SpringerBriefs in Electrical and Computer Engineering

Speech Enhancement in the STFT Domain

Bearbeitet von
Jacob Benesty, Jingdong Chen, Emanuël A.P. Habets

1. Auflage 2011. Taschenbuch. VII, 109 S. Paperback
ISBN 978 3 642 23249 7
Format (B x L): 15,5 x 23,5 cm
Gewicht: 193 g

Chapter 2
Single-Channel Speech Enhancement with a Gain

There are different ways to perform speech enhancement in the frequency domain from a single microphone signal. The simplest way is to estimate the desired signal from the noisy observation with a simple complex gain. This approach is investigated in this chapter and all well-known optimal gains are derived. We start by explaining the single-channel signal model for speech enhancement in the time and frequency domains.

2.1 Signal Model

The noise reduction or speech enhancement problem considered in this study is one of recovering the desired signal (or clean speech) $x(t)$, t being the time index, of zero mean from the noisy observation (microphone signal) [1–3]

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

where $v(t)$ is the unwanted additive noise, which is assumed to be a zero-mean random process white or colored but uncorrelated with $x(t)$. All signals are considered to be real and broadband.

Using the short-time Fourier transform (STFT),[1] (2.1) can be rewritten in the frequency domain as

$$Y(k,m) = X(k,m) + V(k,m), \tag{2.2}$$

where the zero-mean complex random variables $Y(k,m)$, $X(k,m)$, and $V(k,m)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at frequency-bin $k \in \{0, 1, \ldots, K-1\}$ and time-frame m. Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the variance of $Y(k,m)$ is

[1] Note that the concepts presented in this work can be applied to any other transformed domain.

J. Benesty et al., *Speech Enhancement in the STFT Domain*,
SpringerBriefs in Electrical and Computer Engineering,
DOI: 10.1007/978-3-642-23250-3_2, © The Author(s) 2012

$$\begin{aligned}\phi_Y(k,m) &= E\left[|Y(k,m)|^2\right] \\ &= \phi_X(k,m) + \phi_V(k,m),\end{aligned} \tag{2.3}$$

where $E[\cdot]$ denotes mathematical expectation, and

$$\phi_X(k,m) = E\left[|X(k,m)|^2\right], \tag{2.4}$$

$$\phi_V(k,m) = E\left[|V(k,m)|^2\right], \tag{2.5}$$

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

2.2 Microphone Signal Processing with a Gain

In this chapter, we try to estimate the desired signal, $X(k,m)$, from the noisy observation, $Y(k,m)$, i.e.,

$$Z(k,m) = H(k,m)Y(k,m), \tag{2.6}$$

where $Z(k,m)$ is supposed to be the estimate of $X(k,m)$ and $H(k,m)$ is a complex gain that needs to be determined. This procedure is called the single-channel speech enhancement in the STFT domain with a complex gain.

We can express (2.6) as

$$\begin{aligned}Z(k,m) &= H(k,m)\left[X(k,m) + V(k,m)\right] \\ &= X_{\text{fd}}(k,m) + V_{\text{rn}}(k,m),\end{aligned} \tag{2.7}$$

where

$$X_{\text{fd}}(k,m) = H(k,m)X(k,m) \tag{2.8}$$

is the filtered desired signal and

$$V_{\text{rn}}(k,m) = H(k,m)V(k,m) \tag{2.9}$$

is the residual noise.

Since the estimate of the desired signal is the sum of two terms that are uncorrelated, the variance of $Z(k,m)$ is

$$\begin{aligned}\phi_Z(k,m) &= |H(k,m)|^2\,\phi_Y(k,m) \\ &= \phi_{X_{\text{fd}}}(k,m) + \phi_{V_{\text{rn}}}(k,m),\end{aligned} \tag{2.10}$$

where

$$\phi_{X_{\mathrm{fd}}}(k,m) = |H(k,m)|^2 \phi_X(k,m), \tag{2.11}$$
$$\phi_{V_{\mathrm{rn}}}(k,m) = |H(k,m)|^2 \phi_V(k,m), \tag{2.12}$$

are the variances of $X_{\mathrm{fd}}(k,m)$ and $V_{\mathrm{rn}}(k,m)$, respectively.

2.3 Performance Measures

The first attempts to derive relevant and rigorous performance measures in the context of speech enhancement can be found in [1, 4, 5]. These references are the main inspiration for the derivation of measures in the studied context throughout this work.

In this section, we are going to define the most useful performance measures for single-channel speech enhancement with a gain in the STFT domain. We can divide these measures into two categories. The first category evaluates the noise reduction performance while the second one evaluates speech distortion. We are also going to discuss the very convenient mean-square error (MSE) criterion and show how it is related to the performance measures.

2.3.1 Noise Reduction

One of the most fundamental measures in all aspects of speech enhancement is the signal-to-noise ratio (SNR). The input SNR is a second-order measure which quantifies the level of noise present relative to the level of the desired signal.

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

$$\mathrm{iSNR}(k,m) = \frac{\phi_X(k,m)}{\phi_V(k,m)}, \ k = 0, 1, \ldots, K-1, \tag{2.13}$$
$$\mathrm{iSNR}(m) = \frac{\sum_{k=0}^{K-1} \phi_X(k,m)}{\sum_{k=0}^{K-1} \phi_V(k,m)}. \tag{2.14}$$

It is easy to show that [1]

$$\mathrm{iSNR}(m) \leq \sum_{k=0}^{K-1} \mathrm{iSNR}(k,m). \tag{2.15}$$

To quantify the level of the noise remaining after the noise reduction processing via the complex gain, we define the output SNR as the ratio of the variance of the

filtered desired signal over the variance of the residual noise. We easily deduce the subband output SNR

$$\begin{aligned}\text{oSNR}\left[H(k,m)\right] &= \frac{|H(k,m)|^2\,\phi_X(k,m)}{|H(k,m)|^2\,\phi_V(k,m)}\\ &= \frac{\phi_X(k,m)}{\phi_V(k,m)},\ k=0,1,\ldots,K-1\end{aligned} \tag{2.16}$$

and the fullband output SNR

$$\text{oSNR}\left[H(:,m)\right] = \frac{\sum_{k=0}^{K-1}|H(k,m)|^2\,\phi_X(k,m)}{\sum_{k=0}^{K-1}|H(k,m)|^2\,\phi_V(k,m)}. \tag{2.17}$$

We notice that the subband output SNR is equal to the subband input SNR, so the subband SNR cannot be improved with just a gain but the fullband output SNR can. It can be verified that [1]

$$\text{oSNR}\left[H(:,m)\right] \leq \sum_{k=0}^{K-1}\text{iSNR}(k,m). \tag{2.18}$$

The previous inequality shows that the fullband output SNR is always upper bounded no matter the choices of the $H(k,m)$.

For the particular gain $H(k,m)=1$, we have

$$\text{oSNR}\left[1(k,m)\right] = \text{iSNR}(k,m),\ k=0,1,\ldots,K-1, \tag{2.19}$$

$$\text{oSNR}\left[1(:,m)\right] = \text{iSNR}(m). \tag{2.20}$$

With the identity gain, 1, the SNR cannot be improved.

The noise reduction factor [4, 5] quantifies the amount of noise whose is rejected by the complex gain. This quantity is defined as the ratio of the variance of the noise at the microphone over the variance of the residual noise. The subband and fullband noise reduction factors are then

$$\begin{aligned}\xi_{\text{nr}}\left[H(k,m)\right] &= \frac{\phi_V(k,m)}{|H(k,m)|^2\,\phi_V(k,m)}\\ &= \frac{1}{|H(k,m)|^2},\ k=0,1,\ldots,K-1,\end{aligned} \tag{2.21}$$

$$\begin{aligned}\xi_{\text{nr}}\left[H(:,m)\right] &= \frac{\sum_{k=0}^{K-1}\phi_V(k,m)}{\sum_{k=0}^{K-1}|H(k,m)|^2\,\phi_V(k,m)}\\ &= \frac{\sum_{k=0}^{K-1}\phi_V(k,m)}{\sum_{k=0}^{K-1}\xi_{\text{nr}}^{-1}\left[H(k,m)\right]\phi_V(k,m)},\end{aligned} \tag{2.22}$$

and we always have

$$\xi_{\mathrm{nr}}\left[H(:,m)\right] \leq \sum_{k=0}^{K-1} \xi_{\mathrm{nr}}\left[H(k,m)\right]. \tag{2.23}$$

The noise reduction factors are expected to be lower bounded by 1 for appropriate choices of the $H(k,m)$. So the more the noise is reduced, the higher are the values of the noise reduction factors.

2.3.2 Speech Distortion

In practice, the complex gain distorts the desired signal. In order to evaluate the level of this distortion, we define the speech reduction factor [1] as the variance of the desired signal over the variance of the filtered desired signal. Therefore, the subband and fullband speech reduction factors are defined as

$$\begin{aligned}\xi_{\mathrm{sr}}\left[H(k,m)\right] &= \frac{\phi_X(k,m)}{|H(k,m)|^2\,\phi_X(k,m)} \\ &= \frac{1}{|H(k,m)|^2},\ k=0,1,\ldots,K-1, \end{aligned} \tag{2.24}$$

$$\begin{aligned}\xi_{\mathrm{sr}}\left[H(:,m)\right] &= \frac{\sum_{k=0}^{K-1}\phi_X(k,m)}{\sum_{k=0}^{K-1}|H(k,m)|^2\,\phi_X(k,m)} \\ &= \frac{\sum_{k=0}^{K-1}\phi_X(k,m)}{\sum_{k=0}^{K-1}\xi_{\mathrm{sr}}^{-1}\left[H(k,m)\right]\phi_X(k,m)}, \end{aligned} \tag{2.25}$$

and we always have

$$\xi_{\mathrm{sr}}\left[H(:,m)\right] \leq \sum_{k=0}^{K-1} \xi_{\mathrm{sr}}\left[H(k,m)\right]. \tag{2.26}$$

The speech reduction factor is equal to 1 if there is no distortion and expected to be greater than 1 when distortion occurs.

By making the appropriate substitutions, one can derive the relationships:

$$\frac{\mathrm{oSNR}\left[H(k,m)\right]}{\mathrm{iSNR}(k,m)} = \frac{\xi_{\mathrm{nr}}\left[H(k,m)\right]}{\xi_{\mathrm{sr}}\left[H(k,m)\right]},\ k=0,1,\ldots,K-1, \tag{2.27}$$

$$\frac{\mathrm{oSNR}\left[H(:,m)\right]}{\mathrm{iSNR}(m)} = \frac{\xi_{\mathrm{nr}}\left[H(:,m)\right]}{\xi_{\mathrm{sr}}\left[H(:,m)\right]}. \tag{2.28}$$

These expressions indicate the equivalence between gain/loss in SNR and distortion for both the subband and fullband cases.

Another way to measure the distortion of the desired speech signal due to the complex gain is the speech distortion index [1, 4, 5], which is defined as the mean-square error between the desired signal and the filtered desired signal, normalized by the variance of the desired signal, i.e.,

$$\begin{aligned}\upsilon_{\text{sd}}\left[H(k,m)\right] &= \frac{E\left\{\left|H(k,m)X(k,m)-X(k,m)\right|^2\right\}}{\phi_X(k,m)}\\ &= \left|H(k,m)-1\right|^2,\ k=0,1,\ldots,K-1\end{aligned} \tag{2.29}$$

in the subband case and

$$\begin{aligned}\upsilon_{\text{sd}}\left[H(:,m)\right] &= \frac{\sum_{k=0}^{K-1} E\left\{\left|H(k,m)X(k,m)-X(k,m)\right|^2\right\}}{\sum_{k=0}^{K-1}\phi_X(k,m)}\\ &= \frac{\sum_{k=0}^{K-1}\upsilon_{\text{sd}}\left[H(k,m)\right]\phi_X(k,m)}{\sum_{k=0}^{K-1}\phi_X(k,m)}\end{aligned} \tag{2.30}$$

in the fullband case. It can be verified that

$$\upsilon_{\text{sd}}\left[H(:,m)\right] \le \sum_{k=0}^{K-1}\upsilon_{\text{sd}}\left[H(k,m)\right]. \tag{2.31}$$

However, the speech distortion indices are usually upper bounded by 1 for optimal gains.

2.3.3 Mean-Square Error Criterion

Error criteria play a critical role in deriving optimal gains. The MSE [6] is, by far, the most practical one.

In the STFT domain, the error signal between the estimated and desired signals at the frequency-bin k and time-frame m is

$$\begin{aligned}\mathcal{E}(k,m) &= Z(k,m)-X(k,m)\\ &= H(k,m)Y(k,m)-X(k,m),\end{aligned} \tag{2.32}$$

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

$$\mathcal{E}(k,m) = \mathcal{E}_{\text{d}}(k,m)+\mathcal{E}_{\text{r}}(k,m), \tag{2.33}$$

where

$$\mathcal{E}_{\text{d}}(k,m) = \left[H(k,m)-1\right]X(k,m) \tag{2.34}$$

is the speech distortion due to the gain and

$$\mathcal{E}_{\mathrm{r}}(k,m) = H(k,m)V(k,m) \tag{2.35}$$

represents the residual noise.

The subband MSE criterion is then

$$\begin{aligned} J\left[H(k,m)\right] &= E\left[|\mathcal{E}(k,m)|^2\right] \\ &= \phi_X(k,m) + |H(k,m)|^2\,\phi_Y(k,m) - 2\mathcal{R}\left[H(k,m)\phi_{YX}(k,m)\right] \\ &= \phi_X(k,m) + |H(k,m)|^2\,\phi_Y(k,m) - 2\mathcal{R}\left[H(k,m)\phi_X(k,m)\right] \\ &= J_{\mathrm{d}}\left(k,m\right) + J_{\mathrm{r}}\left(k,m\right), \end{aligned} \tag{2.36}$$

where $\mathcal{R}[\cdot]$ is the real part of a complex number,

$$\begin{aligned} \phi_{YX}(k,m) &= E\left[Y(k,m)X^*(k,m)\right] \\ &= \phi_X(k,m) \end{aligned}$$

is the cross-correlation between the signals $Y(k,m)$ and $X(k,m)$, superscript * denotes complex conjugation,

$$\begin{aligned} J_{\mathrm{d}}\left[H(k,m)\right] &= E\left[|\mathcal{E}_{\mathrm{d}}(k,m)|^2\right] \\ &= |H(k,m)-1|^2\,\phi_X(k,m) \\ &= \upsilon_{\mathrm{sd}}\left[H(k,m)\right]\phi_X(k,m), \end{aligned} \tag{2.37}$$

and

$$\begin{aligned} J_{\mathrm{r}}\left[H(k,m)\right] &= E\left[|\mathcal{E}_{\mathrm{r}}(k,m)|^2\right] \\ &= |H(k,m)|^2\,\phi_V(k,m) \\ &= \frac{\phi_V(k,m)}{\xi_{\mathrm{nr}}\left[H(k,m)\right]}. \end{aligned} \tag{2.38}$$

Two particular gains are of great interest: $H(k,m) = 1$ and $H(k,m) = 0$. With the first one (identity gain), we have neither noise reduction nor speech distortion and with the second one (zero gain), we have maximum noise reduction and maximum speech distortion. For both gains, however, it can be verified that the output SNR is equal to the input SNR. For these two particular gains, the subband MSEs are

$$J\left[1(k,m)\right] = J_{\mathrm{r}}\left[1(k,m)\right] = \phi_V(k,m), \tag{2.39}$$

$$J\left[0(k,m)\right] = J_{\mathrm{d}}\left[0(k,m)\right] = \phi_X(k,m). \tag{2.40}$$

As a result,

$$\mathrm{iSNR}(k,m) = \frac{J\left[0(k,m)\right]}{J\left[1(k,m)\right]}. \tag{2.41}$$

We define the subband normalized MSE (NMSE) with respect to $J\left[1(k,m)\right]$ as

$$\begin{aligned}\widetilde{J}\left[H(k,m)\right] &= \frac{J\left[H(k,m)\right]}{J\left[1(k,m)\right]} \\ &= \mathrm{iSNR}(k,m)\cdot \upsilon_{\mathrm{sd}}\left[H(k,m)\right] + \frac{1}{\xi_{\mathrm{nr}}\left[H(k,m)\right]} \\ &= \mathrm{iSNR}(k,m)\left\{\upsilon_{\mathrm{sd}}\left[H(k,m)\right] + \frac{1}{\mathrm{oSNR}\left[H(k,m)\right]\cdot \xi_{\mathrm{sr}}\left[H(k,m)\right]}\right\},\end{aligned} \tag{2.42}$$

where

$$\upsilon_{\mathrm{sd}}\left[H(k,m)\right] = \frac{J_{\mathrm{d}}\left[H(k,m)\right]}{J_{\mathrm{d}}\left[0(k,m)\right]}, \tag{2.43}$$

$$\mathrm{iSNR}(k,m)\cdot \upsilon_{\mathrm{sd}}\left[H(k,m)\right] = \frac{J_{\mathrm{d}}\left[H(k,m)\right]}{J_{\mathrm{r}}\left[1(k,m)\right]}, \tag{2.44}$$

$$\xi_{\mathrm{nr}}\left[H(k,m)\right] = \frac{J_{\mathrm{r}}\left[1(k,m)\right]}{J_{\mathrm{r}}\left[H(k,m)\right]}, \tag{2.45}$$

$$\mathrm{oSNR}\left[H(k,m)\right]\cdot \xi_{\mathrm{sr}}\left[H(k,m)\right] = \frac{J_{\mathrm{d}}\left[0(k,m)\right]}{J_{\mathrm{r}}\left[H(k,m)\right]}. \tag{2.46}$$

This shows how this subband NMSE and the different subband MSEs are related to the performance measures.

We define the subband NMSE with respect to $J\left[0(k,m)\right]$ as

$$\begin{aligned}\overline{J}\left[H(k,m)\right] &= \frac{J\left[H(k,m)\right]}{J\left[0(k,m)\right]} \\ &= \upsilon_{\mathrm{sd}}\left[H(k,m)\right] + \frac{1}{\mathrm{oSNR}\left[H(k,m)\right]\cdot \xi_{\mathrm{sr}}\left[H(k,m)\right]}\end{aligned} \tag{2.47}$$

and, obviously,

$$\widetilde{J}\left[H(k,m)\right] = \mathrm{iSNR}(k,m)\cdot \overline{J}\left[H(k,m)\right]. \tag{2.48}$$

We are only interested in gains for which

$$J_{\mathrm{d}}\left[1(k,m)\right] \le J_{\mathrm{d}}\left[H(k,m)\right] < J_{\mathrm{d}}\left[0(k,m)\right], \tag{2.49}$$

$$J_{\mathrm{r}}\left[0(k,m)\right] < J_{\mathrm{r}}\left[H(k,m)\right] < J_{\mathrm{r}}\left[1(k,m)\right]. \tag{2.50}$$

From the two previous expressions, we deduce that

$$0 \leq \upsilon_{\mathrm{sd}}\left[H(k,m)\right] < 1, \tag{2.51}$$

$$1 < \xi_{\mathrm{nr}}\left[H(k,m)\right] < \infty. \tag{2.52}$$

It is clear that the objective of noise reduction in the STFT domain is to find optimal gains that would either minimize $J\left[H(k,m)\right]$ or minimize $J_{\mathrm{d}}\left[H(k,m)\right]$ or $J_{\mathrm{r}}\left[H(k,m)\right]$ subject to some constraint.

In the same way, we define the fullband MSE at time-frame m as

$$\begin{aligned} J\left[H(:,m)\right] &= \frac{1}{K}\sum_{k=0}^{K-1} J\left[H(k,m)\right] \\ &= \frac{1}{K}\sum_{k=0}^{K-1} J_{\mathrm{d}}\left[H(k,m)\right] + \frac{1}{K}\sum_{k=0}^{K-1} J_{\mathrm{r}}\left[H(k,m)\right] \\ &= J_{\mathrm{d}}\left[H(:,m)\right] + J_{\mathrm{r}}\left[H(:,m)\right]. \end{aligned} \tag{2.53}$$

We then deduce the fullband NMSEs at time-frame m:

$$\begin{aligned} \widetilde{J}\left[H(:,m)\right] &= K\frac{J\left[H(:,m)\right]}{\sum_{k=0}^{K-1}\phi_V(k,m)} \\ &= \mathrm{iSNR}(m)\cdot\upsilon_{\mathrm{sd}}\left[H(:,m)\right] + \frac{1}{\xi_{\mathrm{nr}}\left[H(:,m)\right]}, \end{aligned} \tag{2.54}$$

$$\begin{aligned} \overline{J}\left[H(:,m)\right] &= K\frac{J\left[H(:,m)\right]}{\sum_{k=0}^{K-1}\phi_X(k,m)} \\ &= \upsilon_{\mathrm{sd}}\left[H(:,m)\right] + \frac{1}{\mathrm{oSNR}\left[H(:,m)\right]\cdot\xi_{\mathrm{sr}}\left[H(:,m)\right]}. \end{aligned} \tag{2.55}$$

It is straightforward to see that minimizing the subband MSE at each frequency-bin k is equivalent to minimizing the fullband MSE.

2.4 Optimal Gains

In this section, we are going to derive the most important gains that can help mitigate the level of the noise picked up by the microphone.

2.4.1 Wiener

By minimizing $J\left[H(k,m)\right]$ [Eq. (2.36)] with respect to $H(k,m)$, we easily find the Wiener gain

$$\begin{aligned} H_{\mathrm{W}}(k,m) &= \frac{E\left[|X(k,m)|^2\right]}{E\left[|Y(k,m)|^2\right]} \\ &= 1 - \frac{E\left[|V(k,m)|^2\right]}{E\left[|Y(k,m)|^2\right]} \\ &= \frac{\phi_X(k,m)}{\phi_X(k,m)+\phi_V(k,m)} \\ &= \frac{\mathrm{iSNR}(k,m)}{1+\mathrm{iSNR}(k,m)}. \end{aligned} \tag{2.56}$$

We see that the noncausal Wiener gain is always real and positive. Furthermore, $0 \leq H_{\mathrm{W}}(k,m) \leq 1, \forall k, m$, and

$$\lim_{\mathrm{iSNR}(k,m)\to\infty} H_{\mathrm{W}}(k,m) = 1, \tag{2.57}$$

$$\lim_{\mathrm{iSNR}(k,m)\to 0} H_{\mathrm{W}}(k,m) = 0. \tag{2.58}$$

We deduce the different subband performance measures:

$$\widetilde{J}\left[H_{\mathrm{W}}(k,m)\right] = \frac{\mathrm{iSNR}(k,m)}{1+\mathrm{iSNR}(k,m)} \leq 1, \tag{2.59}$$

$$\begin{aligned} \xi_{\mathrm{nr}}\left[H_{\mathrm{W}}(k,m)\right] &= \left[1+\frac{1}{\mathrm{iSNR}(k,m)}\right]^2 \geq 1 \\ &= \xi_{\mathrm{sr}}\left[H_{\mathrm{W}}(k,m)\right], \end{aligned} \tag{2.60}$$

$$\upsilon_{\mathrm{sd}}\left[H_{\mathrm{W}}(k,m)\right] = \frac{1}{\left[1+\ \mathrm{iSNR}(k,m)\right]^2} \leq 1. \tag{2.61}$$

The fullband output SNR is

$$\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right] = \frac{\sum_{k=0}^{K-1}\phi_X(k,m)\left[\dfrac{\mathrm{iSNR}(k,m)}{1+\ \mathrm{iSNR}(k,m)}\right]^2}{\sum_{k=0}^{K-1}\phi_V(k,m)\left[\dfrac{\mathrm{iSNR}(k,m)}{1+\mathrm{iSNR}(k,m)}\right]^2}. \tag{2.62}$$

We observe from the previous expression that if the subband input SNR is constant across frequencies then the fullband SNR cannot be improved.

Property 2.1 *With the optimal STFT-domain Wiener gain given in* (2.56)*, the fullband output SNR is always greater than or equal to the fullband input SNR, i.e.,* $\mathrm{oSNR}\left[H(:,m)\right] \geq \mathrm{iSNR}(m)$.

Proof We can use exactly the same techniques as the ones exposed in [1] to show this property.

Property 2.2 *We have*

$$\frac{\mathrm{iSNR}(m)}{1+\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]} \leq \widetilde{J}\left[H_{\mathrm{W}}(:,m)\right] \leq \frac{\mathrm{iSNR}(m)}{1+\mathrm{iSNR}(m)}, \tag{2.63}$$

$$\begin{aligned}\frac{\{1+\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]\}^2}{\mathrm{iSNR}(m)\cdot\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]} &\leq \xi_{\mathrm{nr}}\left[H_{\mathrm{W}}(:,m)\right] \\ &\leq \frac{[1+\mathrm{iSNR}(m)]\{1+\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]\}}{\mathrm{iSNR}^2(m)},\end{aligned} \tag{2.64}$$

$$\begin{aligned}\frac{1}{\{1+\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]\}^2} &\leq \upsilon_{\mathrm{sd}}\left[H_{\mathrm{W}}(:,m)\right] \\ &\leq \frac{1+\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]-\mathrm{iSNR}(m)}{[1+\mathrm{iSNR}(m)]\{1+\mathrm{oSNR}\left[H_{\mathrm{W}}(:,m)\right]\}}.\end{aligned} \tag{2.65}$$

Proof We can use exactly the same techniques as the ones exposed in [1] to show these different inequalities.

2.4.2 Tradeoff

The tradeoff gain is obtained by minimizing the speech distortion with the constraint that the residual noise level is equal to a value smaller than the level of the original noise. This is equivalent to solving the problem

$$\min_{H(k,m)} J_{\mathrm{d}}\left[H(k,m)\right] \quad \text{subject to} \quad J_{\mathrm{r}}\left[H(k,m)\right] = \beta\phi_V(k,m), \tag{2.66}$$

where

$$J_{\mathrm{d}}\left[H(k,m)\right] = \left|H(k,m)-1\right|^2\phi_X(k,m), \tag{2.67}$$

$$J_{\mathrm{r}}\left[H(k,m)\right] = \left|H(k,m)\right|^2\phi_V(k,m), \tag{2.68}$$

and $0<\beta<1$ in order to have some noise reduction at the frequency-bin k. If we use a Lagrange multiplier, $\mu \geq 0$, to adjoin the constraint to the cost function, we get the tradeoff gain

$$\begin{aligned}H_{\mathrm{T},\mu}(k,m) &= \frac{\phi_X(k,m)}{\phi_X(k,m)+\mu\phi_V(k,m)} \\ &= \frac{\phi_Y(k,m)-\phi_V(k,m)}{\phi_Y(k,m)+(\mu-1)\phi_V(k,m)} \\ &= \frac{\mathrm{iSNR}(k,m)}{\mu+\mathrm{iSNR}(k,m)}.\end{aligned} \tag{2.69}$$

This gain can be seen as a STFT-domain Wiener gain with adjustable input noise level $\mu\phi_V(k,m)$. The particular cases of $\mu=1$ and $\mu=0$ correspond to the Wiener and distortionless gains, respectively.

The fullband output SNR is

$$\text{oSNR}\left[H_{\text{T},\mu}(:,m)\right]=\frac{\sum_{k=0}^{K-1}\phi_X(k,m)\left[\frac{\text{iSNR}(k,m)}{\mu+\text{iSNR}(k,m)}\right]^2}{\sum_{k=0}^{K-1}\phi_V(k,m)\left[\frac{\text{iSNR}(k,m)}{\mu+\text{iSNR}(k,m)}\right]^2}. \tag{2.70}$$

Property 2.3 *With the STFT-domain tradeoff gain given in* (2.69)*, the fullband output SNR is always greater than or equal to the fullband input SNR, i.e.,* $\text{oSNR}\left[H_{\text{T},\mu}(:,m)\right]\geq\text{iSNR}(m),\forall\mu\geq 0$.

Proof We can use exactly the same techniques as the ones exposed in [1] to show this property.

From (2.70), we deduce that

$$\lim_{\mu\to\infty}\text{oSNR}\left[H_{\text{T},\mu}(:,m)\right]=\frac{\sum_{k=0}^{K-1}\phi_X(k,m)\text{iSNR}^2(k,m)}{\sum_{k=0}^{K-1}\phi_V(k,m)\text{iSNR}^2(k,m)}\leq\sum_{k=0}^{K-1}\text{iSNR}(k,m). \tag{2.71}$$

This shows the trend of the fullband output SNR of the tradeoff gain.

The fullband speech distortion index is

$$\upsilon_{\text{sd}}\left[H_{\text{T},\mu}(:,m)\right]=\frac{\sum_{k=0}^{K-1}\frac{\mu^2\phi_X(k,m)}{[\mu+\text{iSNR}(k,m)]^2}}{\sum_{k=0}^{K-1}\phi_X(k,m)}. \tag{2.72}$$

Property 2.4 *The fullband speech distortion index of the STFT-domain tradeoff gain is an increasing function of the parameter* μ.

Proof It is straightforward to verify that

$$\frac{d\upsilon_{\text{sd}}\left[H_{\text{T},\mu}(:,m)\right]}{d\mu}\geq 0, \tag{2.73}$$

which ends the proof.

It is clear that

$$0\leq\upsilon_{\text{sd}}\left[H_{\text{T},\mu}(:,m)\right]\leq 1,\ \forall\mu\geq 0. \tag{2.74}$$

Therefore, as μ increases, the fullband output SNR increases at the price of more distortion to the desired signal.

The tradeoff gain can be more general if we make the factor β dependent on the frequency, i.e., $\beta(k)$. By doing so, the control between noise reduction and speech distortion can be more effective since each frequency-bin k can be controlled independently of the others. With this consideration, we can easily see that the optimal gain derived from the criterion (2.66) is now

$$H_{\mathrm{T},\mu}(k,m) = \frac{\mathrm{iSNR}(k,m)}{\mu(k) + \mathrm{iSNR}(k,m)}, \tag{2.75}$$

where $\mu(k)$ is the frequency-dependent Lagrange multiplier. This approach can now provide some noise spectral shaping for masking by the speech signal [7–12].

2.4.3 Maximum Signal-to-Noise Ratio

Let us define the $K \times 1$ vector

$$\mathbf{h}(m) = [H(0,m)\ H(1,m) \cdots H(K-1,m)]^T, \tag{2.76}$$

where the superscript ^{T} denotes transpose of a vector or a matrix. The filter $\mathbf{h}(m)$ contains all the subband gains. The fullband output SNR can be rewritten as

$$\begin{aligned} \mathrm{oSNR}\left[H(:,m)\right] &= \mathrm{oSNR}\left[\mathbf{h}(m)\right] \\ &= \frac{\mathbf{h}^H(m)\mathbf{D}_{\phi_X}(m)\mathbf{h}(m)}{\mathbf{h}^H(m)\mathbf{D}_{\phi_V}(m)\mathbf{h}(m)}, \end{aligned} \tag{2.77}$$

where the superscript ^{H} denotes transpose-conjugate and

$$\mathbf{D}_{\phi_X}(m) = \mathrm{diag}\left[\phi_X(0,m), \phi_X(1,m), \ldots, \phi_X(K-1,m)\right], \tag{2.78}$$

$$\mathbf{D}_{\phi_V}(m) = \mathrm{diag}\left[\phi_V(0,m), \phi_V(1,m), \ldots, \phi_V(K-1,m)\right], \tag{2.79}$$

are two diagonal matrices. We assume here that $\phi_V(k,m) \neq 0, \forall k, m$.

In the maximum SNR approach, we find the filter, $\mathbf{h}(m)$, that maximizes the fullband output SNR defined in (2.77). The solution to this problem that we denote by $\mathbf{h}_{\max}(m)$ is simply the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{D}_{\phi_V}^{-1}(m)\mathbf{D}_{\phi_X}(m)$. Since this matrix is diagonal, its maximum eigenvalue is its largest diagonal element, i.e.,

$$\max_k \frac{\phi_X(k,m)}{\phi_V(k,m)} = \max_k\ \mathrm{iSNR}(k,m). \tag{2.80}$$

Assume that this maximum is the k_0th diagonal element of the matrix $\mathbf{D}_{\phi_V}^{-1}(m)\mathbf{D}_{\phi_X}(m)$. In this case, the k_0th component of $\mathbf{h}_{\max}(m)$ is 1 and all its other components are 0. As a result,

$$\begin{aligned} \text{oSNR}\left[\mathbf{h}_{\max}(m)\right] &= \max_k \ \text{iSNR}(k,m) \\ &= \text{iSNR}(k_0,m). \end{aligned} \tag{2.81}$$

We also deduce that

$$\text{oSNR}\left[\mathbf{h}(m)\right] \leq \max_k \ \text{iSNR}(k,m), \ \ \forall \mathbf{h}(m). \tag{2.82}$$

This means that with the Wiener, tradeoff, or any other gain, the fullband output SNR cannot exceed the maximum subband input SNR, which is a very interesting result on its own.

It is easy to derive the fullband speech distortion index:

$$\upsilon_{\text{sd}}\left[\mathbf{h}_{\max}(m)\right] = 1 - \frac{\phi_X(k_0,m)}{\sum_{k=0}^{K-1}\phi_X(k,m)}, \tag{2.83}$$

which can be very close to 1, implying very large distortions of the desired signal.

Needless to say that this maximum SNR filter is never used in practice since all subband signals but one are suppressed. But this filter is still interesting from a theoretical point of view.

References

1. J. Benesty, J. Chen, Y. Huang, I. Cohen, *Noise Reduction in Speech Processing* (Springer, Berlin, 2009)
2. P. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007)
3. P. Vary, R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment* (Wiley, Chichester, 2006)
4. J. Benesty, J. Chen, Y. Huang, S. Doclo, Study of the Wiener filter for noise reduction, in *Speech Enhancement*, ed. by J. Benesty, S. Makino, J. Chen (Springer, Berlin, 2005) pp. 9–41
5. J. Chen, J. Benesty, Y. Huang, S. Doclo, New insights into the noise reduction Wiener filter. IEEE Trans. Audio Speech Lang. Process **14**, 1218–1234 (2006)
6. S. Haykin, *Adaptive Filter Theory*. 4th edn. (Prentice-Hall, Upper Saddle River, 2002)
7. Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process **3**, 251–266 (1995)
8. N. Virag, Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process **7**, 126–137 (1999)
9. R. Vetter, Single channel speech enhancement using MDL-based subspace approach in Bark domain, in *Proceedings IEEE ICASSP*, pp. 641–644 (2001)
10. Y. Hu, P.C. Loizou, A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process **11**, 334–341 (2003)
11. Y. Hu, P.C. Loizou, A perceptually motivated approach for speech enhancement. IEEE Trans. Speech Audio Process **11**, 457–465 (2003)
12. F. Jabloun, B. Champagne, Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process **11**, 700–708 (2003)