SPRINGER BRIEFS IN
ELECTRICAL AND COMPUTER ENGINEERING

Jacob Benesty

Fundamentals of Speech Enhancement

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 to 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

Fundamentals of Speech Enhancement

Springer

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

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Electrical and Computer Engineering
ISBN 978-3-319-74523-7 ISBN 978-3-319-74524-4 (eBook)
https://doi.org/10.1007/978-3-319-74524-4

Library of Congress Control Number: 2017963977

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Abstract

The content of this book is essentially theoretical. We present and develop some very important concepts of speech enhancement in a simple but rigorous way. Many ideas are new; not only they shed light on this old problem but also give good hints on how to make things work better than some well-known conventional approaches. With the proposed presentation, all aspects of speech enhancement, from single channel, multichannel, beamforming, time domain, frequency domain, time-frequency domain, to binaural, are unified in a clear and flexible framework. We start with an exhaustive discussion on the fundamental best (linear and nonlinear) estimators, from which we show how they are connected to some important measures such as the coefficient of determination, the correlation coefficient, the conditional correlation coefficient, and the SNR. Then, in the subsequent chapters, we show how to exploit these measures in order to derive all kinds of noise reduction algorithms that can compromise in a very accurate and versatile way between noise reduction and speech distortion.

Contents

Chapter 1
Introduction

In this chapter, we briefly explain what is speech enhancement and describe its general formulation. Then, we present the organization of this study.

1.1 General Formulation of the Speech Enhancement Problem

We are routinely surrounded by undesired signals, i.e., noise and interferences. In all applications that are related to speech, from sound recording, cellular phones, hands-free communication, teleconferencing, hearing aids, to human-machine interfaces, a speech signal of interest captured by microphones is always contaminated by noise and interferences. Therefore, speech enhancement algorithms are required in order to clean the noisy signals from their disturbances. A solution to this problem was first proposed and developed five decades ago by Schroeder at Bell Laboratories [1], [2]. Since then, a lot of progress has been made and many approaches have been derived to solve this fundamental problem with a single microphone, multiple microphones, and in different domains; see [3], [4], [5], [6] and references therein to have a pretty good idea on how this topic has evolved.

The very general way to formulate the speech enhancement problem is

$$\mathbf{y} = \mathbf{x} + \mathbf{v}, \tag{1.1}$$

where the three vectors $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{v}$, of the same length, are the observed (or noisy), speech, and additive noise signals, respectively. All signals are zero mean, and $\mathbf{x}$ and $\mathbf{v}$ are assumed to be independent. The disturbance is due, obviously, to the signal vector $\mathbf{v}$, which affects both the quality and intelligibility of the signal vector of interest $\mathbf{x}$. Depending on the context and how we want things to be processed, the desired signal can be the first element, x_1, of $\mathbf{x}$, a part of $\mathbf{x}$, or the whole vector $\mathbf{x}$. Then, the objective of

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4_1

speech enhancement is to estimate this desired signal from the observed signal vector, $\mathbf{y}$. For that, we need at least to estimate the second-order statistics of $\mathbf{y}$, i.e., its covariance matrix $\mathbf{\Phi}_{\mathbf{y}}$.

With a single sensor and in the time domain, (1.1) is expressed as

$$\begin{aligned}\mathbf{y}(t) &= \left[\, y(t)\ y(t-1)\ \cdots\ y(t-L+1) \,\right]^T \\ &= \mathbf{x}(t) + \mathbf{v}(t),\end{aligned} \tag{1.2}$$

where t is the discrete-time index, the superscript ^{T} is the transpose operator, and $\mathbf{x}(t)$ and $\mathbf{v}(t)$ are defined similarly to $\mathbf{y}(t)$. The goal is then to estimate $x(t)$, the first component of $\mathbf{x}(t)$, from the observed signal vector, $\mathbf{y}(t)$, which contains L successive time samples picked up by the microphone.

Continuing with the single-channel case but in the time-frequency domain, we can write (1.1), thanks to the short-time Fourier transform, as

$$Y(k,n) = X(k,n) + V(k,n), \tag{1.3}$$

where k and n are the frequency bin and the time frame, respectively. Again, the objective is to estimate $X(k,n)$ from $Y(k,n)$.

In the multichannel scenario, i.e., with multiple (M) microphones, and in the frequency domain, (1.1) becomes

$$\mathbf{y}(f) = X(f)\mathbf{d}(f) + \mathbf{v}(f), \tag{1.4}$$

where $\mathbf{y}(f)$ is a vector of length M containing all the microphone signals at the frequency index f and $\mathbf{d}$ is the known steering (or transfer function ratio) vector whose first element is 1. Then, the objective of multichannel speech enhancement or beamforming is to estimate $X(f)$ from $\mathbf{y}(f)$.

In this book, we discuss these different (and more) aspects of speech enhancement in a unified way.

1.2 Organization of the Work

This work is organized into six chapters including this one. The best (linear and nonlinear) estimators are great tools in statistical signal processing. In Chapter 2, we show how they are applied to the frequency-domain speech enhancement problem. We also write these best estimators as a function of some important performance measures, which will be very useful in the rest of this study. In Chapter 3, we continue our investigation of the best estimators but in the time domain. We also deal with the best binaural speech enhancement estimator. In Chapter 4, we focus on the linear case and show how the most relevant noise reduction filters as well as new ones can be easily derived from the correlation coefficient. In Chapter 5, we discuss the impor-

tance of the output SNR and show how it can be used to find fundamental noise reduction filters and beamformers. Finally, in Chapter 6, we explain why the fullmode output SNR is of great interest and from this measure we derive a whole family of filters that can compromise very smoothly between noise reduction and speech distortion.

References

1. M. R. Schroeder, "Apparatus for suppressing noise and distortion in communication signals," U.S. Patent No. 3,180,936, Filed 1 Dec. 1960, Issued 27 Apr. 1965.
2. M. R. Schroeder, "Processing of communication signals to reduce effects of noise," U.S. Patent No. 3,403,224, Filed 28 May 1965, Issued 24 Sept. 1968.
3. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
4. J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters–A Theoretical Study*. Springer Briefs in Electrical and Computer Engineering, 2011.
5. J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Springer Briefs in Electrical and Computer Engineering, 2011.
6. J. Benesty and I. Cohen, *Canonical Correlation Analysis in Speech Enhancement*. Springer Briefs in Electrical and Computer Engineering, 2018.

Chapter 2
Best Speech Enhancement Estimator in the Frequency Domain

This chapter gives a fresh perspective on the best speech enhancement estimator in the frequency domain. We consider the general nonlinear case. In the first part, we deal with the single-channel scenario, where an example is studied with gamma distributions and a best quadratic estimator is derived. Then, in the second part, we focus on the multichannel scenario. Along this study, a great emphasis is put on some important performance measures (such as the coefficient of determination in the general nonlinear case and the correlation coefficient in the particular linear case) that can accurately tell us how the best estimators behave.

2.1 Signal Model and Problem Formulation

In all this chapter, we drop the dependence on the frequency, f, to simplify the notation. Therefore, for example, when we mention the random variable A, we mean $A(f)$.

Let X and V be two zero-mean, independent, and circular complex random variables belonging to the same probability space. The signal model considered in a large part of this chapter is [1], [2], [3]

$$Y = X + V, \tag{2.1}$$

where Y, X, and V are the observed, desired (speech), and noise signals, respectively. Our objective is to estimate X in the best possible way in some sense given Y. We can, equivalently, estimate V given Y first and then subtract this estimate from the observed signal to obtain the estimate of the speech signal.

Assuming that all variances are finite, the variance of Y is

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4_2

$$\begin{aligned} \text{var}\,(Y) &= E\left(|Y|^2\right) \\ &= \text{var}\,(X) + \text{var}\,(V)\,, \end{aligned} \tag{2.2}$$

where $E(\cdot)$ is the mathematical expectation, and $\text{var}\,(X)$ and $\text{var}\,(V)$ are the variances of X and V, respectively. From (2.2), it is easy to see that the input signal-to-noise ratio (SNR) is

$$\text{iSNR} = \frac{\text{var}\,(X)}{\text{var}\,(V)}. \tag{2.3}$$

We recall that the SNR is one of the most meaningful measures in speech enhancement.

2.2 Laws of Total Expectation and Total Variance

We start this section by giving two important properties: the law of total expectation (or the law of iterated expectations) and the law of total variance (or Eve's law).

Let A and B be two circular complex random variables. The conditional expectation of A given B is the random variable $E\,(A\,|B\,)$, whose randomness is inherited from B. The law of total expectation says that A and $E\,(A\,|B\,)$ have the same mean, i.e., [4]

$$E\,(A) = E\,[E\,(A\,|B\,)]\,, \tag{2.4}$$

where the outer expectation on the right-hand side of (2.4) is over the distribution of B. In fact, we can easily show the more general result:

$$\begin{aligned} E\,[f(B)A] &= E\,\{E\,[f(B)A\,|B\,]\} \\ &= E\,[f(B)E\,(A\,|B\,)]\,, \end{aligned} \tag{2.5}$$

where $f(B)$ is any function of B.

The law of total variance states that [4]

$$\text{var}\,(A) = E\,[\text{var}\,(A\,|B\,)] + \text{var}\,[E\,(A\,|B\,)]\,, \tag{2.6}$$

where the outer expectation and the outer variance on the right-hand side of (2.6) are over the distribution of B. This property can be proved by using the law of total expectation. Basically, (2.6) says that the conditional variance on average is smaller than the variance, which actually makes perfect sense since the uncertainty is reduced.

Since variances are always nonnegative, we deduce from (2.6) that

$$0 \leq \frac{\mathrm{var}\left[E\left(A\,|B\right)\right]}{\mathrm{var}\left(A\right)} \leq 1. \tag{2.7}$$

As a result, the classical magnitude squared coherence function can be generalized to

$$\begin{aligned} \left|\gamma_{A|B}\right|^2 &= 1 - \frac{E\left[\mathrm{var}\left(A\,|B\right)\right]}{\mathrm{var}\left(A\right)} \\ &= \frac{\mathrm{var}\left[E\left(A\,|B\right)\right]}{\mathrm{var}\left(A\right)}. \end{aligned} \tag{2.8}$$

This expression is often called the coefficient of determination in the literature of statistics [5]. If A and B are independent, then $\mathrm{var}\left(A\,|B\right) = \mathrm{var}\left(A\right)$. As a consequence, $\left|\gamma_{A|B}\right|^2 = 0$. Conversely, if $\left|\gamma_{A|B}\right|^2 = 0$, then $\mathrm{var}\left[E\left(A\,|B\right)\right] = 0$, which implies that A and B are independent. At the other limiting case, if $A = B$, then $\mathrm{var}\left(B\,|B\right) = 0$, which leads to $\left|\gamma_{A|B}\right|^2 = 1$. In fact, for $A = f(B)$, we have $E\left(A\,|B\right) = E\left[f(B)\,|B\right] = f(B)$; as a result, $\left|\gamma_{A|B}\right|^2 = 1$. This coefficient of determination, which is a direct consequence of the law of total variance and measures how close A is to $E\left(A\,|B\right)$, plays a key role in the best estimator in general and in speech enhancement in particular. It can be used as a powerful performance measure in all aspects of speech enhancement as explained in great details in the rest.

2.3 Best Estimator

Returning to our signal model in (2.1), it is well known that the best estimator of X in the minimum mean-squared error (MMSE) sense is the conditional expectation of X given Y [6], i.e.,

$$E\left(X\,|Y\right) = Z_X(Y). \tag{2.9}$$

Indeed, let $f_X(Y)$ be any (linear or nonlinear) function of Y, we always have

$$E\left[\left|X - Z_X(Y)\right|^2\right] = E\left(\left|\mathcal{E}_X\right|^2\right) \leq E\left[\left|X - f_X(Y)\right|^2\right], \tag{2.10}$$

where $\mathcal{E}_X = X - E\left(X\,|Y\right)$ is the error signal between the desired signal and its best estimator. By virtue of the law of total expectation, the two random variables X and $Z_X(Y)$ have the same mean, i.e.,

$$E\left[Z_X(Y)\right] = E\left(X\right) = 0. \tag{2.11}$$

It can also be verified that the MMSE is

$$
\begin{aligned}
E\left(\left|\mathcal{E}_X\right|^2\right) &= E\left[\operatorname{var}\left(X\left|Y\right.\right)\right] \\
&= \operatorname{var}\left(X\right) - \operatorname{var}\left[Z_X(Y)\right] \\
&= \operatorname{var}\left(X\right)\left(1 - \left|\gamma_{X|Y}\right|^2\right),
\end{aligned} \tag{2.12}
$$

which clearly depends on the coefficient of determination.

In the same way, the best estimator of V in the MMSE sense is the conditional expectation of V given Y, i.e.,

$$
E\left(V\left|Y\right.\right) = Z_V(Y) \tag{2.13}
$$

and for any (linear or nonlinear) function of Y, $f_V(Y)$, we always have

$$
E\left[\left|V - Z_V(Y)\right|^2\right] = E\left(\left|\mathcal{E}_V\right|^2\right) \leq E\left[\left|V - f_V(Y)\right|^2\right], \tag{2.14}
$$

where $\mathcal{E}_V = V - E\left(V\left|Y\right.\right)$ is the error signal between the noise and its best estimator. By virtue of the law of total expectation, the two random variables V and $Z_V(Y)$ have the same mean, i.e.,

$$
E\left[Z_V(Y)\right] = E\left(V\right) = 0. \tag{2.15}
$$

We also deduce that the MMSE is

$$
E\left(\left|\mathcal{E}_V\right|^2\right) = \operatorname{var}\left(V\right)\left(1 - \left|\gamma_{V|Y}\right|^2\right). \tag{2.16}
$$

By adding together the best estimator of X and the best estimator of V, we obtain the observed signal, i.e.,

$$
\begin{aligned}
Y &= E\left(Y\left|Y\right.\right) \\
&= E\left(X\left|Y\right.\right) + E\left(V\left|Y\right.\right).
\end{aligned} \tag{2.17}
$$

The above property is very interesting. As expected, it shows that the estimation errors of both estimators cancel out. In other words, the best estimator of X can be found, equivalently, from the best estimator of V. In the best estimator of X, $E\left(X\left|Y\right.\right)$ gives the speech distortion perspective while $Y - E\left(V\left|Y\right.\right)$ gives the noise reduction perspective. From (2.17), we easily see that $\mathcal{E}_X = -\mathcal{E}_V$ and, as a result, $E\left(\left|\mathcal{E}_X\right|^2\right) = E\left(\left|\mathcal{E}_V\right|^2\right)$. Then, equating (2.12) and (2.16), we obtain

$$
\mathrm{iSNR} + \left|\gamma_{V|Y}\right|^2 = 1 + \mathrm{iSNR} \times \left|\gamma_{X|Y}\right|^2. \tag{2.18}
$$

From the previous expression, we have

$$\lim_{\text{iSNR}\to 0} \left|\gamma_{V|Y}\right|^2 = 1, \tag{2.19}$$

$$\lim_{\text{iSNR}\to \infty} \left|\gamma_{X|Y}\right|^2 = 1. \tag{2.20}$$

In words, the best estimator is able to completely remove the noise when the input SNR is close to 0 and fully recover the desired signal when the input SNR approaches infinity. However, (2.18) does not give us any information about speech distortion in the first case and noise reduction in the second one. These statements make intuitively sense and confirm what we have always observed for the best estimator. We also see from (2.18) that there is not such a thing such as distortionless (i.e., $\left|\gamma_{X|Y}\right|^2 = 1$) with the best estimator in general, unless the noise is completely removed (i.e., $\left|\gamma_{V|Y}\right|^2 = 1$) at the same time.

In the pathological scenario where X and V are independent and identically distributed (i.i.d.), we have

$$E\left(X\,|Y\right) = E\left(V\,|Y\right) = \frac{Y}{2}. \tag{2.21}$$

As a result, single-channel speech enhancement is not feasible with the best estimator. So when the distribution of the noise resembles the one of the speech, we should not expect much noise reduction, not because of the problem of estimating statistics of nonstationary signals but because the distributions of the speech and noise may be similar. In this difficult scenario, the only way to attenuate the level of the noise is to use more than one sensor (see Section 2.6).

Best Linear Estimator

It is of great interest to study the best linear estimator because of its simple form. The study of this very important particular case can also lead to better insights into the best estimator in general.

It is well known that the best linear estimator of X in the MMSE sense is

$$E\left(X\,|Y\right) = H_{X,\text{W}}Y, \tag{2.22}$$

where

$$\begin{aligned} H_{X,\text{W}} &= \frac{\text{var}\left(X\right)}{\text{var}\left(Y\right)} \\ &= \frac{\text{iSNR}}{1+\text{iSNR}} \end{aligned} \tag{2.23}$$

is the celebrated Wiener gain. In this case, the coefficient of determination and the MMSE are, respectively,

$$\left|\gamma_{X|Y}\right|^2 = H_{X,\mathrm{W}} \leq 1 \tag{2.24}$$

and

$$E\left(\left|X - H_{X,\mathrm{W}}Y\right|^2\right) = \mathrm{var}\,(X)\,(1 - H_{X,\mathrm{W}})\,. \tag{2.25}$$

The coefficient of determination, $\left|\gamma_{X|Y}\right|^2$, is a good measure of the desired signal distortion; a value close to 1 (resp. 0) means low (resp. large) distortion.

In the same manner, the best linear estimator of V in the MMSE sense is

$$E\,(V\,|Y\,) = H_{V,\mathrm{W}}Y, \tag{2.26}$$

where

$$\begin{aligned} H_{V,\mathrm{W}} &= \frac{\mathrm{var}\,(V)}{\mathrm{var}\,(Y)} \\ &= \frac{1}{1 + \mathrm{iSNR}}. \end{aligned} \tag{2.27}$$

We deduce that the coefficient of determination and the MMSE are, respectively,

$$\left|\gamma_{V|Y}\right|^2 = H_{V,\mathrm{W}} \leq 1 \tag{2.28}$$

and

$$E\left(\left|V - H_{V,\mathrm{W}}Y\right|^2\right) = \mathrm{var}\,(V)\,(1 - H_{V,\mathrm{W}})\,. \tag{2.29}$$

The coefficient of determination, $\left|\gamma_{V|Y}\right|^2$, is a good measure of noise reduction; a value close to 1 (resp. 0) means large (resp. low) noise reduction.

It is clear that

$$Y = E\,(X\,|Y\,) + E\,(V\,|Y\,) \tag{2.30}$$

or, equivalently,

$$1 = H_{X,\mathrm{W}} + H_{V,\mathrm{W}}. \tag{2.31}$$

We also have

$$1 = \left|\gamma_{X|Y}\right|^2 + \left|\gamma_{V|Y}\right|^2, \tag{2.32}$$

which is only true for the best linear estimator. This relation shows the fundamental compromise between noise reduction and speech distortion, in the single-channel case and with Gaussian signals, as $\left|\gamma_{X|Y}\right|^2$ and $\left|\gamma_{V|Y}\right|^2$ go in opposite directions. For large noise reduction (resp. low speech distortion),

$\left|\gamma_{V|Y}\right|^2$ (resp. $\left|\gamma_{X|Y}\right|^2$) is rather close to 1, so that $\left|\gamma_{X|Y}\right|^2$ (resp. $\left|\gamma_{V|Y}\right|^2$) is close to 0 implying large speech distortion (resp. low noise reduction).

We can state that nonlinear noise reduction in the single-channel case is extremely efficient if the following condition holds

$$\left|\gamma_{X|Y}\right|^2 + \left|\gamma_{V|Y}\right|^2 \approx 2. \tag{2.33}$$

Under this condition, which can be fulfilled depending on the distributions of X and V but certainly not when they are both Gaussians, we can have large noise reduction and low speech distortion. It can be very instructive to find some distributions for which (2.33) is more or less fulfilled. This is certainly possible since $\left|\gamma_{X|Y}\right|^2 = \left|\gamma_{V|Y}\right|^2 = 1$ is a solution of (2.18).

This study suggests that, in the general case, we can combine the speech distortion and noise reduction measures into one convenient measure:

$$\varrho_{\mathrm{SC}} = \left|\gamma_{X|Y}\right|^2 + \left|\gamma_{V|Y}\right|^2, \tag{2.34}$$

where $1 \leq \varrho_{\mathrm{SC}} \leq 2$, with the subscript $_{\mathrm{SC}}$ standing for single channel. For ϱ_{SC} close to 2, we have the almost perfect estimator while for ϱ_{SC} close to 1, we deal with the linear case and the well-known unavoidable compromise. Therefore, the larger is ϱ_{SC}, the less the compromise between speech distortion and noise reduction.

We conclude this part by saying that $\left|\gamma_{X|Y}\right|^2$, $\left|\gamma_{V|Y}\right|^2$, and ϱ_{SC} are accurate, convenient, and most useful performance measures for the evaluation of the single-channel speech enhancement problem with the best estimator. The first measure quantifies distortion of the desired signal, the second one evaluates noise reduction, and the last one tells us about the compromise.

2.4 Example with Gamma Distributions

The study of this section is somewhat an extension of the works presented in [7], [8], [9], [10].

2.4.1 Reformulation of the Problem and Approximation

We can also express (2.1) as

$$|Y|e^{\jmath\theta_Y} = |X|e^{\jmath\theta_X} + |V|e^{\jmath\theta_V}, \tag{2.35}$$

where $\jmath$ is the imaginary unit, and θ_Y, θ_X, and θ_V are the phases of Y, X, and V, respectively. In the rest, we will use the approximation:

$$|Y| = \widetilde{Y} \approx \widetilde{X} + \widetilde{V} = |X| + |V|, \tag{2.36}$$

as it is very often the case in the single-channel speech enhancement problem in the frequency domain. Therefore, we have $\widetilde{Y} \in [0, \infty)$ and $\widetilde{X}, \widetilde{V} \in [0, \widetilde{Y}]$. As a result, when the estimator $\widehat{\widetilde{X}}$ is derived for the magnitude of the speech, the estimator of X is

$$\widehat{X} = \widehat{\widetilde{X}} e^{\jmath \theta_Y}. \tag{2.37}$$

Then, our objective is to derive and evaluate the best estimator of $\widetilde{X}$, from an MMSE perspective, with gamma distributions.

2.4.2 Best Estimator

It is well known that the modulus of the desired speech signal can be well modeled with the gamma distribution:

$$p_{\widetilde{X}}\left(\widetilde{X}\right) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \widetilde{X}^{\alpha-1} e^{-\lambda \widetilde{X}}, \ \widetilde{X} \geq 0, \tag{2.38}$$

where $\alpha > 0$ is the shape parameter, $\lambda > 0$ the scale parameter, and $\Gamma(\cdot)$ the gamma function. The gamma distributed random variable $\widetilde{X}$ is denoted

$$\widetilde{X} \sim \Gamma_{\alpha,\lambda}. \tag{2.39}$$

The mean of $\widetilde{X}$ can be easily calculated; it is given by

$$E\left(\widetilde{X}\right) = \frac{\alpha}{\lambda}. \tag{2.40}$$

The magnitude of the noise can also be modeled with the gamma distribution but with a different shape parameter, $\beta > 0$, i.e.,

$$p_{\widetilde{V}}\left(\widetilde{V}\right) = \frac{\lambda^{\beta}}{\Gamma(\beta)} \widetilde{V}^{\beta-1} e^{-\lambda \widetilde{V}}, \ \widetilde{V} \geq 0, \tag{2.41}$$

which we denote $\widetilde{V} \sim \Gamma_{\beta,\lambda}$. The mean of $\widetilde{V}$ is then

$$E\left(\widetilde{V}\right) = \frac{\beta}{\lambda}. \tag{2.42}$$

It can be verified that

$$\begin{aligned} p_{\widetilde{Y}}\left(\widetilde{Y}\right) &= p_{\widetilde{X}+\widetilde{V}}\left(\widetilde{Y}\right) \\ &= \int_0^{\widetilde{Y}} p_{\widetilde{X}}\left(\widetilde{X}\right) p_{\widetilde{V}}\left(\widetilde{Y}-\widetilde{X}\right) d\widetilde{X} \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma\left(\alpha+\beta\right)} \widetilde{Y}^{\alpha+\beta-1} e^{-\lambda\widetilde{Y}}, \end{aligned} \tag{2.43}$$

meaning that $\widetilde{Y}$ is also a gamma distributed random variable, i.e., $\widetilde{Y} \sim \Gamma_{\alpha+\beta,\lambda}$.

The joint distribution of $\widetilde{Y}$ and $\widetilde{X}$ is

$$\begin{aligned} p_{\widetilde{Y},\widetilde{X}}\left(\widetilde{Y},\widetilde{X}\right) &= p_{\widetilde{V},\widetilde{X}}\left(\widetilde{Y}-\widetilde{X},\widetilde{X}\right) \\ &= p_{\widetilde{V}}\left(\widetilde{Y}-\widetilde{X}\right) p_{\widetilde{X}}\left(\widetilde{X}\right), \end{aligned} \tag{2.44}$$

where the last equation is the consequence of the fact that $\widetilde{X}$ and $\widetilde{V}$ are independent. Therefore, the conditional distribution of $\widetilde{X}$ given $\widetilde{Y}$ is

$$\begin{aligned} p_{\widetilde{X}|\widetilde{Y}}\left(\widetilde{X}\,\middle|\,\widetilde{Y}\right) &= \frac{p_{\widetilde{Y},\widetilde{X}}\left(\widetilde{Y},\widetilde{X}\right)}{p_{\widetilde{Y}}\left(\widetilde{Y}\right)} \\ &= \frac{p_{\widetilde{V}}\left(\widetilde{Y}-\widetilde{X}\right) p_{\widetilde{X}}\left(\widetilde{X}\right)}{p_{\widetilde{Y}}\left(\widetilde{Y}\right)}. \end{aligned} \tag{2.45}$$

Substituting (2.38), (2.41), and (2.43) into (2.45), we easily find that

$$p_{\widetilde{X}|\widetilde{Y}}\left(\widetilde{X}\,\middle|\,\widetilde{Y}\right) = \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \times \frac{1}{\widetilde{X}} \times \left(\frac{\widetilde{X}}{\widetilde{Y}}\right)^{\alpha} \left(1-\frac{\widetilde{X}}{\widetilde{Y}}\right)^{\beta-1}. \tag{2.46}$$

Now, we have everything to find the best estimator in the MMSE sense:

$$\begin{aligned} \widehat{\widetilde{X}} &= E\left(\widetilde{X}\,\middle|\,\widetilde{Y}\right) \\ &= \int_0^{\widetilde{Y}} \widetilde{X} p_{\widetilde{X}|\widetilde{Y}}\left(\widetilde{X}\,\middle|\,\widetilde{Y}\right) d\widetilde{X} \\ &= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \int_0^{\widetilde{Y}} \left(\frac{\widetilde{X}}{\widetilde{Y}}\right)^{\alpha} \left(1-\frac{\widetilde{X}}{\widetilde{Y}}\right)^{\beta-1} d\widetilde{X}. \end{aligned} \tag{2.47}$$

Making the change of variables $U = \widetilde{X}/\widetilde{Y}$, we can write the previous expression as

$$E\left(\widetilde{X}\left|\widetilde{Y}\right.\right)=\widetilde{Y}\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}B\left(\alpha+1,\beta\right), \tag{2.48}$$

where

$$\begin{aligned} B\left(\alpha,\beta\right)&=\int_0^1 U^{\alpha-1}\left(1-U\right)^{\beta-1}dU \\ &=\frac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)} \end{aligned} \tag{2.49}$$

is the beta function. Using the relationship:

$$\Gamma\left(\alpha+1\right)=\alpha\Gamma\left(\alpha\right), \tag{2.50}$$

the best estimator simplifies to

$$\begin{aligned} E\left(\widetilde{X}\left|\widetilde{Y}\right.\right)&=\widetilde{Y}\frac{\alpha}{\alpha+\beta} \\ &=\widetilde{Y}\frac{E\left(\widetilde{X}\right)}{E\left(\widetilde{X}\right)+E\left(\widetilde{V}\right)}. \end{aligned} \tag{2.51}$$

Obviously, the estimator in (2.51) leads to the MMSE, assuming the approximation in (2.36). We see that noise reduction is possible because the two distributions of the speech and noise have different shapes.

Finally, we deduce that our estimator is

$$\widehat{X}_{\mathrm{G}}=H_{\mathrm{G}}Y, \tag{2.52}$$

where

$$H_{\mathrm{G}}=\frac{E\left(|X|\right)}{E\left(|X|\right)+E\left(|V|\right)} \tag{2.53}$$

is a positive gain. It is of interest to compare this approach to the classical Wiener gain technique:

$$\widehat{X}_{\mathrm{W}}=H_{\mathrm{W}}Y, \tag{2.54}$$

where

$$H_{\mathrm{W}}=\frac{E\left(|X|^2\right)}{E\left(|X|^2\right)+E\left(|V|^2\right)}. \tag{2.55}$$

From (2.52), we easily find that the noise reduction factor and the speech distortion index are, respectively,

$$\xi_{\mathrm{nr}}\left(H_{\mathrm{G}}\right)=\frac{1}{H_{\mathrm{G}}^{2}} \tag{2.56}$$

and

$$\upsilon_{\mathrm{sd}}\left(H_{\mathrm{G}}\right)=\left(1-H_{\mathrm{G}}\right)^{2}. \tag{2.57}$$

2.5 A Brief Study of the Best Quadratic Estimator

From Section 2.3, we know that any random variable X can be decomposed as

$$X=E\left(X\,|Y\,\right)+\mathcal{E}_{X}, \tag{2.58}$$

where $\mathcal{E}_X$ is a zero-mean random variable with $E\left(\mathcal{E}_X\,|Y\,\right)=0$ and $E\left[f(Y)\mathcal{E}_X\right]=0$, and $f(Y)$ being any function of Y. Obviously, the same decomposition applies for the noise signal, V. In this section, we assume that at least one of the two signals X and V is not Gaussian. Therefore, we can generalize the linear model to the quadratic one:

$$E\left(X\,|Y\,\right)=\widetilde{H}_{X,1}^{*}Y+\widetilde{H}_{X,2}^{*}Y|Y|=\widetilde{\mathbf{h}}_{X}^{H}\widetilde{\mathbf{y}}, \tag{2.59}$$

$$E\left(V\,|Y\,\right)=\widetilde{H}_{V,1}^{*}Y+\widetilde{H}_{V,2}^{*}Y|Y|=\widetilde{\mathbf{h}}_{V}^{H}\widetilde{\mathbf{y}}, \tag{2.60}$$

where the superscript * and ^{H} are the complex-conjugate and conjugate-transpose operators, $\widetilde{\mathbf{h}}_X$ and $\widetilde{\mathbf{h}}_V$ are two complex-valued filters of length 2, and

$$\widetilde{\mathbf{y}}=\left[\,Y\;Y|Y|\,\right]^{T}.$$

For convenience, we also define the two vectors of length 2:

$$\widetilde{\mathbf{x}}=\left[\,X\;X|X|\,\right]^{T},$$
$$\widetilde{\mathbf{v}}=\left[\,V\;V|V|\,\right]^{T}.$$

The minimization of $E\left(\left|\mathcal{E}_X\right|^2\right)$ and $E\left(\left|\mathcal{E}_V\right|^2\right)$ leads to the best quadratic estimators:

$$\widetilde{\mathbf{h}}_{X,\mathrm{Q}}=\mathrm{cov}^{-1}\left(\widetilde{\mathbf{y}}\right)\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)\mathbf{i}, \tag{2.61}$$

$$\widetilde{\mathbf{h}}_{V,\mathrm{Q}}=\mathrm{cov}^{-1}\left(\widetilde{\mathbf{y}}\right)\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{v}}\right)\mathbf{i}, \tag{2.62}$$

where $\mathrm{cov}\left(\widetilde{\mathbf{y}}\right)=E\left(\widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^{H}\right)$, $\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)=E\left(\widetilde{\mathbf{y}}\widetilde{\mathbf{x}}^{H}\right)$, $\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{v}}\right)=E\left(\widetilde{\mathbf{y}}\widetilde{\mathbf{v}}^{H}\right)$, and $\mathbf{i}=\left[\,1\;0\,\right]^{T}$. It is clear that

$$\widetilde{\mathbf{h}}_{X,\mathrm{Q}}^{H}\widetilde{\mathbf{y}} + \widetilde{\mathbf{h}}_{V,\mathrm{Q}}^{H}\widetilde{\mathbf{y}} = Y. \tag{2.63}$$

We deduce that the MMSEs are

$$E\left(\left|X - \widetilde{\mathbf{h}}_{X,\mathrm{Q}}^{H}\widetilde{\mathbf{y}}\right|^2\right) = \mathrm{var}\left(X\right)\left(1 - \left|\widetilde{\gamma}_{X|Y}\right|^2\right), \tag{2.64}$$

$$E\left(\left|V - \widetilde{\mathbf{h}}_{V,\mathrm{Q}}^{H}\widetilde{\mathbf{y}}\right|^2\right) = \mathrm{var}\left(V\right)\left(1 - \left|\widetilde{\gamma}_{V|Y}\right|^2\right), \tag{2.65}$$

where

$$\left|\widetilde{\gamma}_{X|Y}\right|^2 = \frac{\mathbf{i}^T \mathrm{cov}\left(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}\right)\mathrm{cov}^{-1}\left(\widetilde{\mathbf{y}}\right)\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)\mathbf{i}}{\mathrm{var}\left(X\right)}, \tag{2.66}$$

$$\left|\widetilde{\gamma}_{V|Y}\right|^2 = \frac{\mathbf{i}^T \mathrm{cov}\left(\widetilde{\mathbf{v}},\widetilde{\mathbf{y}}\right)\mathrm{cov}^{-1}\left(\widetilde{\mathbf{y}}\right)\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{v}}\right)\mathbf{i}}{\mathrm{var}\left(V\right)}. \tag{2.67}$$

Property 2.1. We have

$$\left|\widetilde{\gamma}_{X|Y}\right|^2 \geq \left|\gamma_{X|Y}\right|^2, \tag{2.68}$$

$$\left|\widetilde{\gamma}_{V|Y}\right|^2 \geq \left|\gamma_{V|Y}\right|^2, \tag{2.69}$$

where $\left|\gamma_{X|Y}\right|^2 = \mathrm{var}\left(X\right)/\mathrm{var}\left(Y\right)$ and $\left|\gamma_{V|Y}\right|^2 = \mathrm{var}\left(V\right)/\mathrm{var}\left(Y\right)$ are the coefficients of determination for the best linear estimators (see Section 2.3).

Proof. Let us define the normalized covariance matrix:

$$\mathrm{covn}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right) = \frac{\mathrm{cov}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)}{\mathrm{var}\left(X\right)}. \tag{2.70}$$

It is easy to verify that the first element of the vector $\mathrm{covn}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)\mathbf{i}$ is 1. We can express (2.66) as

$$\left|\widetilde{\gamma}_{X|Y}\right|^2 = \mathrm{var}\left(X\right)\mathbf{i}^T \mathrm{covn}\left(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}\right)\mathrm{cov}^{-1}\left(\widetilde{\mathbf{y}}\right)\mathrm{covn}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)\mathbf{i}. \tag{2.71}$$

Using the Cauchy-Schwarz inequality:

$$\begin{aligned}&\left[\mathbf{i}^T \mathrm{covn}\left(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}\right)\mathrm{cov}^{-1}\left(\widetilde{\mathbf{y}}\right)\mathrm{covn}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)\mathbf{i}\right]\left[\mathbf{i}^T \mathrm{cov}\left(\widetilde{\mathbf{y}}\right)\mathbf{i}\right]\\ &\quad\geq \left|\mathbf{i}^T \mathrm{covn}\left(\widetilde{\mathbf{y}},\widetilde{\mathbf{x}}\right)\mathbf{i}\right|^2 = 1\end{aligned} \tag{2.72}$$

and substituting this result into (2.71), we find the inequality in (2.68). The inequality in (2.69) can be shown in the exact same way. Inequalities in (2.68) and (2.69) are also consequences of the MMSE.

From Property 2.1, we can say that the best quadratic estimator reduces more noise and distorts less the desired speech than the best linear estimator. We also have

$$\widetilde{\varrho}_{\mathrm{SC}} = \left|\widetilde{\gamma}_{X|Y}\right|^2 + \left|\widetilde{\gamma}_{V|Y}\right|^2 \geq 1, \tag{2.73}$$

which means that the best quadratic estimator better compromises between noise reduction and speech distortion than the best linear estimator.

2.6 Generalization to the Multichannel Case

In the multichannel context, we assume that we have M sensors and, hence, M observations. Therefore, the observation signal vector is given by [11]

$$\begin{aligned} \mathbf{y} &= \left[\, Y_1 \; Y_2 \; \cdots \; Y_M \,\right]^T \\ &= X\mathbf{d} + \mathbf{v}, \end{aligned} \tag{2.74}$$

where X is the zero-mean random desired signal, $\mathbf{d}$ is the known steering (or transfer function ratio) vector whose first element is 1, and $\mathbf{v}$ is the zero-mean random noise vector. Assuming that X and $\mathbf{v}$ are independent, we deduce that the covariance matrix of $\mathbf{y}$ is

$$\begin{aligned} \mathrm{cov}\,(\mathbf{y}) &= E\left(\mathbf{y}\mathbf{y}^H\right) \\ &= \mathrm{cov}\,(\mathbf{x}) + \mathrm{cov}\,(\mathbf{v}) \\ &= \mathrm{var}\,(X)\,\mathbf{d}\mathbf{d}^H + \mathrm{cov}\,(\mathbf{v})\,, \end{aligned} \tag{2.75}$$

where $\mathrm{var}\,(X)$ is the variance of X, and $\mathrm{cov}\,(\mathbf{x})$ and $\mathrm{cov}\,(\mathbf{v})$ are the covariance matrices of $\mathbf{x}$ and $\mathbf{v}$, respectively.

Considering the random variable X and the random vector $\mathbf{y}$ from the signal model in (2.74), the law of total variance is

$$\mathrm{var}\,(X) = E\left[\mathrm{var}\,(X\,|\mathbf{y}\,)\right] + \mathrm{var}\left[E\,(X\,|\mathbf{y}\,)\right]. \tag{2.76}$$

As a consequence, the coefficient of determination is

$$\begin{aligned} \left|\gamma_{X|\mathbf{y}}\right|^2 &= 1 - \frac{E\left[\mathrm{var}\,(X\,|\mathbf{y}\,)\right]}{\mathrm{var}\,(X)} \\ &= \frac{\mathrm{var}\left[E\,(X\,|\mathbf{y}\,)\right]}{\mathrm{var}\,(X)}. \end{aligned} \tag{2.77}$$

This measure is close to 1 when there is little noise and may get smaller when the noise increases. Let V_1 be the first component of $\mathbf{v}$. The law of total variance and the coefficient of determination are, respectively,

$$\mathrm{var}\,(V_1) = E\left[\mathrm{var}\,(V_1\,|\mathbf{y}\,)\right] + \mathrm{var}\left[E\,(V_1\,|\mathbf{y}\,)\right] \tag{2.78}$$

and

$$\begin{aligned}\left|\gamma_{V_1|\mathbf{y}}\right|^2 &= 1 - \frac{E\left[\mathrm{var}\left(V_1\,|\mathbf{y}\right)\right]}{\mathrm{var}\left(V_1\right)} \\ &= \frac{\mathrm{var}\left[E\left(V_1\,|\mathbf{y}\right)\right]}{\mathrm{var}\left(V_1\right)},\end{aligned} \tag{2.79}$$

where $\mathrm{var}\left(V_1\right)$ is the variance of V_1. We see that when the noise dominates, $\left|\gamma_{V_1|\mathbf{y}}\right|^2$ is close to 1.

Similar to the single-channel case, the best estimator of X in the MMSE sense is

$$E\left(X\,|\mathbf{y}\right) = Z_X(\mathbf{y}) \tag{2.80}$$

and the MMSE is

$$\begin{aligned}E\left[\left|X - Z_X(\mathbf{y})\right|^2\right] &= E\left[\mathrm{var}\left(X\,|\mathbf{y}\right)\right] \\ &= \mathrm{var}\left(X\right) - \mathrm{var}\left[Z_X(\mathbf{y})\right] \\ &= \mathrm{var}\left(X\right)\left(1 - \left|\gamma_{X|\mathbf{y}}\right|^2\right).\end{aligned} \tag{2.81}$$

Also, the best estimator of V_1 in the MMSE sense is

$$E\left(V_1\,|\mathbf{y}\right) = Z_{V_1}(\mathbf{y}) \tag{2.82}$$

and the MMSE is

$$\begin{aligned}E\left[\left|V_1 - Z_{V_1}(\mathbf{y})\right|^2\right] &= E\left[\mathrm{var}\left(X\,|\mathbf{y}\right)\right] \\ &= \mathrm{var}\left(V_1\right) - \mathrm{var}\left[Z_{V_1}(\mathbf{y})\right] \\ &= \mathrm{var}\left(V_1\right)\left(1 - \left|\gamma_{V_1|\mathbf{y}}\right|^2\right).\end{aligned} \tag{2.83}$$

Let $Y_1 = X + V_1$ be the first component of $\mathbf{y}$. We have

$$\begin{aligned}Y_1 &= E\left(Y_1\,|\mathbf{y}\right) \\ &= E\left(X\,|\mathbf{y}\right) + E\left(V_1\,|\mathbf{y}\right).\end{aligned} \tag{2.84}$$

This means that if we know $E\left(X\,|\mathbf{y}\right)$, we can deduce $E\left(V_1\,|\mathbf{y}\right)$, and vice versa. Also, in the best estimator of X, $E\left(X\,|\mathbf{y}\right)$ gives the speech distortion perspective while $Y_1 - E\left(V_1\,|\mathbf{y}\right)$ gives the noise reduction perspective. We have the fundamental relation:

$$\mathrm{iSNR}_{\mathrm{SC}} + \left|\gamma_{V_1|\mathbf{y}}\right|^2 = 1 + \mathrm{iSNR}_{\mathrm{SC}} \times \left|\gamma_{X|\mathbf{y}}\right|^2, \tag{2.85}$$

where

$$\mathrm{iSNR}_{\mathrm{SC}} = \frac{\mathrm{var}\left(X\right)}{\mathrm{var}\left(V_1\right)} \tag{2.86}$$

is the input SNR at the first sensor, which is equivalent to the single-channel input SNR.

Now, let us focus on the best linear estimators for X and V_1 in the MMSE sense. For X, we have

$$E\left(X\left|\mathbf{y}\right.\right) = \mathbf{h}_{X,\mathrm{W}}^{H}\mathbf{y}, \tag{2.87}$$

where

$$\mathbf{h}_{X,\mathrm{W}} = \mathrm{var}\left(X\right)\mathrm{cov}^{-1}\left(\mathbf{y}\right)\mathbf{d} \tag{2.88}$$

is the multichannel Wiener filter. We find that the coefficient of determination is

$$\begin{aligned}\left|\gamma_{X|\mathbf{y}}\right|^2 &= \mathrm{var}\left(X\right)\mathbf{d}^H\mathrm{cov}^{-1}\left(\mathbf{y}\right)\mathbf{d} \\ &= \frac{\mathrm{var}\left(X\right)\mathbf{d}^H\mathrm{cov}^{-1}\left(\mathbf{v}\right)\mathbf{d}}{1+\mathrm{var}\left(X\right)\mathbf{d}^H\mathrm{cov}^{-1}\left(\mathbf{v}\right)\mathbf{d}},\end{aligned} \tag{2.89}$$

which is a good measure of speech distortion. By analogy to the single-channel case, another interesting way to define the input SNR in the multichannel case is the fullmode input SNR (see Chapter 5 for a detailed discussion on this measure):

$$\begin{aligned}\mathrm{iSNR}_{\mathrm{FM}} &= \frac{\mathrm{tr}\left[\mathrm{cov}^{-1}\left(\mathbf{v}\right)\mathrm{cov}\left(\mathbf{x}\right)\right]}{M} \\ &= \frac{\mathrm{var}\left(X\right)\mathbf{d}^H\mathrm{cov}^{-1}\left(\mathbf{v}\right)\mathbf{d}}{M},\end{aligned} \tag{2.90}$$

where tr[·] is the trace of a square matrix. Therefore, we can express (2.89) as

$$\left|\gamma_{X|\mathbf{y}}\right|^2 = \frac{M\times\mathrm{iSNR}_{\mathrm{FM}}}{1+M\times\mathrm{iSNR}_{\mathrm{FM}}} \leq 1, \tag{2.91}$$

which strongly depends on M. As the number of microphones increases, this measure gets closer to 1, which of course makes sense since increasing M improves the estimator. For V_1, we have

$$E\left(V_1\left|\mathbf{y}\right.\right) = \mathbf{h}_{V_1,\mathrm{W}}^{H}\mathbf{y}, \tag{2.92}$$

where

$$\mathbf{h}_{V_1,\mathrm{W}} = \mathrm{cov}^{-1}\left(\mathbf{y}\right)\mathrm{cov}\left(\mathbf{v}\right)\mathbf{i} \tag{2.93}$$

is the multichannel Wiener filter for the estimation of V_1, with $\mathbf{i}$ being the first column of the $M\times M$ identity matrix $\mathbf{I}_M$. The coefficient of determination is then

$$\begin{aligned}\left|\gamma_{V_1|\mathbf{y}}\right|^2 &= \frac{\mathbf{i}^T \mathrm{cov}\left(\mathbf{v}\right) \mathrm{cov}^{-1}\left(\mathbf{y}\right) \mathrm{cov}\left(\mathbf{v}\right) \mathbf{i}}{\mathrm{var}\left(V_1\right)} \\ &= \frac{1 + M \times \mathrm{iSNR}_{\mathrm{FM}} - \mathrm{iSNR}_{\mathrm{SC}}}{1 + M \times \mathrm{iSNR}_{\mathrm{FM}}} \leq 1,\end{aligned} \tag{2.94}$$

which is a good measure of noise reduction.

We now give an important property.

Property 2.2. Let

$$\varrho_{\mathrm{MC}} = \left|\gamma_{X|\mathbf{y}}\right|^2 + \left|\gamma_{V_1|\mathbf{y}}\right|^2 \tag{2.95}$$

be the combined speech distortion and noise reduction measures in the multichannel case. With the best linear estimators, we always have

$$\varrho_{\mathrm{MC}} \geq 1. \tag{2.96}$$

Proof. Using (2.91) and (2.94), it is easy to see that

$$\varrho_{\mathrm{MC}} = 1 + \frac{M \times \mathrm{iSNR}_{\mathrm{FM}} - \mathrm{iSNR}_{\mathrm{SC}}}{1 + M \times \mathrm{iSNR}_{\mathrm{FM}}}. \tag{2.97}$$

We need to show that the quantity $M \times \mathrm{iSNR}_{\mathrm{FM}} - \mathrm{iSNR}_{\mathrm{SC}}$ is positive, i.e.,

$$\begin{aligned}M \times \mathrm{iSNR}_{\mathrm{FM}} - \mathrm{iSNR}_{\mathrm{SC}} &= \mathrm{var}\left(X\right) \left[\mathbf{d}^H \mathrm{cov}^{-1}\left(\mathbf{v}\right) \mathbf{d} - \frac{1}{\mathrm{var}\left(V_1\right)}\right] \\ &\geq 0.\end{aligned} \tag{2.98}$$

From the Cauchy-Schwarz inequality, we have

$$\left|\mathbf{i}^T \mathbf{d}\right|^2 = 1 \leq \left[\mathbf{i}^T \mathrm{cov}\left(\mathbf{v}\right) \mathbf{i}\right] \left[\mathbf{d}^H \mathrm{cov}^{-1}\left(\mathbf{v}\right) \mathbf{d}\right], \tag{2.99}$$

implying that

$$\mathbf{d}^H \mathrm{cov}^{-1}\left(\mathbf{v}\right) \mathbf{d} \geq \frac{1}{\mathrm{var}\left(V_1\right)}. \tag{2.100}$$

As a result, $\varrho_{\mathrm{MC}} \geq 1$.

Therefore, we always have $1 \leq \varrho_{\mathrm{MC}} \leq 2$. This is the fundamental difference with the single-channel case, where $\varrho_{\mathrm{SC}} = 1$, showing the compromise between noise reduction and speech distortion, while in the multichannel scenario, we can limit this distortion and have more noise reduction by using more microphones. In fact, it is easy to show that $\left|\gamma_{X|\mathbf{y}}\right|^2 \geq \left|\gamma_{X|Y}\right|^2$ and $\left|\gamma_{V|\mathbf{y}}\right|^2 \geq \left|\gamma_{V|Y}\right|^2$, meaning that the multichannel best linear estimator distorts less the desired speech and reduces more noise than the single-channel best linear estimator. One can also verify that

$$\lim_{M\to\infty} \varrho_{\mathrm{MC}} = 2. \tag{2.101}$$

It can be checked that

$$Y_1 = E\left(X\,|\mathbf{y}\right) + E\left(V_1\,|\mathbf{y}\right) \tag{2.102}$$

or, equivalently,

$$\mathbf{i} = \mathbf{h}_{X,\mathrm{W}} + \mathbf{h}_{V_1,\mathrm{W}}. \tag{2.103}$$

It is interesting to observe that the coefficients of determination can also be expressed as

$$\left|\gamma_{X|\mathbf{y}}\right|^2 = \mathbf{h}_{X,\mathrm{W}}^H \mathbf{d} \tag{2.104}$$

and

$$\left|\gamma_{V_1|\mathbf{y}}\right|^2 = \frac{\mathbf{h}_{V_1,\mathrm{W}}^H \mathrm{cov}\left(\mathbf{v}\right)\mathbf{i}}{\mathrm{var}\left(V_1\right)}, \tag{2.105}$$

which are well-known measures of the desired signal distortion and noise reduction in the multichannel case with linear estimators. Indeed, the closer $\mathbf{h}_{X,\mathrm{W}}^H \mathbf{d}$ is to 1, the less distorted the speech signal, and the closer $\mathbf{h}_{V_1,\mathrm{W}}^H \mathrm{cov}\left(\mathbf{v}\right)\mathbf{i}$ is to $\mathrm{var}\left(V_1\right)$, the more noise reduction.

From (2.85) and (2.95), we deduce for the best estimator (linear or not) that

$$\left|\gamma_{X|\mathbf{y}}\right|^2 = \frac{\mathrm{iSNR}_{\mathrm{SC}} - 1 + \varrho_{\mathrm{MC}}}{1 + \mathrm{iSNR}_{\mathrm{SC}}} \tag{2.106}$$

and

$$\left|\gamma_{V_1|\mathbf{y}}\right|^2 = \frac{\varrho_{\mathrm{MC}} \times \mathrm{iSNR}_{\mathrm{SC}} + 1 - \mathrm{iSNR}_{\mathrm{SC}}}{1 + \mathrm{iSNR}_{\mathrm{SC}}}. \tag{2.107}$$

The two previous expressions tell us the following. For a low input SNR, $\left|\gamma_{V_1|\mathbf{y}}\right|^2$ is close to 1, meaning that there is a good amount of noise reduction; however, $\left|\gamma_{X|\mathbf{y}}\right|^2$ depends mostly on $\varrho_{\mathrm{MC}} - 1$, meaning that distortion depends on the number of microphones and the distributions of X and V_1. For a large input SNR, $\left|\gamma_{X|\mathbf{y}}\right|^2$ is close to 1, meaning that there is low distortion; however, $\left|\gamma_{V_1|\mathbf{y}}\right|^2$ is close to $\varrho_{\mathrm{MC}} - 1$, meaning that noise reduction depends on the number of microphones and the distributions of X and V_1. For a large number of sensors, the effect of the distributions of X and V on the performance of the multichannel best estimator becomes negligible.

References

1. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
2. P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.
3. J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Springer Briefs in Electrical and Computer Engineering, 2011.
4. N. A. Weiss, P. T. Holmes, and M. Hardy, *A Course in Probability*. Boston: Addison-Wesley, 2005.
5. R. Steyer, "Conditional expectations: an introduction to the concept and its applications in empirical sciences," *Methodika*, vol. 2, issue 1, pp. 53–78, 1988.
6. S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall PTR, 1993.
7. R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE ICASSP*, 2002, pp. I-253–I-256.
8. R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-square error amplitude estimators for speech enhancement under the generalized gamma distribution," in *Proc. IWAENC*, 2006.
9. J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 1741–1752, Aug. 2007.
10. B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma speech priors throughout," in *Proc. IEEE ICASSP*, 2012, pp. 4033–4036.
11. J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

Chapter 3
Best Speech Enhancement Estimator in the Time Domain

In this chapter, we study the best speech enhancement estimator in the time domain. The first part focuses on the single-channel scenario, where important insights are given thanks to different kinds of correlation coefficients; in the linear case, we obtain the well-known Wiener filter whose functioning is explained within this general framework. The second part deals with the best binaural speech enhancement estimator; the approach taken here is by the reformulation of the binaural problem into a monaural one thanks to complex random variables. As a consequence, the linear case results in the widely linear Wiener filter.

3.1 Signal Model and Problem Formulation

In the first part of this chapter, we are concerned with the speech enhancement (or noise reduction) problem, in which the time-domain desired signal, x_t, with t being the discrete-time index, needs to be recovered from the noisy observation [1], [2], [3], [4]:

$$y_t = x_t + v_t, \tag{3.1}$$

where v_t is the unwanted additive noise signal, which is assumed to be independent of x_t. All signals are considered to be real, zero mean, stationary, and broadband.

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.,

$$\begin{aligned} \mathbf{y}_t &= \left[y_t \; y_{t-1} \; \cdots \; y_{t-L+1} \right]^T \\ &= \mathbf{x}_t + \mathbf{v}_t, \end{aligned} \tag{3.2}$$

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4_3

where $\mathbf{y}_t$ is a vector of length L, and $\mathbf{x}_t$ and $\mathbf{v}_t$ are defined in a similar way to $\mathbf{y}_t$.

Since x_t and v_t are independent by assumption, the covariance matrix (of size $L \times L$) of the noisy signal can be written as

$$\begin{aligned} \operatorname{cov}(\mathbf{y}_t) &= E\left(\mathbf{y}_t \mathbf{y}_t^T\right) \\ &= \operatorname{cov}(\mathbf{x}_t) + \operatorname{cov}(\mathbf{v}_t), \end{aligned} \tag{3.3}$$

where $\operatorname{cov}(\mathbf{x}_t)$ and $\operatorname{cov}(\mathbf{v}_t)$ are the covariance matrices of $\mathbf{x}_t$ and $\mathbf{v}_t$, respectively. From (3.3), we deduce that the input SNR is

$$\begin{aligned} \text{iSNR} &= \frac{\operatorname{tr}[\operatorname{cov}(\mathbf{x}_t)]}{\operatorname{tr}[\operatorname{cov}(\mathbf{v}_t)]} \\ &= \frac{\operatorname{var}(x_t)}{\operatorname{var}(v_t)}, \end{aligned} \tag{3.4}$$

where $\operatorname{var}(x_t) = E\left(x_t^2\right)$ and $\operatorname{var}(v_t) = E\left(v_t^2\right)$ are the variances of x_t and v_t, respectively.

Another important measure in the context of speech enhancement is the squared Pearson correlation coefficient (SPCC) [1], [5]. It is easy to see that the SPCC between x_t and y_t is

$$\begin{aligned} \rho_{x_t,y_t}^2 &= \frac{\operatorname{cov}^2(x_t, y_t)}{\operatorname{var}(x_t)\operatorname{var}(y_t)} \\ &= \frac{E^2(x_t y_t)}{E(x_t^2)\,E(y_t^2)} \\ &= \frac{\text{iSNR}}{1+\text{iSNR}}. \end{aligned} \tag{3.5}$$

In the same way, the SPCC between v_t and y_t is

$$\begin{aligned} \rho_{v_t,y_t}^2 &= \frac{E^2(v_t y_t)}{E(v_t^2)\,E(y_t^2)} \\ &= \frac{1}{1+\text{iSNR}}. \end{aligned} \tag{3.6}$$

As a result,

$$1 = \rho_{x_t,y_t}^2 + \rho_{v_t,y_t}^2. \tag{3.7}$$

This shows how the important correlation coefficients are naturally related to the input SNR since they also equivalently tell us how the observed signal is noisy.

As it can be guessed, in this single-channel noise reduction problem, our desired signal is x_t that we would like to estimate from the observation signal vector, $\mathbf{y}_t$, in an optimal way thanks to the best estimator.

3.2 Best Estimator

Considering the random variable x_t and the random vector $\mathbf{y}_t$, and using conditional expectations, we can decompose the variance of x_t as

$$\operatorname{var}(x_t) = E\left[\operatorname{var}(x_t|\mathbf{y}_t)\right] + \operatorname{var}\left[E(x_t|\mathbf{y}_t)\right], \tag{3.8}$$

where the outer expectation and the outer variance on the right-hand side of (3.8) are over the distribution of $\mathbf{y}_t$. This property is called the law of total variance [6]. Since variances are always nonnegative, we deduce from (3.8) that

$$0 \leq \frac{\operatorname{var}\left[E(x_t|\mathbf{y}_t)\right]}{\operatorname{var}(x_t)} \leq 1. \tag{3.9}$$

As a result, the SPCC can be generalized to

$$\begin{aligned} \rho^2_{x_t|\mathbf{y}_t} &= 1 - \frac{E\left[\operatorname{var}(x_t|\mathbf{y}_t)\right]}{\operatorname{var}(x_t)} \\ &= \frac{\operatorname{var}\left[E(x_t|\mathbf{y}_t)\right]}{\operatorname{var}(x_t)}. \end{aligned} \tag{3.10}$$

This expression is often called the coefficient of determination in the literature of statistics [7]. If x_t and $\mathbf{y}_t$ are independent, then $\operatorname{var}(x_t|\mathbf{y}_t) = E\left(x_t^2|\mathbf{y}_t\right) - E^2(x_t|\mathbf{y}_t) = \operatorname{var}(x_t)$. As a consequence, $\rho^2_{x_t|\mathbf{y}_t} = 0$. Conversely, if $\rho^2_{x_t|\mathbf{y}_t} = 0$, then $\operatorname{var}\left[E(x_t|\mathbf{y}_t)\right] = 0$, which implies that x_t and $\mathbf{y}_t$ are independent. At the other limiting case, if $x_t = y_t$, then $\operatorname{var}(y_t|\mathbf{y}_t) = 0$, which leads to $\rho^2_{x_t|\mathbf{y}_t} = 1$. In fact, for $x_t = f(y_t)$, we have $E(x_t|\mathbf{y}_t) = E\left[f(y_t)|\mathbf{y}_t\right] = f(y_t)$; as a result, $\rho^2_{x_t|\mathbf{y}_t} = 1$. This coefficient of determination, which is a direct consequence of the law of total variance and measures how close x_t is to $E(x_t|\mathbf{y}_t)$, plays a key role in the best estimator in general and in speech enhancement in particular. It can be used as a powerful performance measure in all aspects of speech enhancement as explained in great details in the rest.

In the same manner, we can decompose the variance of v_t as

$$\operatorname{var}(v_t) = E\left[\operatorname{var}(v_t|\mathbf{y}_t)\right] + \operatorname{var}\left[E(v_t|\mathbf{y}_t)\right], \tag{3.11}$$

from which we deduce the coefficient of determination:

$$\begin{aligned} \rho^2_{v_t|\mathbf{y}_t} &= 1 - \frac{E\left[\operatorname{var}(v_t|\mathbf{y}_t)\right]}{\operatorname{var}(v_t)} \\ &= \frac{\operatorname{var}\left[E(v_t|\mathbf{y}_t)\right]}{\operatorname{var}(v_t)}. \end{aligned} \tag{3.12}$$

It is well known that the best estimator of x_t in the MMSE sense is the conditional expectation of x_t given $\mathbf{y}_t$ [8], i.e.,

$$E\left(x_t\left|\mathbf{y}_t\right.\right) = z_{x_t}(\mathbf{y}_t). \tag{3.13}$$

Indeed, let

$$e_{x_t} = x_t - E\left(x_t\left|\mathbf{y}_t\right.\right) \tag{3.14}$$

be the error signal between the desired signal and its best estimator, and let $f_{x_t}(\mathbf{y}_t)$ be any (linear or nonlinear) function of $\mathbf{y}_t$. We always have

$$E\left(e_{x_t}^2\right) = E\left\{\left[x_t - z_{x_t}(\mathbf{y}_t)\right]^2\right\} \leq E\left\{\left[x_t - f_{x_t}(\mathbf{y}_t)\right]^2\right\}. \tag{3.15}$$

By virtue of the law of total expectation, the two random variables x_t and $z_{x_t}(\mathbf{y}_t)$ have the same mean, i.e.,

$$E\left[z_{x_t}(\mathbf{y}_t)\right] = E\left(x_t\right) = 0. \tag{3.16}$$

It can also be verified that the MMSE is

$$\begin{aligned} E\left(e_{x_t}^2\right) &= E\left[\mathrm{var}\left(x_t\left|\mathbf{y}_t\right.\right)\right] \\ &= \mathrm{var}\left(x_t\right) - \mathrm{var}\left[z_{x_t}(\mathbf{y}_t)\right] \\ &= \mathrm{var}\left(x_t\right)\left(1 - \rho_{x_t|\mathbf{y}_t}^2\right). \end{aligned} \tag{3.17}$$

The MMSE clearly depends on the coefficient of determination, $\rho_{x_t|\mathbf{y}_t}^2$, which can be seen as a good measure of distortion of the desired signal, x_t. The closer is $\rho_{x_t|\mathbf{y}_t}^2$ to 1, the less distorted the desired signal with the best estimator. Then, we can deduce a distortion measure, which is close to the conventional speech distortion index, i.e.,

$$\upsilon_{\mathrm{sd}} = \frac{E\left(e_{x_t}^2\right)}{\mathrm{var}\left(x_t\right)} = 1 - \rho_{x_t|\mathbf{y}_t}^2. \tag{3.18}$$

We also notice in (3.17) the law of total variance since $E\left(e_{x_t}^2\left|\mathbf{y}_t\right.\right) = \mathrm{var}\left(x_t\left|\mathbf{y}_t\right.\right)$ and $E\left[E\left(e_{x_t}^2\left|\mathbf{y}_t\right.\right)\right] = E\left(e_{x_t}^2\right)$.

In the same way, the best estimator of v_t in the MMSE sense is the conditional expectation of v_t given $\mathbf{y}_t$, i.e.,

$$E\left(v_t\left|\mathbf{y}_t\right.\right) = z_{v_t}(\mathbf{y}_t) \tag{3.19}$$

and for any (linear or nonlinear) function of $\mathbf{y}_t$, $f_{v_t}(\mathbf{y}_t)$, we always have

$$E\left(e_{v_t}^2\right) = E\left\{\left[v_t - z_{v_t}(\mathbf{y}_t)\right]^2\right\} \leq E\left\{\left[v_t - f_{v_t}(\mathbf{y}_t)\right]^2\right\}, \tag{3.20}$$

where

$$e_{v_t} = v_t - E\left(v_t\left|\mathbf{y}_t\right.\right) \tag{3.21}$$

is the error signal between the noise and its best estimator. By virtue of the law of total expectation, the two random variables v_t and $z_{v_t}(\mathbf{y}_t)$ have the same mean, i.e.,

$$E\left[z_{v_t}(\mathbf{y}_t)\right] = E\left(v_t\right) = 0. \tag{3.22}$$

We also deduce that the MMSE is

$$E\left(e_{v_t}^2\right) = \text{var}\left(v_t\right)\left(1 - \rho_{v_t|\mathbf{y}_t}^2\right). \tag{3.23}$$

The coefficient of determination, $\rho_{v_t|\mathbf{y}_t}^2$, is a good measure of noise reduction. The closer is $\rho_{v_t|\mathbf{y}_t}^2$ to 1, the more noise reduction with the best estimator.

In the pathological scenario where x_t and v_t are independent and identically distributed (i.i.d.), we have

$$E\left(x_t\,|\mathbf{y}_t\right) = E\left(v_t\,|\mathbf{y}_t\right) = \frac{y_t}{2}, \tag{3.24}$$

and $\rho_{x_t|\mathbf{y}_t}^2 = \rho_{v_t|\mathbf{y}_t}^2 = 1/2$. As a result, single-channel speech enhancement in the time domain is not feasible with the best estimator.

By adding together the best estimator of x_t and the best estimator of v_t, we obtain the observed signal, i.e.,

$$\begin{aligned} y_t &= E\left(y_t\,|\mathbf{y}_t\right) \\ &= E\left(x_t\,|\mathbf{y}_t\right) + E\left(v_t\,|\mathbf{y}_t\right). \end{aligned} \tag{3.25}$$

The above property shows that the estimation errors of both estimators cancel out. In other words, the best estimator of x_t can be found, equivalently, from the best estimator of v_t. In the best estimator of x_t, $E\left(x_t\,|\mathbf{y}_t\right)$ gives the speech distortion perspective while $y_t - E\left(v_t\,|\mathbf{y}_t\right)$ gives the noise reduction perspective. From (3.25), we easily see that $e_{x_t} = -e_{v_t}$ and, as a result, $E\left(e_{x_t}^2\right) = E\left(e_{v_t}^2\right)$. Then, equating (3.17) and (3.23), we obtain

$$\text{iSNR} + \rho_{v_t|\mathbf{y}_t}^2 = 1 + \text{iSNR} \times \rho_{x_t|\mathbf{y}_t}^2. \tag{3.26}$$

From the previous expression, we have

$$\lim_{\text{iSNR}\to 0} \rho_{v_t|\mathbf{y}_t}^2 = 1, \tag{3.27}$$

$$\lim_{\text{iSNR}\to \infty} \rho_{x_t|\mathbf{y}_t}^2 = 1. \tag{3.28}$$

In words, the best estimator is able to completely remove the noise when the input SNR is close to 0 and fully recover the desired signal when the input SNR approaches infinity. However, (3.26) does not give us any information about speech distortion in the first case and noise reduction in the second one. Using the fact that

$$\text{iSNR} = \frac{\rho^2_{x_t,y_t}}{\rho^2_{v_t,y_t}} = \frac{1-\rho^2_{v_t,y_t}}{1-\rho^2_{x_t,y_t}}, \tag{3.29}$$

we can also express (3.26) as

$$\frac{\rho^2_{x_t,y_t}}{\rho^2_{v_t,y_t}} = \frac{1-\rho^2_{v_t|\mathbf{y}_t}}{1-\rho^2_{x_t|\mathbf{y}_t}} = \text{iSNR}. \tag{3.30}$$

Property 3.1. We have

$$\rho^2_{x_t|\mathbf{y}_t} \geq \rho^2_{x_t,y_t}, \tag{3.31}$$

$$\rho^2_{v_t|\mathbf{y}_t} \geq \rho^2_{v_t,y_t}. \tag{3.32}$$

As a consequence,

$$\rho^2_{x_t|\mathbf{y}_t} + \rho^2_{v_t|\mathbf{y}_t} \geq 1. \tag{3.33}$$

Proof. Let us consider an estimate of the desired signal that is proportional to the observation, i.e.,

$$\overline{z}_{x_t}(\mathbf{y}_t) = \alpha \frac{\sqrt{\text{var}\,(x_t)}}{\sqrt{\text{var}\,(y_t)}} y_t, \tag{3.34}$$

where $\alpha \neq 0$ is an arbitrary real number. In this case the MSE is

$$\begin{aligned} E\left(\overline{e}^2_{x_t}\right) &= E\left\{\left[x_t - \overline{z}_{x_t}(\mathbf{y}_t)\right]^2\right\} \\ &= \text{var}\,(x_t)\left[\left(1+\alpha^2\right) - 2\alpha\rho_{x_t,y_t}\right]. \end{aligned} \tag{3.35}$$

Since $E\left(\overline{e}^2_{x_t}\right) \geq E\left(e^2_{x_t}\right)$, we deduce that

$$\rho^2_{x_t|\mathbf{y}_t} \geq -\alpha^2 + 2\alpha\rho_{x_t,y_t}. \tag{3.36}$$

For the particular value of $\alpha = \rho_{x_t,y_t}$ in the previous expression, we find that $\rho^2_{x_t|\mathbf{y}_t} \geq \rho^2_{x_t,y_t}$. We can use a very similar proof to show the inequality in (3.32).

The above suggests that we can combine the speech distortion and noise reduction measures into one convenient measure:

$$\varrho = \rho^2_{x_t|\mathbf{y}_t} + \rho^2_{v_t|\mathbf{y}_t}, \tag{3.37}$$

where $1 \leq \varrho \leq 2$. Fundamentally, ϱ measures the compromise between speech distortion and noise reduction. For ϱ close to 2, we have the almost perfect estimator with the best estimator in the sense that the noise is almost all removed and speech distortion is almost nonexistent. For $\varrho = 1$, the observed signal is fundamentally not affected; this will happen only when speech and

noise are i.i.d., so that $\rho^2_{x_t|\mathbf{y}_t} = \rho^2_{v_t|\mathbf{y}_t} = 1/2$. From (3.26) and (3.37), we deduce for the best estimator that

$$\begin{aligned}\rho^2_{x_t|\mathbf{y}_t} &= \frac{\varrho - 1 + \mathrm{iSNR}}{1 + \mathrm{iSNR}} \\ &= \rho^2_{v_t,y_t}\,(\varrho - 1) + \rho^2_{x_t,y_t}\end{aligned} \tag{3.38}$$

and

$$\begin{aligned}\rho^2_{v_t|\mathbf{y}_t} &= \frac{(\varrho - 1)\,\mathrm{iSNR} + 1}{1 + \mathrm{iSNR}} \\ &= \rho^2_{x_t,y_t}\,(\varrho - 1) + \rho^2_{v_t,y_t}.\end{aligned} \tag{3.39}$$

The two previous expressions tell us the following. For a low input SNR, $\rho^2_{v_t|\mathbf{y}_t}$ is close to 1, meaning that there is a good amount of noise reduction; however, $\rho^2_{x_t|\mathbf{y}_t}$ depends mostly on $\varrho - 1$, meaning that distortion depends on the distributions of x_t and v_t. For a large input SNR, $\rho^2_{x_t|\mathbf{y}_t}$ is close to 1, meaning that there is low distortion; however, $\rho^2_{v_t|\mathbf{y}_t}$ is close to $\varrho - 1$, meaning that noise reduction depends on the distributions of x_t and v_t.

While the SPCCs $\rho^2_{x_t,y_t}$ and $\rho^2_{v_t,y_t}$ give a very good indication on the state of the noisy signal (since they are related to the input SNR), the coefficients of determination $\rho^2_{x_t|\mathbf{y}_t}$ and $\rho^2_{v_t|\mathbf{y}_t}$, as well as ϱ give a very good indication on the enhanced noisy signal with the best estimator since $\rho^2_{x_t|\mathbf{y}_t}$ and $\rho^2_{v_t|\mathbf{y}_t}$ are good measures of speech distortion and noise reduction, respectively, and ϱ is a good measure on the compromise between the two.

From Property 3.1, we can also define the gain in SNR of the best estimator:

$$\begin{aligned}\mathcal{G} &= \frac{\mathrm{oSNR}}{\mathrm{iSNR}} \\ &= \left(\frac{\rho^2_{x_t|\mathbf{y}_t}}{1 - \rho^2_{v_t|\mathbf{y}_t}}\right)^2 \geq 1,\end{aligned} \tag{3.40}$$

where oSNR is the output SNR of the best estimator[1]. This definition of the gain in SNR is justified by the facts that $\mathcal{G}$ is always greater than or equal to 1 and for i.i.d. speech and noise, $\mathcal{G} = 1$. Using (3.30), we easily see that the output SNR in (3.40) is

[1] It is important to keep in mind that a rigorous definition of the output SNR of the best estimator in general may not be possible since the output SNR is a second-order measure while the best estimator depends on distributions. This is why the coefficients of determination may be the most natural and reliable measures in this context.

$$\begin{aligned}\text{oSNR} &= \text{iSNR} \times \left(\frac{\rho^2_{x_t|\mathbf{y}_t}}{1-\rho^2_{v_t|\mathbf{y}_t}} \right)^2 \\ &= \frac{\rho^4_{x_t|\mathbf{y}_t}}{\left(1-\rho^2_{v_t|\mathbf{y}_t}\right)\left(1-\rho^2_{x_t|\mathbf{y}_t}\right)} \geq \text{iSNR}. \end{aligned} \tag{3.41}$$

One can check from the previous expression that when speech is large as compared to noise, the output SNR is also large, and when speech is small as compared to noise, the output SNR is also small; this is consistent with the definition of the output SNR. In this context, we define the speech reduction factor and the noise reduction factor as, respectively,

$$\xi_{\text{sr}} = \frac{1}{\rho^4_{x_t|\mathbf{y}_t}} \tag{3.42}$$

and

$$\xi_{\text{nr}} = \frac{1}{\left(1-\rho^2_{v_t|\mathbf{y}_t}\right)^2}. \tag{3.43}$$

As a consequence, we deduce the fundamental relationship for the best estimator:

$$\frac{\xi_{\text{nr}}}{\xi_{\text{sr}}} = \frac{\text{oSNR}}{\text{iSNR}}, \tag{3.44}$$

which is well known in the linear case. This is an even more insightful way to explain the compromise between noise reduction and speech distortion with more intuitive measures. It is quite remarkable that these measures, which should resemble the conventional ones with linear filtering, are derived in such a simple way for the whole class of best estimators (linear and nonlinear).

Another interesting measure is the conditional correlation coefficient (CCC). The CCC between x_t and v_t given $\mathbf{y}_t$ is

$$\rho_{x_t,v_t|\mathbf{y}_t} = \frac{\text{cov}\left(x_t, v_t|\mathbf{y}_t\right)}{\sqrt{\text{var}\left(x_t|\mathbf{y}_t\right)\text{var}\left(v_t|\mathbf{y}_t\right)}}, \tag{3.45}$$

where

$$\begin{aligned}\text{cov}\left(x_t, v_t|\mathbf{y}_t\right) &= E\left\{\left[x_t - E\left(x_t|\mathbf{y}_t\right)\right]\left[v_t - E\left(v_t|\mathbf{y}_t\right)\right]|\mathbf{y}_t\right\} \\ &= E\left[e_{x_t} e_{v_t}|\mathbf{y}_t\right] \\ &= -\text{var}\left(x_t|\mathbf{y}_t\right) \\ &= -\text{var}\left(v_t|\mathbf{y}_t\right). \end{aligned} \tag{3.46}$$

Therefore, we deduce that

$$\rho_{x_t,v_t|\mathbf{y}_t} = -1. \tag{3.47}$$

While $\rho_{x_t,v_t} = 0$, the magnitude of the CCC, $|\rho_{x_t,v_t|\mathbf{y}_t}|$, is maximized; this is due to the fact that the best estimators of x_t and v_t are conditionally fully correlated. The minus sign in (3.47) comes from the fact that $e_{x_t} = -e_{v_t}$.

3.3 Best Linear Estimator

The best linear estimator is a very important and extremely useful particular case of the best estimator in general. It is well known that the best linear estimator of x_t in the MMSE sense is

$$E\left(x_t\,|\mathbf{y}_t\right) = \mathbf{h}_{x_t,\mathrm{W}}^T \mathbf{y}_t, \tag{3.48}$$

where

$$\mathbf{h}_{x_t,\mathrm{W}} = \mathrm{cov}^{-1}\left(\mathbf{y}_t\right)\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{i} \tag{3.49}$$

is the classical single-channel Wiener filter in the time domain [1], with $\mathbf{i}$ being the first column of the $L \times L$ identity matrix $\mathbf{I}_L$. We deduce that the square of the coefficient of determination is

$$\begin{aligned}\rho_{x_t|\mathbf{y}_t}^4 &= \left[\frac{\mathbf{i}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathrm{cov}^{-1}\left(\mathbf{y}_t\right)\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{i}}{\mathrm{var}\left(x_t\right)}\right]^2 \\ &= \frac{\mathbf{h}_{x_t,\mathrm{W}}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{i}\mathbf{i}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{h}_{x_t,\mathrm{W}}}{\mathrm{var}^2\left(x_t\right)} = \xi_{\mathrm{sr}}^{-1},\end{aligned} \tag{3.50}$$

which is a good measure of speech distortion. This measure is very similar to the inverse of the conventional speech reduction factor [1]:

$$\xi_{\mathrm{sr}}^{-1}\left(\mathbf{h}_{x_t,\mathrm{W}}\right) = \frac{\mathbf{h}_{x_t,\mathrm{W}}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{h}_{x_t,\mathrm{W}}}{\mathrm{var}\left(x_t\right)}. \tag{3.51}$$

From the Cauchy-Schwarz inequality, i.e.,

$$\left[\mathbf{h}_{x_t,\mathrm{W}}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{i}\right]^2 \leq \mathbf{h}_{x_t,\mathrm{W}}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{h}_{x_t,\mathrm{W}} \times \mathrm{var}\left(x_t\right), \tag{3.52}$$

it results that

$$\xi_{\mathrm{sr}} \geq \xi_{\mathrm{sr}}\left(\mathbf{h}_{x_t,\mathrm{W}}\right). \tag{3.53}$$

The vector $\mathbf{x}_t$ can be decomposed into two orthogonal components; one proportional to the desired signal, x_t, and the other to what we may consider as an interference [4]:

$$\mathbf{x}_t = x_t \boldsymbol{\gamma}_{x_t} + \mathbf{x}_{t,\mathrm{i}}, \tag{3.54}$$

where

$$\boldsymbol{\gamma}_{x_t} = \frac{\mathrm{cov}\,(\mathbf{x}_t)\,\mathbf{i}}{\mathrm{var}\,(x_t)} \tag{3.55}$$

is the normalized correlation vector between between $\mathbf{x}_t$ and x_t,

$$\mathbf{x}_{t,\mathrm{i}} = \mathbf{x}_t - x_t \boldsymbol{\gamma}_{x_t} \tag{3.56}$$

is the interference signal vector, and

$$E\,(\mathbf{x}_{t,\mathrm{i}} x_t) = \mathbf{0}. \tag{3.57}$$

Obviously, we can express the square of the coefficient of determination in (3.50) as

$$\rho^4_{x_t|\mathbf{y}_t} = \left(\mathbf{h}^T_{x_t,\mathrm{W}} \boldsymbol{\gamma}_{x_t}\right)^2, \tag{3.58}$$

which may give a better perspective on distortion since when $\mathbf{h}^T_{x_t,\mathrm{W}} \boldsymbol{\gamma}_{x_t}$ is close to 1, the desired signal, x_t, is well recovered. If $\mathrm{cov}\,(\mathbf{x}_t)$ is of rank 1, i.e., $\mathbf{x}_{t,\mathrm{i}} = \mathbf{0}$, then $\mathrm{cov}\,(\mathbf{x}_t) = \mathrm{var}\,(x_t)\,\boldsymbol{\gamma}_{x_t} \boldsymbol{\gamma}^T_{x_t}$ and $\mathbf{h}_{x_t,\mathrm{W}} = \mathrm{var}\,(x_t)\,\mathrm{cov}^{-1}\,(\mathbf{y}_t)\,\boldsymbol{\gamma}_{x_t}$. As a consequence,

$$\xi_{\mathrm{sr}} = \xi_{\mathrm{sr}}\,(\mathbf{h}_{x_t,\mathrm{W}})\,. \tag{3.59}$$

Also, the best linear estimator of v_t in the MMSE sense is

$$E\,(v_t\,|\mathbf{y}_t\,) = \mathbf{h}^T_{v_t,\mathrm{W}} \mathbf{y}_t, \tag{3.60}$$

where

$$\mathbf{h}_{v_t,\mathrm{W}} = \mathrm{cov}^{-1}\,(\mathbf{y}_t)\,\mathrm{cov}\,(\mathbf{v}_t)\,\mathbf{i} \tag{3.61}$$

is the Wiener filter for the estimation of v_t. Then, the coefficient of determination is

$$\begin{aligned} \rho^2_{v_t|\mathbf{y}_t} &= \frac{\mathbf{i}^T \mathrm{cov}\,(\mathbf{v}_t)\,\mathrm{cov}^{-1}\,(\mathbf{y}_t)\,\mathrm{cov}\,(\mathbf{v}_t)\,\mathbf{i}}{\mathrm{var}\,(v_t)} \\ &= \frac{\mathbf{h}^T_{v_t,\mathrm{W}} \mathrm{cov}\,(\mathbf{v}_t)\,\mathbf{i}}{\mathrm{var}\,(v_t)}, \end{aligned} \tag{3.62}$$

which is a good measure of noise reduction. It is not hard to see that

$$\mathbf{i} = \mathbf{h}_{x_t,\mathrm{W}} + \mathbf{h}_{v_t,\mathrm{W}}. \tag{3.63}$$

Therefore, (3.62) can be rewritten as

$$\left(1-\rho^2_{v_t|\mathbf{y}_t}\right)^2 = \frac{\mathbf{h}^T_{x_t,\mathrm{W}}\mathrm{cov}\left(\mathbf{v}_t\right)\mathbf{i}\mathbf{i}^T\mathrm{cov}\left(\mathbf{v}_t\right)\mathbf{h}_{x_t,\mathrm{W}}}{\mathrm{var}^2\left(v_t\right)} = \xi^{-1}_{\mathrm{nr}}. \tag{3.64}$$

The measure $\left(1-\rho^2_{v_t|\mathbf{y}_t}\right)^2$ is very similar to the inverse of the conventional noise reduction factor [1]:

$$\xi^{-1}_{\mathrm{nr}}\left(\mathbf{h}_{x_t,\mathrm{W}}\right) = \frac{\mathbf{h}^T_{x_t,\mathrm{W}}\mathrm{cov}\left(\mathbf{v}_t\right)\mathbf{h}_{x_t,\mathrm{W}}}{\mathrm{var}\left(v_t\right)} \tag{3.65}$$

and one can verify that

$$\xi_{\mathrm{nr}} \geq \xi_{\mathrm{nr}}\left(\mathbf{h}_{x_t,\mathrm{W}}\right). \tag{3.66}$$

Using our definition of the output SNR in (3.41) of the best estimator, we have

$$\mathrm{oSNR} = \frac{\mathbf{h}^T_{x_t,\mathrm{W}}\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{i}\mathbf{i}^T\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{h}_{x_t,\mathrm{W}}/\mathrm{var}\left(x_t\right)}{\mathbf{h}^T_{x_t,\mathrm{W}}\mathrm{cov}\left(\mathbf{v}_t\right)\mathbf{i}\mathbf{i}^T\mathrm{cov}\left(\mathbf{v}_t\right)\mathbf{h}_{x_t,\mathrm{W}}/\mathrm{var}\left(v_t\right)}, \tag{3.67}$$

which resembles the conventional output SNR:

$$\mathrm{oSNR}\left(\mathbf{h}_{x_t,\mathrm{W}}\right) = \frac{\mathbf{h}^T_{x_t,\mathrm{W}}\mathrm{cov}\left(\mathbf{x}_t\right)\mathbf{h}_{x_t,\mathrm{W}}}{\mathbf{h}^T_{x_t,\mathrm{W}}\mathrm{cov}\left(\mathbf{v}_t\right)\mathbf{h}_{x_t,\mathrm{W}}}. \tag{3.68}$$

Now, if we compute the SPCC between x_t and $\mathbf{h}^T_{x_t,\mathrm{W}}\mathbf{y}_t$, and the the SPCC between v_t and $\mathbf{h}^T_{v_t,\mathrm{W}}\mathbf{y}_t$, it is easy to verify that

$$\rho^2_{x_t,\mathbf{h}^T_{x_t,\mathrm{W}}\mathbf{y}_t} = \rho^2_{x_t|\mathbf{y}_t},$$
$$\rho^2_{v_t,\mathbf{h}^T_{v_t,\mathrm{W}}\mathbf{y}_t} = \rho^2_{v_t|\mathbf{y}_t}.$$

Let us open a short parenthesis on the so-called partial correlation coefficient (PCC), whose function is to evaluate the correlation between two variables after eliminating the effect of another variable on these two variables. The PCC between x_t and v_t with respect to $\mathbf{y}_t$, denoted $\rho_{x_t,v_t\cdot\mathbf{y}_t}$, is computed in two steps. In the first step, we find the two filters $\mathbf{h}_{x_t}$ and $\mathbf{h}_{v_t}$ that minimize the error signals $e_{x_t} = x_t - \mathbf{h}^T_{x_t}\mathbf{y}_t$ and $e_{v_t} = v_t - \mathbf{h}^T_{v_t}\mathbf{y}_t$, respectively. We get the Wiener filters $\mathbf{h}_{x_t,\mathrm{W}}$ and $\mathbf{h}_{v_t,\mathrm{W}}$. Substituting these filters back into the errors, we obtain the two residuals $e_{x_t,\mathrm{W}}$ and $e_{v_t,\mathrm{W}}$. Then, in the second step we compute the correlation coefficient between $e_{x_t,\mathrm{W}}$ and $e_{v_t,\mathrm{W}}$, i.e.,

$$\rho_{x_t,v_t\cdot\mathbf{y}_t} = \frac{E\left(e_{x_t,\mathrm{W}}e_{v_t,\mathrm{W}}\right)}{\sqrt{E\left(e_{x_t,\mathrm{W}}^2\right)E\left(e_{v_t,\mathrm{W}}^2\right)}}. \tag{3.69}$$

It is easy to check that (3.69) simplifies to

$$\rho_{x_t,v_t\cdot\mathbf{y}_t} = -1, \tag{3.70}$$

showing that CCC and PCC are strictly equivalent in the linear case.

3.4 Generalization to the Binaural Case

3.4.1 Problem Formulation

In binaural speech enhancement, we need to extract two "clean" signals from the sensor array that will be delivered to the left and right ears of a human subject.

Without loss of generality, we consider the signal model in which an array consisting of $2M$ sensors capture a source (speech) signal convolved with acoustic impulse responses in some noise field. The signal received at the ith sensor is then expressed as [9]

$$\begin{aligned} y'_{t,i} &= g'_{t,i} * x_t + v'_{t,i} \\ &= x'_{t,i} + v'_{t,i},\ i = 1, 2, \ldots, 2M, \end{aligned} \tag{3.71}$$

where $g'_{t,i}$ is the acoustic impulse response from the unknown desired source, x_t, location to the ith sensor, $*$ stands for linear convolution, and $v'_{t,i}$ is the additive noise at sensor i. We assume that the impulse responses are time invariant and that the signals $x'_{t,i} = g'_{t,i} * x_t$ and $v'_{t,i}$ are mutually independent, zero mean, real, broadband, and stationary.

Since we are interested in binaural estimation, it is more convenient to work in the complex domain in order that the original (binaural) problem is transformed into the conventional (monaural) noise reduction processing with a sensor array [10]. In other words, instead of having two real-valued outputs, we will have one complex-valued output. Indeed, from the $2M$ real-valued microphone signals given in (3.71), we can artificially build M complex-valued sensor signals as

$$\begin{aligned} y_{t,m} &= y'_{t,m} + \jmath y'_{t,M+m} \\ &= x_{t,m} + v_{t,m},\ m = 1, 2, \ldots, M, \end{aligned} \tag{3.72}$$

where

$$x_{t,m} = x'_{t,m} + \jmath x'_{t,M+m},\ m = 1, 2, \ldots, M \tag{3.73}$$

is the complex desired speech signal and

$$v_{t,m} = v'_{t,m} + \jmath v'_{t,M+m},\ m = 1, 2, \ldots, M \tag{3.74}$$

is the complex additive noise at the complex sensor m.

It is customary to work with blocks of L successive time samples, i.e.,

$$\begin{aligned} \mathbf{y}_{t,m} &= \left[y_{t,m}\ y_{t-1,m}\ \cdots\ y_{t-L+1,m} \right]^T \\ &= \mathbf{x}_{t,m} + \mathbf{v}_{t,m},\ m = 1, 2, \ldots, M, \end{aligned} \tag{3.75}$$

where $\mathbf{x}_{t,m}$ and $\mathbf{v}_{t,m}$ are defined in a similar way to $\mathbf{y}_{t,m}$. Concatenating all the observations together, we get the vector of length ML:

$$\begin{aligned} \underline{\mathbf{y}}_t &= \left[\mathbf{y}_{t,1}^T\ \mathbf{y}_{t,2}^T\ \cdots\ \mathbf{y}_{t,M}^T \right]^T \\ &= \underline{\mathbf{x}}_t + \underline{\mathbf{v}}_t, \end{aligned} \tag{3.76}$$

where $\underline{\mathbf{x}}_t$ and $\underline{\mathbf{v}}_t$ are also concatenated vectors of $\mathbf{x}_{t,m}$ and $\mathbf{v}_{t,m}$, respectively. We deduce that the $ML \times ML$ covariance matrix of $\underline{\mathbf{y}}_t$ is

$$\begin{aligned} \mathrm{cov}\left(\underline{\mathbf{y}}_t\right) &= E\left(\underline{\mathbf{y}}_t \underline{\mathbf{y}}_t^H\right) \\ &= \mathrm{cov}\left(\underline{\mathbf{x}}_t\right) + \mathrm{cov}\left(\underline{\mathbf{v}}_t\right), \end{aligned} \tag{3.77}$$

where $\mathrm{cov}\left(\underline{\mathbf{x}}_t\right)$ and $\mathrm{cov}\left(\underline{\mathbf{v}}_t\right)$ are the covariance matrices of $\underline{\mathbf{x}}_t$ and $\underline{\mathbf{v}}_t$, respectively.

Obviously, from the model given in (3.72), we deal with complex random variables (CRVs) and it can be verified that, in general, $x_{t,m}$ and $v_{t,m}$ are highly noncircular CRVs [11]. Let a be a zero-mean CRV, a good measure of the second-order circularity is the circularity quotient [12] defined as the ratio between the pseudo-variance and the variance of a, i.e.,

$$\gamma_a = \frac{E\left(a^2\right)}{E\left(|a|^2\right)}. \tag{3.78}$$

This measure coincides with the coherence function between a and a^*. Since $x_{t,m}$ and/or $v_{t,m}$ are noncircular CRVs, the vector $\underline{\mathbf{y}}_t^*$ should also be included as part of the observations [13], [14]. Therefore, we define the augmented observation vector of length $2ML$ as

$$\begin{aligned} \widetilde{\underline{\mathbf{y}}}_t &= \begin{bmatrix} \underline{\mathbf{y}}_t \\ \underline{\mathbf{y}}_t^* \end{bmatrix} \\ &= \widetilde{\underline{\mathbf{x}}}_t + \widetilde{\underline{\mathbf{v}}}_t, \end{aligned} \tag{3.79}$$

where $\underline{\widetilde{\mathbf{x}}}_t$ and $\underline{\widetilde{\mathbf{v}}}_t$ are defined similarly to $\underline{\widetilde{\mathbf{y}}}_t$. We deduce that the $2ML \times 2ML$ covariance matrix of $\underline{\widetilde{\mathbf{y}}}_t$ is

$$\operatorname{cov}\left(\underline{\widetilde{\mathbf{y}}}_t\right) = \operatorname{cov}\left(\underline{\widetilde{\mathbf{x}}}_t\right) + \operatorname{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right), \tag{3.80}$$

where $\operatorname{cov}\left(\underline{\widetilde{\mathbf{x}}}_t\right)$ and $\operatorname{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right)$ are the covariance matrices of $\underline{\widetilde{\mathbf{x}}}_t$ and $\underline{\widetilde{\mathbf{v}}}_t$, respectively.

In the rest, we consider the first complex sensor signal, i.e., $y_{t,1}$, as the reference. Therefore, our aim is to recover the complex desired speech signal, $x_{t,1}$, from the augmented complex observation vector, $\underline{\widetilde{\mathbf{y}}}_t$, in the best possible way. Using this reference, we define the input SNR as

$$\text{iSNR} = \frac{\operatorname{var}\left(x_{t,1}\right)}{\operatorname{var}\left(v_{t,1}\right)}, \tag{3.81}$$

where $\operatorname{var}\left(x_{t,1}\right) = E\left(\left|x_{t,1}\right|^2\right)$ and $\operatorname{var}\left(v_{t,1}\right) = E\left(\left|v_{t,1}\right|^2\right)$ are the variances of $x_{t,1}$ and $v_{t,1}$, respectively.

Another perspective in the context of binaural speech enhancement is from the magnitude squared Pearson correlation coefficient (MSPCC) [1], [5]. It is easy to see that the MSPCC between $x_{t,1}$ and $y_{t,1}$ is

$$\begin{aligned}\left|\rho_{x_{t,1},y_{t,1}}\right|^2 &= \frac{\left|E\left(x_{t,1}y_{t,1}^*\right)\right|^2}{E\left(\left|x_{t,1}\right|^2\right)E\left(\left|y_{t,1}\right|^2\right)} \\ &= \frac{\text{iSNR}}{1+\text{iSNR}}.\end{aligned} \tag{3.82}$$

In the same way, the MSPCC between $v_{t,1}$ and $y_{t,1}$ is

$$\begin{aligned}\left|\rho_{v_{t,1},y_{t,1}}\right|^2 &= \frac{\left|E\left(v_{t,1}y_{t,1}^*\right)\right|^2}{E\left(\left|v_{t,1}\right|^2\right)E\left(\left|y_{t,1}\right|^2\right)} \\ &= \frac{1}{1+\text{iSNR}}.\end{aligned} \tag{3.83}$$

As a result,

$$1 = \left|\rho_{x_{t,1},y_{t,1}}\right|^2 + \left|\rho_{v_{t,1},y_{t,1}}\right|^2. \tag{3.84}$$

This shows how the important correlations are related to the input SNR.

3.4.2 Best Estimator

Considering the CRV $x_{t,1}$ and the complex random vector $\widetilde{\underline{\mathbf{y}}}_t$, and using conditional expectations, we can express the law of total variance [6] with respect to $x_{t,1}$ as

$$\operatorname{var}(x_{t,1}) = E\left[\operatorname{var}\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right] + \operatorname{var}\left[E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right], \tag{3.85}$$

where the outer expectation and the outer variance on the right-hand side of (3.85) are over the distribution of $\widetilde{\underline{\mathbf{y}}}_t$. Since variances are always nonnegative, we deduce from (3.85) that

$$0 \leq \frac{\operatorname{var}\left[E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right]}{\operatorname{var}(x_{t,1})} \leq 1. \tag{3.86}$$

Therefore, the MSPCC can be generalized to

$$\begin{aligned}\left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 &= 1 - \frac{E\left[\operatorname{var}\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right]}{\operatorname{var}(x_{t,1})} \\ &= \frac{\operatorname{var}\left[E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right]}{\operatorname{var}(x_{t,1})}.\end{aligned} \tag{3.87}$$

This expression is often called the coefficient of determination in the literature of statistics [7]. If $x_{t,1}$ and $\widetilde{\underline{\mathbf{y}}}_t$ are independent, then $\operatorname{var}\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) = \operatorname{var}(x_{t,1})$. As a consequence, $\left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 = 0$. Conversely, if $\left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 = 0$, then $\operatorname{var}\left[E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right] = 0$, which implies that $x_{t,1}$ and $\widetilde{\underline{\mathbf{y}}}_t$ are independent. At the other limiting case, if $x_{t,1} = y_{t,1}$, then $\operatorname{var}\left(y_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) = 0$, which leads to $\left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 = 1$. In fact, for $x_{t,1} = f(y_{t,1}, y^*_{t,1})$, we have $E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) = E\left[f(y_{t,1}, y^*_{t,1})\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right] = f(y_{t,1}, y^*_{t,1})$; as a result, $\left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 = 1$. This coefficient, which is a direct consequence of the law of total variance and measures how close $x_{t,1}$ is to $E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)$, plays a key role in the best estimator in general and in binaural speech enhancement in particular. It can be used as a powerful performance measure in all aspects of binaural speech enhancement as explained in the rest.

In the same manner, we can express the variance of $v_{t,1}$ as

$$\operatorname{var}(v_{t,1}) = E\left[\operatorname{var}\left(v_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right] + \operatorname{var}\left[E\left(v_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)\right], \tag{3.88}$$

from which we deduce the coefficient of determination:

$$\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 = 1 - \frac{E\left[\mathrm{var}\left(v_{t,1}\left|\underline{\widetilde{\mathbf{y}}}_t\right.\right)\right]}{\mathrm{var}\left(v_{t,1}\right)} = \frac{\mathrm{var}\left[E\left(v_{t,1}\left|\underline{\widetilde{\mathbf{y}}}_t\right.\right)\right]}{\mathrm{var}\left(v_{t,1}\right)}. \tag{3.89}$$

The best estimator of $x_{t,1}$ in the MMSE sense is well known to be the conditional expectation of $x_{t,1}$ given $\underline{\widetilde{\mathbf{y}}}_t$ [8], i.e.,

$$E\left(x_{t,1}\left|\underline{\widetilde{\mathbf{y}}}_t\right.\right) = z_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right). \tag{3.90}$$

Indeed, let $f_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)$ be any (linear or nonlinear) function of $\underline{\widetilde{\mathbf{y}}}_t$, we always have

$$\begin{aligned} E\left(\left|e_{x_{t,1}}\right|^2\right) &= E\left[\left|x_{t,1} - z_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)\right|^2\right] \\ &\leq E\left[\left|x_{t,1} - f_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)\right|^2\right], \end{aligned} \tag{3.91}$$

where $e_{x_{t,1}} = x_{t,1} - E\left(x_{t,1}\left|\underline{\widetilde{\mathbf{y}}}_t\right.\right)$ is the error signal between the desired signal and its best estimator. By virtue of the law of total expectation, the two random variables $x_{t,1}$ and $z_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)$ have the same mean, i.e.,

$$E\left[z_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)\right] = E\left(x_{t,1}\right) = 0. \tag{3.92}$$

It can also be verified that the MMSE is

$$\begin{aligned} E\left(\left|e_{x_{t,1}}\right|^2\right) &= E\left[\mathrm{var}\left(x_{t,1}\left|\underline{\widetilde{\mathbf{y}}}_t\right.\right)\right] \\ &= \mathrm{var}\left(x_{t,1}\right) - \mathrm{var}\left[z_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)\right] \\ &= \mathrm{var}\left(x_{t,1}\right)\left(1 - \left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2\right). \end{aligned} \tag{3.93}$$

The MMSE clearly depends on the coefficient of determination, $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$, which can be seen as a good measure of distortion of the desired signal, $x_{t,1}$. The closer is $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ to 1, the less distorted the desired signal with the best binaural estimator.

In the same way, the best estimator of $v_{t,1}$ in the MMSE sense is the conditional expectation of $v_{t,1}$ given $\underline{\widetilde{\mathbf{y}}}_t$, i.e.,

$$E\left(v_{t,1}\left|\underline{\widetilde{\mathbf{y}}}_t\right.\right) = z_{v_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right) \tag{3.94}$$

and for any (linear or nonlinear) function of $\widetilde{\underline{\mathbf{y}}}_t$, $f_{v_{t,1}}\left(\widetilde{\underline{\mathbf{y}}}_t\right)$, we always have

$$\begin{aligned} E\left(\left|e_{v_{t,1}}\right|^2\right) &= E\left[\left|v_{t,1} - z_{v_{t,1}}\left(\widetilde{\underline{\mathbf{y}}}_t\right)\right|^2\right] \\ &\leq E\left[\left|v_{t,1} - f_{v_{t,1}}\left(\widetilde{\underline{\mathbf{y}}}_t\right)\right|^2\right], \end{aligned} \tag{3.95}$$

where $e_{v_{t,1}} = v_{t,1} - E\left(v_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)$ is the error signal between the noise and its best estimator. By virtue of the law of total expectation, the two random variables $v_{t,1}$ and $z_{v_{t,1}}\left(\widetilde{\underline{\mathbf{y}}}_t\right)$ have the same mean, i.e.,

$$E\left[z_{v_{t,1}}\left(\widetilde{\underline{\mathbf{y}}}_t\right)\right] = E\left(v_{t,1}\right) = 0. \tag{3.96}$$

We also deduce that the MMSE is

$$E\left(\left|e_{v_{t,1}}\right|^2\right) = \text{var}\left(v_{t,1}\right)\left(1 - \left|\rho_{v_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2\right). \tag{3.97}$$

The coefficient of determination, $\left|\rho_{v_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2$, is a good measure of noise reduction. The closer is $\left|\rho_{v_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2$ to 1, the more noise reduction with the best binaural estimator.

Now, if we add together the best estimator of $x_{t,1}$ and the best estimator of $v_{t,1}$, we obtain the observed signal, i.e.,

$$\begin{aligned} y_{t,1} &= E\left(y_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) \\ &= E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) + E\left(v_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right). \end{aligned} \tag{3.98}$$

The above property shows that the estimation errors of both estimators cancel out. In other words, the best estimator of $x_{t,1}$ can be found, equivalently, from the best estimator of $v_{t,1}$. In the best estimator of $x_{t,1}$, $E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)$ gives the speech distortion perspective while $y_{t,1} - E\left(v_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right)$ gives the noise reduction perspective. From (3.98), we easily see that $e_{x_{t,1}} = -e_{v_{t,1}}$ and, as a result, $E\left(\left|e_{x_{t,1}}\right|^2\right) = E\left(\left|e_{v_{t,1}}\right|^2\right)$. Then, equating (3.93) and (3.97), we obtain

$$\text{iSNR} + \left|\rho_{v_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 = 1 + \text{iSNR} \times \left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2. \tag{3.99}$$

From the previous expression, we have

$$\lim_{\text{iSNR}\to 0} \left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 = 1, \tag{3.100}$$

$$\lim_{\text{iSNR}\to \infty} \left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 = 1. \tag{3.101}$$

In words, the best binaural estimator is able to completely remove the noise when the input SNR is close to 0 and fully recover the desired signal when the input SNR approaches infinity. However, (3.99) does not give us any information about speech distortion in the first case and noise reduction in the second one. Using the fact that

$$\text{iSNR} = \frac{\left|\rho_{x_{t,1},y_{t,1}}\right|^2}{\left|\rho_{v_{t,1},y_{t,1}}|\right|^2}, \tag{3.102}$$

we can also express (3.99) as

$$\frac{1-\left|\rho_{x_{t,1},y_{t,1}}\right|^2}{1-\left|\rho_{v_{t,1},y_{t,1}}\right|^2} = \frac{1-\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2}{1-\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2}. \tag{3.103}$$

Property 3.2. We have

$$\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 \geq \left|\rho_{x_{t,1},y_{t,1}}\right|^2, \tag{3.104}$$

$$\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 \geq \left|\rho_{x_{t,1},y_{t,1}}\right|^2. \tag{3.105}$$

As a consequence,

$$\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 + \left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 \geq 1. \tag{3.106}$$

Proof. Let us consider an estimate of the desired signal that is proportional to the observed signal at the reference sensor, i.e.,

$$\overline{z}_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right) = \alpha \frac{\sqrt{\text{var}\left(x_{t,1}\right)}}{\sqrt{\text{var}\left(y_{t,1}\right)}} y_{t,1}, \tag{3.107}$$

where $\alpha \neq 0$ is an arbitrary complex number. In this case the MSE is

$$\begin{aligned} E\left(\left|\overline{e}_{x_{t,1}}\right|^2\right) &= E\left[\left|x_{t,1} - \overline{z}_{x_{t,1}}\left(\underline{\widetilde{\mathbf{y}}}_t\right)\right|^2\right] \\ &= \text{var}\left(x_{t,1}\right)\left[\left(1+|\alpha|^2\right) - \left(\alpha+\alpha^*\right)\rho_{x_{t,1},y_{t,1}}\right]. \end{aligned} \tag{3.108}$$

Since $E\left(\left|\overline{e}_{x_{t,1}}\right|^2\right) \geq E\left(\left|e_{x_{t,1}}\right|^2\right)$, we deduce that

$$\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 \geq -\left|\alpha\right|^2 + (\alpha + \alpha^*)\,\rho_{x_{t,1},y_{t,1}}. \tag{3.109}$$

For the particular value of $\alpha = \rho_{x_{t,1},y_{t,1}}$ in the previous expression[2], we find that $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 \geq \left|\rho_{x_{t,1},y_{t,1}}\right|^2$. We can use a very similar proof to show the inequality in (3.105).

Property 3.2 suggests that we can combine the speech distortion and noise reduction measures into one convenient measure:

$$\varrho = \left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 + \left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2, \tag{3.110}$$

where $1 \leq \varrho \leq 2$. Fundamentally, ϱ measures the compromise between speech distortion and noise reduction. For ϱ close to 2, we have the almost perfect estimator with the best binaural estimator in the sense that the noise is almost all removed and speech distortion is almost nonexistent. From (3.99) and (3.110), we deduce for the best binaural estimator that

$$\begin{aligned}\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 &= \frac{\varrho - 1 + \mathrm{iSNR}}{1 + \mathrm{iSNR}} \\ &= \left|\rho_{v_{t,1},y_{t,1}}\right|^2 (\varrho - 1) + \left|\rho_{x_{t,1},y_{t,1}}\right|^2\end{aligned} \tag{3.111}$$

and

$$\begin{aligned}\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 &= \frac{(\varrho - 1)\,\mathrm{iSNR} + 1}{1 + \mathrm{iSNR}} \\ &= \left|\rho_{x_{t,1},y_{t,1}}\right|^2 (\varrho - 1) + \left|\rho_{v_{t,1},y_{t,1}}\right|^2.\end{aligned} \tag{3.112}$$

The two previous expressions tell us the following. For a low input SNR, $\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ is close to 1, meaning that there is a good amount of noise reduction; however, $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ depends mostly on $\varrho - 1$, meaning that distortion depends on the distributions of $x_{t,1}$ and $v_{t,1}$. For a large input SNR, $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ is close to 1, meaning that there is low distortion; however, $\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ is close to $\varrho - 1$, meaning that noise reduction depends on the distributions of $x_{t,1}$ and $v_{t,1}$.

While the MSPCCs $\left|\rho_{x_{t,1},y_{t,1}}\right|^2$ and $\left|\rho_{v_{t,1},y_{t,1}}\right|^2$ give a very good indication on the state of the noisy signal (since they are related to the input SNR), the coefficients of determination $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ and $\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$, as well as ϱ give a very good indication on the enhanced noisy complex signal with the best binaural estimator since $\left|\rho_{x_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ and $\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ are good measures of speech

[2] One can check that $\rho_{x_{t,1},y_{t,1}}$ is a real number, i.e., $\rho_{x_{t,1},y_{t,1}} = \sqrt{\frac{\mathrm{iSNR}}{1+\mathrm{iSNR}}}$.

distortion and noise reduction, respectively, and ϱ is a good measure on the compromise between the two.

3.4.3 Best Widely Linear Estimator

From the recent literature [10], [13], [14], it is known that the best widely linear estimator of $x_{t,1}$ in the MMSE sense is

$$E\left(x_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) = \mathbf{h}_{x_{t,1},\mathrm{W}}^H \widetilde{\underline{\mathbf{y}}}_t, \tag{3.113}$$

where

$$\mathbf{h}_{x_{t,1},\mathrm{W}} = \mathrm{cov}^{-1}\left(\widetilde{\underline{\mathbf{y}}}_t\right)\mathrm{cov}\left(\widetilde{\underline{\mathbf{x}}}_t\right)\mathbf{i} \tag{3.114}$$

is the widely linear Wiener filter in the time domain [10], with $\mathbf{i}$ being the first column of the $2ML \times 2ML$ identity matrix $\mathbf{I}_{2ML}$. We deduce that the coefficient of determination is

$$\begin{aligned}\left|\rho_{x_{t,1}|\widetilde{\underline{\mathbf{y}}}_t}\right|^2 &= \frac{\mathbf{i}^T\mathrm{cov}\left(\widetilde{\underline{\mathbf{x}}}_t\right)\mathrm{cov}^{-1}\left(\widetilde{\underline{\mathbf{y}}}_t\right)\mathrm{cov}\left(\widetilde{\underline{\mathbf{x}}}_t\right)\mathbf{i}}{\mathrm{var}\left(x_{t,1}\right)} \\ &= \frac{\mathbf{h}_{x_{t,1},\mathrm{W}}^H\mathrm{cov}\left(\widetilde{\underline{\mathbf{x}}}_t\right)\mathbf{i}}{\mathrm{var}\left(x_{t,1}\right)},\end{aligned} \tag{3.115}$$

which is a good measure of speech distortion. This measure is very much related to the inverse of the conventional speech reduction factor [1]:

$$\xi_{\mathrm{sr}}^{-1}\left(\mathbf{h}_{x_{t,1},\mathrm{W}}\right) = \frac{\mathbf{h}_{x_{t,1},\mathrm{W}}^H\mathrm{cov}\left(\widetilde{\underline{\mathbf{x}}}_t\right)\mathbf{h}_{x_{t,1},\mathrm{W}}}{\mathrm{var}\left(x_{t,1}\right)}. \tag{3.116}$$

Also, the best widely linear estimator of $v_{t,1}$ in the MMSE sense is

$$E\left(v_{t,1}\left|\widetilde{\underline{\mathbf{y}}}_t\right.\right) = \mathbf{h}_{v_{t,1},\mathrm{W}}^H \widetilde{\underline{\mathbf{y}}}_t, \tag{3.117}$$

where

$$\mathbf{h}_{v_{t,1},\mathrm{W}} = \mathrm{cov}^{-1}\left(\widetilde{\underline{\mathbf{y}}}_t\right)\mathrm{cov}\left(\widetilde{\underline{\mathbf{v}}}_t\right)\mathbf{i} \tag{3.118}$$

is the widely linear Wiener filter for the estimation of $v_{t,1}$. Then, the coefficient of determination is

$$\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 = \frac{\mathbf{i}^T \text{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right) \text{cov}^{-1}\left(\underline{\widetilde{\mathbf{y}}}_t\right) \text{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right) \mathbf{i}}{\text{var}\left(v_{t,1}\right)} = \frac{\mathbf{h}_{v_{t,1},\text{W}}^H \text{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right) \mathbf{i}}{\text{var}\left(v_{t,1}\right)}, \tag{3.119}$$

which is a good measure of noise reduction. It is easy to check that

$$\mathbf{i} = \mathbf{h}_{x_{t,1},\text{W}} + \mathbf{h}_{v_{t,1},\text{W}}. \tag{3.120}$$

Therefore, (3.119) can be rewritten as

$$\left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2 = 1 - \frac{\mathbf{h}_{x_{t,1},\text{W}}^H \text{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right) \mathbf{i}}{\text{var}\left(v_{t,1}\right)}. \tag{3.121}$$

The measure $1 - \left|\rho_{v_{t,1}|\underline{\widetilde{\mathbf{y}}}_t}\right|^2$ is very much related to the inverse of the conventional noise reduction factor [1]:

$$\xi_{\text{nr}}^{-1}\left(\mathbf{h}_{x_{t,1},\text{W}}\right) = \frac{\mathbf{h}_{x_{t,1},\text{W}}^H \text{cov}\left(\underline{\widetilde{\mathbf{v}}}_t\right) \mathbf{h}_{x_{t,1},\text{W}}}{\text{var}\left(v_{t,1}\right)}. \tag{3.122}$$

References

1. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
2. P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Chichester, England: John Wiley & Sons Ltd, 2006.
3. P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.
4. J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters–A Theoretical Study*. Springer Briefs in Electrical and Computer Engineering, 2011.
5. J. Benesty, J. Chen, and Y. Huang, "On the importance of the Pearson correlation coefficient in noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 757–765, May 2008.
6. N. A. Weiss, P. T. Holmes, and M. Hardy, *A Course in Probability*. Boston: Addison-Wesley, 2005.
7. R. Steyer, "Conditional expectations: an introduction to the concept and its applications in empirical sciences," *Methodika*, vol. 2, issue 1, pp. 53–78, 1988.
8. S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall PTR, 1993.
9. J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
10. 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.
11. 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.

12. E. Ollila, "On the circularity of a complex random variable," *IEEE Signal Process. Lett.*, vol. 15, pp. 841–844, 2008.
13. D. P. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. Wiley, 2009.
14. B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Process.*, vol. 43, pp. 2030–2033, Aug. 1995.

Chapter 4
Speech Enhancement Via Correlation Coefficients

In the previous two chapters, we showed the importance of different kinds of correlation coefficients in the formulation and analysis of the best estimators for speech enhancement. In this chapter, we focus on the linear case and show how the most relevant noise reduction filters as well as new ones can be easily derived from the Pearson correlation coefficient. We work in the time domain but the extension of these ideas to the more convenient short-time Fourier transform domain is straightforward. Also, to simplify derivations and make things as clear as possible, we only focus on the single-channel case; generalization to the multichannel scenario is immediate.

4.1 Signal Model and Problem Formulation

The contribution in this chapter is an extension and a generalization of the work presented in [1], [2], [3], [4].

We consider the single-channel noise reduction problem in the time domain described in Chapter 3 (Section 3.1), i.e.,

$$y(t) = x(t) + v(t), \tag{4.1}$$

where $y(t)$, $x(t)$, and $v(t)$ are the microphone, desired, and noise signals, respectively[1]. In a vector form, (4.1) is

$$\begin{aligned}\mathbf{y}(t) &= \left[\, y(t)\ y(t-1)\ \cdots\ y(t-L+1) \,\right]^T \\ &= \mathbf{x}(t) + \mathbf{v}(t).\end{aligned} \tag{4.2}$$

Thus, the covariance matrix (of size $L \times L$) of the noisy signal is

[1] In this chapter, we slightly change the notation for convenience.

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4_4

$$\begin{aligned}\mathbf{R}_{\mathbf{y}} &= E\left[\mathbf{y}(t)\mathbf{y}^T(t)\right] \\ &= \mathbf{R}_{\mathbf{x}} + \mathbf{R}_{\mathbf{v}},\end{aligned} \tag{4.3}$$

where $\mathbf{R}_{\mathbf{x}} = E\left[\mathbf{x}(t)\mathbf{x}^T(t)\right]$ and $\mathbf{R}_{\mathbf{v}} = E\left[\mathbf{v}(t)\mathbf{v}^T(t)\right]$ are the covariance matrices of $\mathbf{x}(t)$ and $\mathbf{v}(t)$, respectively. Then, our objective is to estimate $x(t)$ from the observations, in different ways and different levels of compromises, thanks to the many forms of the squared Pearson correlation coefficient (SPCC) among all signals of interest.

We end this section by recalling the definition of the input SNR:

$$\text{iSNR} = \frac{\sigma_x^2}{\sigma_v^2}, \tag{4.4}$$

where $\sigma_x^2 = E\left[x^2(t)\right]$ and $\sigma_v^2 = E\left[v^2(t)\right]$ are the variances of $x(t)$ and $v(t)$, respectively.

4.2 Linear Filtering and Correlation Coefficients

In this chapter, we estimate the desired signal sample, $x(t)$, or the noise signal sample, $v(t)$, by applying a real-valued filter, $\mathbf{h}$, of length L, to the observation signal vector, $\mathbf{y}(t)$, i.e.,

$$\begin{aligned}z(t) &= \mathbf{h}^T\mathbf{y}(t) \\ &= x_{\text{fd}}(t) + v_{\text{fn}}(t),\end{aligned} \tag{4.5}$$

where $z(t)$ can be either the estimate of $x(t)$ or $v(t)$,

$$x_{\text{fd}}(t) = \mathbf{h}^T\mathbf{x}(t) \tag{4.6}$$

is the filtered desired signal, and

$$v_{\text{fn}}(t) = \mathbf{h}^T\mathbf{v}(t) \tag{4.7}$$

is the filtered noise signal. If $z(t)$ is the estimate of $v(t)$, then the estimate of $x(t)$ is

$$\begin{aligned}\widehat{x}(t) &= y(t) - z(t) \\ &= y(t) - \mathbf{h}^T\mathbf{y}(t) \\ &= (\mathbf{i} - \mathbf{h})^T\,\mathbf{y}(t),\end{aligned} \tag{4.8}$$

where $\mathbf{i}$ is the first column of the $L \times L$ identity matrix $\mathbf{I}_L$. In the rest, we will also use the notation $\mathbf{h}_x$ and $\mathbf{h}_v$. The first filter, $\mathbf{h}_x$, corresponds to the estimation of $x(t)$ while the second filter, $\mathbf{h}_v$, corresponds to the estimation

of $v(t)$. Obviously, from (4.8), we have the relationship:

$$\mathbf{h}_x + \mathbf{h}_v = \mathbf{i}, \tag{4.9}$$

which will extensively be used in all this chapter. Therefore, when $v(t)$ is estimated with $\mathbf{h}_v$, we can estimate $x(t)$ with $\mathbf{h}_x$, thanks to the relation in (4.9).

It is of great interest to know how much of $x(t)$ [resp. $x_{\mathrm{fd}}(t)$] or $v(t)$ [resp. $v_{\mathrm{fn}}(t)$] is contained in the estimator $z(t)$. The best second-order statistics based measure to evaluate this is via the SPCC [1]. Next, we propose four different forms of the SPCC.

We define the SPCC between $z(t)$ and $x(t)$ as

$$\begin{aligned}\rho^2_{z,x}(\mathbf{h}) &= \frac{E^2[z(t)x(t)]}{E[z^2(t)]\,E[x^2(t)]} \\ &= \frac{\sigma_x^2\left(\mathbf{h}^T\boldsymbol{\gamma}_x\right)^2}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}} \\ &= \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{x}_1}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}},\end{aligned} \tag{4.10}$$

where

$$\boldsymbol{\gamma}_x = \frac{E[\mathbf{x}(t)x(t)]}{\sigma_x^2} \tag{4.11}$$

is the normalized correlation vector between $\mathbf{x}(t)$ and $x(t)$, and

$$\mathbf{R}_{\mathbf{x}_1} = \sigma_x^2\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T \tag{4.12}$$

is a rank-1 matrix. In fact, we know from Chapter 3 that we can decompose $\mathbf{R}_{\mathbf{x}}$ as

$$\begin{aligned}\mathbf{R}_{\mathbf{x}} &= \sigma_x^2\boldsymbol{\gamma}_x\boldsymbol{\gamma}_x^T + E\left[\mathbf{x}_{\mathrm{i}}(t)\mathbf{x}_{\mathrm{i}}^T(t)\right] \\ &= \mathbf{R}_{\mathbf{x}_1} + \mathbf{R}_{\mathbf{x}_{\mathrm{i}}},\end{aligned} \tag{4.13}$$

where $\mathbf{R}_{\mathbf{x}_1}$ is defined in (4.12) and $\mathbf{R}_{\mathbf{x}_{\mathrm{i}}}$ is the covariance matrix of the so-called interference signal, $\mathbf{x}_{\mathrm{i}}(t)$, with $E[\mathbf{x}_{\mathrm{i}}(t)x(t)] = \mathbf{0}$.

In the same manner, we define the SPCC between $z(t)$ and $v(t)$ as

$$\begin{aligned}\rho^2_{z,v}(\mathbf{h}) &= \frac{E^2\left[z(t)v(t)\right]}{E\left[z^2(t)\right]E\left[v^2(t)\right]} \\ &= \frac{\sigma_v^2\left(\mathbf{h}^T\boldsymbol{\gamma}_v\right)^2}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}} \\ &= \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{v}_1}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}},\end{aligned} \tag{4.14}$$

where $\boldsymbol{\gamma}_v$ is the normalized correlation vector between $\mathbf{v}(t)$ and $v(t)$, and

$$\mathbf{R}_{\mathbf{v}_1} = \sigma_v^2\boldsymbol{\gamma}_v\boldsymbol{\gamma}_v^T \tag{4.15}$$

is a rank-1 matrix. We also have

$$\begin{aligned}\mathbf{R}_{\mathbf{v}} &= \sigma_v^2\boldsymbol{\gamma}_v\boldsymbol{\gamma}_v^T + E\left[\mathbf{v}_{\mathrm{u}}(t)\mathbf{v}_{\mathrm{u}}^T(t)\right] \\ &= \mathbf{R}_{\mathbf{v}_1} + \mathbf{R}_{\mathbf{v}_{\mathrm{u}}},\end{aligned} \tag{4.16}$$

where $\mathbf{R}_{\mathbf{v}_{\mathrm{u}}}$ is the covariance matrix of $\mathbf{v}_{\mathrm{u}}(t) = \mathbf{v}(t) - v(t)\boldsymbol{\gamma}_v$, and this latter vector is uncorrelated with $v(t)$.

The SPCC between $z(t)$ and $x_{\mathrm{fd}}(t)$ is also of great interest. It is given by

$$\rho^2_{z,x_{\mathrm{fd}}}(\mathbf{h}) = \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{x}}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}}. \tag{4.17}$$

Using the decomposition in (4.13), (4.17) can be expressed as

$$\begin{aligned}\rho^2_{z,x_{\mathrm{fd}}}(\mathbf{h}) &= \rho^2_{z,x}(\mathbf{h}) + \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{x}_{\mathrm{i}}}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}} \\ &= \rho^2_{z,x}(\mathbf{h}) + \rho^2_{z,x_{\mathrm{fi}}}(\mathbf{h}) \\ &\geq \rho^2_{z,x}(\mathbf{h}),\end{aligned} \tag{4.18}$$

where $\rho^2_{z,x_{\mathrm{fi}}}(\mathbf{h})$ is the SPCC between $z(t)$ and the filtered interference, i.e., $x_{\mathrm{fi}}(t) = \mathbf{h}^T\mathbf{x}_{\mathrm{i}}(t)$. Expression (4.18) tells us that $z(t)$ and $x_{\mathrm{fd}}(t)$ are more correlated than $z(t)$ and $x(t)$ are.

Finally, the last SPCC of interest is the one between $z(t)$ and $v_{\mathrm{fn}}(t)$, i.e.,

$$\rho^2_{z,v_{\mathrm{fn}}}(\mathbf{h}) = \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{v}}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}}. \tag{4.19}$$

With the help of (4.16), we can decompose (4.19) as

$$\begin{aligned}\rho^2_{z,v_{\mathrm{fn}}}(\mathbf{h}) &= \rho^2_{z,v}(\mathbf{h}) + \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{v}_{\mathrm{u}}}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}} \\ &= \rho^2_{z,v}(\mathbf{h}) + \rho^2_{z,v_{\mathrm{fu}}}(\mathbf{h}) \\ &\geq \rho^2_{z,v}(\mathbf{h}),\end{aligned} \tag{4.20}$$

where $\rho^2_{z,v_{\rm fu}}(\mathbf{h})$ is the SPCC between $z(t)$ and the filtered uncorrelated noise, i.e., $v_{\rm fu}(t) = \mathbf{h}^T\mathbf{v}_{\rm u}(t)$. We can observe from (4.20) that $z(t)$ and $v_{\rm fn}(t)$ are more correlated than $z(t)$ and $v(t)$ are.

It can easily be checked that

$$\rho^2_{z,x_{\rm fd}}(\mathbf{h}) + \rho^2_{z,v_{\rm fn}}(\mathbf{h}) = 1, \tag{4.21}$$

but

$$\begin{aligned}\rho^2_{z,x}(\mathbf{h}) + \rho^2_{z,v}(\mathbf{h}) &= 1 - \rho^2_{z,x_{\rm fi}}(\mathbf{h}) - \rho^2_{z,v_{\rm fu}}(\mathbf{h}) \\ &\leq 1.\end{aligned} \tag{4.22}$$

We see that the four SPCCs defined above depend explicitly on the filter, $\mathbf{h}$, and measure different kinds of correlation. So it makes intuitively sense to optimize them in order to get different kinds of noise reduction filters.

4.3 Optimal Filters

4.3.1 SPCC Between Filter Output and Desired Signal

In this subsection, we consider the SPCC between $z(t)$ and $x(t)$. A maximal (resp. minimal) value of the SPCC implies that $z(t)$ could be the estimate of $x(t)$ [resp. $v(t)$].

4.3.1.1 Maximization of the SPCC

It is obvious that the maximization of (4.10) leads to the estimate of the desired signal since, in this case, $x(t)$ will be maximally correlated with its estimate, $z(t)$. In (4.10), we recognize the generalized Rayleigh quotient [5]. It is well known that this quotient is maximized with the eigenvector, $\mathbf{a}_1$, corresponding to the maximum eigenvalue of the matrix $\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}_1}$[2]. Let us denote $\lambda_{\mathbf{a}_1}$ this maximum eigenvalue. Since the rank of the mentioned matrix is equal to 1, we have

$$\mathbf{a}_1 = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x}{\sqrt{\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x}}, \tag{4.23}$$

$$\lambda_{\mathbf{a}_1} = \sigma_x^2\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x, \tag{4.24}$$

[2] In the rest of this chapter, we will use some well-known properties of joined diagonalized matrices in order to simplify some of the expressions of the derived noise reduction filters.

and the maximum SPCC is

$$\rho_{z,x}^2\left(\mathbf{a}_1\right)=\lambda_{\mathbf{a}_1}. \tag{4.25}$$

As a result, the optimal filter is proportional to $\mathbf{a}_1$, i.e.,

$$\mathbf{h}_x=\alpha\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x, \tag{4.26}$$

where $\alpha\neq 0$ is an arbitrary real number, whose value is important in practice when we deal with nonstationary signals such as speech[3]; its value is even more important when $\mathbf{h}_x$ is implemented in another domain such as the STFT domain, where a frequency-dependent scaling does not affect the subband performance measures but greatly affects the fullband ones. Hence, with $\mathbf{h}_x$ in (4.26), the estimate of $x(t)$ is

$$\widehat{x}(t)=\mathbf{h}_x^T\mathbf{y}(t) \tag{4.27}$$

and the output SNR is given by

$$\mathrm{oSNR}\left(\mathbf{h}_x\right)=\frac{\mathbf{h}_x^T\mathbf{R}_{\mathbf{x}}\mathbf{h}_x}{\mathbf{h}_x^T\mathbf{R}_{\mathbf{v}}\mathbf{h}_x}\geq \mathrm{iSNR}. \tag{4.28}$$

Now, we need to determine α. There are at least three ways to find this parameter. The first one is from the mean-squared error (MSE) criterion between $x(t)$ and $\widehat{x}(t)$, i.e.,

$$\begin{aligned}J\left(\alpha\right)&=E\left\{\left[x(t)-\mathbf{h}_x^T\mathbf{y}(t)\right]^2\right\}\\&=E\left\{\left[x(t)-\alpha\boldsymbol{\gamma}_x^T\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{y}(t)\right]^2\right\}.\end{aligned} \tag{4.29}$$

The minimization of $J\left(\alpha\right)$ with respect to α leads to

$$\alpha=\sigma_x^2. \tag{4.30}$$

Substituting this value into (4.26), we get the conventional Wiener filter [2]:

$$\begin{aligned}\mathbf{h}_{\mathrm{W}}&=\sigma_x^2\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_x\\&=\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}_1}\mathbf{i}\\&=\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}\mathbf{i}\\&=\left(\mathbf{I}_L-\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\right)\mathbf{i}.\end{aligned} \tag{4.31}$$

[3] Obviously, for stationary signals, the value of α is not relevant at all as long as it is different from zero.

Obviously, this filter maximizes the SPCC in (4.10) but it does not maximize the output SNR. We will see later which kind of the SPCC whose maximization is equivalent to maximizing the output SNR.

The second possibility is from the distortion-based MSE, i.e.,

$$\begin{aligned} J_{\mathrm{d}}(\alpha) &= E\left\{\left[x(t) - \mathbf{h}_x^T \mathbf{x}(t)\right]^2\right\} \\ &= E\left\{\left[x(t) - \alpha \boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{x}(t)\right]^2\right\}. \end{aligned} \tag{4.32}$$

By minimizing $J_{\mathrm{d}}(\alpha)$ with respect to α, we obtain

$$\alpha = \frac{\lambda_{\mathbf{a}_1}}{\boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{R}_{\mathbf{x}} \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x} \tag{4.33}$$

and substituting the previous result into (4.26) gives the minimum distortion (MD) filter:

$$\mathbf{h}_{\mathrm{MD}} = \frac{\lambda_{\mathbf{a}_1} \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \mathbf{R}_{\mathbf{x}} \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}. \tag{4.34}$$

Clearly, as far as the output SNR is concerned, the two filters $\mathbf{h}_{\mathrm{W}}$ and $\mathbf{h}_{\mathrm{MD}}$ are equivalent but when implemented in the STFT domain, they will give much different values of the fullband output SNR.

Finally, the last manner to find α is by plugging $\mathbf{h}_v = \mathbf{i} - \alpha \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x$ into (4.10). We get

$$\begin{aligned} \rho_{z,x}^2(\alpha) &= \frac{\left(\mathbf{i} - \alpha \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x\right)^T \mathbf{R}_{\mathbf{x}_1} \left(\mathbf{i} - \alpha \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x\right)}{\left(\mathbf{i} - \alpha \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x\right)^T \mathbf{R}_{\mathbf{y}} \left(\mathbf{i} - \alpha \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x\right)} \\ &= \frac{\sigma_x^2 \left(1 - \alpha \boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x\right)^2}{\sigma_y^2 - 2\alpha + \alpha^2 \boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}. \end{aligned} \tag{4.35}$$

Since $\mathbf{h}_v$ is involved in the SPCC, we need to minimize this latter. Minimizing the previous expression is equivalent to minimizing its numerator. Therefore, we have

$$\alpha = \frac{1}{\boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}. \tag{4.36}$$

As a result, we deduce the so-called minimum variance distortionless response (MVDR) filter [6]:

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}. \tag{4.37}$$

Indeed, one can check that $\mathbf{h}_{\text{MVDR}}^T\boldsymbol{\gamma}_x = 1$, which means that the desired signal is recovered if $\mathbf{x}_\text{i}(t)$ is considered as an interference. This filter works very well in the STFT domain [7], [8], [9], [10].

4.3.1.2 Minimization of the SPCC

Another perspective is to find the filter that minimizes (4.10). Therefore, the filter output will be the estimate of $v(t)$. The matrix $\mathbf{R}_\mathbf{y}^{-1}\mathbf{R}_{\mathbf{x}_1}$ has $L-1$ eigenvalues equal to 0, since its rank is equal to 1. Let $\mathbf{a}_2, \mathbf{a}_3, \ldots, \mathbf{a}_L$ be the corresponding eigenvectors and let us consider the filter, which is a linear combination of these eigenvectors:

$$\begin{aligned}\mathbf{h}_v &= \sum_{i=2}^{L} \alpha_i \mathbf{a}_i \\ &= \mathbf{A}_2\boldsymbol{\alpha},\end{aligned} \tag{4.38}$$

where

$$\mathbf{A}_2 = \left[\, \mathbf{a}_2 \; \mathbf{a}_3 \; \cdots \; \mathbf{a}_L \,\right] \tag{4.39}$$

is a matrix of size $L \times (L-1)$ and

$$\boldsymbol{\alpha} = \left[\, \alpha_2 \; \alpha_3 \; \cdots \; \alpha_L \,\right]^T \neq \mathbf{0} \tag{4.40}$$

is a vector of length $L-1$. It is clear that $\mathbf{h}_v$ in (4.38) minimizes (4.10), since

$$\rho_{zx}^2\left(\mathbf{h}_v\right) = 0. \tag{4.41}$$

Therefore, the estimates of $v(t)$ and $x(t)$ are, respectively,

$$\widehat{v}(t) = \mathbf{h}_v^T\mathbf{y}(t) \tag{4.42}$$

and

$$\begin{aligned}\widehat{x}(t) &= y(t) - \widehat{v}(t) \\ &= \mathbf{h}_x^T\mathbf{y}(t),\end{aligned} \tag{4.43}$$

where

$$\mathbf{h}_x = \mathbf{i} - \mathbf{h}_v \tag{4.44}$$

is the equivalent filter for the estimation of $x(t)$.

There are at least two interesting ways to find $\boldsymbol{\alpha}$. The first one is from the power of the residual noise, i.e.,

$$\begin{aligned} J_{\mathrm{r}}(\boldsymbol{\alpha}) &= \mathbf{h}_x^T \mathbf{R}_{\mathbf{v}} \mathbf{h}_x \\ &= (\mathbf{i} - \mathbf{h}_v)^T \mathbf{R}_{\mathbf{v}} (\mathbf{i} - \mathbf{h}_v) \\ &= \sigma_v^2 - 2\boldsymbol{\alpha}^T \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{i} + \boldsymbol{\alpha}^T \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{A}_2 \boldsymbol{\alpha} \end{aligned} \tag{4.45}$$

and the second one is from the MSE between $x(t)$ and $\widehat{x}(t)$, i.e.,

$$\begin{aligned} J(\boldsymbol{\alpha}) &= E\left\{ \left[x(t) - \mathbf{h}_x^T \mathbf{y}(t) \right]^2 \right\} \\ &= E\left\{ \left[x(t) - (\mathbf{i} - \mathbf{h}_v)^T \mathbf{y}(t) \right]^2 \right\} \\ &= E\left\{ \left[v(t) - \boldsymbol{\alpha}^T \mathbf{A}_2^T \mathbf{y}(t) \right]^2 \right\}. \end{aligned} \tag{4.46}$$

The minimization of $J_{\mathrm{r}}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ gives

$$\boldsymbol{\alpha} = \left(\mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{A}_2 \right)^{-1} \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{i}. \tag{4.47}$$

As a result, we obtain the minimum noise (MN) filter for the estimation of $x(t)$:

$$\mathbf{h}_{\mathrm{MN}} = \left[\mathbf{I}_L - \mathbf{A}_2 \left(\mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{A}_2 \right)^{-1} \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \right] \mathbf{i}. \tag{4.48}$$

While this filter may reduce quite a lot of noise, it may introduce an unacceptable amount of distortion to the desired signal.

By minimizing the MSE, we find that

$$\begin{aligned} \boldsymbol{\alpha} &= \left(\mathbf{A}_2^T \mathbf{R}_{\mathbf{y}} \mathbf{A}_2 \right)^{-1} \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{i} \\ &= \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{i}. \end{aligned} \tag{4.49}$$

We deduce the MVDR filter for the estimation of $x(t)$:

$$\begin{aligned} \mathbf{h}_{\mathrm{MVDR}} &= \mathbf{i} - \mathbf{A}_2 \mathbf{A}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{i} \\ &= \mathbf{i} - \left(\mathbf{R}_{\mathbf{y}}^{-1} - \mathbf{a}_1 \mathbf{a}_1^T \right) \mathbf{R}_{\mathbf{v}} \mathbf{i} \\ &= \mathbf{h}_{\mathrm{W}} + \left(\mathbf{a}_1^T \mathbf{R}_{\mathbf{v}} \mathbf{i} \right) \mathbf{a}_1 \\ &= \mathbf{h}_{\mathrm{W}} + \left(\mathbf{a}_1^T \mathbf{R}_{\mathbf{y}} \mathbf{i} - \mathbf{a}_1^T \mathbf{R}_{\mathbf{x}_1} \mathbf{i} \right) \mathbf{a}_1 \\ &= \left(\mathbf{a}_1^T \mathbf{R}_{\mathbf{y}} \mathbf{i} \right) \mathbf{a}_1 \\ &= \frac{\mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}{\boldsymbol{\gamma}_x^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_x}. \end{aligned} \tag{4.50}$$

As far as the output SNR is concerned, the two filters $\mathbf{h}_{\mathrm{W}}$ and $\mathbf{h}_{\mathrm{MVDR}}$ are equivalent. However, in the STFT domain, $\mathbf{h}_{\mathrm{W}}$ and $\mathbf{h}_{\mathrm{MVDR}}$ will behave differently. Another insightful way to derive $\mathbf{h}_{\mathrm{MVDR}}$ is by substituting $\mathbf{h}_x = \mathbf{i} - \mathbf{A}_2 \boldsymbol{\alpha}$ into the SPCC in (4.10). We get

$$\rho_{z,x}^2(\boldsymbol{\alpha}) = \frac{(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha})^T \mathbf{R}_{\mathbf{x}_1} (\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha})}{(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha})^T \mathbf{R}_{\mathbf{y}} (\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha})} = \frac{\sigma_x^2}{\sigma_y^2 - 2\boldsymbol{\alpha}^T\mathbf{A}_2^T\mathbf{R}_{\mathbf{y}}\mathbf{i} + \boldsymbol{\alpha}^T\boldsymbol{\alpha}}. \tag{4.51}$$

Maximizing the previous expression is equivalent to minimizing its denominator. We easily obtain

$$\begin{aligned}\boldsymbol{\alpha} &= \mathbf{A}_2^T\mathbf{R}_{\mathbf{y}}\mathbf{i} \\ &= \mathbf{A}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \mathbf{A}_2^T\mathbf{R}_{\mathbf{v}}\mathbf{i} \\ &= \mathbf{A}_2^T\mathbf{R}_{\mathbf{x}_1}\mathbf{i} + \mathbf{A}_2^T\mathbf{R}_{\mathbf{v}}\mathbf{i} \\ &= \mathbf{A}_2^T\mathbf{R}_{\mathbf{v}}\mathbf{i},\end{aligned} \tag{4.52}$$

which leads to $\mathbf{h}_{\mathrm{MVDR}}$. It is clear from the above that

$$\sigma_v^2 \geq \boldsymbol{\alpha}^T\boldsymbol{\alpha}. \tag{4.53}$$

Therefore, with $\mathbf{h}_{\mathrm{W}}$ or $\mathbf{h}_{\mathrm{MVDR}}$, we can express the SPCC between $z(t)$ and $x(t)$ as

$$\begin{aligned}\rho_{z,x}^2(\mathbf{h}_{\mathrm{W}}) &= \lambda_{\mathbf{a}_1} \\ &= \frac{\sigma_x^2}{\sigma_y^2 - \boldsymbol{\alpha}^T\boldsymbol{\alpha}} \\ &= \frac{\mathrm{iSNR}}{\frac{\sigma_v^2 - \boldsymbol{\alpha}^T\boldsymbol{\alpha}}{\sigma_v^2} + \mathrm{iSNR}},\end{aligned} \tag{4.54}$$

which shows a very interesting relationship between the eigenvalue of interest and the input SNR. We always have

$$\rho_{z,x}^2(\mathbf{h}_{\mathrm{W}}) \geq \rho_{x,y}^2 = \frac{\mathrm{iSNR}}{1+\mathrm{iSNR}}, \tag{4.55}$$

where $\rho_{x,y}^2$ is the SPCC between $x(t)$ and $y(t)$. The previous expression tells us that $z(t)$ (with the Wiener filter) and $x(t)$ are more correlated than $y(t)$ and $x(t)$ are, which basically means that the SNR of $z(t)$ is better than that of $y(t)$.

4.3.2 SPCC Between Filter Output and Noise Signal

In this subsection, we consider the SPCC between $z(t)$ and $v(t)$. A maximal (resp. minimal) value of the SPCC implies that $z(t)$ could be the estimate of $v(t)$ [resp. $x(t)$].

4.3.2.1 Maximization of the SPCC

The rank of the matrix $\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}_1}$ is equal to 1, so its only non-null and positive eigenvalue is

$$\lambda_{\mathbf{b}_1} = \sigma_v^2 \boldsymbol{\gamma}_v^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_v, \tag{4.56}$$

whose corresponding eigenvector is

$$\mathbf{b}_1 = \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}{\sqrt{\boldsymbol{\gamma}_v^T \mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}}. \tag{4.57}$$

As a result, the filter that maximizes (4.14) is

$$\mathbf{h}_v = \beta \mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v, \tag{4.58}$$

where $\beta \neq 0$ is an arbitrary real number, and the maximum SPCC is

$$\rho_{z,v}^2(\mathbf{b}_1) = \lambda_{\mathbf{b}_1}. \tag{4.59}$$

This filter output gives the estimate of $v(t)$, i.e.,

$$\widehat{v}(t) = \mathbf{h}_v^T \mathbf{y}(t). \tag{4.60}$$

We deduce that the estimate of the desired signal is

$$\begin{aligned} \widehat{x}(t) &= y(t) - \widehat{v}(t) \\ &= \mathbf{h}_x^T \mathbf{y}(t), \end{aligned} \tag{4.61}$$

where

$$\mathbf{h}_x = \mathbf{i} - \mathbf{h}_v \tag{4.62}$$

is the equivalent filter for the estimation of $x(t)$.

One way to find β is from the MSE between $x(t)$ and $\widehat{x}(t)$ [or, equivalently, $v(t)$ and $\widehat{v}(t)$], i.e.,

$$J(\beta) = E\left\{\left[v(t) - \mathbf{h}_v^T\mathbf{y}(t)\right]^2\right\}$$
$$= E\left\{\left[v(t) - \beta\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{y}(t)\right]^2\right\}. \tag{4.63}$$

Indeed, the optimization of the previous expression leads to

$$\beta = \sigma_v^2. \tag{4.64}$$

Therefore, we have

$$\mathbf{h}_v = \sigma_v^2\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v \tag{4.65}$$

and from (4.62),

$$\mathbf{h}_{\mathrm{W}} = \left(\mathbf{I}_L - \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\right)\mathbf{i}, \tag{4.66}$$

which is the classical Wiener filter.

The second way to find β is from the power of the residual noise, i.e.,

$$J_{\mathrm{r}}(\beta) = \left(\mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right)^T\mathbf{R}_{\mathbf{v}}\left(\mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right) \tag{4.67}$$
$$= \sigma_v^2 - 2\beta\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\mathbf{i} + \beta^2\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v.$$

After minimizing $J_{\mathrm{r}}(\beta)$ and substituting the obtained value of β into (4.58), we easily find that the MN-type filter for the estimation of $x(t)$ is

$$\mathbf{h}_{\mathrm{MN},2} = \left[\mathbf{I}_L - \frac{\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}_1}}{\mathrm{tr}\left(\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}_1}\right)}\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\right]\mathbf{i}. \tag{4.68}$$

Finally, the last way to find β is by plugging $\mathbf{h}_x = \mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v$ into (4.14). We get

$$\rho_{z,v}^2(\beta) = \frac{\left(\mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right)^T\mathbf{R}_{\mathbf{v}_1}\left(\mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right)}{\left(\mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right)^T\mathbf{R}_{\mathbf{y}}\left(\mathbf{i} - \beta\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right)}$$
$$= \frac{\sigma_v^2\left(1 - \beta\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\right)^2}{\sigma_y^2 - 2\beta + \beta^2\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}. \tag{4.69}$$

Minimizing the previous expression is equivalent to minimizing its numerator. This leads to

$$\beta = \frac{1}{\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}. \tag{4.70}$$

As a result, we find the null constraint (NC) filter for the estimation of $x(t)$:

$$\mathbf{h}_{\mathrm{NC}} = \left(\mathbf{I}_L - \frac{\mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_v \boldsymbol{\gamma}_v^T}{\boldsymbol{\gamma}_v^T \mathbf{R}_{\mathbf{y}}^{-1} \boldsymbol{\gamma}_v} \right) \mathbf{i}. \tag{4.71}$$

Indeed, it can easily be verified that $\mathbf{h}_{\mathrm{NC}}^T \boldsymbol{\gamma}_v = 0$, which means that the correlated noise is completely canceled.

4.3.2.2 Minimization of the SPCC

Let $\mathbf{b}_2, \mathbf{b}_3, \ldots, \mathbf{b}_L$ be the eigenvectors corresponding to the $L-1$ null eigenvalues of the matrix $\mathbf{R}_{\mathbf{y}}^{-1} \mathbf{R}_{\mathbf{v}_1}$. Let us form the filter:

$$\begin{aligned} \mathbf{h}_x &= \sum_{i=2}^{L} \beta_i \mathbf{b}_i \\ &= \mathbf{B}_2 \boldsymbol{\beta}, \end{aligned} \tag{4.72}$$

where $\beta_i,\ i = 2, 3, \ldots, L$ are arbitrary real numbers with at least one of them different from 0,

$$\mathbf{B}_2 = \left[\mathbf{b}_2\ \mathbf{b}_3\ \cdots\ \mathbf{b}_L \right] \tag{4.73}$$

is a matrix of size $L \times (L-1)$, and

$$\boldsymbol{\beta} = \left[\beta_2\ \beta_3\ \cdots\ \beta_L \right]^T \neq \mathbf{0} \tag{4.74}$$

is a vector of length $L-1$. It can be verified that $\mathbf{h}_x$ in (4.72) minimizes (4.14), since

$$\rho_{z,v}^2 \left(\mathbf{h}_x \right) = 0. \tag{4.75}$$

Therefore, the filter output can be considered as the estimate of the desired signal, i.e.,

$$\widehat{x}(t) = \mathbf{h}_x^T \mathbf{y}(t). \tag{4.76}$$

The MSE between $x(t)$ and $\widehat{x}(t)$ is then

$$\begin{aligned} J\left(\boldsymbol{\beta}\right) &= E\left\{ \left[x(t) - \mathbf{h}_x^T \mathbf{y}(t) \right]^2 \right\} \\ &= \sigma_x^2 - 2\boldsymbol{\beta}^T \mathbf{B}_2^T \mathbf{R}_{\mathbf{x}} \mathbf{i} + \boldsymbol{\beta}^T \boldsymbol{\beta} \\ &= \sigma_x^2 - 2\boldsymbol{\beta}^T \mathbf{B}_2^T \mathbf{R}_{\mathbf{x}} \mathbf{i} + \boldsymbol{\beta}^T \mathbf{B}_2^T \mathbf{R}_{\mathbf{x}} \mathbf{B}_2 \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{B}_2^T \mathbf{R}_{\mathbf{v}} \mathbf{B}_2 \boldsymbol{\beta} \\ &= J_{\mathrm{d}}\left(\boldsymbol{\beta}\right) + J_{\mathrm{r}}\left(\boldsymbol{\beta}\right). \end{aligned} \tag{4.77}$$

From (4.77), we observe that we have at least two obvious options to find $\boldsymbol{\beta}$. The first one is to minimize $J\left(\boldsymbol{\beta}\right)$. The second option is to minimize $J_{\mathrm{d}}\left(\boldsymbol{\beta}\right)$.

From the first option, we obtain the NC filter:

$$\begin{aligned}\mathbf{h}_{\mathrm{NC}} &= \mathbf{B}_2\mathbf{B}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{i}\\ &= \left(\mathbf{R}_{\mathbf{y}}^{-1} - \mathbf{b}_1\mathbf{b}_1^T\right)\left(\mathbf{R}_{\mathbf{y}} - \mathbf{R}_{\mathbf{v}}\right)\mathbf{i}\\ &= \mathbf{i} - \mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{v}}\mathbf{i} - \mathbf{b}_1\mathbf{b}_1^T\mathbf{R}_{\mathbf{y}}\mathbf{i} + \mathbf{b}_1\mathbf{b}_1^T\mathbf{R}_{\mathbf{v}_1}\mathbf{i}\\ &= \mathbf{h}_{\mathrm{W}} + \left(\mathbf{b}_1^T\mathbf{R}_{\mathbf{y}}\mathbf{i}\right)\left(\lambda_{\mathbf{b}_1}\mathbf{b}_1 - \mathbf{b}_1\right)\\ &= \mathbf{i} - \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}{\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}\\ &= \left(\mathbf{I}_L - \frac{\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v\boldsymbol{\gamma}_v^T}{\boldsymbol{\gamma}_v^T\mathbf{R}_{\mathbf{y}}^{-1}\boldsymbol{\gamma}_v}\right)\mathbf{i}. \end{aligned} \tag{4.78}$$

The second option gives the MD-type filter:

$$\mathbf{h}_{\mathrm{MD},2} = \mathbf{B}_2\left(\mathbf{B}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{B}_2\right)^{-1}\mathbf{B}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{i}, \tag{4.79}$$

where it is assumed that the rank of $\mathbf{R}_{\mathbf{x}}$ is at least equal to $L-1$.

Now, let us find $\boldsymbol{\beta}$ from the SPCC. Substituting $\mathbf{h}_v = \mathbf{i} - \mathbf{B}_2\boldsymbol{\beta}$ into the SPCC in (4.14), we obtain

$$\begin{aligned}\rho_{z,v}^2\left(\boldsymbol{\beta}\right) &= \frac{\left(\mathbf{i} - \mathbf{B}_2\boldsymbol{\beta}\right)^T\mathbf{R}_{\mathbf{v}_1}\left(\mathbf{i} - \mathbf{B}_2\boldsymbol{\beta}\right)}{\left(\mathbf{i} - \mathbf{B}_2\boldsymbol{\beta}\right)^T\mathbf{R}_{\mathbf{y}}\left(\mathbf{i} - \mathbf{B}_2\boldsymbol{\beta}\right)}\\ &= \frac{\sigma_v^2}{\sigma_y^2 - 2\boldsymbol{\beta}^T\mathbf{B}_2^T\mathbf{R}_{\mathbf{y}}\mathbf{i} + \boldsymbol{\beta}^T\boldsymbol{\beta}}. \end{aligned} \tag{4.80}$$

Maximizing the previous expression is equivalent to minimizing its denominator. We get

$$\begin{aligned}\boldsymbol{\beta} &= \mathbf{B}_2^T\mathbf{R}_{\mathbf{y}}\mathbf{i}\\ &= \mathbf{B}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \mathbf{B}_2^T\mathbf{R}_{\mathbf{v}}\mathbf{i}\\ &= \mathbf{B}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \mathbf{B}_2^T\mathbf{R}_{\mathbf{v}_1}\mathbf{i}\\ &= \mathbf{B}_2^T\mathbf{R}_{\mathbf{x}}\mathbf{i}, \end{aligned} \tag{4.81}$$

which leads to $\mathbf{h}_{\mathrm{NC}}$. It is clear from the above that

$$\sigma_x^2 \geq \boldsymbol{\beta}^T\boldsymbol{\beta}. \tag{4.82}$$

Therefore, we can express the maximum value of the SPCC between $z(t)$ and $v(t)$ as

$$\begin{aligned}\rho^2_{z,v}(\mathbf{b}_1) &= \lambda_{\mathbf{b}_1} \\ &= \frac{\sigma_v^2}{\sigma_y^2 - \boldsymbol{\beta}^T\boldsymbol{\beta}} \\ &= \frac{1}{1 + \text{iSNR} \times \dfrac{\sigma_x^2 - \boldsymbol{\beta}^T\boldsymbol{\beta}}{\sigma_x^2}},\end{aligned} \tag{4.83}$$

which shows a very interesting relationship between the eigenvalue of interest and the input SNR. We always have

$$\rho^2_{z,v}(\mathbf{b}_1) \geq \rho^2_{v,y} = \frac{1}{1 + \text{iSNR}}, \tag{4.84}$$

where $\rho^2_{v,y}$ is the SPCC between $v(t)$ and $y(t)$. The previous expression tells us that $z(t)$ (which is here the estimate of $v(t)$ with a filter proportional to $\mathbf{b}_1$) and $x(t)$ are more correlated than $y(t)$ and $v(t)$ are, which basically means that the SNR of $y(t) - z(t)$ is better than that of $y(t)$.

4.3.3 SPCC Between Filter Output and Filtered Desired Signal

This subsection is concerned with the SPCC between $z(t)$ and $x_{\text{fd}}(t)$. A maximal (resp. minimal) value of the SPCC implies that $z(t)$ is the estimate of $x(t)$ [resp. $v(t)$].

4.3.3.1 Maximization of the SPCC

Let $\lambda_{\mathbf{t}_1}$ be the largest eigenvalue, with multiplicity P, of the matrix $\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}$[4]. We denote $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_P$ the corresponding eigenvectors. It is clear that the filter:

$$\mathbf{h}_x = \sum_{p=1}^{P} \theta_p \mathbf{t}_p, \tag{4.85}$$

where $\theta_p,\ p = 1, 2, \ldots, P$ are arbitrary real numbers with at least one of them different from 0, maximizes (4.17), and the maximum SPCC[5] is

$$\rho^2_{z,x_{\text{fd}}}(\mathbf{h}_x) = \lambda_{\mathbf{t}_1}. \tag{4.86}$$

[4] In practice, we may consider the P largest eigenvalues of $\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}$. In this case, they are denoted $\lambda_{\mathbf{t}_1}, \lambda_{\mathbf{t}_2}, \ldots, \lambda_{\mathbf{t}_P}$.

[5] In case we take the P largest eigenvalues, we have $\rho^2_{z,x_{\text{fd}}}(\mathbf{h}_x) = \sum_{p=1}^{P} \lambda_{\mathbf{t}_p}/P$.

We can rewrite (4.85) as

$$\mathbf{h}_x = \mathbf{T}\boldsymbol{\theta}, \tag{4.87}$$

where

$$\mathbf{T} = \left[\mathbf{t}_1\ \mathbf{t}_2\ \cdots\ \mathbf{t}_P\right] \tag{4.88}$$

is a matrix of size $L \times P$ and

$$\boldsymbol{\theta} = \left[\theta_1\ \theta_2\ \cdots\ \theta_P\right]^T \neq \mathbf{0} \tag{4.89}$$

is a vector of length P. It can be checked that the SPCC can be written as

$$\rho^2_{z,x_{\mathrm{fd}}}\left(\mathbf{h}_x\right) = \frac{\mathrm{oSNR}\left(\mathbf{h}_x\right)}{1+\mathrm{oSNR}\left(\mathbf{h}_x\right)}, \tag{4.90}$$

which means that $\mathbf{h}_x$ in (4.85) or in (4.87) also maximizes the output SNR.

The estimate of $x(t)$ is

$$\widehat{x}(t) = \mathbf{h}_x^T\mathbf{y}(t). \tag{4.91}$$

The MSE between $x(t)$ and $\widehat{x}(t)$ is then

$$\begin{aligned} J\left(\boldsymbol{\theta}\right) &= E\left\{\left[x(t)-\mathbf{h}_x^T\mathbf{y}(t)\right]^2\right\} \\ &= \sigma_x^2 - 2\boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \boldsymbol{\theta}^T\boldsymbol{\theta} \\ &= \sigma_x^2 - 2\boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{T}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{v}}\mathbf{T}\boldsymbol{\theta} \\ &= J_{\mathrm{d}}\left(\boldsymbol{\theta}\right) + J_{\mathrm{r}}\left(\boldsymbol{\theta}\right). \end{aligned} \tag{4.92}$$

From (4.92), we observe that we have at least two obvious options to find $\boldsymbol{\theta}$. The first one is to minimize $J\left(\boldsymbol{\theta}\right)$. The second option is to minimize $J_{\mathrm{d}}\left(\boldsymbol{\theta}\right)$.

From the first option, we obtain the Wiener-type filter:

$$\mathbf{h}_{\mathrm{W},2} = \mathbf{T}\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{i}. \tag{4.93}$$

The second option gives the MD-type filter:

$$\mathbf{h}_{\mathrm{MD},3} = \mathbf{T}\left(\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{i}. \tag{4.94}$$

In the assumed case where we have a maximum eigenvalue, $\lambda_{\mathbf{t}_1}$, with multiplicity P, we have $\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{T} = \lambda_{\mathbf{t}_1}\mathbf{I}_P$, where $\mathbf{I}_P$ is the $P \times P$ identity matrix. As a result,

$$\mathbf{h}_{\mathrm{MD},3} = \lambda_{\mathbf{t}_1}\mathbf{h}_{\mathrm{W},2}. \tag{4.95}$$

Now, substituting $\mathbf{h}_v = \mathbf{i} - \mathbf{T}\boldsymbol{\theta}$ into (4.17), we get

$$\rho^2_{z,x_{\text{fd}}}(\boldsymbol{\theta}) = \frac{(\mathbf{i}-\mathbf{T}\boldsymbol{\theta})^T \mathbf{R}_{\mathbf{x}} (\mathbf{i}-\mathbf{T}\boldsymbol{\theta})}{(\mathbf{i}-\mathbf{T}\boldsymbol{\theta})^T \mathbf{R}_{\mathbf{y}} (\mathbf{i}-\mathbf{T}\boldsymbol{\theta})} = \frac{\sigma_x^2 - 2\boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{T}\boldsymbol{\theta}}{\sigma_y^2 - 2\boldsymbol{\theta}^T\mathbf{T}^T\mathbf{R}_{\mathbf{y}}\mathbf{i} + \boldsymbol{\theta}^T\boldsymbol{\theta}}. \tag{4.96}$$

It is clear that minimizing (4.96) is the same as minimizing its numerator. This leads to $\boldsymbol{\theta} = \left(\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{R}_{\mathbf{x}}\mathbf{i}$ and then to $\mathbf{h}_{\text{MD},3}$. This is another way to derive the MD-type filter given in (4.94).

4.3.3.2 Minimization of the SPCC

Let $\lambda_{\mathbf{t}_L}$ be the smallest eigenvalue, with multiplicity Q, of the matrix $\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}$[6]. We denote $\mathbf{t}_{L-Q+1} = \mathbf{t}'_1, \mathbf{t}_{L-Q+2} = \mathbf{t}'_2, \ldots, \mathbf{t}_L = \mathbf{t}'_Q$ the corresponding eigenvectors. The filter:

$$\mathbf{h}_v = \sum_{q=1}^{Q} \theta'_q \mathbf{t}'_q, \tag{4.97}$$

where $\theta'_q,\ q = 1, 2, \ldots, Q$ are arbitrary real numbers with at least one of them different from 0, minimizes (4.17), and the minimum SPCC[7] is

$$\rho^2_{z,x_{\text{fd}}}(\mathbf{h}_v) = \lambda_{\mathbf{t}_L}. \tag{4.98}$$

A more convenient way to write (4.97) is

$$\mathbf{h}_v = \mathbf{T}'\boldsymbol{\theta}', \tag{4.99}$$

where

$$\mathbf{T}' = \left[\mathbf{t}'_1\ \mathbf{t}'_2\ \cdots\ \mathbf{t}'_Q\right] \tag{4.100}$$

is a matrix of size $L \times Q$ and

$$\boldsymbol{\theta}' = \left[\theta'_1\ \theta'_2\ \cdots\ \theta'_Q\right]^T \neq \mathbf{0} \tag{4.101}$$

is a vector of length Q. Therefore, the estimates of $v(t)$ and $x(t)$ are, respectively,

$$\widehat{v}(t) = \mathbf{h}_v^T\mathbf{y}(t) \tag{4.102}$$

and

---

[6] In practice, we may consider the Q smallest eigenvalues of $\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{x}}$. In this case, they are denoted $\lambda_{\mathbf{t}_{L-Q+1}}, \lambda_{\mathbf{t}_{L-Q+2}}, \ldots, \lambda_{\mathbf{t}_L}$.

[7] In case we take the Q smallest eigenvalues, we have $\rho^2_{z,x_{\text{fd}}}(\mathbf{h}_v) = \sum_{q=1}^{Q} \lambda_{\mathbf{t}_{L-q+1}}/Q$.

$$\begin{aligned}\widehat{x}(t) &= y(t) - \widehat{v}(t) \\ &= \mathbf{h}_x^T \mathbf{y}(t),\end{aligned} \tag{4.103}$$

where

$$\mathbf{h}_x = \mathbf{i} - \mathbf{h}_v \tag{4.104}$$

is the equivalent filter for the estimation of $x(t)$.

There are at least two interesting ways to find $\boldsymbol{\theta}'$. The first one is from the power of the residual noise, i.e.,

$$J_{\mathrm{r}}\left(\boldsymbol{\theta}'\right) = E\left\{\left[v(t) - \boldsymbol{\theta}'^T \mathbf{T}'^T \mathbf{v}(t)\right]^2\right\} \tag{4.105}$$

and the second one is from the MSE between $x(t)$ and $\widehat{x}(t)$, i.e.,

$$J\left(\boldsymbol{\theta}'\right) = E\left\{\left[v(t) - \boldsymbol{\theta}'^T \mathbf{T}'^T \mathbf{y}(t)\right]^2\right\}. \tag{4.106}$$

The minimization of $J_{\mathrm{r}}\left(\boldsymbol{\theta}'\right)$ with respect to $\boldsymbol{\theta}'$ gives

$$\boldsymbol{\theta}' = \left(\mathbf{T}'^T \mathbf{R}_{\mathbf{v}} \mathbf{T}'\right)^{-1} \mathbf{T}'^T \mathbf{R}_{\mathbf{v}} \mathbf{i}. \tag{4.107}$$

As a result,

$$\mathbf{h}_v = \mathbf{T}' \left(\mathbf{T}'^T \mathbf{R}_{\mathbf{v}} \mathbf{T}'\right)^{-1} \mathbf{T}'^T \mathbf{R}_{\mathbf{v}} \mathbf{i} \tag{4.108}$$

and the MN-type filter for the estimation of $x(t)$ is

$$\mathbf{h}_{\mathrm{MN},3} = \left[\mathbf{I}_L - \mathbf{T}' \left(\mathbf{T}'^T \mathbf{R}_{\mathbf{v}} \mathbf{T}'\right)^{-1} \mathbf{T}'^T \mathbf{R}_{\mathbf{v}}\right] \mathbf{i}. \tag{4.109}$$

By minimizing the MSE, we find the Wiener-type filter for the estimation of $x(t)$:

$$\begin{aligned}\mathbf{h}_{\mathrm{W},3} &= \left[\mathbf{I}_L - \mathbf{T}' \left(\mathbf{T}'^T \mathbf{R}_{\mathbf{y}} \mathbf{T}'\right)^{-1} \mathbf{T}'^T \mathbf{R}_{\mathbf{v}}\right] \mathbf{i} \\ &= \left(\mathbf{I}_L - \mathbf{T}' \mathbf{T}'^T \mathbf{R}_{\mathbf{v}}\right) \mathbf{i}.\end{aligned} \tag{4.110}$$

Now, let us see what happens from the SPCC perspective. Plugging $\mathbf{h}_x = \mathbf{i} - \mathbf{T}'\boldsymbol{\theta}'$ into (4.17), we get

$$\rho^2_{z,x_{\mathrm{fd}}}\left(\boldsymbol{\theta}'\right) = \frac{\left(\mathbf{i}-\mathbf{T}'\boldsymbol{\theta}'\right)^T \mathbf{R}_{\mathbf{x}}\left(\mathbf{i}-\mathbf{T}'\boldsymbol{\theta}'\right)}{\left(\mathbf{i}-\mathbf{T}'\boldsymbol{\theta}'\right)^T \mathbf{R}_{\mathbf{y}}\left(\mathbf{i}-\mathbf{T}'\boldsymbol{\theta}'\right)} \tag{4.111}$$
$$= \frac{\sigma_x^2 - 2\boldsymbol{\theta}'^T\mathbf{T}'^T\mathbf{R}_{\mathbf{x}}\mathbf{i} + \boldsymbol{\theta}'^T\mathbf{T}'^T\mathbf{R}_{\mathbf{x}}\mathbf{T}'\boldsymbol{\theta}'}{\sigma_y^2 - 2\boldsymbol{\theta}'^T\mathbf{T}'^T\mathbf{R}_{\mathbf{y}}\mathbf{i} + \boldsymbol{\theta}'^T\boldsymbol{\theta}'}$$
$$= \frac{\sigma_x^2 - \lambda_{\mathbf{t}_L}\left(2\boldsymbol{\theta}'^T\mathbf{T}'^T\mathbf{R}_{\mathbf{x}}\mathbf{i} - \boldsymbol{\theta}'^T\boldsymbol{\theta}'\right)}{\sigma_y^2 - \left(2\boldsymbol{\theta}'^T\mathbf{T}'^T\mathbf{R}_{\mathbf{y}}\mathbf{i} - \boldsymbol{\theta}'^T\boldsymbol{\theta}'\right)}.$$

Maximizing the previous expression is equivalent to minimizing the quantity $2\boldsymbol{\theta}'^T\mathbf{T}'^T\mathbf{R}_{\mathbf{y}}\mathbf{i} - \boldsymbol{\theta}'^T\boldsymbol{\theta}'$. Therefore, we get another Wiener-type filter:

$$\mathbf{h}_{\mathrm{W},4} = \left(\mathbf{I}_L - \mathbf{T}'\mathbf{T}'^T\mathbf{R}_{\mathbf{y}}\right)\mathbf{i}. \tag{4.112}$$

Because of the relation (4.21), the optimization of the SPCC between the filter output and the filtered noise signal, i.e., $\rho^2_{zv_{\mathrm{fn}}}(\mathbf{h})$, will lead to the same optimal filters derived in this subsection.

4.3.4 Other Possibilities

Obviously, it is possible to derive other noise reduction filters by combining some of the defined SPCCs. Here, we briefly discuss one valuable possibility.

In this approach, we combine the two SPCCs:

$$\rho^2_{z,v_{\mathrm{fn}}}(\mathbf{h}) + \rho^2_{z,x_{\mathrm{fi}}}(\mathbf{h}) = \frac{\mathbf{h}^T\left(\mathbf{R}_{\mathbf{v}} + \mathbf{R}_{\mathbf{x}_{\mathrm{i}}}\right)\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{\mathbf{y}}\mathbf{h}} \tag{4.113}$$
$$= \rho^2_{z,v_{\mathrm{fn}}+x_{\mathrm{fi}}}(\mathbf{h}),$$

where $\mathbf{x}_{\mathrm{i}}(t)$ is considered as an uncorrelated interference vector. Clearly, the filter that minimizes (4.113) will make $z(t)$ the estimate of $x(t)$. A maximal value of $\rho^2_{z,v_{\mathrm{fn}}+x_{\mathrm{fi}}}(\mathbf{h})$ implies that $z(t)$ will be the estimate of $v(t)+x_{\mathrm{i}}(t)$ and, as a result, the estimate of $x(t)$ will be $y(t) - z(t)$. It is easy to derive all relevant filters by following the same steps as above.

References

1. J. Benesty, J. Chen, and Y. Huang, "On the importance of the Pearson correlation coefficient in noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 757–765, May 2008.
2. J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.

3. J. Yu, J. Benesty, G. Huang, and J. Chen, "Examples of optimal noise reduction filters derived from the squared Pearson correlation coefficient," in *Proc. IEEE ICASSP*, 2014, pp. 1571–1575.
4. J. Yu, J. Benesty, G. Huang, and J. Chen, "Optimal single-channel noise reduction filtering matrices from the Pearson correlation coefficient perspective," in *Proc. IEEE ICASSP*, 2015, pp. 201–205.
5. J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.
6. J. Benesty, J. Chen, Y. Huang, and T. Gaensler, "Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction," *J. Acoust. Soc. Am.*, vol. 132, pp. 452–464, July 2012.
7. J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE ICASSP*, 2011, pp. 273–276.
8. A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, pp. 1355-1365, Sept. 2014.
9. D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE ICASSP*, 2016, pp. 201–205.
10. D. Fischer, S. Doclo, E. A. P. Habets, and T. Gerkmann, "Combined single-microphone Wiener and MVDR filtering based on speech interframe correlations and speech presence probability," in *Proc. ITG Conf. Speech Communication*, 2016, pp. 292–296.

Chapter 5
On the Output SNR in Speech Enhancement and Beamforming

The output SNR is a well-known and accurate measure of the SNR after the filtering/beamforming operation; it is widely used to evaluate all kinds of optimal filters/beamformers for speech enhancement. However, it has never really been fully exploited for the derivation of other noise reduction filters than the classical maximum SNR filter. In this chapter, we first show how the output SNR is related to the fullmode input SNR and, then, derive very interesting filters/beamformers by alternating between two related filters in the maximization and minimization of the output SNR.

5.1 Signal Model and Problem Formulation

We consider the signal model in which we have M observed signals in a vector form [1]:

$$\begin{aligned}\mathbf{y} &= \left[\, Y_1 \; Y_2 \; \cdots \; Y_M \,\right]^T \\ &= \mathbf{x} + \mathbf{v},\end{aligned} \tag{5.1}$$

where $\mathbf{y}$ is the noisy (observed) signal vector, $\mathbf{x}$ is the speech signal vector, $\mathbf{v}$ is the noise signal vector, and vectors $\mathbf{x}$ and $\mathbf{v}$ are defined similarly to vector $\mathbf{y}$. All signals are assumed to be random, complex, circular, zero mean, and stationary. Furthermore, the vectors $\mathbf{x}$ and $\mathbf{v}$ are assumed to be uncorrelated, i.e., $E\left(\mathbf{x}\mathbf{v}^H\right) = \mathbf{0}$. It can be verified that the signal model in (5.1) encompasses all aspects of speech enhancement and beamforming, from the single-channel to the multichannel scenario, in the time, frequency, and time-frequency domains. In the particular case of beamforming and taking the first microphone as the reference, (5.1) is expressed as [1], [2]

$$\mathbf{y} = X_1\mathbf{d} + \mathbf{v}, \tag{5.2}$$

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4_5

where $\mathbf{d}$ is the (deterministic) steering vector of length M, whose first entry is equal to 1.

Then, with the first element of $\mathbf{y}$ being the reference, which will always be true here, our objective in the general case is to estimate X_1, i.e., the first element of $\mathbf{x}$, given $\mathbf{y}$ uniquely from the output SNR, which is a good measure of the SNR after linear processing. We want to show that the output SNR is also an excellent criterion from which optimal filters/beamformers can be derived.

From (5.1), we deduce that the covariance matrix (of size $M \times M$) of $\mathbf{y}$ is

$$\begin{aligned} \mathbf{\Phi}_{\mathbf{y}} &= E\left(\mathbf{y}\mathbf{y}^H\right) \\ &= \mathbf{\Phi}_{\mathbf{x}} + \mathbf{\Phi}_{\mathbf{v}}, \end{aligned} \tag{5.3}$$

where $\mathbf{\Phi}_{\mathbf{x}} = E\left(\mathbf{x}\mathbf{x}^H\right)$ and $\mathbf{\Phi}_{\mathbf{v}} = E\left(\mathbf{v}\mathbf{v}^H\right)$ are the covariance matrices of $\mathbf{x}$ and $\mathbf{v}$, respectively. It will always be assumed that $\mathbf{\Phi}_{\mathbf{v}}$ has full rank. For the covariance matrix $\mathbf{\Phi}_{\mathbf{x}}$, we are interested in three cases that often appear in the problem of speech enhancement. They are the following. Case 1: $\text{rank}\left(\mathbf{\Phi}_{\mathbf{x}}\right) = 1$ [and corresponds to the signal model in (5.2)], which implies that $\mathbf{\Phi}_{\mathbf{x}} = \phi_{X_1}\mathbf{d}\mathbf{d}^H$, where $\phi_{X_1} = E\left(|X_1|^2\right)$ is the variance of X_1. Case 2: $\text{rank}\left(\mathbf{\Phi}_{\mathbf{x}}\right) = P$, where $1 \leq P < M$. Case 3: $\text{rank}\left(\mathbf{\Phi}_{\mathbf{x}}\right) = M$, i.e., $\mathbf{\Phi}_{\mathbf{x}}$ has full rank. From (5.3), we can define the input SNR as

$$\begin{aligned} \text{iSNR} &= \frac{\text{tr}\left(\mathbf{\Phi}_{\mathbf{x}}\right)}{\text{tr}\left(\mathbf{\Phi}_{\mathbf{v}}\right)} \\ &= \frac{\phi_{X_1}}{\phi_{V_1}}, \end{aligned} \tag{5.4}$$

where $\phi_{V_1} = E\left(|V_1|^2\right)$ is the variance of V_1, i.e., the first component of $\mathbf{v}$. In (5.4), it is explicitly assumed that $\phi_{X_1} \approx \text{tr}\left(\mathbf{\Phi}_{\mathbf{x}}\right)/M$ and $\phi_{V_1} \approx \text{tr}\left(\mathbf{\Phi}_{\mathbf{v}}\right)/M$, which is almost always the case in practice. With this conventional definition of the SNR, we conclude this section.

5.2 Linear Filtering, Output and Fullmode Input SNRs

In this chapter, we estimate the desired speech signal, X_1, or the noise signal, V_1, by applying a complex-valued filter, $\mathbf{h}$, of length M, to the noisy signal vector, $\mathbf{y}$, i.e.,

$$\begin{aligned} Z &= \mathbf{h}^H\mathbf{y} \\ &= X_{\text{fd}} + V_{\text{fn}}, \end{aligned} \tag{5.5}$$

where Z can be either the estimate of X_1 or V_1,

$$X_{\text{fd}} = \mathbf{h}^H \mathbf{x} \tag{5.6}$$

is the filtered desired signal, and

$$V_{\text{fn}} = \mathbf{h}^H \mathbf{v} \tag{5.7}$$

is the filtered noise signal. If Z is the estimate of V_1, then the estimate of X_1 is

$$\begin{aligned} \widehat{X}_1 &= Y_1 - Z \\ &= Y_1 - \mathbf{h}^H \mathbf{y} \\ &= (\mathbf{i} - \mathbf{h})^H \mathbf{y}, \end{aligned} \tag{5.8}$$

where $\mathbf{i}$ is the first column of the $M \times M$ identity matrix $\mathbf{I}_M$. In the rest, we will also use the notation $\mathbf{h}_X$ and $\mathbf{h}_V$. The first filter, $\mathbf{h}_X$, corresponds to the estimation of X_1 while the second filter, $\mathbf{h}_V$, corresponds to the estimation of V_1. Obviously, from (5.8), we have the relationship:

$$\mathbf{h}_X + \mathbf{h}_V = \mathbf{i}. \tag{5.9}$$

Therefore, when V_1 is estimated with $\mathbf{h}_V$, we can estimate X_1 with $\mathbf{h}_X$, thanks to the relation in (5.9). From (5.5), we see that the variance of Z is

$$\begin{aligned} \phi_Z &= E\left(|Z|^2\right) \\ &= \phi_{X_{\text{fd}}} + \phi_{V_{\text{fn}}}, \end{aligned} \tag{5.10}$$

where

$$\phi_{X_{\text{fd}}} = \mathbf{h}^H \mathbf{\Phi_x} \mathbf{h}, \tag{5.11}$$

$$\phi_{V_{\text{fn}}} = \mathbf{h}^H \mathbf{\Phi_v} \mathbf{h}. \tag{5.12}$$

As a result, the output SNR can be defined as

$$\text{oSNR}(\mathbf{h}) = \frac{\mathbf{h}^H \mathbf{\Phi_x} \mathbf{h}}{\mathbf{h}^H \mathbf{\Phi_v} \mathbf{h}}. \tag{5.13}$$

When the output SNR is maximized (resp. minimized), we write $\text{oSNR}(\mathbf{h}_X)$ [resp. $\text{oSNR}(\mathbf{h}_V)$] since in this case, the filter $\mathbf{h}_X$ (resp. $\mathbf{h}_V$) corresponds to the estimation of X_1 (resp. V_1). We will see that by alternating between the two filters $\mathbf{h}_X$ and $\mathbf{h}_V$ in the optimization (i.e., maximization or minimization) of the output SNR, we can derive very interesting filters/beamformers.

Given the structure of the output SNR, which is simply the generalized Rayleigh quotient, joint diagonalization is going to be a very natural tool to exploit here. The two Hermitian matrices $\mathbf{\Phi_x}$ and $\mathbf{\Phi_v}$ can be jointly diagonalized as follows [3]:

$$\mathbf{A}^H \mathbf{\Phi}_{\mathbf{x}} \mathbf{A} = \mathbf{\Lambda}, \tag{5.14}$$

$$\mathbf{A}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{A} = \mathbf{I}_M, \tag{5.15}$$

where

$$\mathbf{A} = \left[\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_M \right] \tag{5.16}$$

is a full-rank square matrix (of size $M \times M$) and

$$\mathbf{\Lambda} = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_M\right) \tag{5.17}$$

is a diagonal matrix whose main elements are real and nonnegative. The eigenvalues of $\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{x}}$ are ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0$. We also denote by $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$, the corresponding eigenvectors.

The procedure for jointly diagonalizing $\mathbf{\Phi}_{\mathbf{x}}$ and $\mathbf{\Phi}_{\mathbf{v}}$ consists of two steps [4].

(i) Calculate $\mathbf{\Lambda}$ and $\mathbf{A}'$, the eigenvalue and (unnormalized) eigenvector matrices, respectively, of $\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{x}}$, i.e.,

$$\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{x}}\mathbf{A}' = \mathbf{A}'\mathbf{\Lambda}. \tag{5.18}$$

(ii) Normalize the eigenvectors of $\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{x}}$ such that (5.15) is satisfied. Denoting by $\mathbf{a}'_m$, $m = 1, 2, \ldots, M$ the (unnormalized) eigenvectors of $\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{\Phi}_{\mathbf{x}}$, then we need to find the constants C_m's such that $\mathbf{a}_m = C_m \mathbf{a}'_m$ satisfy $\mathbf{a}_m^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{a}_m = 1$. Hence,

$$C_m = \frac{1}{\sqrt{\mathbf{a}_m'^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{a}'_m}}, \ m = 1, 2, \ldots, M. \tag{5.19}$$

Thus, we have

$$\mathbf{A} = \mathbf{A}'\mathbf{C}, \tag{5.20}$$

where $\mathbf{C}$ is a diagonal normalization matrix with the elements $\{C_1, C_2, \ldots, C_M\}$ on its main diagonal.

In the particular case of $\mathrm{rank}\left(\mathbf{\Phi}_{\mathbf{x}}\right) = 1$, we have

$$\mathbf{A} = \left[\mathbf{a}_1 \ \mathbf{A}_2 \right], \tag{5.21}$$

$$\mathbf{\Lambda} = \mathrm{diag}\left(\lambda_1, 0, \ldots, 0\right), \tag{5.22}$$

where

$$\mathbf{a}_1 = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}}{\sqrt{\mathbf{d}^H \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}}} \tag{5.23}$$

and

$$\lambda_1 = \phi_{X_1} \mathbf{d}^H \mathbf{\Phi}_{\mathbf{v}}^{-1} \mathbf{d}. \tag{5.24}$$

It is always possible to write $\mathbf{h}$ in a basis formed from the vectors $\mathbf{a}_m,\ m = 1, 2, \ldots, M$, i.e.,

$$\mathbf{h} = \mathbf{A}\boldsymbol{\alpha}, \tag{5.25}$$

where the components, $\alpha_m,\ m = 1, 2, \ldots, M$, of the vector $\boldsymbol{\alpha}$ are the coordinates of $\mathbf{h}$ in the new basis. As a consequence, the output SNR in (5.13) can be rewritten, equivalently, as

$$\text{oSNR}\left(\boldsymbol{\alpha}\right) = \frac{\boldsymbol{\alpha}^H \mathbf{\Lambda} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^H \boldsymbol{\alpha}}. \tag{5.26}$$

Another possible measure of the SNR, which can be close to the input SNR, is the fullmode input SNR defined as

$$\begin{aligned} \text{iSNR}_{\text{FM}} &= \frac{\text{tr}\left(\mathbf{\Phi}_{\mathbf{v}}^{-1} \mathbf{\Phi}_{\mathbf{x}}\right)}{M} \\ &= \frac{\text{tr}\left(\mathbf{A}\mathbf{A}^H \mathbf{\Phi}_{\mathbf{x}}\right)}{M} \\ &= \frac{\text{tr}\left(\mathbf{\Lambda}\right)}{M}. \end{aligned} \tag{5.27}$$

From the previous expression, we define the mth ($m = 1, 2, \ldots, M$) spectral mode input SNR:

$$\text{iSNR}_m = \lambda_m. \tag{5.28}$$

The number of nonnull spectral modes is obviously equal to the rank of $\mathbf{\Phi}_{\mathbf{x}}$. So in the case of $\text{rank}\left(\mathbf{\Phi}_{\mathbf{x}}\right) = 1$, the first spectral mode input SNR is equal to M times the fullmode input SNR, i.e., $\text{iSNR}_1 = M \times \text{iSNR}_{\text{FM}}$ and $\text{iSNR}_i = 0,\ i = 2, 3, \ldots, M$. As a result, we can express the fullmode input and output SNRs as, respectively,

$$\text{iSNR}_{\text{FM}} = \frac{\sum_{m=1}^{M} \text{iSNR}_m}{M} \tag{5.29}$$

and

$$\text{oSNR}\left(\boldsymbol{\alpha}\right) = \frac{\sum_{m=1}^{M} \left|\alpha_m\right|^2 \text{iSNR}_m}{\sum_{m=1}^{M} \left|\alpha_m\right|^2}. \tag{5.30}$$

Since

$$\text{tr}\left(\mathbf{\Phi}_{\mathbf{v}}^{-1} \mathbf{\Phi}_{\mathbf{x}}\right) \geq \frac{\text{tr}\left(\mathbf{\Phi}_{\mathbf{x}}\right)}{\text{tr}\left(\mathbf{\Phi}_{\mathbf{v}}\right)},$$

it follows that

$$\mathrm{iSNR}_{\mathrm{FM}} \geq \frac{\mathrm{iSNR}}{M}, \tag{5.31}$$

or, equivalently,

$$\sum_{m=1}^{M} \mathrm{iSNR}_m \geq \mathrm{iSNR}. \tag{5.32}$$

Property 5.1. Let

$$\mathrm{cond}\left(\mathbf{\Phi_v}\right) = \frac{\lambda_1\left(\mathbf{\Phi_v}\right)}{\lambda_M\left(\mathbf{\Phi_v}\right)} \tag{5.33}$$

be the condition number of the matrix $\mathbf{\Phi_v}$, where $\lambda_1\left(\mathbf{\Phi_v}\right)$ and $\lambda_M\left(\mathbf{\Phi_v}\right)$ are, respectively, the largest and smallest eigenvalues of $\mathbf{\Phi_v}$. We have

$$\frac{\mathrm{iSNR}}{\mathrm{cond}\left(\mathbf{\Phi_v}\right)} \leq \mathrm{iSNR}_{\mathrm{FM}} \leq \mathrm{cond}\left(\mathbf{\Phi_v}\right) \times \mathrm{iSNR}, \tag{5.34}$$

with $\mathrm{iSNR}_{\mathrm{FM}} = \mathrm{iSNR}$ if and only if $\mathrm{cond}\left(\mathbf{\Phi_v}\right) = 1$.

Proof. Since $\mathbf{\Phi_v}$ is a positive definite matrix and $\mathbf{\Phi_x}$ is a positive semidefinite matrix, it can be shown that

$$\frac{\mathrm{tr}\left(\mathbf{\Phi_x}\right)}{\lambda_1\left(\mathbf{\Phi_v}\right)} \leq \mathrm{tr}\left(\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\right) \leq \frac{\mathrm{tr}\left(\mathbf{\Phi_x}\right)}{\lambda_M\left(\mathbf{\Phi_v}\right)}. \tag{5.35}$$

But

$$\begin{aligned}\frac{\mathrm{tr}\left(\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\right)}{M} &\leq \frac{\mathrm{tr}\left(\mathbf{\Phi_v}\right)}{M\lambda_M\left(\mathbf{\Phi_v}\right)}\mathrm{iSNR} \\ &\leq \frac{M\lambda_1\left(\mathbf{\Phi_v}\right)}{M\lambda_M\left(\mathbf{\Phi_v}\right)}\mathrm{iSNR} \\ &\leq \mathrm{cond}\left(\mathbf{\Phi_v}\right) \times \mathrm{iSNR}\end{aligned} \tag{5.36}$$

and

$$\begin{aligned}\frac{\mathrm{tr}\left(\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\right)}{M} &\geq \frac{\mathrm{tr}\left(\mathbf{\Phi_v}\right)}{M\lambda_1\left(\mathbf{\Phi_v}\right)}\mathrm{iSNR} \\ &\geq \frac{M\lambda_M\left(\mathbf{\Phi_v}\right)}{M\lambda_1\left(\mathbf{\Phi_v}\right)}\mathrm{iSNR} \\ &\geq \frac{\mathrm{iSNR}}{\mathrm{cond}\left(\mathbf{\Phi_v}\right)}.\end{aligned} \tag{5.37}$$

What does the fullmode input SNR mean? We can see that it has the potential to be quite large and much larger than the conventional input SNR, depending on the condition number of $\mathbf{\Phi}_{\mathbf{v}}$. The most interesting and insightful part of the fullmode input SNR is its decomposition into different spectral modes, which clearly shows the repartition of the SNR at different spectral bands. So when $\mathrm{cond}\left(\mathbf{\Phi}_{\mathbf{v}}\right)$ is large, this means that the fullmode input SNR is mostly governed by its largest modes. A great consequence of this is that it tells us what amount of the output SNR we can expect with a linear filter since this amount is always upper bounded by the maximum spectral mode input SNR. In other words, the fullmode input SNR gives us great insights into the potential of noise reduction while the conventional input SNR definition does not lead to much interpretation except for its main purpose.

From the formulation of the output SNR in (5.30), which weights the different spectral modes of the fullmode input SNR, three obvious particular cases of $\boldsymbol{\alpha}$ appear naturally. The first one is the equal-coordinate filter, i.e., $\boldsymbol{\alpha} = \alpha\mathbf{1} = \alpha\left[1\ 1\ \cdots\ 1\right]^T$, where $\alpha \neq 0$, which equally weights the different modes; therefore

$$\mathrm{oSNR}\left(\alpha\mathbf{1}\right) = \mathrm{oSNR}\left(\mathbf{1}\right) = \mathrm{iSNR}_{\mathrm{FM}}.$$

The second particular case is the maximum SNR filter, i.e., $\boldsymbol{\alpha}_{\max} = \left[\alpha_1\ 0\ \cdots\ 0\right]^T$, where $\alpha_1 \neq 0$, which gives the maximum value of the output SNR, i.e.,

$$\mathrm{oSNR}\left(\boldsymbol{\alpha}_{\max}\right) = \mathrm{iSNR}_1 \geq \mathrm{oSNR}\left(\boldsymbol{\alpha}\right),\ \forall\boldsymbol{\alpha} \neq \mathbf{0},$$

or, equivalently,

$$\mathrm{oSNR}\left(\mathbf{h}_{\max}\right) = \lambda_1 \geq \mathrm{oSNR}\left(\mathbf{h}\right),\ \forall\mathbf{h} \neq \mathbf{0},$$

where $\mathbf{h}_{\max} = \mathbf{A}\boldsymbol{\alpha}_{\max}$. Finally, the last one is the minimum SNR filter, i.e., $\boldsymbol{\alpha}_{\min} = \left[0\ \cdots\ 0\ \alpha_M\right]^T$, where $\alpha_M \neq 0$, which gives the minimum value of the output SNR, i.e.,

$$\mathrm{oSNR}\left(\boldsymbol{\alpha}_{\min}\right) = \mathrm{iSNR}_M \leq \mathrm{oSNR}\left(\boldsymbol{\alpha}\right),\ \forall\boldsymbol{\alpha} \neq \mathbf{0},$$

or, equivalently,

$$\mathrm{oSNR}\left(\mathbf{h}_{\min}\right) = \lambda_M \leq \mathrm{oSNR}\left(\mathbf{h}\right),\ \forall\mathbf{h} \neq \mathbf{0},$$

where $\mathbf{h}_{\min} = \mathbf{A}\boldsymbol{\alpha}_{\min}$. Also, by playing on the values of the α_m's, we can precisely manipulate the different spectral modes of the fullmode input SNR as we wish for speech enhancement. In other words, improving the SNR with a linear filter is just a matter of adjusting the different spectral mode input SNRs, showing the importance of the fullmode input SNR definition. From the above, we see that we always have

$$\begin{aligned}&\mathrm{iSNR}_M \leq \mathrm{iSNR}_{\mathrm{FM}} \leq \mathrm{iSNR}_1,\\ &\mathrm{iSNR}_M \leq \mathrm{oSNR}\left(\boldsymbol{\alpha}\right) \leq \mathrm{iSNR}_1,\ \forall \boldsymbol{\alpha} \neq \mathbf{0}.\end{aligned}$$

Of course, for the estimation of the desired signal, X_1, we must always ensure that

$$\mathrm{oSNR}\left(\mathbf{h}_X\right) > \mathrm{oSNR}\left(\mathbf{i}\right) = \mathrm{iSNR}. \tag{5.38}$$

5.3 Optimal Filters

In this section, we develop a large class of optimal filters from the output SNR depending on the rank of the speech covariance matrix.

5.3.1 Rank-One Speech Covariance Matrix

When the rank of $\boldsymbol{\Phi}_{\mathbf{x}}$ is equal to 1, it is clear that the filter that maximizes the output SNR is proportional to $\mathbf{a}_1$ [see (5.23)], i.e.,

$$\mathbf{h}_X = \alpha \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}, \tag{5.39}$$

where $\alpha \neq 0$ is an arbitrary complex number.

Now, we need to determine α. This can be done by observing that while $\mathbf{h}_X$ maximizes the output SNR and gives the estimate of X_1, the output SNR with the filter $\mathbf{h}_V = \mathbf{i} - \alpha \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}$ can also be minimized in order to get the estimate of V_1. Substituting $\mathbf{h}_V$ into (5.13), we get

$$\mathrm{oSNR}\left(\alpha\right) = \frac{\phi_{X_1}\left(\mathbf{i} - \alpha \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}\right)^H \mathbf{d}\mathbf{d}^H \left(\mathbf{i} - \alpha \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}\right)}{\left(\mathbf{i} - \alpha \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}\right)^H \boldsymbol{\Phi}_{\mathbf{v}} \left(\mathbf{i} - \alpha \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}\right)}. \tag{5.40}$$

Minimizing the previous expression is equivalent to minimizing its numerator. Therefore, we have

$$\alpha = \frac{1}{\mathbf{d}^H \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}}. \tag{5.41}$$

Substituting α back into (5.40), we see that $\mathrm{oSNR}\left(\alpha\right) = 0$, proving that the output SNR is indeed minimized. As a result, we deduce the celebrated MVDR filter:

$$\mathbf{h}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}}{\mathbf{d}^H \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}}. \tag{5.42}$$

Let us turn our attention to the estimation of V_1 in the first step. It is cleat that the filter:

$$\mathbf{h}_V = \mathbf{A}_2\boldsymbol{\alpha}_2, \tag{5.43}$$

where $\boldsymbol{\alpha}_2 \neq \mathbf{0}$ is a vector of length $M-1$, minimizes the output SNR since $\text{oSNR}(\mathbf{h}_V) = 0$. To obtain the estimate of X_1, we plug $\mathbf{h}_X = \mathbf{i} - \mathbf{A}_2\boldsymbol{\alpha}_2$ in the definition of the output SNR, resulting in

$$\begin{aligned}\text{oSNR}(\boldsymbol{\alpha}_2) &= \frac{\phi_{X_1}(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha}_2)^H \mathbf{d}\mathbf{d}^H(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha}_2)}{(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha}_2)^H \boldsymbol{\Phi}_{\mathbf{v}}(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha}_2)} \\ &= \frac{\phi_{X_1}}{(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha}_2)^H \boldsymbol{\Phi}_{\mathbf{v}}(\mathbf{i}-\mathbf{A}_2\boldsymbol{\alpha}_2)}. \end{aligned} \tag{5.44}$$

The maximization of $\text{oSNR}(\boldsymbol{\alpha}_2)$ is equivalent to the minimization of its denominator. We easily get

$$\boldsymbol{\alpha}_2 = \mathbf{A}_2^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{i} \tag{5.45}$$

and the optimal filter for the estimation of X_1 is

$$\begin{aligned}\mathbf{h}_X &= \mathbf{i} - \mathbf{A}_2\mathbf{A}_2^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{i} && (5.46)\\ &= \mathbf{i} - \left(\boldsymbol{\Phi}_{\mathbf{v}}^{-1} - \mathbf{a}_1\mathbf{a}_1^H\right)\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{i} \\ &= \left(\mathbf{a}_1^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{i}\right)\mathbf{a}_1 \\ &= \frac{\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}}{\mathbf{d}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{d}} = \mathbf{h}_{\text{MVDR}}, && (5.47)\end{aligned}$$

which is again the MVDR filter.

5.3.2 Rank-Deficient Speech Covariance Matrix

In this subsection, we focus on the case where $\text{rank}(\boldsymbol{\Phi}_{\mathbf{x}}) = P$ with $1 \leq P < M$. We already know that the filter that maximizes the output SNR is

$$\mathbf{h}_X = \alpha_1\mathbf{a}_1, \tag{5.48}$$

where $\alpha_1 \neq 0$ is an arbitrary complex number. To find α_1, we use the filter $\mathbf{h}_V = \mathbf{i} - \alpha_1\mathbf{a}_1$ in the output SNR, which leads to

$$\begin{aligned}\mathrm{oSNR}\left(\alpha_1\right) &= \frac{\left(\mathbf{i}-\alpha_1\mathbf{a}_1\right)^H \mathbf{\Phi}_{\mathbf{x}} \left(\mathbf{i}-\alpha_1\mathbf{a}_1\right)}{\left(\mathbf{i}-\alpha_1\mathbf{a}_1\right)^H \mathbf{\Phi}_{\mathbf{v}} \left(\mathbf{i}-\alpha_1\mathbf{a}_1\right)} \\ &= \frac{\phi_{X_1} - \lambda_1 \left[2\Re\left(\alpha_1 \mathbf{i}^T \mathbf{\Phi}_{\mathbf{v}} \mathbf{a}_1\right) - \left|\alpha_1\right|^2\right]}{\phi_{V_1} - \left[2\Re\left(\alpha_1 \mathbf{i}^T \mathbf{\Phi}_{\mathbf{v}} \mathbf{a}_1\right) - \left|\alpha_1\right|^2\right]},\end{aligned} \tag{5.49}$$

and whose minimization gives

$$\begin{aligned}\alpha_1 &= \mathbf{a}_1^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i} \\ &= \frac{\mathbf{a}_1^H \mathbf{\Phi}_{\mathbf{x}} \mathbf{i}}{\lambda_1},\end{aligned} \tag{5.50}$$

where $\Re(\cdot)$ is the real part of a complex number. We deduce the maximum SNR filter with minimum distortion (MD):

$$\begin{aligned}\mathbf{h}_{\mathrm{mMD}} &= \frac{\mathbf{a}_1 \mathbf{a}_1^H \mathbf{\Phi}_{\mathbf{x}} \mathbf{i}}{\lambda_1} \\ &= \mathbf{a}_1 \mathbf{a}_1^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i}.\end{aligned} \tag{5.51}$$

Obviously, this filter is very much different from the MVDR filter in (5.42) since $\mathbf{a}_1$ does not have the form in (5.23), in general. In fact, the larger the value of P, the more different are the two filters. While $\mathbf{h}_{\mathrm{mMD}}$ gives the maximum possible output SNR, speech distortion worsens as P increases. However, for $P = 1$, $\mathbf{h}_{\mathrm{MVDR}}$ and $\mathbf{h}_{\mathrm{mMD}}$ are identical.

Now, let us derive the optimal filter when V_1 is estimated first. Define the matrix of size $M \times (M - P)$:

$$\mathbf{A}_{P+1} = \left[\mathbf{a}_{P+1}\ \mathbf{a}_{P+2} \cdots\ \mathbf{a}_M\right]. \tag{5.52}$$

One can verify that the filter:

$$\mathbf{h}_V = \mathbf{A}_{P+1} \boldsymbol{\alpha}_{P+1}, \tag{5.53}$$

where $\boldsymbol{\alpha}_{P+1} \neq \mathbf{0}$ is a vector of length $M-P$, minimizes the output SNR since $\mathrm{oSNR}\left(\mathbf{h}_V\right) = 0$. To get the estimate of X_1, we insert $\mathbf{h}_X = \mathbf{i} - \mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}$ in the definition of the output SNR, resulting in

$$\begin{aligned}\mathrm{oSNR}\left(\boldsymbol{\alpha}_{P+1}\right) &= \frac{\left(\mathbf{i}-\mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}\right)^H \mathbf{\Phi}_{\mathbf{x}} \left(\mathbf{i}-\mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}\right)}{\left(\mathbf{i}-\mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}\right)^H \mathbf{\Phi}_{\mathbf{v}} \left(\mathbf{i}-\mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}\right)} \\ &= \frac{\phi_{X_1}}{\left(\mathbf{i}-\mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}\right)^H \mathbf{\Phi}_{\mathbf{v}} \left(\mathbf{i}-\mathbf{A}_{P+1}\boldsymbol{\alpha}_{P+1}\right)}.\end{aligned} \tag{5.54}$$

The maximization of the previous expression gives

$$\boldsymbol{\alpha}_{P+1} = \mathbf{A}_{P+1}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i}. \tag{5.55}$$

As a result, we obtain the distortionless (DL) filter:

$$\mathbf{h}_{\mathrm{DL}} = \mathbf{i} - \mathbf{A}_{P+1}\mathbf{A}_{P+1}^{H}\mathbf{\Phi}_{\mathbf{v}}\mathbf{i}. \tag{5.56}$$

This filter is, indeed, distortionless since

$$\begin{aligned}\mathbf{h}_{\mathrm{DL}}^{T}\mathbf{x} &= X_1 - \mathbf{i}^{T}\mathbf{\Phi}_{\mathbf{v}}\mathbf{A}_{P+1}\mathbf{A}_{P+1}^{H}\mathbf{x} \\ &= X_1,\end{aligned} \tag{5.57}$$

where we used the fact that $\mathbf{A}_{P+1}^{H}\mathbf{x} = \mathbf{0}$ [derived from (5.14)]. For $P = 1$, one can check that $\mathbf{h}_{\mathrm{DL}}$ and $\mathbf{h}_{\mathrm{MVDR}}$ are identical. As P increases, the output SNR of $\mathbf{h}_{\mathrm{DL}}$ decreases.

5.3.3 Full-Rank Speech Covariance Matrix

When $\mathbf{\Phi}_{\mathbf{x}} = M$, we can also derive the maximum SNR filter with minimum distortion, i.e., $\mathbf{h}_{\mathrm{mMD}}$. However, this filter may lead to very large distortions since it considers only the main direction of the desired signal as compared to the noise, i.e., the maximum spectral mode of the fullmode input SNR. In order to reduce distortion, we need to consider more than one spectral mode but at the price of a lower output SNR. This is the classical compromise between noise reduction and speech distortion that we clearly see from this formulation, which can lead to much more accurate compromises than those obtained from some conventional approaches.

Let us consider the Q ($1 \leq Q \leq M$) largest spectral modes of the fullmode input SNR. For that, we define the matrix of size $M \times Q$:

$$\mathbf{A}_{1:Q} = \left[\, \mathbf{a}_1 \; \mathbf{a}_2 \; \cdots \; \mathbf{a}_Q \,\right]. \tag{5.58}$$

We choose filters of the form:

$$\mathbf{h}_{X,Q} = \mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}, \tag{5.59}$$

where $\boldsymbol{\alpha}_{1:Q} \neq \mathbf{0}$ is a vector of length Q. To find $\boldsymbol{\alpha}_{1:Q}$, we use the filter $\mathbf{h}_{V,Q} = \mathbf{i} - \mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}$ in the output SNR, which leads to

$$\begin{aligned}\mathrm{oSNR}\left(\boldsymbol{\alpha}_{1:Q}\right) &= \frac{\left(\mathbf{i} - \mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right)^{H}\mathbf{\Phi}_{\mathbf{x}}\left(\mathbf{i} - \mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right)}{\left(\mathbf{i} - \mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right)^{H}\mathbf{\Phi}_{\mathbf{v}}\left(\mathbf{i} - \mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right)} \\ &= \frac{\phi_{X_1} - \left[2\Re\left(\mathbf{i}^{T}\mathbf{\Phi}_{\mathbf{v}}\mathbf{A}_{1:Q}\mathbf{\Lambda}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right) - \boldsymbol{\alpha}_{1:Q}^{H}\mathbf{\Lambda}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right]}{\phi_{V_1} - \left[2\Re\left(\mathbf{i}^{T}\mathbf{\Phi}_{\mathbf{v}}\mathbf{A}_{1:Q}\boldsymbol{\alpha}_{1:Q}\right) - \boldsymbol{\alpha}_{1:Q}^{H}\boldsymbol{\alpha}_{1:Q}\right]},\end{aligned} \tag{5.60}$$

where

$$\mathbf{\Lambda}_{1:Q} = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_Q\right). \tag{5.61}$$

The minimization of (5.60) gives

$$\begin{aligned} \boldsymbol{\alpha}_{1:Q} &= \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i} \\ &= \mathbf{\Lambda}_{1:Q}^{-1} \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{x}} \mathbf{i}. \end{aligned} \tag{5.62}$$

We deduce the first class of compromising filters:

$$\begin{aligned} \mathbf{h}_{X,1,Q} &= \mathbf{A}_{1:Q} \mathbf{\Lambda}_{1:Q}^{-1} \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{x}} \mathbf{i} \\ &= \mathbf{A}_{1:Q} \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i}. \end{aligned} \tag{5.63}$$

For $Q = 1$, we have the maximum SNR filter with MD, i.e., $\mathbf{h}_{X,1,Q} = \mathbf{h}_{\mathrm{mMD}}$, and for $Q = M$, we have the identity filter, i.e., $\mathbf{h}_{X,1,M} = \mathbf{i}$. We should always have

$$\mathrm{oSNR}\left(\mathbf{h}_{X,1,1}\right) \geq \mathrm{oSNR}\left(\mathbf{h}_{X,1,2}\right) \geq \cdots \geq \mathrm{oSNR}\left(\mathbf{h}_{X,1,M}\right) = \mathrm{iSNR}. \tag{5.64}$$

A very interesting particular case of (5.59) is

$$\mathbf{h}_{X,\mathbf{1},Q} = \alpha \mathbf{A}_{1:Q} \mathbf{1}_{1:Q}, \tag{5.65}$$

where $\alpha \neq 0$ and $\mathbf{1}_{1:Q}$ is a vector of length Q whose all elements are 1's. the parameter α is obtained as explained above. We get

$$\alpha = \frac{\mathbf{1}_{1:Q}^T \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i}}{Q} \tag{5.66}$$

As a result, the filter in (5.65) is

$$\mathbf{h}_{X,\mathbf{1},Q} = \frac{\mathbf{A}_{1:Q} \mathbf{1}_{1:Q} \mathbf{1}_{1:Q}^T \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{i}}{Q}. \tag{5.67}$$

What makes this filter so interesting is that its output SNR is

$$\begin{aligned} \mathrm{oSNR}\left(\mathbf{h}_{X,\mathbf{1},Q}\right) &= \frac{\mathbf{1}_{1:Q}^T \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{x}} \mathbf{A}_{1:Q} \mathbf{1}_{1:Q}}{\mathbf{1}_{1:Q}^T \mathbf{A}_{1:Q}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{A}_{1:Q} \mathbf{1}_{1:Q}} \\ &= \frac{\sum_{q=1}^{Q} \mathrm{iSNR}_q}{Q}. \end{aligned} \tag{5.68}$$

Therefore,

$$\mathrm{oSNR}\left(\mathbf{h}_{X,\mathbf{1},1}\right) \geq \mathrm{oSNR}\left(\mathbf{h}_{X,\mathbf{1},2}\right) \geq \cdots \geq \mathrm{oSNR}\left(\mathbf{h}_{X,\mathbf{1},M}\right) = \mathrm{iSNR}_{\mathrm{FM}}. \tag{5.69}$$

However, this filter may distort more the speech signal than $\mathbf{h}_{X,1,Q}$.

Now, let us consider the $M-R$ $(0 \leq R \leq M-1)$ smallest spectral modes of the fullmode input SNR and define the filters:

$$\mathbf{h}_{V,R} = \mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}, \tag{5.70}$$

where

$$\mathbf{A}_{R+1} = \left[\, \mathbf{a}_{R+1}\ \mathbf{a}_{R+2}\ \cdots\ \mathbf{a}_{M} \,\right] \tag{5.71}$$

is a matrix of size $M \times (M-R)$ and $\boldsymbol{\alpha}_{R+1} \neq \mathbf{0}$ is a vector of length $M-R$. Substituting $\mathbf{h}_{X,R} = \mathbf{i} - \mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}$ into the output SNR, we get

$$\begin{aligned} \mathrm{oSNR}\left(\boldsymbol{\alpha}_{R+1}\right) &= \frac{\left(\mathbf{i}-\mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}\right)^{H}\boldsymbol{\Phi}_{\mathbf{x}}\left(\mathbf{i}-\mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}\right)}{\left(\mathbf{i}-\mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}\right)^{H}\boldsymbol{\Phi}_{\mathbf{v}}\left(\mathbf{i}-\mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}\right)} \\ &= \frac{\phi_{X_1} - \left[2\Re\left(\mathbf{i}^{T}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{A}_{R+1}\boldsymbol{\Lambda}_{R+1}\boldsymbol{\alpha}_{R+1}\right) - \boldsymbol{\alpha}_{R+1}^{H}\boldsymbol{\Lambda}_{1:Q}\boldsymbol{\alpha}_{R+1}\right]}{\phi_{V_1} - \left[2\Re\left(\mathbf{i}^{T}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{A}_{R+1}\boldsymbol{\alpha}_{R+1}\right) - \boldsymbol{\alpha}_{R+1}^{H}\boldsymbol{\alpha}_{R+1}\right]}, \end{aligned} \tag{5.72}$$

where

$$\boldsymbol{\Lambda}_{R+1} = \mathrm{diag}\left(\lambda_{R+1}, \lambda_{R+1}, \ldots, \lambda_{M}\right). \tag{5.73}$$

From the maximization of (5.72), we obtain

$$\begin{aligned} \boldsymbol{\alpha}_{R+1} &= \mathbf{A}_{R+1}^{H}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{i} \\ &= \boldsymbol{\Lambda}_{R+1}^{-1}\mathbf{A}_{R+1}^{H}\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{i}. \end{aligned} \tag{5.74}$$

We deduce the second class of compromising filters:

$$\begin{aligned} \mathbf{h}_{X,2,R} &= \mathbf{i} - \mathbf{A}_{R+1}\boldsymbol{\Lambda}_{R+1}^{-1}\mathbf{A}_{R+1}^{H}\boldsymbol{\Phi}_{\mathbf{x}}\mathbf{i} \\ &= \mathbf{i} - \mathbf{A}_{R+1}\mathbf{A}_{R+1}^{H}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{i}, \end{aligned} \tag{5.75}$$

which is equivalent to the first class.

5.4 Application to Fixed and Superdirective Beamforming

We consider a plane wave, in the farfield, that propagates in an anechoic acoustic environment at the speed of sound, i.e., $c = 340$ m/s, and impinges on a uniform linear array (ULA) consisting of M omnidirectional microphones, where the distance between two successive sensors is equal to δ. The direction of the source signal to the array is parameterized by the azimuth angle θ. In this context, the steering vector (of length M) is given by

$$\mathbf{d}_\theta = \left[1 \; e^{-\jmath 2\pi f \tau_0 \cos\theta} \; \cdots \; e^{-\jmath (M-1) 2\pi f \tau_0 \cos\theta} \right]^T, \tag{5.76}$$

where $\jmath = \sqrt{-1}$ is the imaginary unit, $f > 0$ is the temporal frequency, and $\tau_0 = \delta/c$ is the delay between two successive sensors at the angle $\theta = 0$. Like in superdirective beamforming [5], [6], we assume that the main lobe is at the angle $\theta = 0$ (endfire direction) and the desired signal propagates from the same angle, so that the corresponding steering vector is $\mathbf{d}_0$. It will also be assumed that δ is small.

From the gain in SNR, two important measures, which do not depend on the statistics of the signals but on some noise models, are derived for fixed beamforming. They are the white noise gain (WNG):

$$\mathcal{W}(\mathbf{h}) = \frac{\left|\mathbf{h}^H \mathbf{d}_0\right|^2}{\mathbf{h}^H \mathbf{h}} \tag{5.77}$$

and the directivity factor (DF):

$$\mathcal{D}(\mathbf{h}) = \frac{\left|\mathbf{h}^H \mathbf{d}_0\right|^2}{\mathbf{h}^H \mathbf{\Gamma}_\mathrm{d} \mathbf{h}}, \tag{5.78}$$

where the elements of $\mathbf{\Gamma}_\mathrm{d}$ are given by

$$\begin{aligned} \left[\mathbf{\Gamma}_\mathrm{d}\right]_{ij} &= \frac{\sin\left[2\pi f (j-i)\tau_0\right]}{2\pi f (j-i)\tau_0} \\ &= \mathrm{sinc}\left[2\pi f (j-i)\tau_0\right]. \end{aligned} \tag{5.79}$$

The WNG is a measure of the sensitivity of the microphone array to some of its imperfections, such as sensor noise, while the DF quantifies how the same array performs in the presence of reverberation.

From the maximization of the WNG, we find the well-known delay-and-sum (DS) beamformer [4]:

$$\mathbf{h}_\mathrm{DS} = \frac{\mathbf{d}_0}{M}, \tag{5.80}$$

with $\mathcal{W}(\mathbf{h}_\mathrm{DS}) = M = \mathcal{W}_\mathrm{max}$. While the DS beamformer maximizes the WNG, it never amplifies the diffuse noise since $\mathcal{D}(\mathbf{h}_\mathrm{DS}) \geq 1$. However, this DF is not very large and the beampattern of the DS beamformer is very frequency dependent. If we maximize the DF, we easily get the superdirective beamformer [4], [5], [6]:

$$\mathbf{h}_\mathrm{S} = \frac{\mathbf{\Gamma}_\mathrm{d}^{-1} \mathbf{d}_0}{\mathbf{d}_0^H \mathbf{\Gamma}_\mathrm{d}^{-1} \mathbf{d}_0}, \tag{5.81}$$

with $\mathcal{D}(\mathbf{h}_S) = \mathbf{d}_0^H \boldsymbol{\Gamma}_d^{-1} \mathbf{d}_0 = \mathcal{D}_{max}$. While the superdirective beamformer maximizes the DF (leading to supergains), its WNG may be smaller than 1, which implies white noise amplification, especially at low frequencies.

Now, let us develop things from our perspective. Let us start by defining the set $\mathcal{S} = \{\mathbf{d}_0, \mathbf{i}_2, \ldots, \mathbf{i}_M\}$ containing M linearly independent vectors that span the M-dimensional Euclidean space, where $\mathbf{i}_i$ is the ith column of $\mathbf{I}_M$. Thanks to the Gram-Schmidt orthonormalization process, we can easily generate from $\mathcal{S}$ another set $\mathcal{S}_o = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M\}$ whose orthonormal vectors span the same space. It is cleat that

$$\mathbf{u}_1 = \frac{\mathbf{d}_0}{\sqrt{\mathbf{d}_0^H \mathbf{d}_0}} = \frac{\mathbf{d}_0}{\sqrt{M}} \tag{5.82}$$

and

$$\mathbf{u}_i^H \mathbf{d}_0 = 0, \ i = 2, 3, \ldots, M. \tag{5.83}$$

From $\mathcal{S}_o$, we can form the $M \times M$ unitary matrix:

$$\begin{aligned} \mathbf{U} &= \left[\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_M \right] \\ &= \left[\mathbf{u}_1 \ \mathbf{U}_2 \right], \end{aligned} \tag{5.84}$$

where $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}_M$.

The beamformer that minimizes both the WNG and the DF has the form:

$$\mathbf{h}_V = \mathbf{U}_2 \boldsymbol{\alpha}_2, \tag{5.85}$$

where $\boldsymbol{\alpha}_2 \neq \mathbf{0}$ is an arbitrary complex-valued vector of length $M-1$. Indeed, one can check that $\mathcal{W}(\mathbf{h}_V) = \mathcal{D}(\mathbf{h}_V) = 0$. Therefore, with $\mathbf{h}_V$, we can have the estimate of the diffuse-plus-white noise at the reference sensor. To have the estimate of the desired signal, we use the beamformer:

$$\mathbf{h}_X = \mathbf{i} - \mathbf{U}_2 \boldsymbol{\alpha}_2. \tag{5.86}$$

Substituting (5.86) into the definition of the WNG, we get

$$\begin{aligned} \mathcal{W}(\boldsymbol{\alpha}_2) &= \frac{(\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2)^H \mathbf{d}_0 \mathbf{d}_0^H (\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2)}{(\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2)^H (\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2)} \\ &= \frac{1}{1 - 2\Re\left(\boldsymbol{\alpha}_2^H \mathbf{U}_2^H \mathbf{i}\right) + \boldsymbol{\alpha}_2^H \boldsymbol{\alpha}_2}, \end{aligned} \tag{5.87}$$

whose maximization leads to

$$\boldsymbol{\alpha}_2 = \mathbf{U}_2^H \mathbf{i}. \tag{5.88}$$

As a consequence, the beamformer in (5.86) becomes

$$\begin{aligned}\mathbf{h}_X &= \mathbf{i} - \mathbf{U}_2\mathbf{U}_2^H\mathbf{i} \\ &= \mathbf{i} - \left(\mathbf{I}_M - \mathbf{u}_1\mathbf{u}_1^H\right)\mathbf{i} \\ &= \left(\mathbf{u}_1^H\mathbf{i}\right)\mathbf{u}_1 \\ &= \mathbf{h}_{\mathrm{DS}},\end{aligned} \tag{5.89}$$

which is another way to derive the DS beamformer.

Using again (5.86) but in the definition of the DF gives

$$\begin{aligned}\mathcal{D}\left(\boldsymbol{\alpha}_2\right) &= \frac{\left(\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2\right)^H \mathbf{d}_0\mathbf{d}_0^H\left(\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2\right)}{\left(\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2\right)^H \boldsymbol{\Gamma}_{\mathrm{d}}\left(\mathbf{i} - \mathbf{U}_2\boldsymbol{\alpha}_2\right)} \\ &= \frac{1}{1 - 2\Re\left(\boldsymbol{\alpha}_2^H\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{i}\right) + \boldsymbol{\alpha}_2^H\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{U}_2\boldsymbol{\alpha}_2}.\end{aligned} \tag{5.90}$$

From the maximization of the previous expression, we get

$$\boldsymbol{\alpha}_2 = \left(\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{U}_2\right)^{-1}\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{i}. \tag{5.91}$$

As a result, the beamformer in (5.86) becomes

$$\begin{aligned}\mathbf{h}_X &= \mathbf{i} - \mathbf{U}_2\left(\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{U}_2\right)^{-1}\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{i} \\ &= \boldsymbol{\Gamma}_{\mathrm{d}}^{-1/2}\left[\mathbf{I}_M - \boldsymbol{\Gamma}_{\mathrm{d}}^{1/2}\mathbf{U}_2\left(\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{U}_2\right)^{-1}\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}^{1/2}\right]\boldsymbol{\Gamma}_{\mathrm{d}}^{1/2}\mathbf{i} \\ &= \boldsymbol{\Gamma}_{\mathrm{d}}^{-1/2}\left(\frac{\boldsymbol{\Gamma}_{\mathrm{d}}^{-1/2}\mathbf{u}_1\mathbf{u}_1^H\boldsymbol{\Gamma}_{\mathrm{d}}^{-1/2}}{\mathbf{u}_1^H\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}\mathbf{u}_1}\right)\boldsymbol{\Gamma}_{\mathrm{d}}^{1/2}\mathbf{i} \\ &= \frac{\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}\mathbf{u}_1}{\sqrt{M}\times\mathbf{u}_1^H\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}\mathbf{u}_1} \\ &= \mathbf{h}_{\mathrm{S}},\end{aligned} \tag{5.92}$$

which is another way to derive the superdirective beamformer, where we have used the fact that

$$\mathbf{I}_M = \boldsymbol{\Gamma}_{\mathrm{d}}^{1/2}\mathbf{U}_2\left(\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{U}_2\right)^{-1}\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}^{1/2} + \frac{\boldsymbol{\Gamma}_{\mathrm{d}}^{-1/2}\mathbf{u}_1\mathbf{u}_1^H\boldsymbol{\Gamma}_{\mathrm{d}}^{-1/2}}{\mathbf{u}_1^H\boldsymbol{\Gamma}_{\mathrm{d}}^{-1}\mathbf{u}_1}. \tag{5.93}$$

Now, if we want to compromise between supergains and white noise amplification, we propose to maximize the DF subject to a constraint on the WNG, the same way it was done in [6]. This is equivalent to minimizing $1/\mathcal{D}\left(\boldsymbol{\alpha}_2\right)$ with a constraint on $1/\mathcal{W}\left(\boldsymbol{\alpha}_2\right)$, i.e., minimizing

$$\begin{aligned}\frac{1}{\mathcal{D}\left(\boldsymbol{\alpha}_2\right)} + \epsilon\frac{1}{\mathcal{W}\left(\boldsymbol{\alpha}_2\right)} &= 1 - 2\Re\left(\boldsymbol{\alpha}_2^H\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{i}\right) + \boldsymbol{\alpha}_2^H\mathbf{U}_2^H\boldsymbol{\Gamma}_{\mathrm{d}}\mathbf{U}_2\boldsymbol{\alpha}_2 \\ &\quad + \epsilon\left[1 - 2\Re\left(\boldsymbol{\alpha}_2^H\mathbf{U}_2^H\mathbf{i}\right) + \boldsymbol{\alpha}_2^H\boldsymbol{\alpha}_2\right],\end{aligned} \tag{5.94}$$

where $\epsilon \geq 0$ is a Lagrange multiplier. We easily find that

$$\boldsymbol{\alpha}_2 = \left(\mathbf{U}_2^H \boldsymbol{\Gamma}_{\mathrm{d},\epsilon} \mathbf{U}_2\right)^{-1} \mathbf{U}_2^H \boldsymbol{\Gamma}_{\mathrm{d},\epsilon} \mathbf{i}, \tag{5.95}$$

where

$$\boldsymbol{\Gamma}_{\mathrm{d},\epsilon} = \boldsymbol{\Gamma}_{\mathrm{d}} + \epsilon \mathbf{I}_M. \tag{5.96}$$

Therefore, the robust superdirective beamformer is

$$\mathbf{h}_{\mathrm{R},\epsilon} = \mathbf{i} - \mathbf{U}_2 \left(\mathbf{U}_2^H \boldsymbol{\Gamma}_{\mathrm{d},\epsilon} \mathbf{U}_2\right)^{-1} \mathbf{U}_2^H \boldsymbol{\Gamma}_{\mathrm{d},\epsilon} \mathbf{i}. \tag{5.97}$$

It is clear that $\mathbf{h}_{\mathrm{R},0} = \mathbf{h}_{\mathrm{S}}$ and $\mathbf{h}_{\mathrm{R},\infty} = \mathbf{h}_{\mathrm{DS}}$.

References

1. J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
2. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, pp. 4-24, Apr. 1988.
3. J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.
4. J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE, 2018.
5. H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 393–398, June 1986.
6. H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, pp. 1365–1376, Oct. 1987.

Chapter 6
Speech Enhancement from the Fullband Output SNR Perspective

Most of the speech enhancement algorithms are implemented in the time-frequency domain, i.e., the short-time Fourier transform (STFT) domain. The two main advantages of the STFT are that the algorithms can be implemented very efficiently and the different frequency bins can apparently be manipulated in a very flexible way in order to better compromise between noise reduction and speech distortion. Therefore, it is important to understand how things work from the fullband output SNR perspective and how gains/filters for noise reduction can be improved by fully exploiting all facets of this fundamental measure. This is the objective of this chapter, where two cases of the single-channel problem are discussed as well as the multichannel scenario.

6.1 Signal Model and Problem Formulation

The work developed in this chapter is an important generalization and extension of some of the ideas presented in [1].

Let us take the single-channel speech enhancement problem in the time domain of Section 4.1 (Chapter 4), i.e.,

$$y(t) = x(t) + v(t), \tag{6.1}$$

where $y(t)$, $x(t)$, and $v(t)$ are the microphone, desired, and noise signals, respectively. Using the short-time Fourier transform (STFT), (6.1) can be rewritten in the time-frequency domain as [2]

$$Y(k,n) = X(k,n) + V(k,n), \tag{6.2}$$

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at the frequency bin $k \in \{0,1,\ldots,K-1\}$ and the time frame n. In order to simplify the notation, we

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4_6

drop the dependence on the time frame; therefore, (6.2) for example is written as $Y(k) = X(k) + V(k)$. Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the variance of $Y(k)$ is

$$\begin{aligned} \phi_Y(k) &= E\left[|Y(k)|^2\right] \\ &= \phi_X(k) + \phi_V(k), \end{aligned} \tag{6.3}$$

where $\phi_X(k) = E\left[|X(k)|^2\right]$ and $\phi_V(k) = E\left[|V(k)|^2\right]$ are the variances of $X(k)$ and $V(k)$, respectively. From (6.3), we can define the subband input SNR:

$$\text{iSNR}(k) = \frac{\phi_X(k)}{\phi_V(k)} \tag{6.4}$$

and the fullband input:

$$\text{iSNR} = \frac{\sum_{k=0}^{K-1} \phi_X(k)}{\sum_{k=0}^{K-1} \phi_V(k)}. \tag{6.5}$$

It can be seen that

$$\min_k \text{iSNR}(k) \leq \text{iSNR} \leq \max_k \text{iSNR}(k). \tag{6.6}$$

In words, the fullband input SNR can never exceed the maximum subband input SNR and can never go below the minimum subband input SNR.

Then, our objective is the estimation of the desired signal, $X(k)$, from the observed signal, $Y(k)$, in the best possible (or flexible) way from the fullband output SNR that will be defined in the next section.

6.2 Speech Enhancement with Gains

The simplest and most effective way to perform speech enhancement in the STFT domain is by applying a complex gain, $H(k)$, to the observed signal, $Y(k)$, i.e.,

$$\begin{aligned} Z(k) &= H(k)Y(k) \\ &= X_{\text{fd}}(k) + V_{\text{fn}}(k), \end{aligned} \tag{6.7}$$

where $Z(k)$ is either the estimate of $X(k)$ or $V(k)$, $X_{\text{fd}}(k) = H(k)X(k)$ is the filtered desired signal, and $V_{\text{fn}}(k) = H(k)V(k)$ is the filtered noise. If $Z(k)$ is the estimate of $V(k)$, then the estimate of $X(k)$ is $\widehat{X}(k) = Y(k) - Z(k)$. The variance of $Z(k)$ is then

$$\begin{aligned}\phi_Z(k) &= |H(k)|^2 \phi_Y(k) \\ &= \phi_{X_{\text{fd}}}(k) + \phi_{V_{\text{fn}}}(k),\end{aligned} \tag{6.8}$$

where $\phi_{X_{\text{fd}}}(k) = |H(k)|^2 \phi_X(k)$ and $\phi_{V_{\text{fn}}}(k) = |H(k)|^2 \phi_V(k)$ are the variances of $X_{\text{fd}}(k)$ and $V_{\text{fn}}(k)$, respectively.

It is clear that the subband input and output SNRs are equal, i.e.,

$$\begin{aligned}\text{oSNR}\left[H(k)\right] &= \frac{\phi_{X_{\text{fd}}}(k)}{\phi_{V_{\text{fn}}}(k)} \\ &= \text{iSNR}(k).\end{aligned} \tag{6.9}$$

However, the fullband output SNR is

$$\text{oSNR}\left[H(:)\right] = \frac{\sum_{k=0}^{K-1} \phi_{X_{\text{fd}}}(k)}{\sum_{k=0}^{K-1} \phi_{V_{\text{fn}}}(k)}. \tag{6.10}$$

Therefore, our aim is to find the K subband gains, $H(k),\ k = 0, 1, \ldots, K-1$, in such a way that the fullband output SNR is greater than the fullband input SNR, i.e., $\text{oSNR}\left[H(:)\right] > \text{iSNR}$.

For convenience, we propose to use the index $k_i,\ i = 0, 1, \ldots, K-1$ and $k_i \in \{0, 1, \ldots, K-1\}$, which allows us to order the K subband input SNRs from the largest to the smallest, i.e.,

$$\text{iSNR}(k_0) \geq \text{iSNR}(k_1) \geq \cdots \geq \text{iSNR}(k_{K-1}). \tag{6.11}$$

We can also express the fullband output SNR as

$$\begin{aligned}\text{oSNR}\left(\underline{\mathbf{h}}\right) &= \frac{\underline{\mathbf{h}}^H \mathbf{D}_X \underline{\mathbf{h}}}{\underline{\mathbf{h}}^H \mathbf{D}_V \underline{\mathbf{h}}} \\ &= \frac{\sum_{i=0}^{K-1} |H(k_i)|^2 \phi_X(k_i)}{\sum_{i=0}^{K-1} |H(k_i)|^2 \phi_V(k_i)},\end{aligned} \tag{6.12}$$

where

$$\underline{\mathbf{h}} = \left[\, H(k_0)\ H(k_1) \cdots\ H(k_{K-1}) \,\right]^T \tag{6.13}$$

is a filter of length K containing all the subband gains and

$$\mathbf{D}_X = \text{diag}\left[\phi_X(k_0), \phi_X(k_1), \ldots, \phi_X(k_{K-1})\right] \tag{6.14}$$

$$\mathbf{D}_V = \text{diag}\left[\phi_V(k_0), \phi_V(k_1), \ldots, \phi_V(k_{K-1})\right] \tag{6.15}$$

are two diagonal matrices. It is assumed that $\phi_V(k_i) \neq 0,\ \forall k_i \in \{0, 1, \ldots, K-1\}$. Let

$$\lambda(k_i) = \text{iSNR}(k_i),\ i = 0, 1, \ldots, K-1. \tag{6.16}$$

It is worth noticing that

$$\mathbf{D}_V^{-1}\mathbf{D}_X = \mathrm{diag}\left[\lambda(k_0), \lambda(k_1), \ldots, \lambda(k_{K-1})\right] \tag{6.17}$$

is also a diagonal matrix containing all the K subband input SNRs ordered from the largest to the smallest.

Now, we give two important properties.

Property 6.1. Let $\lambda(k_0) \geq \lambda(k_1) \geq \cdots \geq \lambda(k_{K-1}) \geq 0$. We have

$$\begin{aligned}\frac{\sum_{i=0}^{K-1} |\alpha_i|^2 \lambda(k_i)}{\sum_{i=0}^{K-1} |\alpha_i|^2} \leq \frac{\sum_{i=0}^{K-2} |\alpha_i|^2 \lambda(k_i)}{\sum_{i=0}^{K-2} |\alpha_i|^2} \leq \cdots \\ \cdots \leq \frac{\sum_{i=0}^{1} |\alpha_i|^2 \lambda(k_i)}{\sum_{i=0}^{1} |\alpha_i|^2} \leq \lambda(k_0)\end{aligned} \tag{6.18}$$

or, equivalently,

$$\begin{aligned}\frac{\sum_{i=0}^{K-1} |\alpha_i|^2 \phi_X(k_i)}{\sum_{i=0}^{K-1} |\alpha_i|^2 \phi_V(k_i)} \leq \frac{\sum_{i=0}^{K-2} |\alpha_i|^2 \phi_X(k_i)}{\sum_{i=0}^{K-2} |\alpha_i|^2 \phi_V(k_i)} \leq \cdots \\ \cdots \leq \frac{\sum_{i=0}^{1} |\alpha_i|^2 \phi_X(k_i)}{\sum_{i=0}^{1} |\alpha_i|^2 \phi_V(k_i)} \leq \frac{\phi_X(k_0)}{\phi_V(k_0)},\end{aligned} \tag{6.19}$$

where $\alpha_i,\ i = 0, 1, \ldots, K-1$ are arbitrary complex numbers with at least one of them different from 0.

Proof. The previous inequalities can be easily shown by induction.

Property 6.2. Let $\lambda(k_0) \geq \lambda(k_1) \geq \cdots \geq \lambda(k_{K-1}) \geq 0$. We have

$$\begin{aligned}\lambda(k_{K-1}) \leq \frac{\sum_{i=0}^{1} |\beta_{K-1-i}|^2 \lambda(k_{K-1-i})}{\sum_{i=0}^{1} |\beta_{K-1-i}|^2} \leq \cdots \\ \cdots \leq \frac{\sum_{i=0}^{K-2} |\beta_{K-1-i}|^2 \lambda(k_{K-1-i})}{\sum_{i=0}^{K-2} |\beta_{K-1-i}|^2} \leq \frac{\sum_{i=0}^{K-1} |\beta_{K-1-i}|^2 \lambda(k_{K-1-i})}{\sum_{i=0}^{K-1} |\beta_{K-1-i}|^2}\end{aligned} \tag{6.20}$$

or, equivalently,

$$\begin{aligned}\frac{\phi_X(k_{K-1})}{\phi_V(k_{K-1})} \leq \frac{\sum_{i=0}^{1} |\beta_{K-1-i}|^2 \phi_X(k_{K-1-i})}{\sum_{i=0}^{1} |\beta_{K-1-i}|^2 \phi_V(k_{K-1-i})} \leq \cdots \\ \cdots \leq \frac{\sum_{i=0}^{K-2} |\beta_{K-1-i}|^2 \phi_X(k_{K-1-i})}{\sum_{i=0}^{K-2} |\beta_{K-1-i}|^2 \phi_V(k_{K-1-i})} \leq \frac{\sum_{i=0}^{K-1} |\beta_{K-1-i}|^2 \phi_X(k_{K-1-i})}{\sum_{i=0}^{K-1} |\beta_{K-1-i}|^2 \phi_V(k_{K-1-i})},\end{aligned} \tag{6.21}$$

where $\beta_{K-1-i},\ i = 0, 1, \ldots, K-1$ are arbitrary complex numbers with at least one of them different from 0.

Proof. The previous inequalities can be easily shown by induction.

It follows from the previous properties that[1]

$$\text{iSNR}(k_{K-1}) \leq \text{oSNR}\left(\underline{\mathbf{h}}\right) \leq \text{iSNR}(k_0),\ \forall \underline{\mathbf{h}}, \tag{6.22}$$

as well as the inequalities in (6.6). Clearly, both the fullband input and output SNRs can never exceed the maximum subband input SNR.

6.3 Determination of the Optimal Gains

There are two approaches to find the optimal gains from the fullband output SNR in order to perform speech enhancement. The first one considers the largest subband input SNRs. In this case, we get the estimate of the desired signal directly. The second method considers the smallest subband input SNRs. As a result, we get the estimate of the noise signal from which we easily deduce the estimate of the desired signal.

6.3.1 Maximization of the Fullband Output SNR

The filter, $\underline{\mathbf{h}}$, that maximizes the fullband output SNR given in (6.12) is simply the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{D}_V^{-1}\mathbf{D}_X$. Since this matrix is diagonal, its maximum eigenvalue is its largest diagonal element, i.e., $\lambda(k_0)$. As a consequence, the maximum SNR filter is

$$\underline{\mathbf{h}}_{\max} = \alpha(k_0)\mathbf{i}_1, \tag{6.23}$$

where $\alpha(k_0) \neq 0$ is an arbitrary complex number and $\mathbf{i}_1$ is the first column of the $K \times K$ identity matrix, $\mathbf{I}_K$. Equivalently, we can write (6.23) as

$$\begin{cases} H_{\max}(k_0) = \alpha(k_0) \\ H_{\max}(k_i) = 0,\ i = 1, 2, \ldots, K-1 \end{cases}. \tag{6.24}$$

With (6.23), we get the maximum possible fullband output SNR, which is

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\max}\right) = \lambda(k_0) = \max_k \text{iSNR}(k) \geq \text{iSNR}. \tag{6.25}$$

As a result,

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\max}\right) \geq \text{oSNR}\left(\underline{\mathbf{h}}\right),\ \forall \underline{\mathbf{h}}. \tag{6.26}$$

[1] This is also a consequence of the definition of the fullband output SNR in (6.12), whose form is the generalized Rayleigh quotient.

We deduce that the estimate of the desired signal is

$$\begin{cases} \widehat{X}_{\max}(k_0) = H_{\max}(k_0)Y(k_0) \\ \widehat{X}_{\max}(k_i) = 0,\ i = 1, 2, \ldots, K-1 \end{cases}. \tag{6.27}$$

Now, we need to determine $\alpha(k_0)$. There are at least two ways to find this parameter. The first one is from the MSE between $X(k_0)$ and $\widehat{X}_{\max}(k_0)$, i.e.,

$$J\left[\alpha(k_0)\right] = E\left[\left|X(k_0) - \alpha(k_0)Y(k_0)\right|^2\right]. \tag{6.28}$$

The second possibility is to use the distortion-based MSE, i.e.,

$$J_{\mathrm{d}}\left[\alpha(k_0)\right] = E\left[\left|X(k_0) - \alpha(k_0)X(k_0)\right|^2\right]. \tag{6.29}$$

The minimization of $J\left[\alpha(k_0)\right]$ leads to the Wiener gain at the frequency bin k_0, i.e.,

$$\alpha_{\mathrm{W}}(k_0) = \frac{\mathrm{iSNR}(k_0)}{1 + \mathrm{iSNR}(k_0)}, \tag{6.30}$$

while the minimization of $J_{\mathrm{d}}\left[\alpha(k_0)\right]$ gives the unitary gain at the frequency bin k_0, i.e.,

$$\alpha_{\mathrm{U}}(k_0) = 1. \tag{6.31}$$

Even though this method maximizes the fullband output SNR, it is expected to introduce a huge amount of distortion to the desired signal, since all its frequency bins are put to 0 except at k_0. A much better approach when we deal with broadband signals such as speech is to form the filter from a linear combination of the eigenvectors corresponding to the $P(\leq K)$ largest eigenvalues of $\mathbf{D}_V^{-1}\mathbf{D}_X$, i.e.,

$$\underline{\mathbf{h}}_P = \sum_{p=0}^{P-1} \alpha(k_p)\mathbf{i}_{p+1}, \tag{6.32}$$

where $\alpha(k_p),\ p = 0, 1, \ldots, P-1$ are arbitrary complex numbers with at least one of them different from 0 and $\mathbf{i}_{p+1}$ is the $(p+1)$th column of $\mathbf{I}_K$. We can also express (6.32) as

$$\begin{cases} H_P(k_p) = \alpha(k_p),\ p = 0, 1, \ldots, P-1 \\ H_P(k_i) = 0,\ i = P, P+1, \ldots, K-1 \end{cases}. \tag{6.33}$$

Hence, the estimate of the desired signal is

$$\begin{cases} \widehat{X}_P(k_p) = H_P(k_p)Y(k_p),\ p = 0, 1, \ldots, P-1 \\ \widehat{X}_P(k_i) = 0,\ i = P, P+1, \ldots, K-1 \end{cases}. \tag{6.34}$$

To find the $\alpha(k_p)$'s, we can either optimize $J\left[\alpha(k_p)\right]$ or $J_{\mathrm{d}}\left[\alpha(k_p)\right]$. The first one leads to the Wiener gains at the frequency bins $k_p,\ p = 0, 1, \ldots, P-1$, i.e.,

$$\alpha_{\mathrm{W}}(k_p) = \frac{\mathrm{iSNR}(k_p)}{1 + \mathrm{iSNR}(k_p)}, \tag{6.35}$$

while the second one gives the unitary gains at the frequency bins $k_p,\ p = 0, 1, \ldots, P-1$, i.e.,

$$\alpha_{\mathrm{U}}(k_p) = 1. \tag{6.36}$$

The filters (of length K) corresponding to (6.35) and (6.36) are, respectively,

$$\underline{\mathbf{h}}_{P,\mathrm{W}} = \left[\, \alpha_{\mathrm{W}}(k_0) \cdots\ \alpha_{\mathrm{W}}(k_{P-1})\ 0 \cdots 0 \,\right]^T \tag{6.37}$$

and

$$\underline{\mathbf{h}}_{P,\mathrm{U}} = \left[\, 1 \cdots\ 1\ 0 \cdots 0 \,\right]^T. \tag{6.38}$$

For $P = K$, $\underline{\mathbf{h}}_{K,\mathrm{W}}$ corresponds to the classical Wiener approach [2] and $\underline{\mathbf{h}}_{K,\mathrm{U}}$ is the identity filter, which does not affect the observations. Clearly, $\underline{\mathbf{h}}_{P,\mathrm{U}}$ corresponds to the ideal binary mask [3], since the subband observation signals with the P largest subband input SNRs are not affected while the $K - P$ others with the smallest subband input SNRs are put to 0. We should always have

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{P,\mathrm{U}}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{P,\mathrm{W}}\right). \tag{6.39}$$

From Property 6.1, we deduce that

$$\mathrm{iSNR} \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K,\mathrm{W}}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K-1,\mathrm{W}}\right) \leq \cdots \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{1,\mathrm{W}}\right) = \lambda(k_0) \tag{6.40}$$

and

$$\mathrm{iSNR} = \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K,\mathrm{U}}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K-1,\mathrm{U}}\right) \leq \cdots \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{1,\mathrm{U}}\right) = \lambda(k_0). \tag{6.41}$$

6.3.2 Minimization of the Fullband Output SNR

It is clear that the filter denoted $\underline{\mathbf{h}}_V$ that minimizes the fullband output SNR given in (6.12) is the eigenvector corresponding to the minimum eigenvalue of the matrix $\mathbf{D}_V^{-1}\mathbf{D}_X$, which is $\lambda(k_{K-1})$. Therefore, the minimum SNR filter is

$$\underline{\mathbf{h}}_V = \beta(k_{K-1})\mathbf{i}_K, \tag{6.42}$$

where $\beta(k_{K-1}) \neq 0$ is an arbitrary complex number and $\mathbf{i}_K$ is the Kth column of $\mathbf{I}_K$. Equivalently, we can write (6.42) as

$$\begin{cases} H_V(k_i) = 0, \ i = 0, 1, \ldots, K-2 \\ H_V(k_{K-1}) = \beta(k_{K-1}) \end{cases}. \tag{6.43}$$

With (6.42), we get the minimum possible fullband output SNR, which is

$$\text{oSNR}\left(\underline{\mathbf{h}}_V\right) = \lambda(k_{K-1}) = \min_k \text{iSNR}(k) \leq \text{iSNR}. \tag{6.44}$$

As a result,

$$\text{oSNR}\left(\underline{\mathbf{h}}_V\right) \leq \text{oSNR}\left(\underline{\mathbf{h}}\right), \ \forall \underline{\mathbf{h}}. \tag{6.45}$$

We deduce that the estimates of the noise and desired signals are, respectively,

$$\begin{cases} \widehat{V}(k_i) = 0, \ i = 0, 1, \ldots, K-2 \\ \widehat{V}(k_{K-1}) = H_V(k_{K-1})Y(k_{K-1}) \end{cases} \tag{6.46}$$

and

$$\begin{cases} \widehat{X}(k_i) = Y(k_i), \ i = 0, 1, \ldots, K-2 \\ \widehat{X}(k_{K-1}) = H_X(k_{K-1})Y(k_{K-1}) \end{cases}, \tag{6.47}$$

where

$$H_X(k_{K-1}) = 1 - H_V(k_{K-1}) \tag{6.48}$$

is the equivalent gain for the estimation of $X(k_{K-1})$.

The MSE between $X(k_{K-1})$ and $\widehat{X}_{\beta_{K-1}}(k_{K-1})$ is

$$\begin{aligned} J\left[\beta(k_{K-1})\right] &= E\left[\left|V(k_{K-1}) - \beta(k_{K-1})Y(k_{K-1})\right|^2\right] \\ &= \left|\beta(k_{K-1})\right|^2 \phi_X(k_{K-1}) + \left|1 - \beta(k_{K-1})\right|^2 \phi_V(k_{K-1}) \\ &= J_{\text{d}}\left[\beta(k_{K-1})\right] + J_{\text{r}}\left[\beta(k_{K-1})\right]. \end{aligned} \tag{6.49}$$

From the previous expression, we see that there are at least two ways to find $\beta(k_{K-1})$. The minimization of $J\left[\beta(k_{K-1})\right]$ leads to

$$\beta_{\mathrm{W}}(k_{K-1}) = \frac{1}{1+\mathrm{iSNR}(k_{K-1})}, \tag{6.50}$$

which is the Wiener gain at the frequency bin k_{K-1} for the estimation of $V(k_{K-1})$ or, equivalently,

$$\begin{aligned} \alpha_{\mathrm{W}}(k_{K-1}) &= 1-\beta_{\mathrm{W}}(k_{K-1}) \\ &= \frac{\mathrm{iSNR}(k_{K-1})}{1+\mathrm{iSNR}(k_{K-1})}, \end{aligned} \tag{6.51}$$

which is the Wiener gain at the frequency bin k_{K-1} for the estimation of $X(k_{K-1})$. The minimization of the power of the residual noise, $J_{\mathrm{r}}\left[\beta(k_{K-1})\right]$, gives

$$\beta_{\mathrm{U}}(k_{K-1}) = 1, \tag{6.52}$$

which is the unitary gain at the frequency bin k_{K-1} for the estimation of $V(k_{K-1})$ or, equivalently,

$$\begin{aligned} \alpha_{\mathrm{N}}(k_{K-1}) &= 1-\beta_{\mathrm{U}}(k_{K-1}) \\ &= 0, \end{aligned} \tag{6.53}$$

which is the null gain at the frequency bin k_{K-1} for the estimation of $X(k_{K-1})$.

Obviously, the approach presented above is not meaningful for broadband signals, since only one frequency bin is processed while all the others are not affected at all. This is far to be enough as far as noise reduction is concerned, even though very little distortion is expected. A more practical approach is to form the filter from a linear combination of the eigenvectors corresponding to the $Q(\leq K)$ smallest eigenvalues of $\mathbf{D}_V^{-1}\mathbf{D}_X$, i.e.,

$$\underline{\mathbf{h}}_{V,Q} = \sum_{q=0}^{Q-1} \beta(k_{K-Q+q})\mathbf{i}_{K-Q+q+1}, \tag{6.54}$$

where $\beta(k_{K-Q+q}),\ q=0,1,\ldots,Q-1$ are arbitrary complex numbers with at least one of them different from 0 and $\mathbf{i}_{K-Q+q+1}$ is the $(K-Q+q+1)$th column of $\mathbf{I}_K$. Therefore, the equivalent filter for the estimation of the desired signal at the different frequency bins is

$$\underline{\mathbf{h}}_{X,Q} = \mathbf{1} - \underline{\mathbf{h}}_{V,Q}, \tag{6.55}$$

where $\mathbf{1}$ is a vector of length K with all its elements equal to 1. We can also express (6.55) as

$$\begin{cases} H_{X,Q}(k_i) = 1, \ i = 0, 1, \ldots, K - Q - 1 \\ H_{X,Q}(k_{K-Q+q}) = 1 - \beta(k_{K-Q+q}), \ q = 0, 1, \ldots, Q - 1 \end{cases}. \tag{6.56}$$

Hence, the estimate of the desired signal is

$$\begin{cases} \widehat{X}(k_i) = Y(k_i), \ i = 0, 1, \ldots, K - Q - 1 \\ \widehat{X}(k_{K-Q+q}) = H_{X,Q}(k_{K-Q+q}) Y(k_{K-Q+q}), \ q = 0, 1, \ldots, Q - 1 \end{cases}. \tag{6.57}$$

Following the same steps as above, we deduce the two filters of interest:

$$\underline{\mathbf{h}}_{X,Q,\mathrm{W}} = \left[1 \cdots \ 1 \ \alpha_{\mathrm{W}}(k_{K-Q}) \cdots \ \alpha_{\mathrm{W}}(k_{K-1}) \right]^T \tag{6.58}$$

and

$$\underline{\mathbf{h}}_{X,Q,\mathrm{N}} = \left[1 \cdots \ 1 \ 0 \cdots \ 0 \right]^T. \tag{6.59}$$

For $Q = K$, $\underline{\mathbf{h}}_{X,K,\mathrm{W}} = \underline{\mathbf{h}}_{K,\mathrm{W}}$ corresponds to the classical Wiener approach and $\underline{\mathbf{h}}_{X,K,\mathrm{N}} = \mathbf{0}$ is the null filter, which completely cancels the observations. The filter $\underline{\mathbf{h}}_{X,Q,\mathrm{W}}$ can be seen as a combination of the ideal binary mask and Wiener, where the observations with large subband input SNRs are not affected while the ones with small subband input SNRs are processed with the Wiener gains. The filter $\underline{\mathbf{h}}_{X,Q,\mathrm{N}}$ is, obviously, the ideal binary mask. We should always have

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,Q,\mathrm{N}}\right) \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,Q,\mathrm{W}}\right). \tag{6.60}$$

We can also deduce that

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,K,\mathrm{W}}\right) \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,K-1,\mathrm{W}}\right) \geq \cdots \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,1,\mathrm{W}}\right) \geq \mathrm{iSNR} \tag{6.61}$$

and

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,K,\mathrm{N}}\right) \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,K-1,\mathrm{N}}\right) \geq \cdots \geq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{X,1,\mathrm{N}}\right) \geq \mathrm{iSNR}. \tag{6.62}$$

6.4 Taking the Interframe Correlation Into Account

It is well known that a speech signal at successive time frames in the STFT domain is highly correlated. Therefore, if we wish to improve the performance of noise reduction, we need to take this interframe correlation into account.

Let us consider the $L \geq 1$ most recent time frames of $Y(k)$. Then, we can express (6.2) as

$$\begin{aligned}\mathbf{y}(k) &= \left[\, Y(k,n)\; Y(k,n-1) \cdots\; Y(k,n-L+1) \,\right]^T \\ &= \mathbf{x}(k) + \mathbf{v}(k),\end{aligned} \tag{6.63}$$

where $\mathbf{x}(k)$ and $\mathbf{v}(k)$ are defined similarly to $\mathbf{y}(k)$. The $L \times L$ covariance matrix of $\mathbf{y}(k)$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k) &= E\left[\mathbf{y}(k)\mathbf{y}^H(k)\right] \\ &= \mathbf{\Phi}_{\mathbf{x}}(k) + \mathbf{\Phi}_{\mathbf{v}}(k),\end{aligned} \tag{6.64}$$

where $\mathbf{\Phi}_{\mathbf{x}}(k)$ and $\mathbf{\Phi}_{\mathbf{v}}(k)$ are the covariance matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively.

The two Hermitian matrices $\mathbf{\Phi}_{\mathbf{x}}(k)$ and $\mathbf{\Phi}_{\mathbf{v}}(k)$ in (6.64) can be jointly diagonalized as follows [4]:

$$\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{x}}(k)\mathbf{A}(k) = \mathbf{\Lambda}(k), \tag{6.65}$$

$$\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{v}}(k)\mathbf{A}(k) = \mathbf{I}_L, \tag{6.66}$$

where $\mathbf{A}(k)$ is a full-rank square matrix (of size $L \times L$), $\mathbf{\Lambda}(k)$ is a diagonal matrix whose main elements are real and nonnegative, and $\mathbf{I}_L$ is the $L \times L$ identity matrix. Furthermore, $\mathbf{\Lambda}(k)$ and $\mathbf{A}(k)$ are the eigenvalue and eigenvector matrices, respectively, of $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{\Phi}_{\mathbf{x}}(k)$, i.e.,

$$\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{\Phi}_{\mathbf{x}}(k)\mathbf{A}(k) = \mathbf{A}(k)\mathbf{\Lambda}(k). \tag{6.67}$$

The eigenvalues of $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{\Phi}_{\mathbf{x}}(k)$, denoted $\lambda_l(k)$, $l = 1, 2, \ldots, L$, are ordered as $\lambda_1(k) \geq \lambda_2(k) \geq \cdots \geq \lambda_L(k) \geq 0$ and the corresponding eigenvectors are denoted $\mathbf{a}_1(k), \mathbf{a}_2(k), \ldots, \mathbf{a}_L(k)$. Obviously, the noisy signal covariance matrix can also be diagonalized as

$$\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{y}}(k)\mathbf{A}(k) = \mathbf{\Lambda}(k) + \mathbf{I}_L. \tag{6.68}$$

We will see a bit later that this joint diagonalization is going to be very useful.

Since the interframe correlation is now taken into account, $X(k)$ is estimated by applying a complex-valued filter, $\mathbf{h}(k)$ of length L, to the observation signal vector, $\mathbf{y}(k)$, i.e.,

$$\begin{aligned}Z(k) &= \mathbf{h}^H(k)\mathbf{y}(k) \\ &= X_{\mathrm{fd}}(k) + V_{\mathrm{rn}}(k),\end{aligned} \tag{6.69}$$

where $Z(k)$ is the estimate of $X(k)$[2], $X_{\mathrm{fd}}(k) = \mathbf{h}^H(k)\mathbf{x}(k)$ is the filtered desired signal, and $V_{\mathrm{rn}}(k) = \mathbf{h}^H(k)\mathbf{v}(k)$ is the residual noise. Obviously, the case $L = 1$ corresponds to the conventional single-channel noise reduction

[2] In this section, we only focus on the estimation of $X(k)$; the extension of this approach to the estimation of $V(k)$ is straightforward.

approach in the STFT domain with gains [2]. The variance of $Z(k)$ is then

$$\begin{aligned}\phi_Z(k) &= \mathbf{h}^H(k)\mathbf{\Phi_y}(k)\mathbf{h}(k) \\ &= \phi_{X_{\rm fd}}(k) + \phi_{V_{\rm rn}}(k),\end{aligned} \tag{6.70}$$

where $\phi_{X_{\rm fd}}(k) = \mathbf{h}^H(k)\mathbf{\Phi_x}(k)\mathbf{h}(k)$ and $\phi_{V_{\rm rn}}(k) = \mathbf{h}^H(k)\mathbf{\Phi_v}(k)\mathbf{h}(k)$ are the variances of $X_{\rm fd}(k)$ and $V_{\rm rn}(k)$, respectively. We deduce from (6.70) that the subband and fullband output SNRs are, respectively,

$$\begin{aligned}{\rm oSNR}\left[\mathbf{h}(k)\right] &= \frac{\phi_{X_{\rm fd}}(k)}{\phi_{V_{\rm rn}}(k)} \\ &= \frac{\mathbf{h}^H(k)\mathbf{\Phi_x}(k)\mathbf{h}(k)}{\mathbf{h}^H(k)\mathbf{\Phi_v}(k)\mathbf{h}(k)}\end{aligned} \tag{6.71}$$

and

$${\rm oSNR}\left[\mathbf{h}(:)\right] = \frac{\sum_{k=0}^{K-1}\phi_{X_{\rm fd}}(k)}{\sum_{k=0}^{K-1}\phi_{V_{\rm rn}}(k)}. \tag{6.72}$$

As we did in previous sections, we propose to use the index k_i, $i = 0, 1, \ldots, K-1$ and $k_i \in \{0, 1, \ldots, K-1\}$, which allows us to order the K subband eigenvalues $\lambda_1(k)$, $k = 0, 1, \ldots, K-1$ from the largest to the smallest, i.e.,

$$\lambda_1(k_0) \geq \lambda_1(k_1) \geq \cdots \geq \lambda_1(k_{K-1}). \tag{6.73}$$

So, with this indexing, the subband filter is denoted as $\mathbf{h}(k_i)$, which is assumed in the rest to have the form:

$$\mathbf{h}(k_i) = \psi(k_i)\mathbf{a}_1(k_i), \tag{6.74}$$

where $\psi(k_i)$ is an arbitrary complex number and $\mathbf{a}_1(k_i)$ is the eigenvector corresponding to $\lambda_1(k_i)$. For $\psi(k_i) \neq 0$, it is clear that $\mathbf{h}(k_i)$ in (6.74) maximizes the subband output SNR, since

$${\rm oSNR}\left[\mathbf{h}(k_i)\right] = \lambda_1(k_i). \tag{6.75}$$

As a result, with $\mathbf{h}(k_i)$ in (6.74), (6.73) is equivalent to saying that

$${\rm oSNR}\left[\mathbf{h}(k_0)\right] \geq {\rm oSNR}\left[\mathbf{h}(k_1)\right] \geq \cdots \geq {\rm oSNR}\left[\mathbf{h}(k_{K-1})\right]. \tag{6.76}$$

Also, we have

$$\lambda_1(k_i) \geq {\rm iSNR}(k_i) \tag{6.77}$$

and

$$\text{iSNR} \leq \lambda_1(k_0). \tag{6.78}$$

In the particular case of $L = 1$, we have $\lambda_1(k_i) = \lambda(k_i) = \text{iSNR}(k_i)$, which corresponds to the study of previous sections.

Let

$$\underline{\mathbf{h}} = \left[\, \mathbf{h}^T(k_0)\ \mathbf{h}^T(k_1) \cdots\ \mathbf{h}^T(k_{K-1}) \,\right]^T \tag{6.79}$$

be a long filter of length KL containing all the ordered subband filters. We can express the fullband output SNR as

$$\text{oSNR}\,(\underline{\mathbf{h}}) = \frac{\underline{\mathbf{h}}^H \mathbf{D}_{\mathbf{\Phi_x}} \underline{\mathbf{h}}}{\underline{\mathbf{h}}^H \mathbf{D}_{\mathbf{\Phi_v}} \underline{\mathbf{h}}}, \tag{6.80}$$

where

$$\mathbf{D}_{\mathbf{\Phi_x}} = \text{diag}\left[\mathbf{\Phi_x}(k_0), \mathbf{\Phi_x}(k_1), \ldots, \mathbf{\Phi_x}(k_{K-1})\right] \tag{6.81}$$

$$\mathbf{D}_{\mathbf{\Phi_v}} = \text{diag}\left[\mathbf{\Phi_v}(k_0), \mathbf{\Phi_v}(k_1), \ldots, \mathbf{\Phi_v}(k_{K-1})\right] \tag{6.82}$$

are block diagonal matrices. It is worth noticing that

$$\mathbf{D}_{\mathbf{\Phi_v}}^{-1} \mathbf{D}_{\mathbf{\Phi_x}} \mathbf{D}_{\mathbf{A}} = \mathbf{D}_{\mathbf{A}} \mathbf{D}_{\mathbf{\Lambda}}, \tag{6.83}$$

where

$$\mathbf{D}_{\mathbf{\Phi_v}}^{-1} = \text{diag}\left[\mathbf{\Phi_v}^{-1}(k_0), \mathbf{\Phi_v}^{-1}(k_1), \ldots, \mathbf{\Phi_v}^{-1}(k_{K-1})\right], \tag{6.84}$$

$$\mathbf{D}_{\mathbf{A}} = \text{diag}\left[\mathbf{A}(k_0), \mathbf{A}(k_1), \ldots, \mathbf{A}(k_{K-1})\right], \tag{6.85}$$

$$\mathbf{D}_{\mathbf{\Lambda}} = \text{diag}\left[\mathbf{\Lambda}(k_0), \mathbf{\Lambda}(k_1), \ldots, \mathbf{\Lambda}(k_{K-1})\right]. \tag{6.86}$$

Therefore, our objective is to find $\underline{\mathbf{h}}$ in such a way that $\text{oSNR}\,(\underline{\mathbf{h}}) > \text{iSNR}$. Since

$$\text{oSNR}\,(\underline{\mathbf{h}}) = \frac{\sum_{i=0}^{K-1} |\psi(k_i)|^2\, \lambda_1(k_i)}{\sum_{i=0}^{K-1} |\psi(k_i)|^2}, \tag{6.87}$$

we deduce that

$$\lambda_1(k_{K-1}) \leq \text{oSNR}\,(\underline{\mathbf{h}}) \leq \lambda_1(k_0). \tag{6.88}$$

The filter, $\underline{\mathbf{h}}$, that maximizes the fullband output SNR given in (6.80) is simply the eigenvector corresponding to the maximum eigenvalue, $\lambda_1(k_0)$, of the matrix $\mathbf{D}_{\mathbf{\Phi_v}}^{-1} \mathbf{D}_{\mathbf{\Phi_x}}$. As a consequence, the maximum SNR filter (of length KL) is

$$\underline{\mathbf{h}}_{\max} = \left[\, \psi(k_0) \mathbf{a}_1^T(k_0)\ \mathbf{0}^T \,\right]^T, \tag{6.89}$$

where $\psi(k_0) \neq 0$. Equivalently, we can write (6.89) as

$$\begin{cases} \mathbf{h}_{\max}(k_0) = \psi(k_0)\mathbf{a}_1(k_0) \\ \mathbf{h}_{\max}(k_i) = \mathbf{0},\ i = 1, 2, \ldots, K-1 \end{cases}. \tag{6.90}$$

With (6.89), we get the maximum possible fullband output SNR, which is

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\max}\right) = \lambda_1(k_0) \geq \text{iSNR} \tag{6.91}$$

and

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\max}\right) \geq \text{oSNR}\left(\underline{\mathbf{h}}\right),\ \forall \underline{\mathbf{h}}. \tag{6.92}$$

We deduce that the estimate of the desired signal is

$$\begin{cases} \widehat{X}_{\max}(k_0) = \mathbf{h}_{\max}^H(k_0)\mathbf{y}(k_0) \\ \widehat{X}_{\max}(k_i) = 0,\ i = 1, 2, \ldots, K-1 \end{cases}. \tag{6.93}$$

Now, the parameter $\psi(k_0)$ needs to be determined. There are at least two ways to find it. The first one is from the MSE between $X(k_0)$ and $\widehat{X}_{\max}(k_0)$, i.e.,

$$J\left[\psi(k_0)\right] = E\left[\left|X(k_0) - \psi^*(k_0)\mathbf{a}_1^H(k_0)\mathbf{y}(k_0)\right|^2\right]. \tag{6.94}$$

The second possibility is to use the distortion-based MSE, i.e.,

$$J_{\text{d}}\left[\psi(k_0)\right] = E\left[\left|X(k_0) - \psi^*(k_0)\mathbf{a}_1^H(k_0)\mathbf{x}(k_0)\right|^2\right]. \tag{6.95}$$

The minimization of $J\left[\psi(k_0)\right]$ leads to the maximum SNR filter with minimum MSE at the frequency bin k_0, i.e.,

$$\mathbf{h}_{\max,1}(k_0) = \frac{\mathbf{a}_1(k_0)\mathbf{a}_1^H(k_0)\mathbf{\Phi}_{\mathbf{x}}(k_0)\mathbf{i}_1}{1 + \lambda_1(k_0)}, \tag{6.96}$$

while the minimization of $J_{\text{d}}\left[\psi(k_0)\right]$ gives a minimum distortion filter at the frequency bin k_0, i.e.,

$$\mathbf{h}_{\max,2}(k_0) = \frac{\mathbf{a}_1(k_0)\mathbf{a}_1^H(k_0)\mathbf{\Phi}_{\mathbf{x}}(k_0)\mathbf{i}_1}{\lambda_1(k_0)}, \tag{6.97}$$

where $\mathbf{i}_1$ is the first column of $\mathbf{I}_L$.

Clearly, this method maximizes the fullband output SNR but it is expected to introduce a huge amount of distortion to the desired signal, since all its frequency bins are put to 0 except at k_0. A much better approach when we deal with broadband signals such as speech is to form the filter (of length KL) from a concatenation of the eigenvectors corresponding to the $P(\leq K)$

largest eigenvalues from the set $\{\lambda_1(k_i),\ i = 0, 1, \ldots, K-1\}$, i.e.,

$$\underline{\mathbf{h}}_P = \left[\, \psi(k_0)\mathbf{a}_1^T(k_0) \cdots\ \psi_{P-1}(k_{P-1})\mathbf{a}_1^T(k_{P-1})\ \mathbf{0}^T \,\right]^T, \tag{6.98}$$

where $\psi(k_p),\ p = 0, 1, \ldots, P-1$ are arbitrary complex numbers with at least one of them different from 0. We can also express (6.98) as

$$\begin{cases} \mathbf{h}_P(k_p) = \psi(k_p)\mathbf{a}_1(k_p),\ p = 0, 1, \ldots, P-1 \\ \mathbf{h}_P(k_i) = \mathbf{0},\ i = P, P+1, \ldots, K-1 \end{cases}. \tag{6.99}$$

Hence, the estimate of the desired signal is

$$\begin{cases} \widehat{X}_P(k_p) = \mathbf{h}_P^H(k_p)\mathbf{y}(k_p),\ p = 0, 1, \ldots, P-1 \\ \widehat{X}_P(k_i) = 0,\ i = P, P+1, \ldots, K-1 \end{cases}. \tag{6.100}$$

To find the $\psi(k_p)$'s, we can either optimize $J\left[\psi(k_p)\right]$ or $J_{\mathrm{d}}\left[\psi(k_p)\right]$. The first one leads to filters with minimum MSE at the frequency bins $k_p,\ p = 0, 1, \ldots, P-1$, i.e.,

$$\mathbf{h}_{P,1}(k_p) = \frac{\mathbf{a}_1(k_p)\mathbf{a}_1^H(k_p)\mathbf{\Phi}_{\mathbf{x}}(k_p)\mathbf{i}_1}{1 + \lambda_1(k_p)}, \tag{6.101}$$

while the second one gives the minimum distortion filters at the frequency bins $k_p,\ p = 0, 1, \ldots, P-1$, i.e.,

$$\mathbf{h}_{P,2}(k_p) = \frac{\mathbf{a}_1(k_p)\mathbf{a}_1^H(k_p)\mathbf{\Phi}_{\mathbf{x}}(k_p)\mathbf{i}_1}{\lambda_1(k_p)}. \tag{6.102}$$

The filters (of length KL) corresponding to (6.101) and (6.102) are, respectively,

$$\underline{\mathbf{h}}_{P,1} = \left[\, \mathbf{h}_{P,1}^T(k_0) \cdots\ \mathbf{h}_{P,1}^T(k_{P-1})\ \mathbf{0}^T \,\right]^T \tag{6.103}$$

and

$$\underline{\mathbf{h}}_{P,2} = \left[\, \mathbf{h}_{P,2}^T(k_0) \cdots\ \mathbf{h}_{P,2}^T(k_{P-1})\ \mathbf{0}^T \,\right]^T. \tag{6.104}$$

This approach can be seen as a generalization of the ideal binary mask [3], since the subband observation signals of the microphone with the P largest subband output SNRs are processed with filters with minimum MSE or minimum distortion, while the $K - P$ others with the smallest subband output SNRs are put to 0. We should always have

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{P,2}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{P,1}\right). \tag{6.105}$$

We can deduce that

$$\text{iSNR} \leq \text{oSNR}\left(\underline{\mathbf{h}}_{K,1}\right) \leq \text{oSNR}\left(\underline{\mathbf{h}}_{K-1,1}\right) \leq \cdots \leq \text{oSNR}\left(\underline{\mathbf{h}}_{1,1}\right) = \lambda_1(k_0) \tag{6.106}$$

and

$$\text{iSNR} \leq \text{oSNR}\left(\underline{\mathbf{h}}_{K,2}\right) \leq \text{oSNR}\left(\underline{\mathbf{h}}_{K-1,2}\right) \leq \cdots \leq \text{oSNR}\left(\underline{\mathbf{h}}_{1,2}\right) = \lambda_1(k_0). \tag{6.107}$$

6.5 Generalization to the Multichannel Case

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, at the time index t, are expressed as [5], [6]

$$\begin{aligned} y_m(t) &= g_m(t) * x(t) + v_m(t) \\ &= x_m(t) + v_m(t), \ m = 1, 2, \ldots, M, \end{aligned} \tag{6.108}$$

where $g_m(t)$ is the acoustic impulse response from the unknown speech source, $x(t)$, location to the mth microphone, $*$ stands for linear convolution, and $v_m(t)$ is the additive noise at microphone m. We assume that the signals $x_m(t) = g_m(t) * x(t)$ and $v_m(t)$ are uncorrelated, zero mean, stationary, real, and broadband. By definition, the convolved speech signals, $x_m(t),\ m = 1, 2, \ldots, M$, are coherent across the array while the noise signals, $v_m(t),\ m = 1, 2, \ldots, M$, are typically only partially coherent across the array. Using the STFT, (6.108) can be rewritten in the time-frequency domain as

$$\begin{aligned} Y_m(k,n) &= G_m(k)X(k,n) + V_m(k,n) \\ &= X_m(k,n) + V_m(k,n), \ m = 1, 2, \ldots, M, \end{aligned} \tag{6.109}$$

where $Y_m(k,n)$, $G_m(k)$, $X(k,n)$, $V_m(k,n)$, and $X_m(k,n)$ are the STFTs of $y_m(t)$, $g_m(t)$, $x(t)$, $v_m(t)$, and $x_m(t)$, respectively, at the frequency bin $k \in \{0, 1, \ldots, K-1\}$ and the time frame n. Assuming that the first sensor is the reference and dropping the dependence on n, we can write the M STFT-domain microphone signals in a vector notation as

$$\begin{aligned} \mathbf{y}(k) &= \left[Y_1(k)\ Y_2(k) \cdots\ Y_M(k) \right]^T \\ &= \mathbf{d}(k)X_1(k) + \mathbf{v}(k) \\ &= \mathbf{x}(k) + \mathbf{v}(k), \end{aligned} \tag{6.110}$$

where

$$\mathbf{d}(k) = \left[1\ \frac{G_2(k)}{G_1(k)} \cdots\ \frac{G_M(k)}{G_1(k)} \right]^T, \tag{6.111}$$

and $\mathbf{x}(k)$ and $\mathbf{v}(k)$ are defined similarly to $\mathbf{y}(k)$. Since $X_m(k)$ and $V_m(k)$ are uncorrelated by assumption, we deduce that the $M \times M$ covariance matrix of $\mathbf{y}(k)$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k) &= E\left[\mathbf{y}(k)\mathbf{y}^H(k)\right] \\ &= \phi_{X_1}(k)\mathbf{d}(k)\mathbf{d}^H(k) + \mathbf{\Phi}_{\mathbf{v}}(k) \\ &= \mathbf{\Phi}_{\mathbf{x}}(k) + \mathbf{\Phi}_{\mathbf{v}}(k),\end{aligned} \tag{6.112}$$

where $\phi_{X_1}(k) = E\left[|X_1(k)|^2\right]$ is the variance of $X_1(k)$, and $\mathbf{\Phi}_{\mathbf{x}}(k)$ and $\mathbf{\Phi}_{\mathbf{v}}(k)$ are the covariance matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively. It results that the subband and input SNRs are, respectively,

$$\text{iSNR}(k) = \frac{\phi_{X_1}(k)}{\phi_{V_1}(k)} \tag{6.113}$$

and

$$\text{iSNR} = \frac{\sum_{k=0}^{K-1} \phi_{X_1}(k)}{\sum_{k=0}^{K-1} \phi_{V_1}(k)}, \tag{6.114}$$

where $\phi_{V_1}(k) = E\left[|V_1(k)|^2\right]$ is the variance of $V_1(k)$, the additive noise at the first (reference) sensor. It is obvious that

$$\min_k \text{iSNR}(k) \leq \text{iSNR} \leq \max_k \text{iSNR}(k). \tag{6.115}$$

As before, the two Hermitian matrices $\mathbf{\Phi}_{\mathbf{x}}(k)$ and $\mathbf{\Phi}_{\mathbf{v}}(k)$ can be jointly diagonalized as follows [4]:

$$\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{x}}(k)\mathbf{A}(k) = \mathbf{\Lambda}(k), \tag{6.116}$$

$$\mathbf{A}^H(k)\mathbf{\Phi}_{\mathbf{v}}(k)\mathbf{A}(k) = \mathbf{I}_M, \tag{6.117}$$

where $\mathbf{A}(k)$ is a full-rank square matrix (of size $M \times M$), $\mathbf{\Lambda}(k)$ is a diagonal matrix whose main elements are real and nonnegative, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. Furthermore, $\mathbf{\Lambda}(k)$ and $\mathbf{A}(k)$ are the eigenvalue and eigenvector matrices, respectively, of $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{\Phi}_{\mathbf{x}}(k)$. Since the rank of $\mathbf{\Phi}_{\mathbf{x}}(k)$ is equal to 1, the eigenvalues of $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{\Phi}_{\mathbf{x}}(k)$ are $\lambda_1(k) = \phi_{X_1}(k)\mathbf{d}^H(k)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{d}(k)$ and $\lambda_2(k) = \lambda_3(k) = \cdots = \lambda_M(k) = 0$. In other words, the first and last $M-1$ eigenvalues of the matrix product $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{\Phi}_{\mathbf{x}}(k)$ are positive and exactly zero, respectively. We also denote $\mathbf{a}_1(k), \mathbf{a}_2(k), \ldots, \mathbf{a}_M(k)$, the corresponding eigenvectors, where the first one can be expressed as

$$\mathbf{a}_1(k) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{d}(k)}{\sqrt{\mathbf{d}^H(k)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k)\mathbf{d}(k)}}. \tag{6.118}$$

Conventional multichannel speech enhancement in the STFT domain is performed by applying a complex-valued filter, $\mathbf{h}(k)$ of length M, to the observation signal vector, $\mathbf{y}(k)$, i.e.,

$$\begin{aligned} Z(k) &= \mathbf{h}^H(k)\mathbf{y}(k) \\ &= X_{\text{fd}}(k) + V_{\text{rn}}(k), \end{aligned} \tag{6.119}$$

where $Z(k)$ is the estimate of $X_1(k,n)$, $X_{\text{fd}}(k) = \mathbf{h}^H(k)\mathbf{x}(k) = X_1(k)\mathbf{h}^H(k)\mathbf{d}(k)$ is the filtered desired signal, and $V_{\text{rn}}(k) = \mathbf{h}^H(k)\mathbf{v}(k)$ is the residual noise. The variance of $Z(k)$ is then

$$\begin{aligned} \phi_Z(k) &= \mathbf{h}^H(k)\mathbf{\Phi}_{\mathbf{y}}(k)\mathbf{h}(k) \\ &= \phi_{X_{\text{fd}}}(k) + \phi_{V_{\text{rn}}}(k), \end{aligned} \tag{6.120}$$

where $\phi_{X_{\text{fd}}}(k) = \phi_{X_1}(k)\left|\mathbf{h}^H(k)\mathbf{d}(k)\right|^2$ and $\phi_{V_{\text{rn}}}(k) = \mathbf{h}^H(k)\mathbf{\Phi}_{\mathbf{v}}(k)\mathbf{h}(k)$ are the variances of $X_{\text{fd}}(k)$ and $V_{\text{rn}}(k)$, respectively. We deduce from (6.120) that the subband and fullband output SNRs are, respectively,

$$\begin{aligned} \text{oSNR}\left[\mathbf{h}(k)\right] &= \frac{\phi_{X_{\text{fd}}}(k)}{\phi_{V_{\text{rn}}}(k)} \\ &= \frac{\phi_{X_1}(k)\left|\mathbf{h}^H(k)\mathbf{d}(k)\right|^2}{\mathbf{h}^H(k)\mathbf{\Phi}_{\mathbf{v}}(k)\mathbf{h}(k)} \end{aligned} \tag{6.121}$$

and

$$\text{oSNR}\left[\mathbf{h}(:)\right] = \frac{\sum_{k=0}^{K-1}\phi_{X_{\text{fd}}}(k)}{\sum_{k=0}^{K-1}\phi_{V_{\text{rn}}}(k)}. \tag{6.122}$$

Again, we propose to use the index k_i, $i = 0, 1, \ldots, K-1$ and $k_i \in \{0, 1, \ldots, K-1\}$ to order the K subband eigenvalues $\lambda_1(k)$, $k = 0, 1, \ldots, K-1$ from the largest to the smallest, i.e.,

$$\lambda_1(k_0) \geq \lambda_1(k_1) \geq \cdots \geq \lambda_1(k_{K-1}). \tag{6.123}$$

So, with this indexing, the subband filter is denoted as $\mathbf{h}(k_i)$, which is assumed in the rest of this section to have the form:

$$\mathbf{h}(k_i) = \psi(k_i)\mathbf{a}_1(k_i), \tag{6.124}$$

where $\psi(k_i)$ is an arbitrary complex number and $\mathbf{a}_1(k_i)$ is the eigenvector corresponding to $\lambda_1(k_i)$. For $\psi(k_i) \neq 0$, it is clear that $\mathbf{h}(k_i)$ maximizes the subband output SNR, since

$$\begin{aligned} \text{oSNR}\left[\mathbf{h}(k_i)\right] &= \lambda_1(k_i) \\ &= \phi_{X_1}(k_i)\left|\mathbf{a}_1^H(k_i)\mathbf{d}(k_i)\right|^2. \end{aligned} \tag{6.125}$$

As a result, (6.123) is equivalent to

$$\text{oSNR}\left[\mathbf{h}(k_0)\right] \geq \text{oSNR}\left[\mathbf{h}(k_1)\right] \geq \cdots \geq \text{oSNR}\left[\mathbf{h}(k_{K-1})\right]. \tag{6.126}$$

Also, we have

$$\lambda_1(k_i) \geq \text{iSNR}(k_i) \tag{6.127}$$

and

$$\text{iSNR} \leq \lambda_1(k_0). \tag{6.128}$$

Let

$$\underline{\mathbf{h}} = \left[\mathbf{h}^T(k_0)\ \mathbf{h}^T(k_1) \cdots\ \mathbf{h}^T(k_{K-1})\right]^T \tag{6.129}$$

be a long filter of length KM containing all the ordered subband filters. We can express the fullband output in (6.122) as

$$\text{oSNR}\left(\underline{\mathbf{h}}\right) = \frac{\underline{\mathbf{h}}^H \mathbf{D}_{\mathbf{\Phi_x}} \underline{\mathbf{h}}}{\underline{\mathbf{h}}^H \mathbf{D}_{\mathbf{\Phi_v}} \underline{\mathbf{h}}}, \tag{6.130}$$

where

$$\mathbf{D}_{\mathbf{\Phi_x}} = \text{diag}\left[\mathbf{\Phi_x}(k_0), \mathbf{\Phi_x}(k_1), \ldots, \mathbf{\Phi_x}(k_{K-1})\right] \tag{6.131}$$

$$\mathbf{D}_{\mathbf{\Phi_v}} = \text{diag}\left[\mathbf{\Phi_v}(k_0), \mathbf{\Phi_v}(k_1), \ldots, \mathbf{\Phi_v}(k_{K-1})\right] \tag{6.132}$$

are block diagonal matrices. It is worth noticing that

$$\mathbf{D}_{\mathbf{\Phi_v}}^{-1} \mathbf{D}_{\mathbf{\Phi_x}} \mathbf{D}_{\mathbf{A}} = \mathbf{D}_{\mathbf{A}} \mathbf{D}_{\mathbf{\Lambda}}, \tag{6.133}$$

where

$$\mathbf{D}_{\mathbf{\Phi_v}}^{-1} = \text{diag}\left[\mathbf{\Phi_v}^{-1}(k_0), \mathbf{\Phi_v}^{-1}(k_1), \ldots, \mathbf{\Phi_v}^{-1}(k_{K-1})\right], \tag{6.134}$$

$$\mathbf{D}_{\mathbf{A}} = \text{diag}\left[\mathbf{A}(k_0), \mathbf{A}(k_1), \ldots, \mathbf{A}(k_{K-1})\right], \tag{6.135}$$

$$\mathbf{D}_{\mathbf{\Lambda}} = \text{diag}\left[\mathbf{\Lambda}(k_0), \mathbf{\Lambda}(k_1), \ldots, \mathbf{\Lambda}(k_{K-1})\right]. \tag{6.136}$$

Since

$$\text{oSNR}\left(\underline{\mathbf{h}}\right) = \frac{\sum_{i=0}^{K-1} \left|\psi(k_i)\right|^2 \lambda_1(k_i)}{\sum_{i=0}^{K-1} \left|\psi(k_i)\right|^2}, \tag{6.137}$$

we deduce that

$$\lambda_1(k_{K-1}) \leq \text{oSNR}\left(\underline{\mathbf{h}}\right) \leq \lambda_1(k_0). \tag{6.138}$$

Therefore, our objective is to find $\underline{\mathbf{h}}$ in such a way that $\text{oSNR}\left(\underline{\mathbf{h}}\right) > \text{iSNR}$.

The filter, $\underline{\mathbf{h}}$, that maximizes the fullband output SNR given in (6.130) is the eigenvector corresponding to the maximum eigenvalue, $\lambda_1(k_0)$, of the matrix $\mathbf{D}_{\mathbf{\Phi}_\mathbf{v}}^{-1}\mathbf{D}_{\mathbf{\Phi}_\mathbf{x}}$. As a consequence, the maximum SNR filter (of length KM) is

$$\underline{\mathbf{h}}_{\max} = \left[\, \psi(k_0)\mathbf{a}_1^T(k_0)\ \mathbf{0}^T \,\right]^T, \tag{6.139}$$

where $\psi(k_0) \neq 0$ is an arbitrary complex number. Equivalently, we can write (6.139) as

$$\begin{cases} \mathbf{h}_{\max}(k_0) = \psi(k_0)\mathbf{a}_1(k_0) \\ \mathbf{h}_{\max}(k_i) = \mathbf{0},\ i = 1, 2, \ldots, K-1 \end{cases}. \tag{6.140}$$

With (6.139), we get the maximum possible fullband output SNR, which is

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\max}\right) = \lambda_1(k_0) \geq \text{iSNR} \tag{6.141}$$

and

$$\text{oSNR}\left(\underline{\mathbf{h}}_{\max}\right) \geq \text{oSNR}\left(\underline{\mathbf{h}}\right),\ \forall \underline{\mathbf{h}}. \tag{6.142}$$

We deduce that the estimate of the desired signal is

$$\begin{cases} \widehat{X}_{1,\max}(k_0) = \mathbf{h}_{\max}^H(k_0)\mathbf{y}(k_0) \\ \widehat{X}_{1,\max}(k_i) = 0,\ i = 1, 2, \ldots, K-1 \end{cases}. \tag{6.143}$$

Now, we need to find $\psi(k_0)$. The first possibility is from the MSE between $X_1(k_0)$ and $\widehat{X}_{1,\max}(k_0)$, i.e.,

$$J\left[\psi(k_0)\right] = E\left[\left|X_1(k_0) - \psi^*(k_0)\mathbf{a}_1^H(k_0)\mathbf{y}(k_0)\right|^2\right]. \tag{6.144}$$

The second possibility is to use the distortion-based MSE, i.e.,

$$J_{\text{d}}\left[\psi(k_0)\right] = E\left[\left|X_1(k_0) - \psi^*(k_0)\mathbf{a}_1^H(k_0)\mathbf{x}(k_0)\right|^2\right]. \tag{6.145}$$

From the minimization of $J\left[\psi(k_0)\right]$, we get the classical Wiener filter at the frequency bin k_0, i.e.,

$$\begin{aligned} \mathbf{h}_{\max,\text{W}}(k_0) &= \frac{\sqrt{\phi_{X_1}(k_0)\lambda_1(k_0)}}{1+\lambda_1(k_0)}\mathbf{a}_1(k_0) \\ &= \frac{\phi_{X_1}(k_0)\mathbf{\Phi}_\mathbf{v}^{-1}(k_0)\mathbf{d}(k_0)}{1+\phi_{X_1}(k_0)\mathbf{d}^H(k_0)\mathbf{\Phi}_\mathbf{v}^{-1}(k_0)\mathbf{d}(k_0)}, \end{aligned} \tag{6.146}$$

while the minimization of $J_{\text{d}}\left[\psi(k_0)\right]$ gives the well-known MVDR filter at the frequency bin k_0, i.e.,

$$\mathbf{h}_{\max,\mathrm{D}}(k_0) = \frac{\sqrt{\phi_{X_1}(k_0)\lambda_1(k_0)}}{\lambda_1(k_0)}\mathbf{a}_1(k_0) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(k_0)\mathbf{d}(k_0)}{\mathbf{d}^H(k_0)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k_0)\mathbf{d}(k_0)}. \tag{6.147}$$

Even though this method maximizes the fullband output SNR, it is expected to introduce a large distortion to the desired signal, since all its frequency bins are put to 0 except at k_0. A more practical approach when we deal with broadband signals such as speech is to form the filter (of length KM) from a concatenation of the eigenvectors corresponding to the $P(\leq K)$ largest eigenvalues from the set $\{\lambda_1(k_i),\ i = 0, 1, \ldots, K-1\}$, i.e.,

$$\underline{\mathbf{h}}_P = \left[\, \psi(k_0)\mathbf{a}_1^T(k_0) \cdots \psi(k_{P-1})\mathbf{a}_1^T(k_{P-1})\ \mathbf{0}^T \,\right]^T, \tag{6.148}$$

where $\psi(k_p),\ p = 0, 1, \ldots, P-1$ are arbitrary complex numbers with at least one of them different from 0. We can also express (6.148) as

$$\begin{cases} \mathbf{h}_P(k_p) = \psi(k_p)\mathbf{a}_1(k_p),\ p = 0, 1, \ldots, P-1 \\ \mathbf{h}_P(k_i) = \mathbf{0},\ i = P, P+1, \ldots, K-1 \end{cases}. \tag{6.149}$$

Hence, the estimate of the desired signal is

$$\begin{cases} \widehat{X}_{1,P}(k_p) = \mathbf{h}_P^H(k_p)\mathbf{y}(k_p),\ p = 0, 1, \ldots, P-1 \\ \widehat{X}_{1,P}(k_i) = 0,\ i = P, P+1, \ldots, K-1 \end{cases}. \tag{6.150}$$

To find the $\psi(k_p)$'s, we can either optimize $J\left[\psi(k_p)\right]$ or $J_{\mathrm{d}}\left[\psi(k_p)\right]$. The first one leads to the Wiener filters at the frequency bins $k_p,\ p = 0, 1, \ldots, P-1$, i.e.,

$$\mathbf{h}_{P,\mathrm{W}}(k_p) = \frac{\phi_{X_1}(k_p)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k_p)\mathbf{d}(k_p)}{1 + \phi_{X_1}(k_p)\mathbf{d}^H(k_p)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k_p)\mathbf{d}(k_p)}, \tag{6.151}$$

while the second one gives the MVDR filters at the frequency bins $k_p,\ p = 0, 1, \ldots, P-1$, i.e.,

$$\mathbf{h}_{P,\mathrm{D}}(k_p) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}(k_p)\mathbf{d}(k_p)}{\mathbf{d}^H(k_p)\mathbf{\Phi}_{\mathbf{v}}^{-1}(k_p)\mathbf{d}(k_p)}. \tag{6.152}$$

The filters (of length KM) corresponding to (6.151) and (6.152) are, respectively,

$$\underline{\mathbf{h}}_{P,\mathrm{W}} = \left[\, \mathbf{h}_{P,\mathrm{W}}^T(k_0) \cdots \mathbf{h}_{P,\mathrm{W}}^T(k_{P-1})\ \mathbf{0}^T \,\right]^T \tag{6.153}$$

and

$$\underline{\mathbf{h}}_{P,\mathrm{D}} = \left[\, \mathbf{h}_{P,\mathrm{D}}^T(k_0) \cdots \mathbf{h}_{P,\mathrm{D}}^T(k_{P-1})\ \mathbf{0}^T \,\right]^T. \tag{6.154}$$

For $P = K$, $\underline{\mathbf{h}}_{K,\mathrm{W}}$ and $\underline{\mathbf{h}}_{K,\mathrm{D}}$ correspond to the classical multichannel Wiener and MVDR approaches, respectively. The case $\underline{\mathbf{h}}_{P,\mathrm{D}}$ can be seen as a generalization of the ideal binary mask [3] to the multichannel case, since the subband observation signals of the reference microphone with the P largest subband output SNRs are processed in such a way that the desired signals are undistorted while the $K - P$ others with the smallest subband output SNRs are put to 0. We should always have

$$\mathrm{oSNR}\left(\underline{\mathbf{h}}_{P,\mathrm{D}}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{P,\mathrm{W}}\right). \tag{6.155}$$

We also deduce that

$$\mathrm{iSNR} \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K,\mathrm{W}}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K-1,\mathrm{W}}\right) \leq \cdots \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{1,\mathrm{W}}\right) = \lambda_1(k_0) \tag{6.156}$$

and

$$\mathrm{iSNR} \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K,\mathrm{D}}\right) \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{K-1,\mathrm{D}}\right) \leq \cdots \leq \mathrm{oSNR}\left(\underline{\mathbf{h}}_{1,\mathrm{D}}\right) = \lambda_1(k_0). \tag{6.157}$$

References

1. Y. Zhao, J. Benesty, and J. Chen, "Single-channel noise reduction in the STFT domain from the fullband output SNR perspective," in *Proc. EUSIPCO*, 2016, pp. 1956–1959.
2. J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Springer Briefs in Electrical and Computer Engineering, 2011.
3. D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Pierre Divenyi, Ed., pp. 181–197, Kluwer, 2005.
4. J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.
5. J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
6. M. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

Index

J. Benesty, *Fundamentals of Speech Enhancement*, SpringerBriefs in Electrical and Computer Engineering, https://doi.org/10.1007/978-3-319-74524-4