that the above long-term average fullband performance measures are similar to those time-domain measures defined in Section 3.3. The only difference is that the output of the multichannel noise reduction filter in this chapter does not have the residual interference component.

In the first simulation, we examine the performance of both the multichannel Wiener and MVDR filters in the white Gaussian noise. In this case, the noise signals at different microphones are white Gaussian and uncorrelated. The input SNR is 10 dB. Following the study in Section 4.7, we set the two forgetting factors as $\alpha_y = \alpha_v = 0.8$. The partial speech intelligibility index, the speech quality index, and the output SNR [as defined, respectively, in (6.96), (6.97), and (6.98)] of the Wiener and MVDR filters as a function of the number of microphones, M, are plotted in Fig. 6.1. As seen, the partial speech intelligibility index and speech quality index of the Wiener filter decreases with M while its output SNR increases with M. So, the more the microphones, the better is the performance of the multichannel Wiener filter in terms of both the speech intelligibility and quality as well as the output SNR.

When there is only one microphone, the MVDR filter degenerates to the unit gain, which is clearly seen in Fig. 6.1. As the number of microphone increases, we observe that the speech quality index decreases and the output SNR increases, while the partial speech intelligibility index is approximately 0. So, same as the Wiener filter, the MVDR filter also yields better performance as the number of microphone increases. In comparison, one can see that the values of the partial speech intelligibility index and the output SNR of the Wiener filter are higher than those of the MVDR filter, while the value of the speech quality index of the Wiener filter is smaller than that of the MVDR filter. This corroborates with the theoretical analysis in Section 6.5.

In the second simulation, we consider the point-source noise case. As explained before, the noise at each sensor is generated by convolving the prerecorded NYSE noise with the acoustic impulse response measured from position (5.3, 3.0) to the sensor's position. The input SNR at each sensor is 10 dB. Again, we set $\alpha_y = \alpha_v = 0.8$. The results of this simulation are plotted in Fig. 6.2.

With one microphone, the MVDR filter is the unit gain and, therefore, does not have any noise reduction, which is seen in Fig. 6.2. But the Wiener filter can achieve more than 6-dB SNR improvement. With two microphones, one can see that both the Winer and MVDR filters achieve significant noise reduction and the SNR improvement is more than 20 dB for both filters. However, the performance of the Wiener and MVDR filters do not change much with the number of microphones if the number is greater than two. The underlying reason can be explained as follows. With two microphones, both the Wiener and MVDR filters can generate a null pointing to the noise source, leading to significant noise reduction. With point-source noise and perfect knowledge of the noise statistics, more (than two) microphones do not

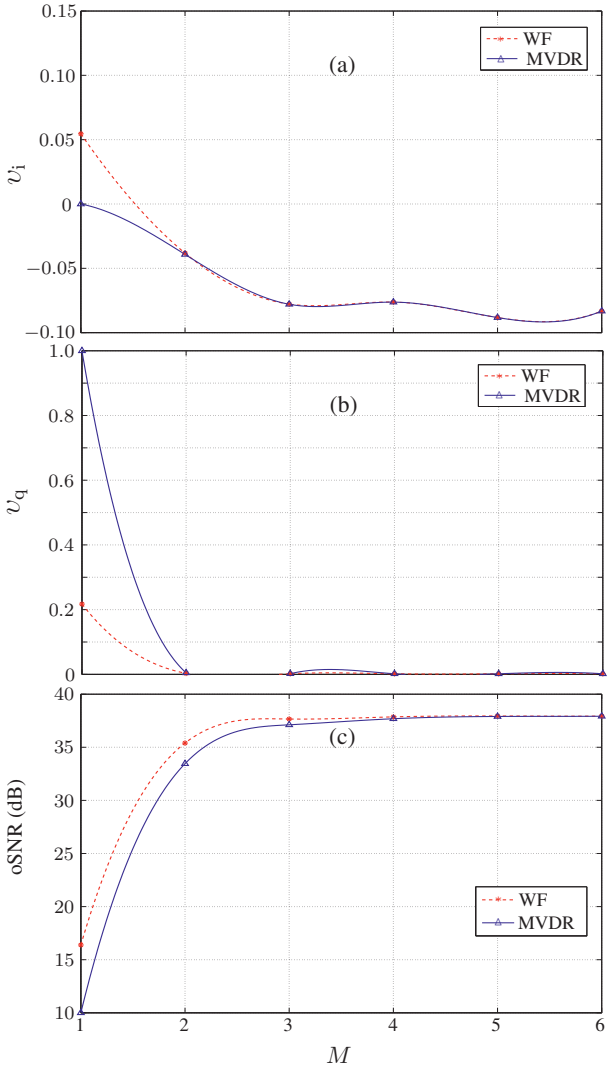

Fig. 6.2 Performance of the multichannel Wiener and MVDR filters as a function of the number of microphones, M, in the point-source noise: (a) partial speech intelligibility index, (b) speech quality index, and (c) output SNR. The window size is $K = 64$ (8 ms) with a 75% overlap, the fullband input SNR is 10 dB, and the forgetting factors are $\alpha_y = \alpha_v = 0.8$.

seem help further improve performance. However, in practice, both additive noise and point-source noise may co-exist and there may be estimation errors in the noise statistics; in this case, increasing the number of microphones can help improve performance.

Theoretically, the value of the partial speech intelligibility index should be in the range between 0 and 1. However, we observe from Fig. 6.2(a) that the value of this index is smaller than 0 if two or more microphones are used, which is different from the theoretical analysis. The reason may be due to the estimation error of the the correlation vector $\boldsymbol{\rho}_{\underline{\mathbf{x}}X_1}(k,n)$. From its definition, the first element of the $\boldsymbol{\rho}_{\underline{\mathbf{x}}X_1}(k,n)$ vector is 1, and the magnitude of any other element is less than 1. In our simulation, the estimate of the $\boldsymbol{\Phi}_{\underline{\mathbf{x}}}(k,n)$ matrix is computed as $\widehat{\boldsymbol{\Phi}}_{\underline{\mathbf{x}}}(k,n) = \widehat{\boldsymbol{\Phi}}_{\underline{\mathbf{y}}}(k,n) - \widehat{\boldsymbol{\Phi}}_{\underline{\mathbf{y}}}(k,n)$ and the estimate of the correlation vector $\boldsymbol{\rho}_{\underline{\mathbf{x}}X_1}(k,n)$ is taken as the first column of this $\widehat{\boldsymbol{\Phi}}_{\underline{\mathbf{x}}}(k,n)$ matrix normalized by its first element. With this estimation, the first element of $\widehat{\boldsymbol{\rho}}_{\underline{\mathbf{x}}X_1}(k,n)$ is one; but occasionally some other elements may have a magnitude larger than 1 and this happens more frequently as the number of microphones increases.

References

1. J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
2. M. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.
3. J. P. Dmochowski and J. Benesty, "Microphone arrays: fundamental concepts," in *Speech Processing in Modern Communication–Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds., Berlin, Germany: Springer-Verlag, 2010, Chapter 8, pp. 199–223.
4. G. W. Elko and J. Meyer, "Microphone arrays," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., Berlin, Germany: Springer-Verlag, 2008, Chapter 48, pp. 1021–1041.
5. J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.

Index

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