

Chapter 2

Conceptual Framework

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

2.1 Signal Model

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

$$y = x + v, \quad (2.1)$$

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

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

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

$$\phi_x = E(|x|^2), \quad (2.3)$$

$$\phi_v = E(|v|^2), \quad (2.4)$$

are the variances of x and v , respectively.

2.2 Principle of the Conceptual Framework

The objective of noise reduction/speech enhancement in any domain is to find a “good” estimate¹, \hat{x} , of the desired signal, x , given y and y^* , where the superscript $*$ denotes complex conjugation, with an appropriate function $f(\cdot)$, i.e.,

$$\hat{x} = f(y, y^*). \quad (2.5)$$

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

$$\begin{aligned} \hat{x} &= x_{\text{ld}} + x_{\text{ri}} + v_{\text{rn}} \\ &= x_{\text{ld}} + u \\ &= \rho^* x + u, \end{aligned} \quad (2.6)$$

where

$$x_{\text{ld}} = \rho^* x \quad (2.7)$$

is a linear version of the desired signal,

$$\begin{aligned} \rho &= \frac{E(x\hat{x}^*)}{\phi_x} \\ &= \frac{\phi_{x\hat{x}}}{\phi_x} \end{aligned} \quad (2.8)$$

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

$$\phi_{x\hat{x}} = E(x\hat{x}^*) \quad (2.9)$$

is the correlation between x and \hat{x} ,

$$\begin{aligned} u &= x_{\text{ri}} + v_{\text{rn}} \\ &= \hat{x} - \rho^* x \end{aligned} \quad (2.10)$$

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

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

$$\phi_{x_{\text{ri}}v_{\text{rn}}} = E(x_{\text{ri}}v_{\text{rn}}^*) = 0, \quad (2.11)$$

$$\phi_{xu} = E(xu^*) = 0. \quad (2.12)$$

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

$$\begin{aligned} \phi_{\hat{x}} &= \phi_{x_{\text{ld}}} + \phi_{x_{\text{ri}}} + \phi_{v_{\text{rn}}} \\ &= |\rho|^2 \phi_x + \phi_u, \end{aligned} \quad (2.13)$$

where

$$\phi_{x_{\text{ld}}} = |\rho|^2 \phi_x, \quad (2.14)$$

$$\phi_{x_{\text{ri}}} = E(|x_{\text{ri}}|^2), \quad (2.15)$$

$$\phi_{v_{\text{rn}}} = E(|v_{\text{rn}}|^2), \quad (2.16)$$

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

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

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

$$\phi_{x_{\text{ld}}} \leq \phi_x, \quad (2.18)$$

which is equivalent to saying that

$$|\rho|^2 \leq 1. \quad (2.19)$$

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

$$\phi_u = E(|\hat{x}|^2) - E(x\hat{x}^*) \quad (2.20)$$

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

We also observe from (2.6) that two different distortions affect the estimated desired signal. The first distortion is due to the scaling factor², ρ^* (and, possibly, to the residual interference, x_{ri}), and the second one is due to the additive residual noise, v_{rn} . We will refer to these two distortions as distortion 1 and distortion 2, respectively. It is reasonable to say that distortion 1 affects the intelligibility of the estimated signal since when ρ^* is small, not much energy of \hat{x} is left in $\phi_{\hat{x}}$, and when ρ^* is close to 1, almost the whole desired signal is in \hat{x} . Distortion 2 affects both the quality and intelligibility of the estimated signal since the smaller is the variance of v_{rn} , the more pleasant it is to hear to \hat{x} and the better is its intelligibility. To summarize, distortion 1 is related to speech intelligibility while distortion 2 is related to both speech quality and intelligibility.

2.3 Performance Measures

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

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

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

$$\begin{aligned} \text{iSNR} &= \frac{\phi_x}{\phi_v} \\ &= \frac{|\gamma_{xy}|^2}{1 - |\gamma_{xy}|^2}, \end{aligned} \quad (2.21)$$

where

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

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

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

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

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

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

$$\frac{\phi_u}{\phi_v} \leq |\rho|^2 \leq 1, \quad (2.24)$$

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

The output SNR can be rewritten as

$$\text{oSNR} = \frac{|\gamma_{x\hat{x}}|^2}{1 - |\gamma_{x\hat{x}}|^2}, \quad (2.25)$$

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

The gain in SNR is defined as

$$\mathcal{G} = \frac{\text{oSNR}}{\text{iSNR}}. \quad (2.26)$$

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

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

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

Let us now open a short parenthesis on a widely used definition in the literature of the SNR after processing, often called SNR improvement³. It is defined as

$$\begin{aligned} \text{SNR}_{\text{imp}} &= \frac{\phi_x}{E(|x - \hat{x}|^2)} \\ &= \frac{\phi_x}{|1 - \rho^*|^2 \phi_x + \phi_u}. \end{aligned} \quad (2.28)$$

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

$$\text{SNR}_{\text{imp}} = \frac{\text{oSNR}}{|1 - \rho^*|^2 \text{oSNR} + |\rho|^2}. \quad (2.29)$$

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

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

$$\begin{aligned} v_i &= \frac{\phi_x - \phi_{x_{\text{id}}}}{\phi_x} \\ &= 1 - |\rho|^2. \end{aligned} \quad (2.30)$$

The larger is v_i , the less intelligible is the estimated desired signal, \hat{x} .

The speech quality index (from distortion 2) is obtained by comparing the variance of the additive noise from the observation signal to the variance of the additive residual noise after processing with $f(y, y^*)$. We have⁴

$$v_q = \frac{\phi_{v_{\text{rn}}}}{\phi_v}. \quad (2.31)$$

For a fixed value of the input SNR, the quality of the signal degrades as v_q increases.

It can be checked that

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

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

$$\frac{\phi_x - \phi_{\hat{x}}}{\phi_x} = v_i - \frac{v_q}{i\text{SNR}} - \frac{\phi_{x_{ri}}}{\phi_x} \quad (2.32)$$

or

$$\phi_{\hat{x}} = (1 - v_i) \phi_x + \phi_{x_{ri}} + v_q \phi_v. \quad (2.33)$$

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

$$v'_i = (1 - \varpi) v_i + \varpi v_q, \quad (2.34)$$

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

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

2.4 Mean-Squared-Error (MSE) Based Criterion

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

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

$$\begin{aligned} e &= \hat{x} - x \\ &= x_{ld} + u - x, \end{aligned} \quad (2.35)$$

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

$$e = e_i + e_q, \quad (2.36)$$

where

$$e_i = (\rho^* - 1) x \quad (2.37)$$

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

$$e_q = u \quad (2.38)$$

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

$$E(e_i e_q^*) = 0. \quad (2.39)$$

The classical MSE criterion is then

$$\begin{aligned} J[f(y, y^*)] &= E(|e|^2) \\ &= \phi_x - \phi_{x\hat{x}} - \phi_{x\hat{x}}^* + \phi_{\hat{x}} \\ &= |1 - \rho^*|^2 \phi_x + \phi_u \\ &= J_i[f(y, y^*)] + J_q[f(y, y^*)], \end{aligned} \quad (2.40)$$

where

$$\begin{aligned} J_i[f(y, y^*)] &= E(|e_i|^2) \\ &= |1 - \rho^*|^2 \phi_x \end{aligned} \quad (2.41)$$

and

$$\begin{aligned} J_q[f(y, y^*)] &= E(|e_q|^2) \\ &= \phi_u. \end{aligned} \quad (2.42)$$

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

$$J[f_1(y, y^*)] = J_q[f_1(y, y^*)] = \phi_v, \quad (2.43)$$

$$J[f_0(y, y^*)] = J_i[f_0(y, y^*)] = \phi_x. \quad (2.44)$$

As a result,

$$\text{iSNR} = \frac{J[f_0(y, y^*)]}{J[f_1(y, y^*)]}. \quad (2.45)$$

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

$$\begin{aligned} J_{n,1}[f(y, y^*)] &= \frac{J[f(y, y^*)]}{J[f_1(y, y^*)]} \\ &= \text{iSNR} \times |1 - \rho^*|^2 + \frac{\phi_u}{\phi_v}. \end{aligned} \quad (2.46)$$

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

$$\begin{aligned} J_{n,2}[f(y, y^*)] &= \frac{J[f(y, y^*)]}{J[f_0(y, y^*)]} \\ &= |1 - \rho^*|^2 + \frac{\phi_u}{\phi_x} \end{aligned} \quad (2.47)$$

and, obviously,

$$J_{n,1}[f(y, y^*)] = \text{iSNR} \times J_{n,2}[f(y, y^*)]. \quad (2.48)$$

We are only interested in functions for which

$$J_i[f_1(y, y^*)] \leq J_i[f(y, y^*)] < J_i[f_0(y, y^*)], \quad (2.49)$$

$$J_q[f_0(y, y^*)] < J_q[f(y, y^*)] < J_q[f_1(y, y^*)]. \quad (2.50)$$

From the two previous expressions, we deduce that

$$0 \leq |1 - \rho^*|^2 < 1, \quad (2.51)$$

$$0 < \frac{\phi_u}{\phi_v} < 1. \quad (2.52)$$

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

$$\phi_{x\hat{x}_W} = \phi_{\hat{x}_W}. \quad (2.53)$$

As a result, the minimum MSE (MMSE) is

$$J_{\min}[f(y, y^*)] = \phi_x - \phi_{\hat{x}_W}. \quad (2.54)$$

We deduce that $\phi_{\hat{x}_W} \leq \phi_x$ [i.e., the function $f(y, y^*)$ does not amplify the estimated desired signal],

$$\rho = \frac{\phi_{\hat{x}_W}}{\phi_x} \leq 1 \quad (2.55)$$

is always real and positive,

$$|\gamma_{x\hat{x}_W}|^2 = \rho, \quad (2.56)$$

$$J_{\min}[f(y, y^*)] = \phi_x \left[1 - |\gamma_{x\hat{x}_W}|^2 \right], \quad (2.57)$$

$$\text{oSNR} = \frac{\rho}{1 - \rho}, \quad (2.58)$$

$$\phi_u = \rho (1 - \rho) \phi_x \leq \phi_v, \quad (2.59)$$

and

$$\frac{J_{\min} [f(y, y^*)]}{\phi_x} = v_i - \frac{v_q}{\text{iSNR}} - \frac{\phi_{x_{ri}}}{\phi_x}. \quad (2.60)$$

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

$$\begin{aligned} J_\mu [f(y, y^*)] &= \mu \frac{J_i [f(y, y^*)]}{\phi_x} + \frac{J_q [f(y, y^*)]}{\phi_v} \\ &= \mu |1 - \rho^*|^2 + \frac{\phi_u}{\phi_v}, \end{aligned} \quad (2.61)$$

where μ is a positive real number allowing to compromise between v_i and v_q .

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

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

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

2.5 Summary

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

References

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

A Conceptual Framework for Noise Reduction

Benesty, J.; Chen, J.

2015, VIII, 89 p. 11 illus. in color., Softcover

ISBN: 978-3-319-12954-9