44. Spectral Enhancement Methods

I. Cohen, S. Gannot

In this chapter, we focus on the statistical methods that constitute a speech spectral enhancement system and describe some of their fundamental components. We begin in Sect. 44.2 by formulating the problem of spectral enhancement. In Sect. 44.3, we address the time–frequency correlation of spectral coefficients for speech and noise signals, and present statistical models that conform with these characteristics. In Sect. 44.4, we present estimators for speech spectral coefficients under speech presence uncertainty based on various fidelity criteria. In Sect. 44.5, we address the problem of speech presence probability estimation. In Sect. 44.6, we present useful estimators for the a priori signal-to-noise ratio (SNR) under speech presence uncertainty. We present the decision-directed approach, which is heuristically motivated, and the recursive estimation approach, which is based on statistical models and follows the rationale of Kalman filtering. In Sect. 44.7, we describe the improved minima-controlled recursive averaging (IMCRA) approach for noise power spectrum estimation. In Sect. 44.8, we provide a detailed example of a speech enhancement algorithm, and demonstrate its performance in environments with various noise types. In Sect. 44.9, we survey the main types of spectral enhancement components, and discuss the significance of the choice of statistical model, fidelity criterion, a priori SNR estimator, and noise spectrum estimator. Some concluding comments are made in Sect. 44.10.

The problem of spectral enhancement of noisy speech signals from a single microphone has attracted considerable research effort for over 30 years. It is a problem with numerous applications ranging from speech recognition, to hearing aids and hands-free mobile communication. In this chapter, we present the fundamental components that constitute a speech spectral enhancement system. We describe statistical models that take into consideration the time correlation between successive spectral components of the speech signal, and present estimators for the speech spectral coefficients based on various fidelity criteria. We address the problem of a priori SNR estimation under speech presence uncertainty, and noise power spectrum estimation. We also provide a detailed design example of a speech enhancement algorithm.

44.1 Spectral Enhancement

Spectral enhancement of noisy speech has been a challenging problem for many researchers for over 30 years, and is still an active research area (see e.g., [44.1–3] and references therein). This problem is often formulated as the estimation of speech spectral components from a speech signal degraded by statistically independent additive noise. In this chapter we consider spectral enhancement methods for single-channel set-ups, assuming that only one-microphone noisy output is available for the estimation. The situation of one-microphone setups is particularly difficult under nonstationary noise and a low signal-to-noise ratio (SNR), since no reference signal is available for the estimation of the background noise.

A variety of different approaches for spectral enhancement of noisy speech signals have been introduced over the years. One of the earlier methods, and perhaps the most well-known approach, is spectral subtraction [44.4, 5], in which an estimate of the short-term power spectral density of the clean signal is obtained by subtracting an estimate of the power spectral density of the background noise from the short-term power spectral density of the degraded signal. The square root of the resulting estimate is considered an estimate of the spectral magnitude of the speech signal. Subsequently, an estimate of the signal is obtained by combining the spectral magnitude estimate with the complex exponential of the phase of the noisy signal. This method generally results in random narrowband fluctuations in the residual noise, also known as musical tones, which is annoying and disturbing to the perception of the enhanced signal. Many variations have been developed to cope with the musical residual noise phenomena [44.4, 6–9], including spectral subtraction techniques based on masking properties of the human auditory system [44.10, 11].

The spectral subtraction method makes minimal assumptions about the signal and noise, and when carefully implemented, produces enhanced signals that may be acceptable for certain applications. Statistical methods [44.12–16] are designed to minimize the expected value of some distortion measure between the clean and estimated signals. This approach requires the presumption of reliable statistical models for the speech and noise signals, the specification of a perceptually meaningful distortion measure, and a mathematically tractable derivation of an efficient signal estimator. A statistical speech model and perceptually meaningful distortion measure, which are the most appropriate for spectral enhancement, have not yet been determined. Hence, statistical methods for spectral enhancement mainly differ in their statistical model [44.12, 14, 15], distortion measure [44.17–19], and the particular implementation of the spectral enhancement algorithm [44.2].

Spectral enhancement based on hidden Markov processes (HMPs) try to circumvent the assumption of specific distributions for the speech and noise processes [44.20–23]. The probability distributions of the two processes are first estimated from long training sequences of clean speech and noise samples, and then used jointly with a given distortion measure to derive an estimator for the speech signal. Normally, vectors generated from a given sequence of states are assumed to be statistically independent. However, the HMP can be extended to take into account the time–frequency correlation of speech signals by using nondiagonal covariance matrices for each subsource, and by assuming that a sequence of vectors generated from a given sequence of states is a nonzero-order autoregressive process [44.21, 24]. HMP-based speech enhancement relies on the type of training data [44.25]. It works best with the type of noise used during training, and often worse with other types of noise. Furthermore, improved performance generally entails more-complex models and greater computational requirements. While hidden Markov models have been successfully applied to automatic recognition of clean speech signals [44.26, 27], they were not found to be sufficiently refined models for speech enhancement applications [44.3].

Subspace methods [44.28–31] attempt to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace. Spectral enhancement is performed by removing the noise subspace and estimating the speech signal from the remaining subspace. The signal subspace decomposition can be achieved by either using the Karhunen–Loève transform (KLT) via eigenvalue decomposition of a Toeplitz covariance estimate of the noisy vector [44.28, 30], or by using the singular value decomposition of a data matrix [44.32, 33]. Linear estimation in the signal-plus-noise subspace is performed with the goal of minimizing signal distortion while masking the residual noise by the signal. A perceptually motivated signal subspace approach takes into account the masking properties of the human auditory system and reduces the perceptual effect of the residual noise [44.34, 35].

44.2 Problem Formulation

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. The observed signal $y(n)$, given by $y(n) = x(n) + d(n)$, is transformed into the time-frequency domain by applying the short-time Fourier transform (STFT). Specifically,

$$Y_{tk} = \sum_{n=0}^{N-1} y(n+tM)h(n)\mathrm{e}^{-\mathrm{i}\frac{2\pi}{N}nk}\,, \tag{44.1}$$

where t is the time frame index ($t = 0, 1, \ldots, k$) is the frequency bin index ($k = 0, 1, \ldots, N-1$), $h(n)$ is an analysis window of size N (e.g., Hamming window), and M is the framing step (number of samples separating two successive frames). Given an estimate $\hat{X}_{tk}$ for the STFT of the clean speech (Fig. 44.1), an estimate for the clean speech signal is obtained by applying the inverse STFT,

$$\hat{x}(n) = \sum_{t}\sum_{k=0}^{N-1} \hat{X}_{tk}\tilde{h}(n-tM)\mathrm{e}^{\mathrm{i}\frac{2\pi}{N}k(n-tM)}\,, \tag{44.2}$$

where $\tilde{h}(n)$ is a synthesis window that is biorthogonal to the analysis window $h(n)$ [44.36], and the inverse STFT is efficiently implemented by using the weighted overlap-add method [44.37] (see also Sect. 44.8).

The spectral enhancement problem is generally formulated as deriving an estimator $\hat{X}_{tk}$ for the speech spectral coefficients, such that the expected value of a certain distortion measure is minimized. Let $d(X_{tk}, \hat{X}_{tk})$ denote a distortion measure between X_{tk} and its estimate $\hat{X}_{tk}$, and let ψ_t represent the information set that can be employed for the estimation at frame t (e.g., the noisy data observed through time t). Let H_1^{tk} and H_0^{tk} denote, respectively, hypotheses of signal presence and absence in the noisy spectral coefficient Y_{tk}:

$$\begin{aligned} H_1^{tk}: &\quad Y_{tk} = X_{tk} + D_{tk} \\ H_0^{tk}: &\quad Y_{tk} = D_{tk}\,. \end{aligned}$$

Let $\hat{p}_{tk} = P(H_1^{tk}|\psi_t)$ denote an estimate for the signal presence probability, $\hat{\lambda}_{tk} = E\{|X_{tk}|^2|H_1^{tk}, \psi_t\}$ denote an estimate for the variance of a speech spectral coefficient X_{tk} under H_1^{tk}, and $\widehat{\sigma_{tk}^2} = E\{|Y_{tk}|^2|H_0^{tk}, \psi_t\}$ denote an estimate for the variance of a noise spectral coefficient D_{tk}. Then, we consider an estimator for X_{tk} which minimizes the expected distortion given $\hat{p}_{tk}$, $\hat{\lambda}_{tk}$, $\widehat{\sigma_{tk}^2}$ and the noisy spectral coefficient Y_{tk}:

$$\min_{\hat{X}_{tk}} E\left\{d\left(X_{tk}, \hat{X}_{tk}\right)\middle|\hat{p}_{tk}, \hat{\lambda}_{tk}, \widehat{\sigma_{tk}^2}, Y_{tk}\right\}\,. \tag{44.3}$$

In particular, restricting ourselves to a squared error distortion measure of the form

$$d\left(X_{tk}, \hat{X}_{tk}\right) = \left|g(\hat{X}_{tk}) - \tilde{g}(X_{tk})\right|^2\,, \tag{44.4}$$

where $g(X)$ and $\tilde{g}(X)$ are specific functions of X (e.g., X, $|X|$, $\log|X|$, $\mathrm{e}^{\mathrm{i}\angle X}$), the estimator $\hat{X}_{tk}$ is calculated from

$$\begin{aligned} g(\hat{X}_{tk}) &= E\left\{\tilde{g}(X_{tk})\middle|\hat{p}_{tk}, \hat{\lambda}_{tk}, \widehat{\sigma_{tk}^2}, Y_{tk}\right\} \\ &= \hat{p}_{tk}E\left\{\tilde{g}(X_{tk})\middle|H_1^{tk}, \hat{\lambda}_{tk}, \widehat{\sigma_{tk}^2}, Y_{tk}\right\} \\ &\quad + (1-\hat{p}_{tk})E\left\{\tilde{g}(X_{tk})\middle|H_0^{tk}, Y_{tk}\right\}\,. \end{aligned} \tag{44.5}$$

Hence, the design of a particular estimator for X_{tk} requires the following specifications:

- Functions $g(X)$ and $\tilde{g}(X)$, which determine the fidelity criterion of the estimator
- A conditional probability density function (pdf) $p(X_{tk}|\lambda_{tk}, H_1^{tk})$ for X_{tk} under H_1^{tk} given its variance λ_{tk}, which determines the statistical model
- An estimator $\hat{\lambda}_{tk}$ for the speech spectral variance
- An estimator σ_{tk}^2 for the noise spectral variance
- An estimator $\hat{p}_{tk|t-1} = P(H_1^{tk}|\psi_{t-1})$ for the a priori signal presence probability, where ψ_{t-1} represents the information set known prior to having the measurement Y_{tk}

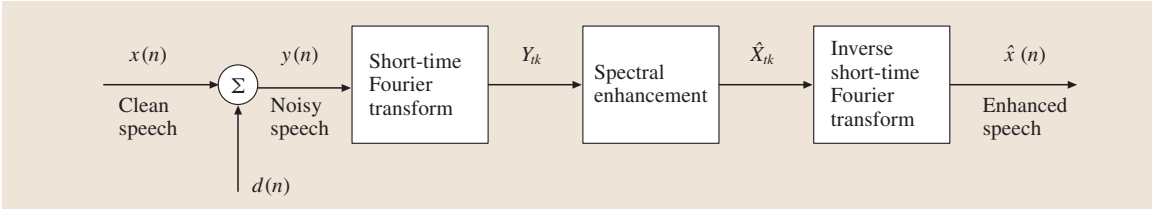

Fig. 44.1 Spectral enhancement approach

Given the a priori signal presence probability $\hat{p}_{tk|t-1}$, the (a posteriori) signal presence probability can be obtained from Bayes' rule:

$$\hat{p}_{tk} = P\left(H_1^{tk}\middle|Y_{tk}, \psi_{t-1}\right) = \left[1 + \frac{\left(1-\hat{p}_{tk|t-1}\right)p\left(Y_{tk}|H_0^{tk}, \psi_{t-1}\right)}{\hat{p}_{tk|t-1}\,p\left(Y_{tk}|H_1^{tk}, \psi_{t-1}\right)}\right]^{-1} . \quad (44.6)$$

In the following sections we present statistical models for speech signals in the STFT domain, and address the estimation problem of the speech spectral coefficient X_{tk} given $\hat{\lambda}_{tk}$, $\widehat{\sigma_{tk}^2}$, and $\hat{p}_{tk}$. Then we consider the estimation of the speech spectral variance λ_{tk}, the noise spectral variance σ_{tk}^2, and the speech presence probability $P(H_1^{tk})$, and describe an example of a speech enhancement algorithm.

44.3 Statistical Models

In this section, we present statistical models that take into account the time correlation between successive spectral components of the speech signal. To see graphically the relation between successive spectral components of a speech signal, in comparison with a noise signal, we present scatter plots for successive spectral magnitudes, and investigate the sample autocorrelation coefficient sequences (ACSs) of the STFT coefficients along time trajectories (the frequency bin index k is held fixed). We consider a speech signal that is constructed from six different utterances, without intervening pauses. The utterances, half from male speakers and half from female speakers, are taken from the TIMIT database [44.38]. (A corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. The speech was recorded at Texas Instruments (TI) and transcribed at Massachusetts Institute of Technology (MIT), hence the corpus' name.) The speech signal is sampled at 16 kHz, and transformed into the STFT domain using Hamming analysis windows of 512 samples (32 ms) length, and 256 samples framing step (50% overlap between successive frames).

Figure 44.2 shows an example of scatter plots for successive spectral magnitudes of white Gaussian noise (WGN) and speech signals. It implies that 50% overlap between successive frames does not yield a significant correlation between the spectral magnitudes of the WGN signal. However, successive spectral magnitudes of the speech signal are highly correlated. Figure 44.3 shows the ACSs of the speech spectral components along time trajectories, for various frequency bins and framing steps. The 95% confidence limits [44.39] are depicted as horizontal dotted lines. In order to prevent an upward bias of the autocovariance estimates due to irrelevant (nonspeech) spectral components, the ACSs are computed from spectral components whose magnitudes are within 30 dB of the maximal magnitude. Specifically, the sample autocorrelation coefficients of the spectral magnitudes are calculated by

$$\rho_m = \frac{\sum_{t\in\mathcal{T}}\left(A_{tk}-\overline{A_k}\right)\left(A_{t+m,k}-\overline{A_k}\right)}{\sum_{t\in\mathcal{T}}\left(A_{tk}-\overline{A_k}\right)^2} , \quad (44.7)$$

where $A_{tk} \triangleq |X_{tk}|$ denotes the magnitude of X_{tk},

$$\overline{A_k} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} A_{tk}$$

denotes the sample mean, m is the lag in frames, and $\mathcal{T}$ represents the set of relevant spectral components

$$\mathcal{T} = \left\{t \,\middle|\, A_{tk} \geq 10^{-30/20}\max_t\{A_{tk}\}\right\} .$$

The corresponding sample autocorrelation coefficients of the spectral phases are obtained by

$$\varrho_m = \frac{\sum_{t\in\mathcal{T}}\varphi_{tk}\varphi_{t+m,k}}{\sum_{t\in\mathcal{T}}\varphi_{tk}^2} , \quad (44.8)$$

where φ_{tk} denotes the phase of X_{tk}. Figure 44.4 shows the variation of the correlation between successive spectral magnitudes on frequency and on overlap between successive frames. Figures 44.3 and 44.4 demonstrate that, for speech signals, successive spectral magnitudes are highly correlated, while the correlation is generally larger at lower frequencies, and increases as the overlap between successive frames increases. As a comparison, the variation of ρ_1 on the overlap between frames is also shown for a realization of WGN (Fig. 44.4b, dotted line). It implies that, for a sufficiently large framing step ($M \geq N/2$, i. e., overlap

between frames $\leq 50\%$), successive spectral components of the *noise* signal, but clearly not of the *speech* signal, can be assumed uncorrelated. For smaller framing steps, the correlation between successive spectral noise components also has to be taken into consideration. Furthermore, since the length of the analysis window cannot be too large (its typical length is 20–40 ms [44.12]), adjacent Fourier expansion coefficients of the noise signal, D_{tk} and $D_{t,k+1}$, as well as adjacent coefficients of the speech signal, X_{tk} and $X_{t,k+1}$, are also correlated to a certain degree. Nevertheless, it is commonly assumed that expansion coefficients in different frequency bins are statistically independent [44.12, 15, 16, 40]. This allows one to formulate independent estimation problems for each frequency bin k, which greatly simplifies the resulting algorithms. In view of this discussion, we employ statistical models in the STFT domain that rely on the following set of assumptions [44.41].

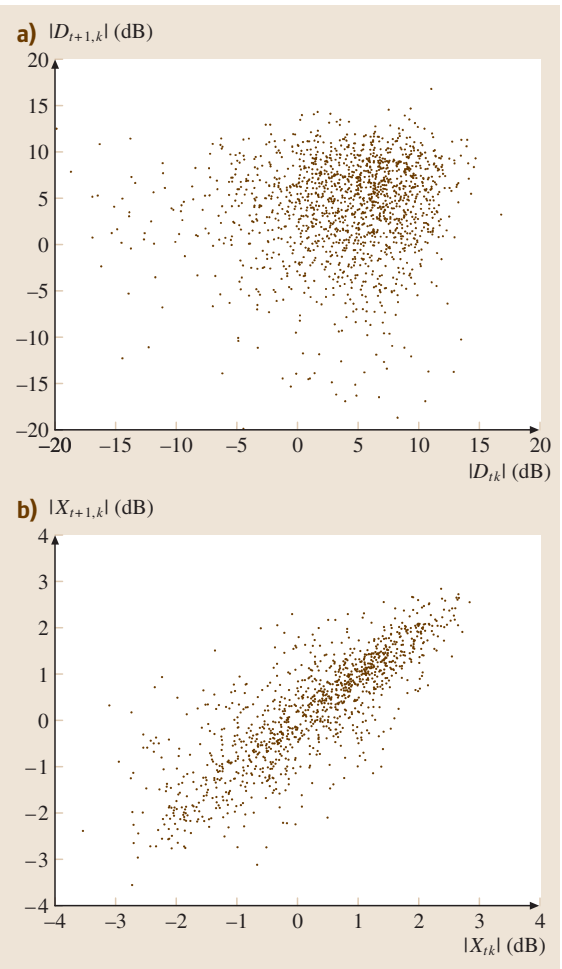

Fig. 44.2a,b Scatter plots for successive spectral magnitudes of **(a)** a white Gaussian noise signal, and **(b)** a speech signal at a center frequency of 500 Hz ($k = 17$). The overlap between successive frames is 50% (after [44.15])

1. The noise spectral coefficients $\{D_{tk}\}$ are zero-mean statistically independent Gaussian random variables. The real and imaginary parts of D_{tk} are independent and identically distributed (iid) random variables $\mathcal{N}(0, \sigma_{tk}^2/2)$.
2. Given $\{\lambda_{tk}\}$ and the state of speech presence in each time–frequency bin (H_1^{tk} or H_0^{tk}), the speech spectral coefficients $\{X_{tk}\}$ are generated by

$$X_{tk} = \sqrt{\lambda_{tk}}\, V_{tk} \,, \tag{44.9}$$

where $\{V_{tk}|H_0^{tk}\}$ are identically zero, and $\{V_{tk}|H_1^{tk}\}$ are statistically independent complex random variables with zero mean, unit variance, and iid real and imaginary parts:

$$\begin{aligned} H_1^{tk}: &\quad E\{V_{tk}\} = 0,\, E\left\{|V_{tk}|^2\right\} = 1 \,, \\ H_0^{tk}: &\quad V_{tk} = 0 \,. \end{aligned} \tag{44.10}$$

3. The pdf of V_{tk} under H_1^{tk} is determined by the specific statistical model. Let $V_{\mathrm{R}tk} = \mathrm{Re}\{V_{tk}\}$ and $V_{\mathrm{I}tk} = \mathrm{Im}\{V_{tk}\}$ denote, respectively, the real and imaginary parts of V_{tk}. Let $p(V_{\rho tk}|H_1^{tk})$ denote the pdf of $V_{\rho tk}$ ($\rho \in \{\mathrm{R}, \mathrm{I}\}$) under H_1^{tk}. Then, for a Gaussian model

$$p\left(V_{\rho tk}|H_1^{tk}\right) = \frac{1}{\sqrt{\pi}} \exp\left(-V_{\rho tk}^2\right) , \tag{44.11}$$

for a gamma model

$$\begin{aligned} &p\left(V_{\rho tk}|H_1^{tk}\right) \\ &= \frac{\sqrt[4]{3}}{2\sqrt{\pi}\sqrt[4]{2}} |V_{\rho tk}|^{-1/2} \exp\left(-\frac{\sqrt{3}|V_{\rho tk}|}{\sqrt{2}}\right) , \end{aligned} \tag{44.12}$$

and for a Laplacian model

$$p\left(V_{\rho tk}|H_1^{tk}\right) = \exp(-2|V_{\rho tk}|) \,. \tag{44.13}$$

4. The sequence of speech spectral variances $\{\lambda_{tk}|t = 0, 1, \ldots\}$ is a random process, which is generally correlated with the sequence of speech spectral magnitudes $\{A_{tk}|t = 0, 1, \ldots\}$. However, given λ_{tk}, A_{tk} is statistically independent of $A_{t'k'}$ for all $t \neq t'$ and $k \neq k'$.

Clearly, the first assumption does not hold when the overlap between successive frames is too large (Fig. 44.4b,

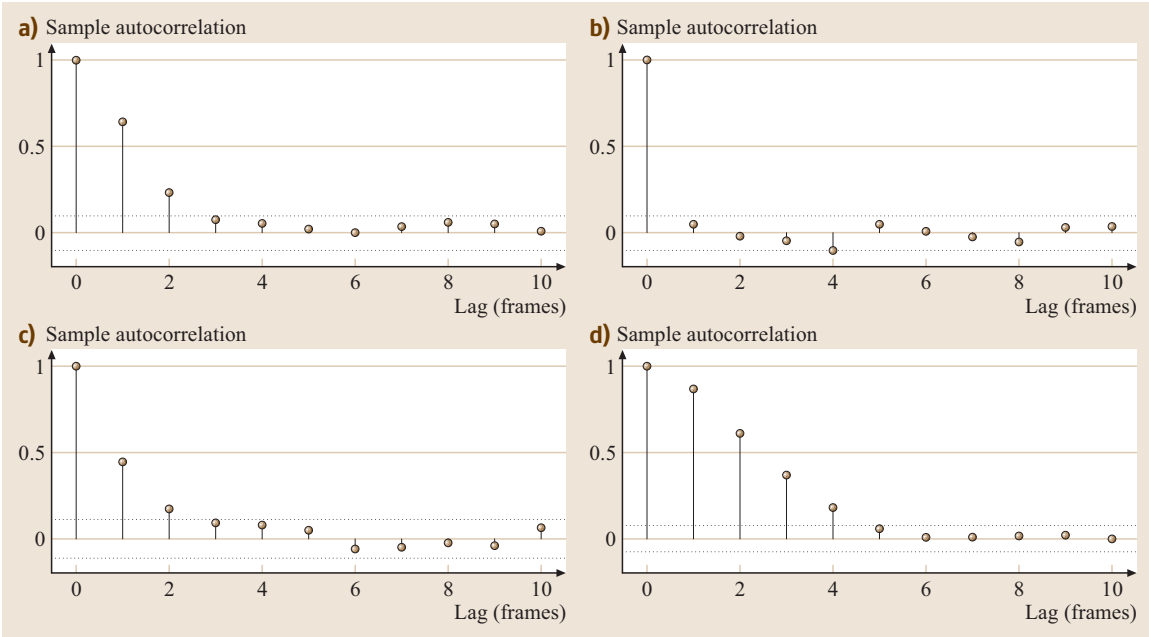

Fig. 44.3a–d Sample autocorrelation coefficient sequences (ACSs) of clean-speech STFT coefficients along time trajectories, for various frequency bins and framing steps. The *dotted lines* represent 95% confidence limits. **(a)** ACS of the spectral magnitude at frequency bin $k = 17$ (center frequency 500 Hz), framing step $M = N/2$ (50% overlap between frames). **(b)** ACS of the spectral phase, $k = 17$, $M = N/2$. **(c)** ACS of the spectral magnitude, $k = 65$ (center frequency 2 kHz), $M = N/2$. **(d)** ACS of the spectral magnitude, $k = 17$, $M = N/4$ (75% overlap between frames) (after [44.15])

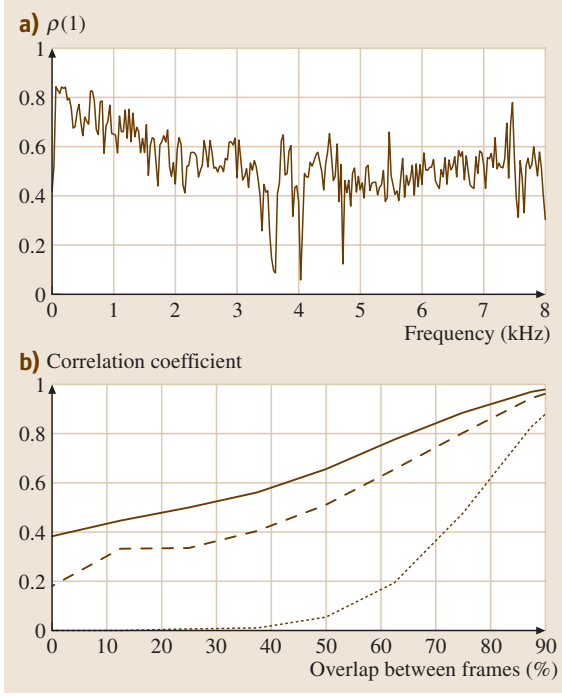

Fig. 44.4a,b Variation of the correlation coefficient between successive spectral magnitudes. **(a)** Typical variation of ρ_1 with frequency for a speech signal and 50% overlap between frames. **(b)** Typical variation of ρ_1 with overlap between frames for a speech signal at center frequencies of 1 kHz (*solid line*) and 2 kHz (*dashed line*), and for a realization of white Gaussian noise (*dotted line*) (after [44.15])

implemented in accordance with this assumption (e.g., the overlap is not greater than 50%). The second assumption implies that the speech spectral coefficients $\{X_{tk}|H_1^{tk}\}$ are conditionally zero-mean statistically independent random variables given their variances $\{\lambda_{tk}\}$. The real and imaginary parts of X_t under H_1^t are conditionally iid random variables given λ_{tk}, satisfying

$$p\left(X_{\rho tk}|\lambda_{tk}, H_1^{tk}\right) = \frac{1}{\sqrt{\lambda_{tk}}} p\left(V_{\rho tk} = \frac{X_{\rho tk}}{\sqrt{\lambda_{tk}}}\middle| H_1^{tk}\right), \quad (44.14)$$

where $\rho \in \{\mathrm{R}, \mathrm{I}\}$. The last assumption allows one to take into account the time correlation between

successive spectral coefficients of the speech signal, while still considering the scalar estimation problem formulated in (44.3). Note that successive spectral coefficients are correlated, since the random processes $\{X_{tk}|t=0,1,\dots\}$ and $\{\lambda_{tk}|t=0,1,\dots\}$ are not independent.

44.4 Signal Estimation

In this section, we derive estimators for X_{tk} using various fidelity criteria, assuming that $\hat{p}_{tk}$, $\hat{\lambda}_{tk}$, and $\widehat{\sigma_{tk}^2}$ are given. Fidelity criteria that are of particular interest for speech enhancement applications are minimum mean-squared error (MMSE) [44.5], MMSE of the spectral amplitude (MMSE-SA) [44.12], and MMSE of the log-spectral amplitude (MMSE-LSA) [44.17,42]. The MMSE estimator is derived by substituting into (44.5) the functions

$$g(\hat{X}_{tk}) = \hat{X}_{tk}$$
$$\tilde{g}(X_{tk}) = \begin{cases} X_{tk}\,, & \text{under } H_1^{tk} \\ G_{\min} Y_{tk}\,, & \text{under } H_0^{tk}\,, \end{cases} \tag{44.15}$$

where $G_{\min} \ll 1$ represents a constant attenuation factor, which retains the noise naturalness during speech absence [44.2,42].

The MMSE-SA estimator is obtained by using the functions

$$g(\hat{X}_{tk}) = |\hat{X}_{tk}|\,,$$
$$\tilde{g}(X_{tk}) = \begin{cases} |X_{tk}|\,, & \text{under } H_1^{tk} \\ G_{\min}|Y_{tk}|\,, & \text{under } H_0^{tk}\,. \end{cases} \tag{44.16}$$

The MMSE-LSA estimator is obtained by using the functions

$$g(\hat{X}_{tk}) = \log|\hat{X}_{tk}|\,,$$
$$\tilde{g}(X_{tk}) = \begin{cases} \log|X_{tk}|\,, & \text{under } H_1^{tk} \\ \log(G_{\min}|Y_{tk}|)\,, & \text{under } H_0^{tk}\,. \end{cases} \tag{44.17}$$

The last two estimators are insensitive to the estimation error of φ_{tk}, the phase of X_{tk}. Therefore, they are combined with the following constrained optimization problem [44.12]:

$$\min_{\hat{\varphi}_{tk}} E\{|\mathrm{e}^{\mathrm{i}\varphi_{tk}} - \mathrm{e}^{\mathrm{i}\hat{\varphi}_{tk}}|^2\} \quad \text{subject to } |\mathrm{e}^{\mathrm{i}\hat{\varphi}_{tk}}| = 1\,. \tag{44.18}$$

This yields an estimator for the complex exponential of the phase, constrained not to affect the spectral magnitude estimate. Alternatively, an estimate for the spectral phase $\hat{\varphi}_{tk}$ can be obtained by minimizing the expected value of the following distortion measure [44.12]

$$d_{\varphi}(\varphi_{tk}, \hat{\varphi}_{tk}) \triangleq 1 - \cos(\varphi_{tk} - \hat{\varphi}_{tk})\,. \tag{44.19}$$

This measure is invariant under modulo 2π transformation of the estimation error $\varphi_{tk} - \hat{\varphi}_{tk}$, and for small estimation errors it closely resembles the squared-error distortion measure, since $1 - \cos\beta \approx \beta^2/2$ for $\beta \ll 1$. The constrained optimization problem (44.18) and the distortion measure (44.19) both yield an estimator $\mathrm{e}^{\mathrm{i}\hat{\varphi}_{tk}} = Y_{tk}/|Y_{tk}|$, which is simply the complex exponential of the noisy signal [44.12].

44.4.1 MMSE Spectral Estimation

Let

$$\xi_{tk} \triangleq \frac{\lambda_{tk}}{\sigma_{tk}^2}\,, \quad \gamma_{\rho tk} \triangleq \frac{Y_{\rho tk}^2}{\sigma_{tk}^2}\,, \tag{44.20}$$

represent the a priori and a posteriori SNRs, respectively ($\rho \in \{\mathrm{R}, \mathrm{I}\}$), and let $G_{\mathrm{MSE}}(\xi, \gamma\rho)$ denote a gain function that satisfies

$$E\{X_{\rho tk}|H_1^{tk}, \lambda_{tk}, \sigma_{tk}^2, Y_{\rho tk}\} = G_{\mathrm{MSE}}(\xi_{tk}, \gamma_{\rho tk}) Y_{\rho tk}\,. \tag{44.21}$$

Then, substituting (44.15) and (44.21) into (44.5), we have

$$\hat{X}_{tk} = \hat{p}_{tk}\big[G_{\mathrm{MSE}}(\hat{\xi}_{tk}, \hat{\gamma}_{\mathrm{R}tk}) Y_{\mathrm{R}tk} + \mathrm{i} G_{\mathrm{MSE}}(\hat{\xi}_{tk}, \hat{\gamma}_{\mathrm{I}tk}) Y_{\mathrm{I}tk}\big] + (1 - \hat{p}_{tk}) G_{\min} Y_{tk}\,. \tag{44.22}$$

The specific expression for $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ depends on the particular statistical model:

$$G_{\mathrm{MSE}}(\xi, \gamma_\rho) = \frac{1}{Y_\rho} \int X_\rho p(X_\rho|H_1, \lambda, \sigma^2, Y_\rho)\,\mathrm{d}X_\rho = \frac{1}{Y_\rho} \int X_\rho \frac{p(Y_\rho|X_\rho, \sigma^2)\, p(X_\rho|\lambda, H_1)}{p(Y_\rho|\lambda, \sigma^2)}\,\mathrm{d}X_\rho\,.$$

For a Gaussian model, the gain function is independent of the a posteriori SNR. It is often referred to as Wiener

filter, given by [44.5]

$$G_{\mathrm{MSE}}(\xi) = \frac{\xi}{1+\xi} \,. \tag{44.23}$$

For a gamma model, the gain function is given by [44.40]

$$\begin{aligned} G_{\mathrm{MSE}}(\xi, \gamma_\rho) = \frac{1}{\sqrt{8\gamma_\rho}} & \left[\exp\left(\frac{C_{\rho-}^2}{4}\right) D_{-1.5}(C_{\rho-}) \right. \\ & \left. - \exp\left(\frac{C_{\rho+}^2}{4}\right) D_{-1.5}(C_{\rho+}) \right] \\ & \times \left[\exp\left(\frac{C_{\rho-}^2}{4}\right) D_{-0.5}(C_{\rho-}) \right. \\ & \left. + \exp\left(\frac{C_{\rho+}^2}{4}\right) D_{-0.5}(C_{\rho+}) \right]^{-1} , \end{aligned} \tag{44.24}$$

where $C_{\rho+}$ and $C_{\rho-}$ are defined by

$$C_{\rho\pm} \triangleq \frac{\sqrt{3}}{2\sqrt{\xi}} \pm \sqrt{2\gamma_\rho} \,, \tag{44.25}$$

and $D_p(z)$ denotes the parabolic cylinder function [44.44, (9.240)]. For a Laplacian speech model, the gain function is given by [44.45]

$$\begin{aligned} & G_{\mathrm{MSE}}(\xi, \gamma_\rho) \\ & = \frac{1}{\sqrt{\gamma_\rho}} [L_{\rho+}\mathrm{erfcx}(L_{\rho+}) - L_{\rho-}\mathrm{erfcx}(L_{\rho-})] \\ & \quad \times [\mathrm{erfcx}(L_{\rho+}) + \mathrm{erfcx}(L_{\rho-})]^{-1} \,, \end{aligned} \tag{44.26}$$

where $L_{\rho+}$ and $L_{\rho-}$ are defined by

$$L_{\rho\pm} \triangleq \frac{1}{\sqrt{\xi}} \pm \sqrt{\gamma_\rho} \,, \tag{44.27}$$

and $\mathrm{erfcx}(x)$ is the scaled complementary error function, defined by

$$\mathrm{erfcx}(x) \triangleq \mathrm{e}^{x^2} \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-t^2} \, \mathrm{d}t \,. \tag{44.28}$$

Note that when the signal is surely absent (i. e., when $\hat{p}_{tk} = 0$), the resulting estimator $\hat{X}_{tk}$ reduces to a constant attenuation of Y_{tk} (i. e., $\hat{X}_{tk} = G_{\min} Y_{tk}$). This retains the noise naturalness, and is closely related to the *spectral floor* proposed by *Berouti* et al. [44.6].

Figure 44.5 displays parametric gain curves describing $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ for several values of γ_ρ, which result from (44.23), (44.24), and (44.26). It shows that the spectral gains are monotonically increasing functions of the a priori SNR when the a posteriori SNR is kept constant. For gamma and Laplacian models, the spectral gains are also monotonically increasing functions of the a posteriori SNR, when the a priori SNR is kept constant.

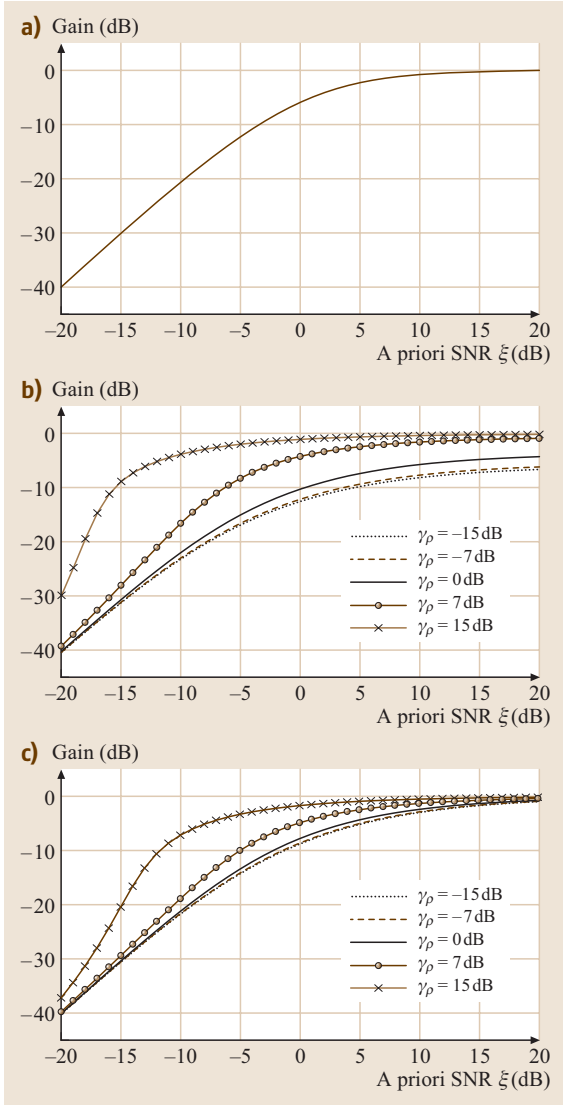

Fig. 44.5a–c Parametric gain curves describing the MMSE gain function $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ for different speech models. **(a)** Gain curve for a Gaussian model, obtained by (44.23). **(b)** Gain curves for a gamma model, obtained by (44.24). **(c)** Gain curves for a Laplacian model, obtained by (44.26) (after [44.43])

44.4.2 MMSE Log-Spectral Amplitude Estimation

In speech enhancement applications, estimators that minimize the mean-squared error of the log-spectral amplitude have been found advantageous to MMSE spectral estimators [44.12, 17, 46]. An MMSE-LSA estimator is obtained by substituting (44.17) into (44.5). It is difficult, or even impossible, to find analytical expressions for an MMSE-LSA estimator under a gamma or Laplacian model. However, assuming a Gaussian model and combing the resulting amplitude estimate with the phase of the noisy spectral coefficient Y_{tk} yields [44.42]

$$\hat{X}_{tk} = \left[G_{\mathrm{LSA}}(\hat{\xi}_{tk}, \hat{\gamma}_{tk})\right]^{\hat{p}_{tk}} G_{\min}^{1-\hat{p}_{tk}} Y_{tk} \,, \tag{44.29}$$

where $\hat{\gamma}_{tk}$ denotes an estimate for the a posteriori SNR

$$\hat{\gamma}_{tk} = \hat{\gamma}_{\mathrm{R}tk} + \hat{\gamma}_{\mathrm{I}tk} \,, \tag{44.30}$$

and $G_{\mathrm{LSA}}(\xi, \gamma)$ represents the LSA gain function under H_1^{tk} which was derived by *Ephraim* and *Malah* [44.17]

$$G_{\mathrm{LSA}}(\xi, \gamma) \triangleq \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_{\vartheta}^{\infty} \frac{\mathrm{e}^{-x}}{x}\,\mathrm{d}x\right) \,, \tag{44.31}$$

where ϑ is defined by $\vartheta \triangleq \xi\gamma/(1+\xi)$. Similar to the MMSE spectral estimator, the MMSE-LSA estimator reduces to a constant attenuation of Y_{tk} when the signal is surely absent (i. e., $\hat{p}_{tk} = 0$ implies $\hat{X}_{tk} = G_{\min} Y_{tk}$). However, the characteristics of these estimators when the signal is present are readily distinctive. Figure 44.6 displays parametric gain curves describing $G_{\mathrm{LSA}}(\xi, \gamma)$ for several values of γ. For a fixed value of the a posteriori SNR, the LSA gain is a monotonically increasing function of ξ. However, for a fixed value of ξ, the LSA gain is a monotonically *decreasing* function of γ. Note that the gain function $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ for a Gaussian model is independent of the a posteriori SNR, while for gamma and Laplacian speech models $G_{\mathrm{MSE}}(\xi, \gamma_\rho)$ is an *increasing* function of the a posteriori SNR (Fig. 44.5). The behavior of $G_{\mathrm{LSA}}(\xi, \gamma)$ is related to the useful mechanism that counters the musical noise phenomenon [44.47]. Local bursts of the a posteriori SNR, during noise-only frames, are *pulled down* to the average noise level, thus avoiding local buildup of noise whenever it exceeds its average characteristics. As a result, the MMSE-LSA estimator generally produces lower levels of residual musical noise, when compared with the MMSE spectral estimators.

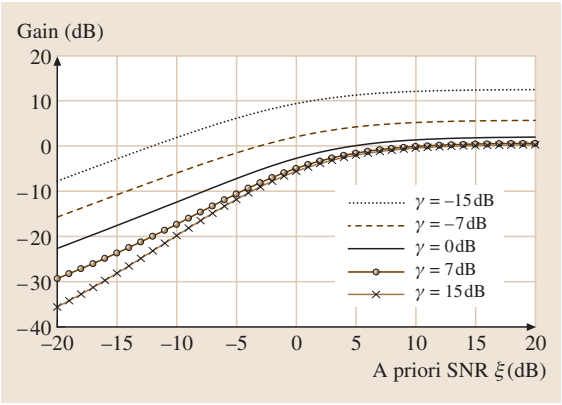

Fig. 44.6 Parametric gain curves describing the MMSE log-spectral amplitude gain function $G_{\mathrm{LSA}}(\xi, \gamma)$ for a Gaussian model, obtained by (44.31)

44.5 Signal Presence Probability Estimation

In this section, we derive an efficient estimator $\hat{p}_{tk|t-1}$ for the a priori speech presence probability. This estimator employs a soft-decision approach to compute three parameters based on the time–frequency distribution of the estimated a priori SNR $\hat{\xi}_{tk}$. The parameters exploit the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

Let ζ_{tk} denote a recursive average of the a priori SNR with a time constant α_ζ,

$$\zeta_{tk} = \alpha_\zeta \zeta_{t-1,k} + (1-\alpha_\zeta)\hat{\xi}_{t-1,k} \,. \tag{44.32}$$

By applying *local* and *global* averaging windows in the frequency domain, we obtain respectively local and global averages of the a priori SNR

$$\zeta_{tk}^{\chi} = \sum_{i=-w_\chi}^{w_\chi} h_\chi(i)\zeta_{t,k-i} \,, \tag{44.33}$$

where the superscript χ designates either *local* or *global*, and h_χ is a normalized window of size $2w_\chi + 1$. We define two parameters, P_{tk}^{local} and P_{tk}^{global}, which represent the relation between the above averages and the like-

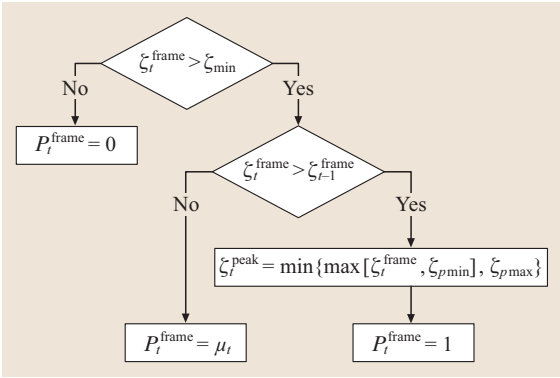

Fig. 44.7 Block diagram for computing P^{frame} (a parameter representing the likelihood of speech in a given frame) (after [44.42])

lihood of speech in the k-th frequency bin of the t-th frame. These parameters are given by

$$P_{tk}^{\chi} = \begin{cases} 0\,, & \text{if } \zeta_{tk}^{\chi} \le \zeta_{\min} \\ 1\,, & \text{if } \zeta_{tk}^{\chi} \ge \zeta_{\max} \\ \frac{\log(\zeta_{tk}^{\chi}/\zeta_{\min})}{\log(\zeta_{\max}/\zeta_{\min})}\,, & \text{otherwise}\,, \end{cases} \tag{44.34}$$

where $\zeta_{\min}$ and $\zeta_{\max}$ are empirical constants, maximized to attenuate noise while maintaining weak speech components.

In order to attenuate noise further in noise-only frames, we define a third parameter, P_t^{frame}, which is based on the speech energy in neighboring frames. An averaging of ζ_{tk} in the frequency domain (possibly over a certain frequency band) yields

$$\zeta_t^{\text{frame}} = \underset{1\le k\le N/2}{\text{mean}}\{\zeta_{tk}\}\,. \tag{44.35}$$

To prevent clipping of speech onsets or weak components, speech is assumed whenever ζ_t^{frame} increases over time. Moreover, the transition from H_1 to H_0 is delayed,

Table 44.1 Values of the parameters used in the implementation of the speech presence probability estimator, for a sampling rate of 16 kHz

$\alpha_\zeta = 0.7$	$\zeta_{\min} = -10\,\text{dB}$	$\zeta_{p\,\min} = 0\,\text{dB}$
$w_{\text{local}} = 1$	$\zeta_{\max} = -5\,\text{dB}$	$\zeta_{p\,\max} = 10\,\text{dB}$
$w_{\text{global}} = 15$	$p_{\min} = 0.005$	
h_{local}, h_{global}: Hann windows		

which reduces the misdetection of weak speech tails, by allowing for a certain decrease in the value of ζ_t^{frame}. Figure 44.7 describes a block diagram for computing P_t^{frame}, where

$$\mu_t \triangleq \begin{cases} 0\,, & \text{if } \zeta_t^{\text{frame}} \le \zeta_t^{\text{peak}}\zeta_{\min} \\ 1\,, & \text{if } \zeta_t^{\text{frame}} \ge \zeta_t^{\text{peak}}\zeta_{\max} \\ \frac{\log(\zeta_t^{\text{frame}}/\zeta_t^{\text{peak}}/\zeta_{\min})}{\log(\zeta_{\max}/\zeta_{\min})}\,, & \text{otherwise}\,, \end{cases} \tag{44.36}$$

represents a soft transition from *speech* to *noise*, ζ_t^{peak} is a confined peak value of ζ_t^{frame}, and $\zeta_{p\,\min}$ and $\zeta_{p\,\max}$ are empirical constants that determine the delay of the transition. Typical values of parameters used for a sampling rate of 16 kHz are summarized in Table 44.1.

The proposed estimate for the a priori speech presence probability is obtained by

$$\hat{p}_{tk|t-1} = P_{tk}^{\text{local}} P_{tk}^{\text{global}} P_t^{\text{frame}}\,. \tag{44.37}$$

Accordingly, $\hat{p}_{tk|t-1}$ is smaller if either previous frames, or recent neighboring frequency bins, do not contain speech. When $\hat{p}_{tk|t-1} \to 0$, the conditional speech presence probability $\hat{p}_{tk} \to 0$ by (44.6), and consequently the signal estimator $\hat{X}_{tk}$ reduces to $\hat{X}_{tk} = G_{\min} Y_{tk}$. Therefore, to reduce the possibility of speech distortion we generally restrict $\hat{p}_{tk|t-1}$ to be larger than a threshold $p_{\min}$ ($p_{\min} > 0$).

44.6 A Priori SNR Estimation

In this section, we address the problem of estimating the speech spectral variance λ_{tk} assuming that $\hat{p}_{tk|t-1}$ and $\widehat{\sigma_{tk}^2}$ are given. We present the decision-directed, causal and noncausal estimators for the a priori SNR $\xi_{tk} = \lambda_{tk}/\sigma_{tk}^2$ under speech presence uncertainty. The a priori SNR ξ_{tk} is estimated for each spectral component and each analysis frame due to the nonstationarity of the speech signal.

44.6.1 Decision-Directed Estimation

Ephraim and *Malah* [44.12] proposed a decision-directed approach, which provides a very useful

estimation method for the a priori SNR [44.47,48]. Accordingly, if speech presence is assumed ($p_{tk} \equiv 1$), then the expression

$$\Xi_{tk} = \alpha \frac{|\hat{X}_{t-1,k}|^2}{\hat{\sigma}^2_{t-1,k}} + (1-\alpha) \max\left\{\hat{\gamma}_{tk} - 1, 0\right\} \tag{44.38}$$

can be substituted for the a priori SNR. The first term, $|\hat{X}_{t-1,k}|^2/\hat{\sigma}^2_{t-1,k}$, represents the a priori SNR resulting from the processing of the previous frame. The second term, $\max\{\hat{\gamma}_{tk} - 1, 0\}$, is a maximum-likelihood estimate for the a priori SNR, based entirely on the current frame. The parameter α ($0 < \alpha < 1$) is a weighting factor that controls the trade-off between noise reduction and transient distortion brought into the signal [44.12, 47]. A larger value of α results in a greater reduction of the musical noise phenomena, but at the expense of attenuated speech onsets and audible modifications of transient components. As a compromise, a value 0.98 of α was determined by simulations and informal listening tests [44.12].

Under speech presence uncertainty, according to [44.12, 49], the expression in (44.38) estimates a *nonconditional a priori* SNR $\eta_{tk} \triangleq E\{|X_{tk}|^2\}/\sigma^2_{tk}$. The a priori SNR $\xi_{tk} = E\{|X_{tk}|^2|H_1^{tk}\}/\sigma^2_{tk}$ is related to η_{tk} by

$$\eta_{tk} = \frac{E\{|X_{tk}|^2|H_1^{tk}\}P(H_1^{tk})}{\sigma^2_{tk}} = \xi_{tk}\, p_{tk|t-1}\,. \tag{44.39}$$

Therefore the estimate for ξ_{tk} should supposedly be given by

$$\hat{\xi}_{tk} = \frac{\Xi_{tk}}{\hat{p}_{tk|t-1}}\,. \tag{44.40}$$

However, the division by $\hat{p}_{tk|t-1}$ may deteriorate the performance of the speech enhancement system [44.50,51]. In some cases, it introduces interaction between the estimated $p_{tk|t-1}$ and the a priori SNR, that adversely affects the total gain for noise-only bins, resulting in an unnaturally structured residual noise [44.52]. To some extent, the noise structuring can be eliminated by utilizing a voice activity detector (VAD) and applying a uniform attenuation factor to frames that do not contain speech [44.49]. Yet, VADs are difficult to tune and their reliability is often insufficient for weak speech components and low input SNR.

Let $\hat{X}_{tk|H_1} = \hat{X}_{tk}|_{\hat{p}_{tk}=1}$ denote an estimate for $\hat{X}_{tk}$ under the hypothesis of speech presence. Then an alternative a priori SNR estimator under speech presence uncertainty is given by [44.42]

$$\hat{\xi}_{tk} = \alpha \frac{|\hat{X}_{t-1,k|H_1}|^2}{\hat{\sigma}^2_{t-1,k}} + (1-\alpha) \max\left\{\hat{\gamma}_{tk} - 1, 0\right\}. \tag{44.41}$$

Notice that for $\hat{p}_{t-1,k|t-2} \neq 1$, this yields a different estimate than either Ξ_{tk} or $\Xi_{tk}/\hat{p}_{tk|t-1}$. In [44.50, 51], it was suggested to simply estimate the a priori SNR by Ξ_{tk}, rather than $\Xi_{tk}/\hat{p}_{tk|t-1}$. However, the use of $\hat{X}_{t-1,k|H_1}$ in (44.41) boosts the gain up when speech is present, which provides a compensation for not dividing by $\hat{p}_{tk|t-1}$.

To show that under speech presence uncertainty it is advantageous to estimate the a priori SNR by the expression in (44.41) rather than by $\Xi_{tk}/\hat{p}_{tk|t-1}$, we assume that an estimate $\hat{p}_{tk|t-1}$ for the a priori speech presence probability is given, and that Ξ_{tk} and $\hat{\xi}_{tk}$ have been calculated by (44.38) and (44.41), respectively. By definition, if H_1^{tk} is true, then the spectral estimate $\hat{X}_{tk}$ should degenerate to $\hat{X}_{tk|H_1}$, and the a priori SNR estimate should coincide with Ξ_{tk}. On the contrary, if H_0^{tk} is true, then $\hat{X}_{tk}$ should reduce to $G_{\min} Y_{tk}$, or equivalently the a priori SNR estimate should be as small as possible. Indeed, if H_1^{tk} is true then

$$\hat{\xi}_{tk}|_{H_1} \approx \Xi_{tk}|_{H_1} \leq \left.\frac{\Xi_{tk}}{\hat{p}_{tk|t-1}}\right|_{H_1}, \tag{44.42}$$

where we have used that under H_1^{tk} the spectral estimate $\hat{X}_{t-1,k}$ is approximately the same as $\hat{X}_{t-1,k|H_1}$ (if H_1^{tk} is true then $H_1^{t-1,k}$ is likely to be true as well, due to the strong correlation of speech presence in successive frames). On the other hand, if H_0^{tk} is true, then $\hat{p}_{tk|t-1}$ is expected to approach zero, and $\hat{\xi}_{tk}$ is likely to be much smaller than $\Xi_{tk}/\hat{p}_{tk|t-1}$:

$$\hat{\xi}_{tk}|_{H_0} \approx \alpha G^2_{\min} \ll \left.\frac{\Xi_{tk}}{\hat{p}_{tk|t-1}}\right|_{H_0} \approx \frac{\alpha G^2_{\min}}{\hat{p}_{tk|t-1}}\,. \tag{44.43}$$

Therefore, under speech presence uncertainty the decision-directed a priori SNR estimator is more favorably modified as in (44.41), rather than dividing Ξ_{tk} by the a priori speech presence probability $\hat{p}_{tk|t-1}$.

44.6.2 Causal Recursive Estimation

In this section, we present a causal conditional estimator $\hat{\xi}_{tk|t} = \hat{\lambda}_{tk|t}/\widetilde{\sigma^2_{tk}}$ for the a priori SNR given the noisy measurements up to frame t. The estimator combines two steps, a *propagation* step and an *update* step, following the rational of Kalman filtering, to predict and update the estimate for λ_{tk} recursively as new data arrive.

Let $\mathcal{X}_0^\tau = \{X_{tk} | t = 0, \ldots, \tau, k = 0, \ldots, N-1\}$ represent the set of clean-speech spectral coefficients up to frame τ, and let $\lambda_{tk|\tau} \triangleq E\{|X_{tk}|^2 | H_1^{tk}, \mathcal{X}_0^\tau\}$ denote the *conditional* variance of X_{tk} under H_1^{tk} given the clean spectral coefficients up to frame τ. Assuming that an estimate $\hat{\lambda}_{tk|t-1}$ for the one-frame-ahead conditional variance of X_{tk} is available, an estimate for $\lambda_{tk|t}$ can be obtained by calculating its conditional mean under H_1^{tk} given Y_{tk} and $\hat{\lambda}_{tk|t-1}$. By definition, $\lambda_{tk|t} = |X_{tk}|^2$. Hence,

$$\begin{aligned}\hat{\lambda}_{tk|t} &= E\left\{|X_{tk}|^2 \middle| H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{tk}\right\} \\ &= E\left\{X_{\mathrm{R}tk}^2 \middle| H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\mathrm{R}tk}\right\} \\ &\quad + E\left\{X_{\mathrm{I}tk}^2 \middle| H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\mathrm{I}tk}\right\}, \end{aligned} \tag{44.44}$$

where we have used that $X_{\mathrm{R}tk}$ is independent of $Y_{\mathrm{I}tk}$, and $X_{\mathrm{I}tk}$ is independent of $Y_{\mathrm{R}tk}$. Let $G_{\mathrm{SP}}(\xi, \gamma_\rho)$ represent the MMSE gain function in the spectral power domain for $Y_{\rho tk} \neq 0$ [44.43]:

$$E\left\{X_{\rho tk}^2 \middle| H_1^{tk}, \hat{\lambda}_{tk|t-1}, Y_{\rho tk}\right\} = G_{\mathrm{SP}}\left(\hat{\xi}_{tk|t-1}, \gamma_{\rho tk}\right) Y_{\rho tk}^2 , \tag{44.45}$$

where $\rho \in \{\mathrm{R}, \mathrm{I}\}$. Then the specific expression for $G_{\mathrm{SP}}(\xi, \gamma_\rho)$ depends on the particular statistical model. For a Gaussian model, the spectral power gain function is given by

$$G_{\mathrm{SP}}(\xi, \gamma_\rho) = \frac{\xi}{1+\xi}\left(\frac{1}{2\gamma_\rho} + \frac{\xi}{1+\xi}\right) . \tag{44.46}$$

For a gamma model [44.43],

$$\begin{aligned} G_{\mathrm{SP}}(\xi, \gamma_\rho) = \frac{3}{8\gamma_\rho} &\left[\exp\left(\frac{C_{\rho-}^2}{4}\right) D_{-2.5}(C_{\rho-}) \right. \\ &\left. + \exp\left(\frac{C_{\rho+}^2}{4}\right) D_{-2.5}(C_{\rho+})\right] \\ &\times \left[\exp\left(\frac{C_{\rho-}^2}{4}\right) D_{-0.5}(C_{\rho-}) \right. \\ &\left. + \exp\left(\frac{C_{\rho+}^2}{4}\right) D_{-0.5}(C_{\rho+})\right]^{-1} , \end{aligned} \tag{44.47}$$

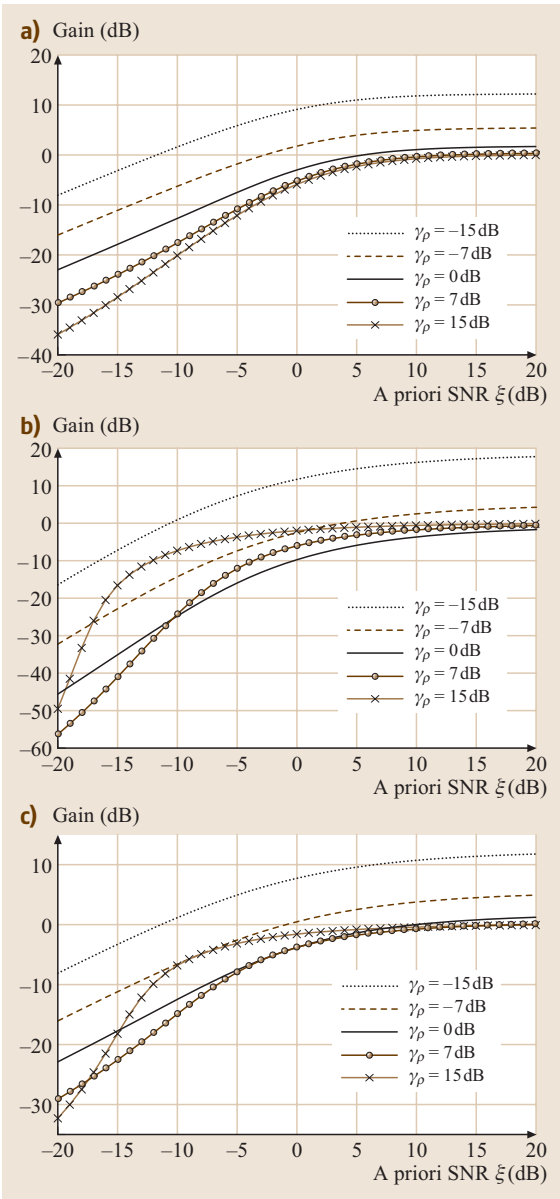

Fig. 44.8a–c Parametric gain curves describing the MMSE spectral power gain function $G_{\mathrm{SP}}(\xi, \gamma_\rho)$ for different speech models. **(a)** Gain curves for a Gaussian model, obtained by (44.46). **(b)** Gain curves for a gamma model, obtained by (44.47). **(c)** Gain curves for a Laplacian model, obtained by (44.48) (after [44.43])

where $C_{\rho\pm}$ are defined by (44.25). For a Laplacian model [44.43],

$$\begin{aligned} G_{\mathrm{SP}}(\xi,\gamma_\rho) &= \frac{1}{\gamma_\rho}[\mathrm{erfcx}(L_{\rho+})+\mathrm{erfcx}(L_{\rho-})]^{-1} \\ &\times\left[\left(L_{\rho+}^2+0.5\right)\mathrm{erfcx}(L_{\rho+})\right. \\ &+\left(L_{\rho-}^2+0.5\right)\mathrm{erfcx}(L_{\rho-}) \\ &\left.-\frac{(L_{\rho+}+L_{\rho-})}{\sqrt{\pi}}\right], \end{aligned} \tag{44.48}$$

where $L_{\rho\pm}$ are defined by (44.27). Figure 44.8 shows parametric gain curves, which describe the functions $G_{\mathrm{SP}}(\xi,\gamma_\rho)$ for several values of γ_ρ, resulting from (44.46), (44.47), and (44.48). When γ_ρ is kept constant, $G_{\mathrm{SP}}(\xi,\gamma_\rho)$ is a monotonically increasing function of ξ. When ξ is kept constant, $G_{\mathrm{SP}}(\xi,\gamma_\rho)$ is a monotonically decreasing function of γ_ρ for a Gaussian model, but is not a monotonic function of γ_ρ for gamma or Laplacian models.

Equation (44.45) does not hold in the case $Y_{\rho tk}\to 0$, since $G_{\mathrm{SP}}(\xi,\gamma_\rho)\to\infty$ as $\gamma_\rho\to 0$, and the conditional variance of $X_{\rho tk}$ is generally not zero. For $Y_{\rho tk}=0$ (or practically for $Y_{\rho tk}$ smaller in magnitude than a predetermined threshold) we use the following expressions [44.43]: for a Gaussian model

$$E\left\{X_{\rho tk}^2\middle|H_1^{tk},\hat{\lambda}_{tk|t-1},Y_{\rho tk}=0\right\}=\frac{\hat{\xi}_{tk|t-1}}{1+\hat{\xi}_{tk|t-1}}\sigma_{tk}^2\,, \tag{44.49}$$

for a gamma model

$$\begin{aligned} &E\left\{X_{\rho tk}^2\middle|H_1^{tk},\hat{\lambda}_{tk|t-1},Y_{\rho tk}=0\right\} \\ &=\frac{3D_{-2.5}\left(\frac{\sqrt{3}}{2\sqrt{\hat{\xi}_{tk|t-1}}}\right)}{8D_{-0.5}\left(\frac{\sqrt{3}}{2\sqrt{\hat{\xi}_{tk|t-1}}}\right)}\sigma_{tk}^2 \end{aligned} \tag{44.50}$$

and for a Laplacian model

$$\begin{aligned} &E\left\{X_{\rho tk}^2\middle|H_1^{tk},\hat{\lambda}_{tk|t-1},Y_{\rho tk}=0\right\} \\ &=\sqrt{\frac{2}{\pi}}\frac{\exp\left(\frac{1}{2\hat{\xi}_{tk|t-1}}\right)D_{-3}\left(\sqrt{\frac{2}{\hat{\xi}_{tk|t-1}}}\right)}{\mathrm{erfcx}\left(\frac{1}{\sqrt{\hat{\xi}_{tk|t-1}}}\right)}\sigma_{tk}^2\,. \end{aligned} \tag{44.51}$$

From (44.45–44.51), we can define a function $f(\xi,\sigma^2,Y_\rho^2)$ such that

$$\begin{aligned} &\frac{1}{\sigma_{tk}^2}E\left\{X_{\rho tk}^2\middle|H_1^{tk},\hat{\lambda}_{tk|t-1},Y_{\rho tk}\right\} \\ &=f\left(\hat{\xi}_{tk|t-1},\sigma_{tk}^2,Y_{\rho tk}^2\right) \end{aligned} \tag{44.52}$$

for all $Y_{\rho tk}$. Substituting (44.52) into (44.44), we obtain an estimate for $\xi_{tk|t}$ given by

$$\hat{\xi}_{tk|t}=f\left(\hat{\xi}_{tk|t-1},\widehat{\sigma_{tk}^2},Y_{\mathrm{R}tk}^2\right)+f\left(\hat{\xi}_{tk|t-1},\widehat{\sigma_{tk}^2},Y_{\mathrm{I}tk}^2\right)\,. \tag{44.53}$$

Equation (44.53) is the update step of the recursive estimation, since we start with an estimate $\hat{\xi}_{tk|t-1}$ that relies on the noisy observations up to frame $t-1$, and then update the estimate by using the additional information Y_{tk}.

To formulate the propagation step, we assume that we are given at frame $t-1$ estimates for the speech spectral coefficient $X_{t-1,k}$ and its spectral variance $\lambda_{t-1,k}$, conditioned on $\mathcal{Y}_0^{t-1}=\{Y_{\tau k}|\tau=0,\ldots,t-1,k=0,\ldots,N-1\}$. Then, these estimates can be *propagated* in time to obtain an estimate for λ_{tk}. Since λ_{tk} is correlated with both $\lambda_{t-1,k}$ and $|X_{t-1,k}|$, it was proposed in [44.15] to use an estimate of the form

$$\begin{aligned} &\hat{\lambda}_{tk|t-1} \\ &=\max\left\{(1-\mu)\hat{\lambda}_{t-1,k|t-1}+\mu\left|\hat{X}_{t-1,k|H_1}\right|^2,\lambda_{\min}\right\}, \end{aligned} \tag{44.54}$$

where μ ($0\le\mu\le 1$) is related to the degree of nonstationarity of the random process $\{\lambda_{tk}|t=0,1,\ldots\}$, and $\lambda_{\min}$ is a lower bound on the variance of X_{tk}. In the case of a pseudostationary process, μ is set to a small value, since $\hat{\lambda}_{tk|t-1}\approx\hat{\lambda}_{t-1,k|t-1}$. In the case of a nonstationary process, μ is set to a larger value, since the variances at successive frames are less correlated, and the relative importance of $\hat{\lambda}_{t-1,k|t-1}$ to predict $\hat{\lambda}_{tk|t-1}$ decreases. Dividing both sides of (44.54) by $\hat{\sigma}_{t-1,k}^2$, we obtain the *propagation* step

$$\begin{aligned} &\hat{\xi}_{tk|t-1} \\ &=\max\left\{(1-\mu)\hat{\xi}_{t-1,k|t-1}+\mu\frac{|\hat{X}_{t-1,k|H_1}|^2}{\hat{\sigma}_{t-1,k}^2},\xi_{\min}\right\}, \end{aligned} \tag{44.55}$$

where $\xi_{\min}$ is a lower bound on the a priori SNR.

The steps of the causal recursive a priori SNR estimation are summarized in Table 44.2. The algorithm is initialized at frame $t=-1$ with $\hat{X}_{-1,k|H_1}=0$ and

Table 44.2 Summary of the causal recursive a priori SNR estimation

Initialization:
$\hat{X}_{-1,k\|H_1} = 0,\ \hat{\xi}_{-1,k\|-1} = \xi_{\min}$, for all k.
For all short-time frames $t = 0, 1, \ldots$
For all frequency bins $k = 0, \ldots, N-1$
Propagation step:
Compute $\hat{\xi}_{tk\|t-1}$ using (44.55)
Update step:
Compute $\hat{\xi}_{tk\|t}$ using (44.53)
Spectral estimation:
Compute $\hat{X}_{tk\|H_1} = \hat{X}_{tk}\big\|_{\hat{p}_{tk}=1}$ using, e.g., (44.22) or (44.29)

$\hat{\xi}_{-1,k|-1} = \xi_{\min}$ for all k. Then, for $t = 0, 1, \ldots$, the propagation and update steps are iterated to obtain estimates for the nonstationary a priori SNR. The spectral gain function employed for the computation of $\hat{X}_{tk|H_1}$ is determined by the particular choice of the distortion measure.

44.6.3 Relation Between Causal Recursive Estimation and Decision-Directed Estimation

The causal conditional estimator $\hat{\xi}_{tk|t}$ for the a priori SNR is closely related to the decision-directed estimator of *Ephraim* and *Malah* [44.12]. The decision-directed estimator under speech presence uncertainty is given by (44.41) where α $(0 \le \alpha \le 1)$ is a weighting factor that controls the trade-off between the noise reduction and the transient distortion introduced into the signal.

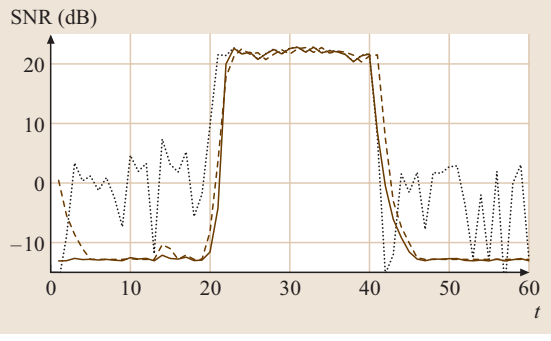

Fig. 44.9 SNRs in successive short-time frames: a posteriori SNR γ_{tk} (*dotted line*), decision-directed a priori SNR $\hat{\xi}^{\mathrm{DD}}_{tk|t}$ (*dashed line*), and causal recursive a priori SNR estimate $\hat{\xi}^{\mathrm{RE}}_{tk|t}$ (*solid line*) (after [44.15])

The update step (44.53) of the causal recursive estimator under a Gaussian model can be written as [44.15]

$$\hat{\xi}_{tk|t} = \alpha_{tk}\hat{\xi}_{tk|t-1} + (1-\alpha_{tk})(\hat{\gamma}_{tk} - 1)\,, \tag{44.56}$$

where α_{tk} is given by

$$\alpha_{tk} \triangleq 1 - \frac{\hat{\xi}^2_{tk|t-1}}{\left(1+\hat{\xi}_{tk|t-1}\right)^2}\,. \tag{44.57}$$

Substituting (44.55) into (44.56) and (44.57) with $\mu \equiv 1$, and applying the lower-bound constraint to $\hat{\xi}_{tk|t}$ rather than $\hat{\xi}_{tk|t-1}$, we have

$$\hat{\xi}_{tk|t} = \max\left\{\alpha_{tk}\frac{\left|\hat{X}_{t-1,k|H_1}\right|^2}{\hat{\sigma}^2_{t-1,k}} + (1-\alpha_{tk})\left(\hat{\gamma}_{tk}-1\right), \xi_{\min}\right\}\,, \tag{44.58}$$

$$\alpha_{tk} = 1 - \frac{\left|\hat{X}_{t-1,k|H_1}\right|^4}{\left(\hat{\sigma}^2_{t-1,k} + \left|\hat{X}_{t-1,k|H_1}\right|^2\right)^2}\,. \tag{44.59}$$

The expression (44.58) with $\alpha_{tk} \equiv \alpha$ is actually a practical form of the decision-directed estimator,

$$\hat{\xi}^{\mathrm{DD}}_{tk|t} = \max\left\{\alpha\frac{\left|\hat{X}_{t-1,k|H_1}\right|^2}{\hat{\sigma}^2_{t-1,k}} + (1-\alpha)\left(\hat{\gamma}_{tk}-1\right), \xi_{\min}\right\}\,, \tag{44.60}$$

that includes a lower-bound constraint to further reduce the level of residual musical noise [44.47]. Accordingly, a special case of the causal recursive estimator with $\mu \equiv 1$ degenerates to a *decision-directed* estimator with a *time-varying frequency-dependent* weighting factor α_{tk}.

It is interesting to note that the weighting factor α_{tk}, given by (44.59), is monotonically decreasing as a function of the instantaneous SNR, $|\hat{X}_{t-1,k|H_1}|^2/\hat{\sigma}^2_{t-1,k}$. A decision-directed estimator with a larger weighting factor is indeed preferable during speech absence (to reduce musical noise phenomena), while a smaller weighting factor is more advantageous during speech presence (to reduce signal distortion) [44.47]. The above special case of the causal recursive estimator conforms to such a desirable behavior. Moreover, the general form of the causal recursive estimator provides an additional degree of freedom for adjusting the value of μ in (44.55) to the degree of spectral nonstationarity. This may produce even further improvement in the performance.

The different behaviors of the causal recursive estimator $\hat{\xi}^{\mathrm{RE}}_{tk|t}$ (Table 44.2) and the decision-directed estimator $\hat{\xi}^{\mathrm{DD}}_{tk|t}$ (44.60) are illustrated in the example

of Fig. 44.9. The analyzed signal contains only white Gaussian noise during the first and last 20 frames, and in between it contains an additional sinusoidal component at the displayed frequency with 0 dB SNR. (Note that the SNR is computed in the time domain, whereas the a priori and a posteriori SNRs are computed in the time–frequency domain. Therefore, the latter SNRs may increase at the displayed frequency well above the average SNR.) The signal is transformed into the STFT domain using half-overlapping Hamming windows. The a priori SNR estimates, $\hat{\xi}_{tk|t}^{\mathrm{RE}}$ and $\hat{\xi}_{tk|t}^{\mathrm{DD}}$, are obtained by using the parameters $\xi_{\min} = -25\,\mathrm{dB}$, $\alpha = 0.98$, $\mu = 0.9$. The spectral estimate $\hat{X}_{tk|H_1}$ is recursively obtained by applying $G_{\mathrm{LSA}}(\hat{\xi}_{tk|t}, \hat{\gamma}_{tk})$ to the noisy spectral measurements (44.29, 31).

Figure 44.9 shows that, when the a posteriori SNR γ_{tk} is sufficiently low, the causal recursive a priori SNR estimate is smoother than the decision-directed estimate, which helps reducing the level of musical noise. When γ_{tk} increases, the response of the a priori SNR $\hat{\xi}_{tk|t}^{\mathrm{RE}}$ is initially slower than $\hat{\xi}_{tk|t}^{\mathrm{DD}}$, but then builds up faster to the a posteriori SNR. When γ_{tk} is sufficiently high, $\hat{\xi}_{tk|t}^{\mathrm{DD}}$ follows the a posteriori SNR with a delay of one frame, whereas $\hat{\xi}_{tk|t}^{\mathrm{RE}}$ follows the a posteriori SNR instantaneously. When γ_{tk} decreases, the response of $\hat{\xi}_{tk|t}^{\mathrm{RE}}$ is immediate, while that of $\hat{\xi}_{tk|t}^{\mathrm{DD}}$ is delayed by one frame. As a consequence, when compared with the decision-directed estimator, the causal recursive estimator produces a lower level of musical noise while not increasing the audible distortion in the enhanced signal [44.15].

44.6.4 Noncausal Recursive Estimation

In some important applications, e.g., digital voice recording, surveillance, speech recognition and speaker identification, a delay of a few short-term frames between the enhanced speech and the noisy observation is tolerable. In such cases, a noncausal estimation approach may produce less signal distortion and less musical residual noise than a causal estimation approach. In this section, we present a noncausal conditional estimator $\hat{\xi}_{tk|t+L}$ for the a priori SNR, given the noisy measurements up to frame $t+L$, where $L > 0$ denotes the admissible time delay in frames. Similar to the causal estimator, the noncausal estimator combines update and propagation steps to recursively estimate λ_{tk} as new data arrive. However, future spectral measurements are also employed in the process to better predict the spectral variances of the clean speech.

Let $\lambda'_{tk|t+L} \triangleq E\{|X_{tk}|^2 | \mathcal{Y}_0^{t-1}, \mathcal{Y}_{t+1}^{t+L}\}$ denote the conditional spectral variance of X_{tk} given $\mathcal{Y}_0^{t+L}$ excluding the noisy measurement at frame t. Let $\lambda_{tk|[t+1,t+L]} \triangleq E\{|X_{tk}|^2 | \mathcal{Y}_{t+1}^{t+L}\}$ denote the conditional spectral variance of X_{tk} given the subsequent noisy measurements $\mathcal{Y}_{t+1}^{t+L}$. Then, similar to (44.53), the estimate for ξ_{tk} given $\hat{\xi}'_{tk|t+L} \triangleq \hat{\lambda}'_{tk|t+L}/\widehat{\sigma_{tk}^2}$ and Y_t can be updated by

$$\hat{\xi}_{tk|t+L} = f\left(\hat{\xi}'_{tk|t+L}, \widehat{\sigma_{tk}^2}, Y_{\mathrm{R}tk}^2\right) + f\left(\hat{\xi}'_{tk|t+L}, \widehat{\sigma_{tk}^2}, Y_{\mathrm{I}tk}^2\right) . \tag{44.61}$$

To obtain an estimate for $\lambda'_{tk|t+L}$, we employ the estimates $\hat{X}_{t-1,k|H_1}$ and $\hat{\lambda}_{t-1,k|t+L-1}$ from the previous frame, and derive an estimate for λ_{tk} from the measurements $\mathcal{Y}_{t+1}^{t+L}$. Suppose an estimate $\hat{\lambda}_{tk|[t+1,t+L]}$ is given, we propagate the estimates from frame $t-1$ to frame t by [44.15, 53]

$$\hat{\lambda}'_{tk|t+L} = \max\left\{\mu\left|\hat{X}_{t-1,k|H_1}\right|^2 + (1-\mu)\left[\mu'\hat{\lambda}_{t-1,k|t+L-1} + (1-\mu')\hat{\lambda}_{tk|[t+1,t+L]}\right], \lambda_{\min}\right\} , \tag{44.62}$$

where μ ($0 \le \mu \le 1$) is related to the stationarity of the random process $\{\lambda_{tk} | t = 0, 1, \dots\}$, and μ' ($0 \le \mu' \le 1$) is associated with the reliability of the estimate $\hat{\lambda}_{tk|[t+1,t+L]}$ in comparison with that of $\hat{\lambda}_{t-1,k|t+L-1}$. Dividing both sides of (44.62) by $\hat{\sigma}_{t-1,k}^2$, we have the following *backward–forward propagation* step:

$$\hat{\xi}'_{tk|t+L} = \max\left\{\mu\frac{\left|\hat{X}_{t-1,k|H_1}\right|^2}{\hat{\sigma}_{t-1,k}^2} + (1-\mu)\left[\mu'\hat{\xi}_{t-1,k|t+L-1} + (1-\mu')\hat{\xi}_{tk|[t+1,t+L]}\right], \xi_{\min}\right\} . \tag{44.63}$$

An estimate for the a priori SNR ξ_{tk} given the measurements $\mathcal{Y}_{t+1}^{t+L}$ is obtained by

$$\hat{\xi}_{tk|[t+1,t+L]} = \begin{cases} \frac{1}{L}\sum_{n=1}^{L}\hat{\gamma}_{t+n,k} - \beta_f , & \text{if positive} , \\ 0 , & \text{otherwise} , \end{cases} \tag{44.64}$$

where β_f ($\beta_f \ge 1$) is an oversubtraction factor to compensate for a sudden increase in the noise level. This estimator is an anticausal version of the maximum-likelihood a priori SNR estimator suggested in [44.12].

Table 44.3 Summary of the noncausal recursive a priori SNR estimation

Initialization:
$\hat{X}_{-1,k\|H_1} = 0,\ \hat{\xi}_{-1\|L-1} = \xi_{\min}$.
For all short-time frames $t = 0, 1, \ldots$
For all frequency bins $k = 0, \ldots, N-1$
Backward estimation:
Compute $\hat{\xi}_{tk\|[t+1,t+L]}$ using (44.64)
Backward–forward propagation:
Compute $\hat{\xi}'_{tk\|t+L}$ using (44.63)
Update step:
Compute $\hat{\xi}_{tk\|t+L}$ using (44.61)
Spectral estimation:
Compute $\hat{X}_{tk\|H_1} = \hat{X}_{tk}\big\|_{\hat{p}_{tk}=1}$ using, e.g., (44.22) or (44.29)

The steps of the noncausal recursive a priori SNR estimation are summarized in Table 44.3. The algorithm is initialized at frame $t = -1$ with $\hat{X}_{-1,k|H_1} = 0$ and $\hat{\xi}_{-1|L-1} = \xi_{\min}$. Then, for $t = 0, 1, \ldots$, the propagation and update steps are iterated to obtain estimates for the a priori SNR and the speech spectral components.

Figure 44.10 demonstrates the behavior of the noncausal recursive estimator in the same example of Fig. 44.9. The noncausal a priori SNR estimate $\hat{\xi}^{\mathrm{RE}}_{tk|t+3}$ is obtained with the parameters $\xi_{\min} = -25\,\mathrm{dB}$, $\mu = \mu' = 0.9$, $\beta_f = 2$, and $L = 3$ frames delay. A comparison of Figs. 44.9 and 44.10 indicates that the differences between the causal and noncausal recursive estimators are primarily noticeable during onsets

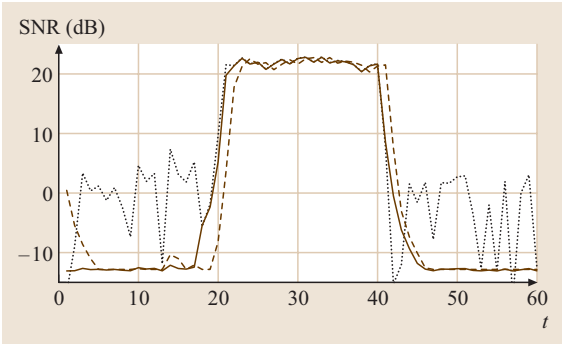

Fig. 44.10 SNRs in successive short-time frames: a posteriori SNR γ_{tk} (*dotted line*), decision-directed a priori SNR $\hat{\xi}^{\mathrm{DD}}_{tk|t}$ (*dashed line*), and noncausal recursive a priori SNR estimate $\hat{\xi}^{\mathrm{RE}}_{tk|t+3}$ with three-frame delay (*solid line*) (after [44.15])

of signal components. Clearly, the *causal* a priori SNR estimator, as well as the decision-directed estimator, cannot respond too fast to an abrupt increase in γ_{tk}, since it necessarily implies an increase in the level of musical residual noise. By contrast, the *noncausal* estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and irregularities in γ_{tk} corresponding to noise only. Therefore, in comparison with the decision-directed estimator, the noncausal a priori SNR estimator produces even lower levels of musical noise and signal distortion [44.15, 53].

44.7 Noise Spectrum Estimation

In this section, we derive an estimator for the noise power spectrum under speech presence uncertainty. The noise estimate is obtained by averaging past spectral power values of the noisy measurement, and multiplying the result by a constant factor that compensates the bias. The recursive averaging is carried out using a time-varying frequency-dependent smoothing parameter that is adjusted by the speech presence probability.

44.7.1 Time-Varying Recursive Averaging

A common noise estimation technique is to average past spectral power values of the noisy measurement recursively during periods of speech absence, and hold the estimate during speech presence. Specifically,

$$H_0^{tk} : \overline{\sigma}^2_{t+1,k} = \alpha_{\mathrm{d}} \overline{\sigma}^2_{tk} + (1-\alpha_{\mathrm{d}})|Y_{tk}|^2$$
$$H_1^{tk} : \overline{\sigma}^2_{t+1,k} = \overline{\sigma}^2_{tk}\,, \qquad (44.65)$$

where α_{d} $(0 < \alpha_{\mathrm{d}} < 1)$ denotes a smoothing parameter. Under speech presence uncertainty, we can employ the conditional speech presence probability, and carry out the recursive averaging by

$$\overline{\sigma}^2_{t+1,k} = \tilde{p}_{tk}\overline{\sigma}^2_{tk} + (1-\tilde{p}_{tk})\left[\alpha_{\mathrm{d}}\overline{\sigma}^2_{tk} + (1-\alpha_{\mathrm{d}})|Y_{tk}|^2\right], \qquad (44.66)$$

where $\tilde{p}_{tk}$ is an estimator for the conditional speech presence probability $p_{tk} = P(H_1^{tk}|Y_{tk})$. Equivalently, the

recursive averaging can be obtained by

$$\overline{\sigma}^2_{t+1,k} = \tilde{\alpha}_{tk}\overline{\sigma}^2_{tk} + \big(1-\tilde{\alpha}_{tk}\big)|Y_{tk}|^2\,, \tag{44.67}$$

where

$$\tilde{\alpha}_{tk} \triangleq \alpha_{\mathrm{d}} + (1-\alpha_{\mathrm{d}})\tilde{p}_{tk} \tag{44.68}$$

is a time-varying frequency-dependent smoothing parameter. The smoothing parameter $\tilde{\alpha}_{tk}$ is adjusted by the speech presence probability, which is estimated based on the noisy measurement.

Here we make a distinction between the estimator $\hat{p}_{tk}$ in (44.3), used for estimating the clean speech, and the estimator $\tilde{p}_{tk}$, which controls the adaptation of the noise spectrum. Clearly, deciding speech is absent (H_0) when speech is present (H_1) is more destructive when estimating the speech than when estimating the noise. Hence, different decision rules are employed [44.42], and generally we tend to employ estimators that satisfy $\hat{p}_{tk} \geq \tilde{p}_{tk}$. Given an estimator $\tilde{p}_{tk|t-1}$ for the a priori speech presence probability, the conditional speech presence probability estimate can be obtained from Bayes' rule, which under a Gaussian model reduces to [44.12]

$$\tilde{p}_{tk} = \left[1 + \frac{\big(1-\tilde{p}_{tk|t-1}\big)\big(1+\hat{\xi}_{tk}\big)}{\tilde{p}_{tk|t-1}\exp\left(\frac{\hat{\xi}_{tk}\hat{\gamma}_{tk}}{1+\hat{\xi}_{tk}}\right)}\right]^{-1}\,. \tag{44.69}$$

In the next subsection we present an estimator $\tilde{p}_{tk|t-1}$ that enables noise spectrum estimation during speech activity. Both $\tilde{p}_{tk|t-1}$ and $\hat{p}_{tk|t-1}$ are biased toward higher values, since deciding that speech is absent when speech is present results ultimately in the attenuation of speech components. Whereas, the alternative false decision, up to a certain extent, merely introduces some level of residual noise. Accordingly, we include a bias compensation factor in the noise estimator

$$\hat{\sigma}^2_{t+1,k} = \beta\overline{\sigma}^2_{t+1,k} \tag{44.70}$$

such that the factor β ($\beta \geq 1$) compensates the bias when speech is absent

$$\beta \triangleq \left.\frac{\sigma^2_{tk}}{E\big\{\overline{\sigma}^2_{tk}\big\}}\right|_{\xi_{tk}=0}\,. \tag{44.71}$$

The value of β is completely determined by the particular estimator for the a priori speech absence probability [44.54]. We note that the noise estimate is based on a variable time segment in each subband, which takes into account the probability of speech presence. The time segment is longer in subbands that contain frequent *speech* portions, and shorter in subbands that contain frequent *silence* portions. This feature has been considered [44.55] a desirable characteristic of the noise estimator, which improves its robustness and tracking capability.

44.7.2 Minima-Controlled Estimation

In this section, we present an estimator $\tilde{p}_{tk|t-1}$ that is controlled by the minima values of a smoothed power spectrum of the noisy signal. In contrast to the minimum statistics (MS) and related methods [44.56, 57], the smoothing of the noisy power spectrum is carried out in both time and frequency. This takes into account the strong correlation of speech presence in neighboring frequency bins of consecutive frames [44.42]. Furthermore, the procedure comprises two iterations of smoothing and minimum tracking. The first iteration provides a rough voice activity detection in each frequency band. Then, the smoothing in the second iteration excludes relatively strong speech components, which makes the minimum tracking during speech activity robust, even when using a relatively large smoothing window. A larger smoothing window decreases the variance of the minima values, but also widens the peaks of the speech activity power. An alternative solution is to modify the smoothing in time and frequency based on a smoothed a posteriori SNR [44.56].

Let α_{s} ($0 < \alpha_{\mathrm{s}} < 1$) be a smoothing parameter, and let b denote a normalized window function of length $2w+1$, i. e., $\sum_{i=-w}^{w} b_i = 1$. The frequency smoothing of the noisy power spectrum in each frame is defined by

$$S^{\mathrm{f}}_{tk} = \sum_{i=-w}^{w} b_i |Y_{t,k-i}|^2\,. \tag{44.72}$$

Subsequently, smoothing in time is performed by a first-order recursive averaging:

$$S_{tk} = \alpha_{\mathrm{s}} S_{t-1,k} + (1-\alpha_{\mathrm{s}}) S^{\mathrm{f}}_{tk}\,. \tag{44.73}$$

In accordance with the MS method, the minima values of S_{tk} are picked within a finite window of length D, for each frequency bin:

$$S^{\min}_{tk} \triangleq \min\{S_{\tau k} \mid t-D+1 \leq \tau \leq t\}\,. \tag{44.74}$$

It follows [44.56] that there exists a constant factor $B_{\min}$, independent of the noise power spectrum, such that

$$E\big\{S^{\min}_{tk}\big|\xi_{tk}=0\big\} = B^{-1}_{\min}\sigma^2_{tk}\,. \tag{44.75}$$

The factor $B_{\min}$ represents the bias of a minimum noise estimate, and generally depends on the values of D, α_{s},

w and the spectral analysis parameters (type, length and overlap of the analysis windows). The value of $B_{\min}$ can be estimated by generating a white Gaussian noise, and computing the inverse of the mean of $S_{tk}^{\min}$. This also takes into account the time–frequency correlation of the noisy periodogram $|Y_{tk}|^2$. Notice that the value of $B_{\min}$ is fixed, whereas in [44.56] it is estimated for each frequency band and each frame.

Let $\gamma_{tk}^{\min}$ and ζ_{tk} be defined by

$$\gamma_{tk}^{\min} \triangleq \frac{|Y_{tk}|^2}{B_{\min} S_{tk}^{\min}} ,$$
$$\zeta_{tk} \triangleq \frac{S_{tk}}{B_{\min} S_{tk}^{\min}} . \tag{44.76}$$

Under a Gaussian model, the probability density functions of $\gamma_{tk}^{\min}$ and ζ_{tk}, in the absence of speech, can be approximated by exponential and chi-square densities, respectively [44.54]:

$$p\left(\gamma_{tk}^{\min} \mid H_0^{tk}\right) \approx \mathrm{e}^{-\gamma_{tk}^{\min}} u\left(\gamma_{tk}^{\min}\right) , \tag{44.77}$$

$$p\left(\zeta_{tk} \mid H_0^{tk}\right) \approx \frac{1}{\left(\frac{2}{m}\right)^{m/2} \Gamma\left(\frac{m}{2}\right)} \zeta_{tk}^{m/2-1} \times \exp\left(-\frac{m\zeta_{tk}}{2}\right) u(\zeta_{tk}) , \tag{44.78}$$

where $\Gamma(\cdot)$ is the gamma function, and m is the equivalent degrees of freedom. Based on the first iteration smoothing and minimum tracking, a rough decision about speech presence is given by

$$I_{tk} = \begin{cases} 1 , & \text{if } \gamma_{tk}^{\min} < \gamma_0 \text{ and } \zeta_{tk} < \zeta_0 \\ & \text{(speech is absent)} \\ 0 , & \text{otherwise} \\ & \text{(speech is present)} . \end{cases} \tag{44.79}$$

The thresholds γ_0 and ζ_0 are set to satisfy a certain significance level ϵ:

$$\mathcal{P}\left(\gamma_{tk}^{\min} \geq \gamma_0 \mid H_0^{tk}\right) < \epsilon , \tag{44.80}$$

$$\mathcal{P}\left(\zeta_{tk} \geq \zeta_0 \mid H_0^{tk}\right) < \epsilon . \tag{44.81}$$

From (44.77) and (44.78) we have

$$\gamma_0 = -\log(\epsilon) , \tag{44.82}$$

$$\zeta_0 = \frac{1}{m} F_{\chi^2;m}^{-1}(1-\epsilon) , \tag{44.83}$$

where $F_{\chi^2;m}(x)$ denotes the standard chi-square cumulative distribution function, with m degrees of freedom. Typically, we use $\epsilon = 0.01$ and $m = 32$, so $\gamma_0 = 4.6$ and $\zeta_0 = 1.67$.

The second iteration of smoothing is conditional on the rough speech activity detection of the first iteration. It includes only the power spectral components, which have been identified as containing primarily noise. We set the initial condition for the first frame by $\tilde{S}_{0,k} = S_{0,k}^{\mathrm{f}}$. Then, for $t > 0$ the smoothing in frequency, employing the above voice activity detector, is obtained by

$$\tilde{S}_{tk}^{\mathrm{f}} = \begin{cases} \dfrac{\sum_{i=-w}^{w} b_i I_{t,k-i} |Y_{t,k-i}|^2}{\sum_{i=-w}^{w} b_i I_{t,k-i}} , & \text{if } \sum_{i=-w}^{w} I_{t,k-i} \neq 0 \\ \tilde{S}_{t-1,k} , & \text{otherwise} . \end{cases} \tag{44.84}$$

Smoothing in time is given, as before, by a first-order recursive averaging

$$\tilde{S}_{tk} = \alpha_{\mathrm{s}} \tilde{S}_{t-1,k} + (1-\alpha_{\mathrm{s}}) \tilde{S}_{tk}^{\mathrm{f}} . \tag{44.85}$$

The minima values of $\tilde{S}_{tk}$ are picked within a finite window of length D, for each frequency bin

$$\tilde{S}_{tk}^{\min} \triangleq \min\left\{\tilde{S}_{\tau k} \mid t - D + 1 \leq \tau \leq t\right\} .$$

Accordingly, $\tilde{S}_{tk}^{\min}$ represents minima tracking that is conditional on the rough speech activity detection of the first iteration. We note that keeping the strong speech components out of the smoothing process enables improved minimum tracking. In particular, a larger smoothing parameter (α_{s}) and smaller minima search window (D) can be used. This reduces the variance of the minima values [44.56], and shortens the delay when responding to a rising noise power, which eventually improves the tracking capability of the noise estimator.

Let $\tilde{\gamma}_{tk}^{\min}$ and $\tilde{\zeta}_{tk}$ be defined by

$$\tilde{\gamma}_{tk}^{\min} \triangleq \frac{|Y_{tk}|^2}{B_{\min} \tilde{S}_{tk}^{\min}} ,$$
$$\tilde{\zeta}_{tk} \triangleq \frac{S_{tk}}{B_{\min} \tilde{S}_{tk}^{\min}} . \tag{44.86}$$

Since we use a relatively small significance level in the first iteration ($\epsilon = 0.01$), the influence of the voice activity detector in noise-only periods can be neglected. That is, the effect of excluding strong *noise* components from the smoothing process is negligible. Accordingly, the conditional distributions of $\tilde{\gamma}_{tk}^{\min}$ and $\tilde{\zeta}_{tk}$, in the absence of speech, are approximately the same as those of $\gamma_{tk}^{\min}$ and ζ_{tk} (44.77, 78).

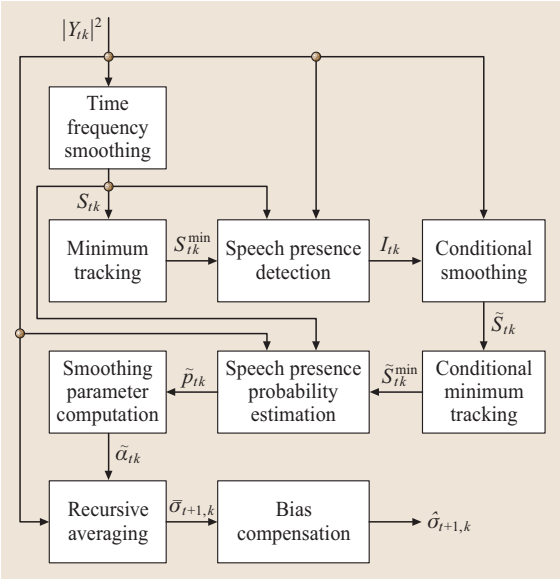

Fig. 44.11 Block diagram of the IMCRA noise spectrum estimation

An estimator for the a priori speech presence probability is given by

$$\tilde{p}_{tk|t-1} = \begin{cases} 0\,, & \text{if } \tilde{\gamma}_{tk}^{\min} \le 1 \\ & \text{and } \tilde{\zeta}_{tk} < \zeta_0 \\ \left(\tilde{\gamma}_{tk}^{\min} - 1\right)/(\gamma_1 - 1)\,, & \text{if } 1 < \tilde{\gamma}_{tk}^{\min} < \gamma_1 \\ & \text{and } \tilde{\zeta}_{tk} < \zeta_0 \\ 1\,, & \text{otherwise}\,. \end{cases} \tag{44.87}$$

The threshold γ_1 is set to satisfy a certain significance level ϵ_1 ($\epsilon_1 > \epsilon$):

$$P\left(\tilde{\gamma}_{tk}^{\min} > \gamma_1 \middle| H_0^{tk}\right) < \epsilon_1 \quad \Rightarrow \quad \gamma_1 \approx -\log(\epsilon_1)\,. \tag{44.88}$$

Typically $\epsilon_1 = 0.05$, and $\gamma_1 = 3$.

The a priori speech presence probability estimator assumes that speech is present ($\tilde{p}_{tk|t-1} = 1$) whenever $\tilde{\zeta}_{tk} \ge \zeta_0$ or $\tilde{\gamma}_{tk}^{\min} \ge \gamma_1$. That is, whenever the local measured power, S_{tk}, or the instantaneous measured power, $|Y_{tk}|^2$, are relatively high compared to the noise power $B_{\min}\tilde{S}_{tk}^{\min} \approx \sigma_{tk}^2$. The estimator assumes that speech is absent ($\tilde{p}_{tk|t-1} = 0$) whenever both the local and instantaneous measured powers are relatively low compared to the noise power ($\tilde{\gamma}_{tk}^{\min} \le 1$ and $\tilde{\zeta}_{tk} < \zeta_0$). In between, the estimator provides a soft transition between speech absence and speech presence, based on the value of $\tilde{\gamma}_{tk}^{\min}$.

The main objective of combining conditions on both $\tilde{\gamma}_{tk}^{\min}$ and $\tilde{\zeta}_{tk}$ is to prevent an increase in the estimated noise during weak speech activity, especially when the input SNR is low. Weak speech components can often be extracted using the condition on $\tilde{\zeta}_{tk}$. Sometimes, speech components are so weak that $\tilde{\zeta}_{tk}$ is smaller than ζ_0. In that case, most of the speech power is still excluded from the averaging process using the condition on $\tilde{\gamma}_{tk}^{\min}$. The remaining speech components can hardly affect the noise estimator, since their power is relatively low compared to that of the noise.

A block diagram of the improved minima-controlled recursive averaging (IMCRA) [44.54] noise spectrum estimation is described in Fig. 44.11. Typical values of parameters used in the implementation of the IMCRA noise estimator for a sampling rate of 16 kHz are summarized in Table 44.4. The noise spectrum estimate, $\hat{\sigma}_{tk}^2$, is initialized at the first frame by $\hat{\sigma}_{0,k}^2 = |Y_{0,k}|^2$. Then, at each frame t ($t \ge 0$), it is used, together with the current observation Y_{tk}, for estimating the noise power spectrum at the next frame, $t+1$. The bias compensation factor β is given by [44.54]

$$\beta = \frac{\gamma_1 - 1 - \mathrm{e}^{-1} + \mathrm{e}^{-\gamma_1}}{\gamma_1 - 1 - 3\mathrm{e}^{-1} + (\gamma_1 + 2)\mathrm{e}^{-\gamma_1}}\,. \tag{44.89}$$

In particular, for $\gamma_1 = 3$, we have $\beta = 1.47$. The value of $\tilde{\alpha}_{tk}$ is updated for each frequency bin and time frame, using the speech presence probability $\tilde{p}_{tk}$, and (44.68).

Table 44.4 Values of the parameters used in the implementation of the IMCRA noise estimator, for a sampling rate of 16 kHz

$w = 1$	$\alpha_s = 0.9$	$\alpha_d = 0.85$	$\beta = 1.47$
$B_{\min} = 1.66$	$\zeta_0 = 1.67$	$\gamma_0 = 4.6$	$\gamma_1 = 3$
$D = 120$	b: Hann window		

44.8 Summary of a Spectral Enhancement Algorithm

In this section, we present an example of a speech enhancement algorithm, which is based on an MMSE log-spectral amplitude estimation under a Gaussian model, IMCRA noise estimation, and decision-directed

Table 44.5 Summary of a speech enhancement algorithm

Initialization at the first frame for all frequency-bins $k = 1, \ldots, N/2$:
$\hat{\sigma}_{0k}^2 = \lvert Y_{0k}\rvert^2;\ \overline{\sigma}_{0k}^2 = \lvert Y_{0k}\rvert^2;\ S_{0k} = S_{0k}^{\mathrm{f}};\ S_{0k}^{\mathrm{min}} = S_{0k}^{\mathrm{f}};\ S_k^{\mathrm{min\_sw}} = S_{0k}^{\mathrm{f}};\ \tilde{S}_{0k} = S_{0k}^{\mathrm{f}};\ \tilde{S}_{0k}^{\mathrm{min}} = S_{0k}^{\mathrm{f}};\ \tilde{S}_k^{\mathrm{min\_sw}} = S_{0k}^{\mathrm{f}};\ \zeta_{0k} = 0.$
Let $\ell = 0$. % ℓ is a counter for frames within a subwindow ($0 \le \ell \le V$).
For all short-time frames $t = 0, 1, \ldots$
For all frequency bins $k = 1, \ldots, N/2$
Compute the a posteriori SNR $\hat{\gamma}_{tk}$ using (44.20) and (44.30), and the a priori SNR $\hat{\xi}_{tk}$ using (44.41), with the initial condition $\hat{\xi}_{0k} = \alpha + (1-\alpha)\max\{\hat{\gamma}_{0k} - 1, 0\}$.
Compute the conditional spectral estimate under the hypothesis of speech presence $\hat{X}_{tk\lvert H_1} = G_{\mathrm{LSA}}(\hat{\xi}_{tk}, \hat{\gamma}_{tk}) Y_{tk}$ using (44.29) and (44.31).
Compute the smoothed power spectrum S_{tk} using (44.72) and (44.73), and update its running minimum: $S_{tk}^{\mathrm{min}} = \min\{S_{t-1,k}^{\mathrm{min}}, S_{tk}\};\ S_k^{\mathrm{min\_sw}} = \min\{S_k^{\mathrm{min\_sw}}, S_{tk}\}$.
Compute the indicator function I_{tk} for the voice activity detection using (44.76) and (44.79).
Compute the conditional smoothed power spectrum $\tilde{S}_{tk}$ using (44.84) and (44.85), and update its running minimum: $\tilde{S}_{tk}^{\mathrm{min}} = \min\{\tilde{S}_{t-1,k}^{\mathrm{min}}, \tilde{S}_{tk}\};\ \tilde{S}_k^{\mathrm{min\_sw}} = \min\{\tilde{S}_k^{\mathrm{min\_sw}}, \tilde{S}_{tk}\}$.
Compute the a priori speech presence probability $\tilde{p}_{tk\lvert t-1}$ using (44.86) and (44.87), the speech presence probability $\tilde{p}_{tk}$ using (44.69), and the smoothing parameter $\tilde{\alpha}_{tk}$ using (44.68).
Update the noise spectrum estimate $\hat{\sigma}_{t+1,k}^2$ using (44.67) and (44.70).
Compute P_{tk}^{local} and P_{tk}^{global} using (44.32–44.34), and P_t^{frame} using the block diagram in Fig. 44.7.
Compute the a priori speech presence probability $\hat{p}_{tk\lvert t-1}$ using (44.37), and the speech presence probability $\hat{p}_{tk}$ using (44.69) by substituting $\tilde{p}_{tk\lvert t-1}$ with $\hat{p}_{tk\lvert t-1}$.
Compute the speech spectral estimate $\hat{X}_{tk}$ using (44.29).
Let $\ell = \ell + 1$.
If $\ell = V$
For all frequency bins k
Store $S_k^{\mathrm{min\_sw}}$, set S_{tk}^{min} to the minimum of the last U stored values of $S_k^{\mathrm{min\_sw}}$, and let $S_k^{\mathrm{min\_sw}} = S_{tk}$.
Store $\tilde{S}_k^{\mathrm{min\_sw}}$, set $\tilde{S}_{tk}^{\mathrm{min}}$ to the minimum of the last U stored values of $\tilde{S}_k^{\mathrm{min\_sw}}$, and let $\tilde{S}_k^{\mathrm{min\_sw}} = \tilde{S}_{tk}$.
Let $\ell = 0$.

a priori SNR estimation. The performance of the algorithm is demonstrated on speech signals degraded by various additive noise types.

The implementation of the speech enhancement algorithm is summarized in Table 44.5. For each time frame t we recursively estimate the STFT coefficients of the clean speech $\{X_{tk} | k = 1, \ldots, N/2\}$ from the noisy STFT coefficients $\{Y_{tk} | k = 1, \ldots, N/2\}$, where N is the length of the analysis window. We typically use a Hamming window of 32 ms length and a framing step of 8 ms (i. e., $N = 512$ and $M = 128$ for a sampling rate of 16 kHz). In the first frame ($t = 0$) we compute $\{Y_{0k} | k = 0, \ldots, N-1\}$ by applying the discrete Fourier transform to a short-time section of the noisy data

$$\boldsymbol{y}_0 = \Big[y(0)h(0)\ \ y(1)h(1)\ \ \ldots\ \ y(N-1)h(N-1)\Big]^{\mathrm{T}},$$

where $h(n)$ is the analysis window. In the following frames ($t > 0$), the section of noisy data is updated with M additional samples

$$\boldsymbol{y}_t = \Big[y(tM)h(0)\quad y(1+tM)h(1)\quad \ldots\quad y(N-1+tM)h(N-1)\Big]^{\mathrm{T}},$$

and subsequently $\{Y_{tk} | k = 0, \ldots, N-1\}$ is computed by applying the discrete Fourier transform to $\boldsymbol{y}_t$. Since the speech signal $x(n)$ is assumed to be real, once we estimate $\{X_{tk} | k = 1, \ldots, N/2\}$, the spectral coefficients for $N/2 < k \le N-1$ are obtained by $\hat{X}_{tk} = \hat{X}_{t,N-k}^*$, where * denotes complex conjugation. The DC component $\hat{X}_{to}$ is set to zero, and a sequence $\{\hat{x}_t(n) | n = 0, \ldots, N-1\}$ is obtained by applying the inverse discrete Fourier transform to $\{\hat{X}_{tk} | k = 0, \ldots, N-1\}$:

$$\hat{x}_t(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}_{tk}\, \mathrm{e}^{\mathrm{i}\frac{2\pi}{N}nk}. \tag{44.90}$$

Employing the weighted overlap-add method [44.37], we compute the following sequence

$$o_t(n) = \begin{cases} o_{t-1}(n+M) + N\tilde{h}(n)\hat{x}_t(n)\,, \\ \qquad\qquad \text{for } 0 \le n \le N-M-1 \\ N\tilde{h}(n)\hat{x}_t(n)\,, \quad \text{for } N-M \le n \le N-1\,, \end{cases} \tag{44.91}$$

where $\tilde{h}(n)$ is the synthesis window. Then, according to (44.2), for each frame t, we obtain M additional samples of the enhanced speech signal:

$$\hat{x}(n+tM) = o_t(n)\,, \quad n = 0, \ldots, M-1\,. \tag{44.92}$$

The synthesis window $\tilde{h}(n)$ should satisfy the completeness condition [44.36]

$$\sum_t \tilde{h}(n-tM)h(n-tM) = \frac{1}{N} \quad \text{for all } n\,. \tag{44.93}$$

Given analysis and synthesis windows that satisfy (44.93), any signal $x(n) \in \ell_2(\mathbb{Z})$ can be perfectly reconstructed from its STFT coefficients X_{tk}. However, for $M < N$ (over-redundant STFT representation) and for a given analysis window $h(n)$, there might be an infinite number of solutions to (44.93). A reasonable choice of a synthesis window is the one with minimum energy [44.36, 58], given by

$$\tilde{h}(n) = \frac{h(n)}{N\sum_\ell h^2(n-\ell M)}\,. \tag{44.94}$$

The estimator for the a priori speech presence probability, $\tilde{p}_{tk|t-1}$ in (44.87), requires two iterations of time–frequency smoothing (S_{tk}, $\tilde{S}_{tk}$) and minimum tracking ($S_{tk}^{\min}$, $\tilde{S}_{tk}^{\min}$). The minimum tracking is implemented by the method proposed in [44.56, 59], which provides a flexible balance between the computational complexity and the update rate of the minima values. Accordingly, we divide the window of D samples into U subwindows of V samples ($UV = D$). Whenever V samples are read, the minimum of the current subwindow is determined and stored for later use. The overall minimum is obtained as the minimum of past samples within the current subwindow and the U previous subwindow minima. Typical values of D and V correspond to 960 ms and 120 ms, respectively. That is, for a framing step of 8 ms (i. e., $M = 128$ for a sampling rate of 16 kHz) we set $D = 120$, $V = 15$, and $U = 8$.

To demonstrate the performance of the speech enhancement algorithm, utterances are taken from the TIMIT database [44.38], degraded by various noise types from the Noisex92 database [44.60], and enhanced by the algorithm in Table 44.5. A clean utterance from a female speaker is shown in Fig. 44.12. The speech signal is sampled at 16 kHz and degraded by the various noise types, which include white Gaussian noise, car interior noise, F16 cockpit noise, and babble noise. Figure 44.13 shows the noisy speech signals with SNR of 5 dB. The corresponding enhanced speech signals are shown in Fig. 44.14.

The performance is evaluated by three objective quality measures and informal listening tests. The first quality measure is the segmental SNR (SegSNR), in dB, defined by [44.61]

$$\text{SegSNR} = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{C}(\text{SNR}_t)\,, \tag{44.95}$$

where T denotes the number of frames in the signal, and

$$\text{SNR}_t = 10\log_{10}\frac{\sum_{n=tM}^{tM+N-1} x^2(n)}{\sum_{n=tM}^{tM+N-1}[x(n)-\hat{x}(n)]^2} \tag{44.96}$$

represents the SNR in the t-th frame. The operator $\mathcal{C}$ confines the SNR at each frame to the perceptually meaningful range between 35 dB and −10 dB ($\mathcal{C}x \triangleq \min[\max(x, -10), 35]$). The operator $\mathcal{C}$ prevents the segmental SNR measure from being biased in either a positive or negative direction due to a few silent or unusually high-SNR frames, which do not contribute

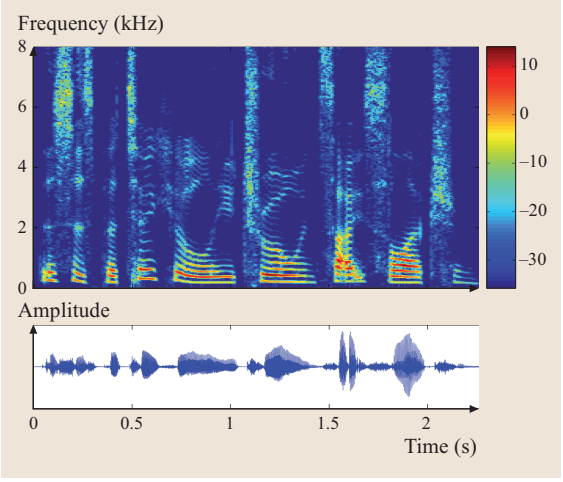

Fig. 44.12 Speech spectrogram and waveform of a clean speech signal: 'This is particularly true in site selection'

significantly to the overall speech quality [44.62, 63]. The second quality measure is log-spectral distortion (LSD), in dB, which is defined by

$$\mathrm{LSD} = \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{2}{N}\sum_{k=1}^{N/2}\left(\mathcal{L}X_{tk} - \mathcal{L}\hat{X}_{tk}\right)^2\right]^{\frac{1}{2}}, \tag{44.97}$$

where $\mathcal{L}X_{tk} \triangleq \max\{20\log_{10}|X_{tk}|, \delta\}$ is the log spectrum confined to about 50 dB dynamic range (that is, $\delta = \max_{tk}\{20\log_{10}|X_{tk}|\} - 50$). The third quality measure is the perceptual evaluation of speech quality (PESQ) score (ITU-T P.862).

The experimental results for the noisy and enhanced signals are given in the captions of Figs. 44.13 and 44.14. The improvement in SegSNR, reduction in LSD, and increase in PESQ scores are summarized in Table 44.6. Generally, the improvement in SegSNR and reduction in LSD are influenced by the variability of the noise characteristics in time and the initial SegSNR and LSD of the noisy signal. The faster the noise spectrum varies in time, the less reliable the noise spectrum estimator, and consequently the lower the quality gain that can be achieved by the speech enhancement system. Furthermore, the lower the SegSNR, respectively the LSD, for the noisy signal, the higher is the SegSNR improvement, respectively the lower is the LSD reduc-

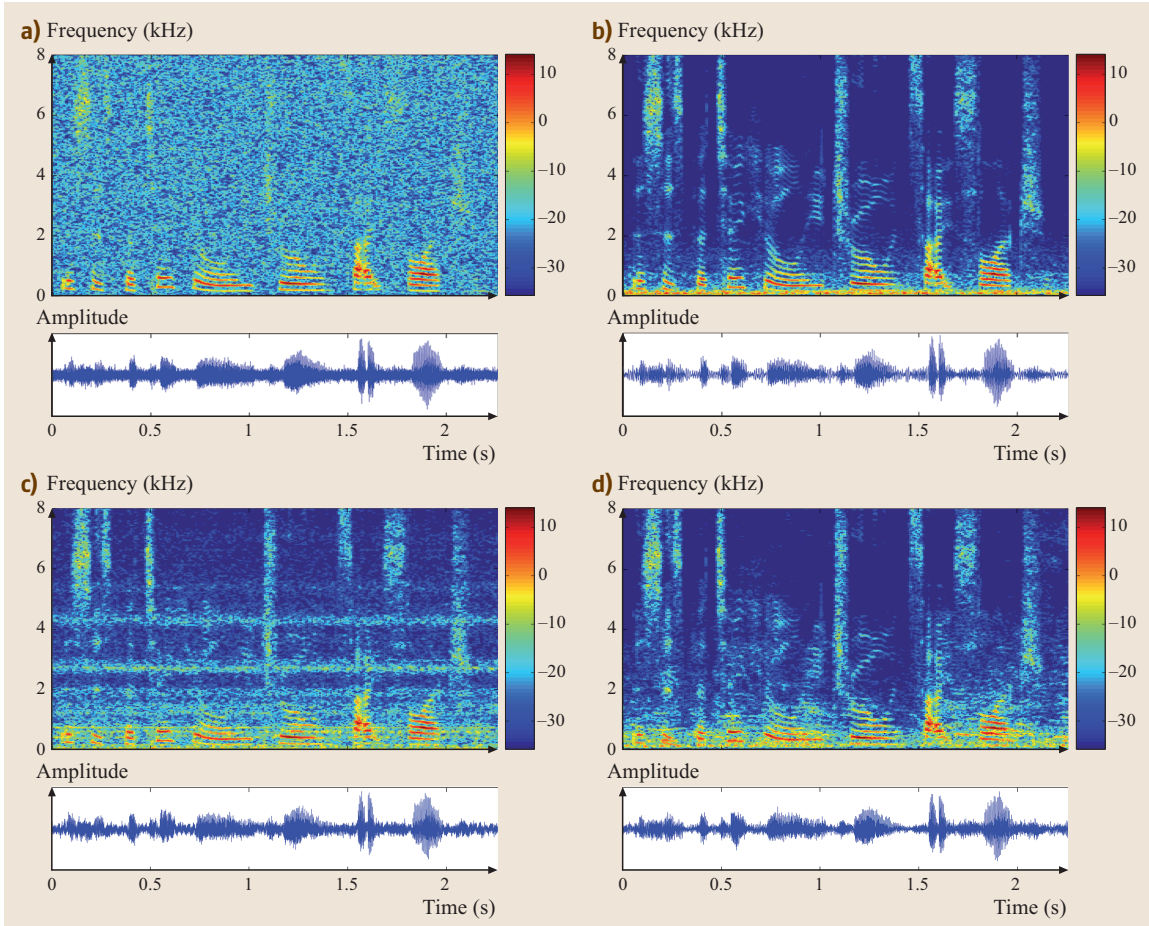

Fig. 44.13a–d Speech spectrograms and waveforms of the speech signal shown in Fig. 44.12 degraded by various noise types with SNR = 5 dB. **(a)** White Gaussian noise (SegSNR = −0.46 dB, LSD = 12.67 dB, PESQ = 1.74); **(b)** car interior noise (SegSNR = 0.30 dB, LSD = 3.48 dB, PESQ = 2.47); **(c)** F16 cockpit noise (SegSNR = −0.33 dB, LSD = 7.99 dB, PESQ = 1.76); **(d)** babble noise (SegSNR = 0.09 dB, LSD = 5.97 dB, PESQ = 1.87)

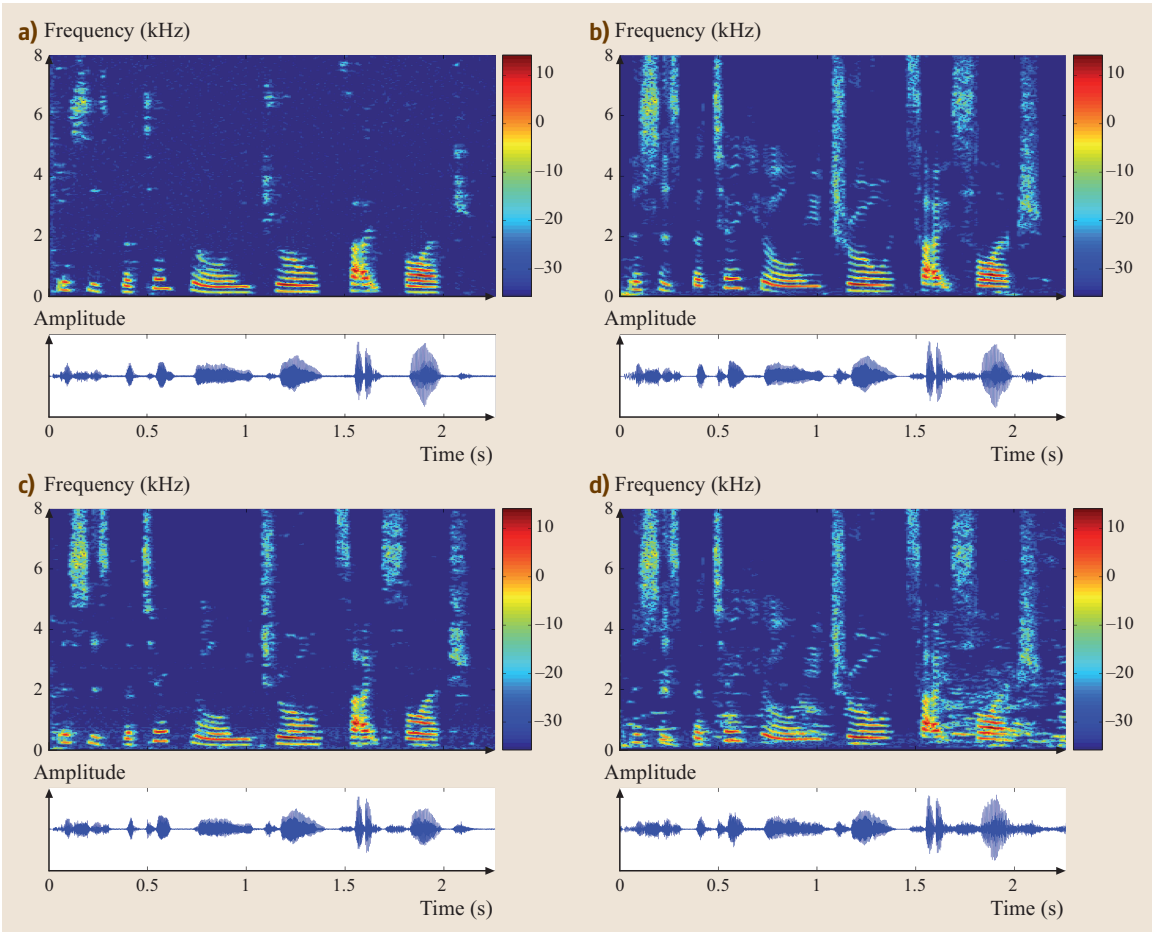

Fig. 44.14a–d Speech spectrograms and waveforms of the signals shown in Fig. 44.13 after enhancement with the algorithm in Table 44.5. **(a)** White Gaussian noise (SegSNR = 5.78 dB, LSD = 5.10 dB, PESQ = 2.34); **(b)** car interior noise (SegSNR = 9.52 dB, LSD = 2.67 dB, PESQ = 3.00); **(c)** F16 cockpit noise (SegSNR = 5.21 dB, LSD = 4.27 dB, PESQ = 2.29); **(d)** babble noise (SegSNR = 4.23 dB, LSD = 4.30 dB, PESQ = 2.13)

tion, that can be achieved by the speech enhancement system.

For car interior noise, most of the noise energy is concentrated in the lower frequencies. Therefore, the noise reduction is large in the low frequencies, and small in the high frequencies. Accordingly, in each frame, the total noise reduction is higher than that obtainable for the case of WGN. Since the SegSNR is mainly affected by the amount of noise reduction in each frame, the improvement in SegSNR is more significant for car interior noise. However, the reduction in LSD for the car interior noise is less substantial, since the initial LSD for the noisy signal is small.

The characteristics of babble noise vary more quickly in time when compared to the other noise

Table 44.6 Segmental SNR improvement, log-spectral distortion reduction and PESQ score improvement for various noise types, obtained by using the speech enhancement algorithm in Table 44.5

Noise type	SegSNR improvement	LSD reduction	PESQ score improvement
WGN	6.24	7.57	0.60
Car	9.22	0.81	0.53
F16	5.54	3.72	0.53
Babble	4.14	1.67	0.26

types. Therefore, the IMCRA noise spectrum estimator is least reliable for babble noise, and the speech enhancement performance is inferior to that achievable in slowly time-varying noise environments. Accordingly, the improvement in SegSNR is smaller for babble noise, when compared with the improvement achieved in more-stationary noise environment with the same initial SegSNR. Similarly, the reduction in LSD, and improvement in PESQ score, are smaller for babble noise, when compared with those achieved in more-stationary noise environment with the same initial LSD, respectively PESQ score, levels.

44.9 Selection of Spectral Enhancement Algorithms

In this section, we discuss some of the fundamental components that constitute a speech spectral enhancement system. Specifically, we address the choice of a statistical model, fidelity criterion, a priori SNR estimator, and noise spectrum estimator.

44.9.1 Choice of a Statistical Model and Fidelity Criterion

The Gaussian model underlies the design of many speech enhancement algorithms, e.g., [44.12, 17, 18, 42, 64–66]. This model is motivated by the central limit theorem, as each Fourier expansion coefficient is a weighted sum of random variables resulting from the random sequence [44.12]. When the span of correlation within the signal is sufficiently short compared to the size of the frames, the probability distribution function of the spectral coefficients asymptotically approaches Gaussian as the frame's size increases. The Gaussian approximation is in the central region of the Gaussian curve near the mean. However, the approximation can be very inaccurate in the tail regions away from the mean [44.67]. *Porter* and *Boll* [44.46] pointed out that a priori speech spectra do not have a Gaussian distribution, but gamma-like distribution. They proposed to compute the optimal estimator directly from the speech data, rather than from a parametric model of the speech statistics.

Martin [44.40] considered a gamma speech model, in which the real and imaginary parts of the clean speech spectral components are modeled as iid gamma random variables. He assumed that distinct spectral components are statistically independent, and derived MMSE estimators for the complex speech spectral coefficients under Gaussian and Laplacian noise modeling. He showed that, under Gaussian noise modeling, the gamma speech model yields a greater improvement in the segmental SNR than the Gaussian speech model. Under Laplacian noise modeling, the gamma speech model results in lower residual musical noise than the Gaussian speech model. *Martin* and *Breithaupt* [44.45] showed that when modeling the real and imaginary parts of the clean speech spectral components as Laplacian random variables, the MMSE estimators for the complex speech spectral coefficients have similar properties to those estimators derived under gamma modeling, but are easier to compute and implement.

Breithaupt and *Martin* [44.68] derived, using the same statistical modeling, MMSE estimators for the magnitude-squared spectral coefficients, and compared their performance to that obtained by using a Gaussian speech model. They showed that improvement in the segmental SNR comes at the expense of additional residual musical noise. *Lotter* and *Vary* [44.69] derived a maximum a posteriori (MAP) estimator for the speech spectral amplitude, based on a Gaussian noise model and a superGaussian speech model. They proposed a parametric pdf for the speech spectral amplitude, which approximates, with a proper choice of the parameters, the gamma and Laplacian densities. Compared with the MMSE spectral amplitude estimator of Ephraim-Malah, the MAP estimator with Laplacian speech modeling demonstrates improved noise reduction.

The Gaussian, gamma, and Laplacian models presented in Sect. 44.3 take into account the time correlation between successive speech spectral components. Spectral components in the STFT domain are assumed to be statistically correlated along the frequency axis, as well as along time trajectories, due to the finite length of the analysis frame in the STFT and the overlap between successive frames [44.15]. Experimental results of speech enhancement performance show [44.16, 43] that the appropriateness of the Gaussian, gamma, and Laplacian speech models are greatly affected by the particular choice of the a priori SNR estimator. When the a priori SNR is estimated by the decision-directed method, the gamma model is more advantageous than the Gaussian model. However, when the a priori SNR is estimated by the noncausal recursive estimation method, the Laplacian speech model yields a higher segmental SNR and a lower LSD than the other speech models,

while the level of residual musical noise is minimal when using a Gaussian speech model. Furthermore, the differences between the Gaussian, gamma, and Laplacian speech models are smaller when using the noncausal a priori estimators than when using the decision-directed method.

It is worthwhile noting that estimators that minimize the MSE distortion of the spectral amplitude or log-spectral amplitude are more suitable for speech enhancement than MMSE estimators. Moreover, it is difficult, or even impossible, to derive analytical expressions for MMSE-LSA estimators under gamma or Laplacian models. Therefore, the MMSE-LSA estimator derived under a Gaussian model is often preferred over the MMSE estimators derived under the other speech models [44.2].

44.9.2 Choice of an A Priori SNR Estimator

Ephraim and *Malah* [44.12, 70] proposed three different methods for the a priori SNR estimation. First, maximum-likelihood estimation, which relies on the assumption that the speech spectral variances are slowly time-varying parameters. This results in musical residual noise, which is annoying and disturbing to the perception of the enhanced signal. Second, decision-directed approach which is particularly useful when combined with the MMSE spectral, or log-spectral, magnitude estimators [44.12, 17, 47]. This results in perceptually colorless residual noise, but is heuristically motivated and its theoretical performance is unknown due to its highly nonlinear nature. Third, maximum a posteriori estimation, which relies on a first-order Markov model for generating a sequence of speech spectral variances. It involves a set of nonlinear equations, which are solved recursively by using the Viterbi algorithm. The computational complexity of the MAP estimator is relatively high, while it does not provide a significant improvement in the enhanced speech quality over the decision-directed estimator [44.70].

The decision-directed approach has become over the last two decades the most acceptable estimation method for the variances of the speech spectral coefficients. However, the parameters of the decision-directed estimator have to be determined by simulations and subjective listening tests for each particular setup of time–frequency transformation and speech enhancement algorithm. Furthermore, since the decision-directed approach is not supported by a statistical model, the parameters are not adapted to the speech components, but are set to specific values in advance. *Ephraim* and *Malah* recognized the limits of their variance estimation methods, and concluded that better speech enhancement performance may be obtained if the variance estimation could be improved [44.12, 70].

The causal estimator for the a priori SNR combines two steps, a *propagation* step and an *update* step, following the rational of Kalman filtering, to predict and update the estimate for the speech spectral variance recursively as new data arrive. The causal a priori SNR estimator is closely related to the decision-directed estimator of Ephraim and Malah. A special case of the causal estimator degenerates to a *decision-directed* estimator with a *time-varying frequency-dependent* weighting factor. The weighting factor is monotonically decreasing as a function of the instantaneous SNR, resulting effectively in a larger weighting factor during speech absence, and a smaller weighting factor during speech presence. This slightly reduces both the musical noise and the signal distortion. Nevertheless, the improvement in speech enhancement performance obtained by using the causal recursive over using the decision-directed method is not substantial. Therefore, if the delay between the enhanced speech and the noisy observation needs to be minimized, the decision-directed method is perhaps preferable due to its computational simplicity. However, in applications where a few-frames delay is tolerable, e.g., digital voice recording, surveillance, and speaker identification, the *noncausal* recursive estimation approach is more advantageous than the decision-directed approach.

The noncausal a priori SNR estimator employs future spectral measurements to predict the spectral variances of clean speech better. A comparison of the causal and noncausal estimators indicates that the differences are primarily noticeable during speech onset. The *causal* a priori SNR estimator, as well as the decision-directed estimator, cannot respond too quickly to an abrupt increase in the instantaneous SNR, since this necessarily implies an increase in the level of musical residual noise. In contrast, the *noncausal* estimator, having a few subsequent spectral measurements at hand, is capable of discriminating between speech onsets and noise irregularities. Experimental results show that the advantages of the noncausal estimator are particularly perceived during onsets of speech and noise only frames. Onsets of speech are better preserved, while a further reduction of musical noise is achieved [44.15, 53]. Furthermore, the differences between the Gaussian, gamma, and Laplacian speech models are smaller when using the noncausal recursive estimation approach than when using the decision-directed method [44.43].

44.9.3 Choice of a Noise Estimator

Traditional noise estimation methods are based on recursive averaging during sections that do not contain speech and holding the estimates during sections which contain speech. However, these methods generally require VADs and the update of the noise estimate is restricted to periods of speech absence. Additionally, VADs are difficult to tune and their reliability severely deteriorates for weak speech components and low input SNR [44.65, 71, 72]. Alternative techniques, based on histograms in the power spectral domain [44.55, 73, 74], are computationally expensive, require much more memory resources, and do not perform well in low-SNR conditions. Furthermore, the signal segments used for building the histograms are typically several hundred milliseconds long, and thus the update rate of the noise estimate is essentially moderate.

A useful noise estimation approach known as the minimum statistics [44.59] tracks the minima values of a smoothed power estimate of the noisy signal, and multiplies the result by a factor that compensates the bias. However, the variance of this noise estimate is about twice as large as the variance of a conventional noise estimator [44.59]. Moreover, this method may occasionally attenuate low-energy phonemes, particularly if the minimum search window is too short [44.75]. These limitations can be overcome, at the price of higher complexity, by adapting the smoothing parameter and the bias compensation factor in time and frequency [44.56].

A computationally efficient minimum tracking scheme is presented in [44.57]. Its main drawbacks are the slow update rate of the noise estimate in the case of a sudden rise in the noise energy level, and its tendency to cancel the signal [44.71]. Other closely related techniques are the *lower-energy envelope tracking* [44.55] and the *quantile-based* [44.76] estimation methods. Rather than picking the minima values of a smoothed periodogram, the noise is estimated based on a temporal quantile of a nonsmoothed periodogram of the noisy signal. Unfortunately, these methods suffer from the high computational complexity associated with the sorting operation, and the extra memory required for keeping past spectral power values.

The IMCRA noise estimator [44.54], presented in Sect. 44.7, combines the robustness of the minimum tracking with the simplicity of recursive averaging. Rather than employing a voice activity detector and restricting the update of the noise estimator to periods of speech absence, the smoothing parameter is adapted in time and frequency according to the speech presence probability. The noise estimate is thereby continuously updated even during weak speech activity. The estimator is controlled by the minima values of a smoothed periodogram of the noisy measurement. It combines conditions on both the instantaneous and local measured power, and provides a soft transition between speech absence and presence. This prevents an occasional increase in the noise estimate during speech activity. Furthermore, carrying out the smoothing and minimum tracking in two iterations allows larger smoothing windows and smaller minimum search windows, while reliably tracking the minima even during strong speech activity. This yields a reduced variance of the minima values and shorter delay when responding to a rising noise power, which eventually improves the tracking capability of the noise estimator. In nonstationary noise environments and under low-SNR conditions, the IMCRA approach is particularly useful [44.54]. It facilitates a lower estimation error, and when integrated into a speech enhancement system, yields improved speech quality and lower residual noise.

44.10 Conclusions

We have described statistical models for speech and noise signals in the STFT domain, and presented estimators for the speech spectral coefficients under speech presence uncertainty. The statistical models take into consideration the time correlation between successive spectral components of the speech signal. The spectral estimators involve estimation of the noise power spectrum, calculation of the speech presence probability, and evaluation of the a priori SNR under speech presence uncertainty. We discussed the behavior of the MMSE-LSA spectral gain function and its advantage for the mechanism that counters the musical noise phenomenon. Local bursts of the a posteriori SNR during noise-only frames are pulled down to the average noise level, thus avoiding local buildup of noise whenever it exceeds its average characteristics. The estimator for the a priori speech presence probability exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames, which enables further attenuation of noise components while avoid-

ing clipping of speech onsets and misdetection of weak speech tails.

We have presented estimators for the a priori SNR under speech presence uncertainty, and showed that a special case of a causal recursive estimator degenerates to a decision-directed estimator with a time-varying frequency-dependent weighting factor. Furthermore, in applications where a delay of a few short-term frames between the enhanced speech and the noisy observation is tolerable, a noncausal estimation approach may produce less signal distortion and less musical residual noise than a causal estimation approach.

We described the IMCRA approach for the noise power spectrum estimation, and provided a detailed example of a speech enhancement algorithm. We showed that the improvement in SegSNR and reduction in LSD are influenced by the variability of the noise characteristics in time and the initial SegSNR and LSD of the noisy signal. The faster the noise spectrum varies in time, the less reliable the noise spectrum estimator, and consequently the lower the quality gain that can be achieved by the speech enhancement system. Furthermore, the lower the initial quality of the noisy signal, the larger the improvement that can be achieved by the speech enhancement system.

References

44.1 J. Benesty, S. Makino, J. Chen (Eds.): *Speech Enhancement* (Springer, Berlin, Heidelberg 2005)

44.2 Y. Ephraim, I. Cohen: Recent advancements in speech enhancement. In: *The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing*, 3rd edn., ed. by R.C. Dorf (CRC, Boca Raton 2006) pp. 15–12–15–26, Chap. 15

44.3 Y. Ephraim, H. Lev-Ari, W.J.J. Roberts: A brief survey of speech enhancement. In: *The Electronic Handbook*, 2nd edn. (CRC-Press, Boca Raton 2005)

44.4 S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)

44.5 J.S. Lim, A.V. Oppenheim: Enhancement and bandwidth compression of noisy speech, Proc. IEEE **67**(12), 1586–1604 (1979)

44.6 M. Berouti, R. Schwartz, J. Makhoul: Enhancement of speech corrupted by acoustic noise, Proc. 4th ICASSP **79**, 208–211 (1979)

44.7 Z. Goh, K.-C. Tan, T.G. Tan: Postprocessing method for suppressing musical noise generated by spectral subtraction, IEEE Trans. Speech Audio Process. **6**(3), 287–292 (1998)

44.8 B.L. Sim, Y.C. Tong, J.S. Chang, C.T. Tan: A parametric formulation of the generalized spectral subtraction method, IEEE Trans. Speech Audio Process. **6**(4), 328–337 (1998)

44.9 H. Gustafsson, S.E. Nordholm, I. Claesson: Spectral subtraction using reduced delay convolution and adaptive averaging, IEEE Trans. Speech Audio Process. **9**(8), 799–807 (2001)

44.10 D.E. Tsoukalas, J.N. Mourjopoulos, G. Kokkinakis: Speech enhancement based on audible noise suppression, IEEE Trans. Speech Audio Process. **5**(6), 497–514 (1997)

44.11 N. Virag: Single channel speech enhancement based on masking properties of the human auditory system, IEEE Trans. Speech Audio Process. **7**(2), 126–137 (1999)

44.12 Y. Ephraim, D. Malah: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. **ASSP-32**(6), 1109–1121 (1984)

44.13 D. Burshtein, S. Gannot: Speech enhancement using a mixture-maximum model, IEEE Trans. Speech Audio Process. **10**(6), 341–351 (2002)

44.14 R. Martin: Speech enhancement based on minimum mean-square error estimation and supergaussian priors, IEEE Trans. Speech Audio Process. **13**(5), 845–856 (2005)

44.15 I. Cohen: Relaxed statistical model for speech enhancement and a priori SNR estimation, IEEE Trans. Speech Audio Process. **13**(5), 870–881 (2005)

44.16 I. Cohen: Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models, Signal Process. **86**(4), 698–709 (2006)

44.17 Y. Ephraim, D. Malah: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. **ASSP-33**(2), 443–445 (1985)

44.18 P.J. Wolfe, S.J. Godsill: Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement, Special Issue EURASIP JASP Digital Audio Multim. Commun. **2003**(10), 1043–1051 (2003)

44.19 P.C. Loizou: Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum, IEEE Trans. Speech Audio Process. **13**(5), 857–869 (2005)

44.20 B.H. Juang, L.R. Rabiner: Mixture autoregressive hidden Markov models for speech signals, IEEE Trans. Acoust. Speech Signal Process. **ASSP-33**(6), 1404–1413 (1985)

44.21 Y. Ephraim: Statistical-model-based speech enhancement systems, Proc. IEEE **80**(10), 1526–1555 (1992)

44.22 H. Sheikhzadeh, L. Deng: Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization, IEEE Trans. Speech Audio Process. **2**, 80–91 (1994)

44.23 Y. Ephraim, N. Merhav: Hidden Markov processes, IEEE Trans. Inform. Theory **48**(6), 1518–1568 (2002)

44.24 C.J. Wellekens: Explicit time correlations in hidden Markov models for speech recognition, Proc. 12th ICASSP **87**, 384–386 (1987)

44.25 H. Sameti, H. Sheikhzadeh, L. Deng, R.L. Brennan: HMM-based strategies for enhancement of speech signals embedded in nonstationary noise, IEEE Trans. Speech Audio Process. **6**(5), 445–455 (1998)

44.26 L.R. Rabiner, B.-H. Juang: *Fundamentals of Speech Recognition* (Prentice-Hall, Upper Saddle River 1993)

44.27 F. Jelinek: *Statistical Methods for Speech Recognition* (MIT Press, Cambridge 1998)

44.28 Y. Ephraim, H.L.V. Trees: A signal subspace approach for speech enhancement, IEEE Trans. Speech Audio Process. **3**(4), 251–266 (1995)

44.29 F. Asano, S. Hayamizu, T. Yamada, S. Nakamura: Speech enhancement based on the subspace method, IEEE Trans. Speech Audio Process. **8**(5), 497–507 (2000)

44.30 U. Mittal, N. Phamdo: Signal/noise KLT based approach for enhancing speech degraded by colored noise, IEEE Trans. Speech Audio Process. **8**(2), 159–167 (2000)

44.31 Y. Hu, P.C. Loizou: A generalized subspace approach for enhancing speech corrupted by colored noise, IEEE Trans. Speech Audio Process. **11**(4), 334–341 (2003)

44.32 S.H. Jensen, P.C. Hansen, S.D. Hansen, J.A. Sørensen: Reduction of broad-band noise in speech by truncated QSVD, IEEE Trans. Speech Audio Process. **3**(6), 439–448 (1995)

44.33 S. Doclo, M. Moonen: GSVD-based optimal filtering for single and multimicrophone speech enhancement, IEEE Trans. Signal Process. **50**(9), 2230–2244 (2002)

44.34 F. Jabloun, B. Champagne: Incorporating the human hearing properties in the signal subspace approach for speech enhancement, IEEE Trans. Speech Audio Process. **11**(6), 700–708 (2003)

44.35 Y. Hu, P.C. Loizou: A perceptually motivated approach for speech enhancement, IEEE Trans. Speech Audio Process. **11**(5), 457–465 (2003)

44.36 J. Wexler, S. Raz: Discrete Gabor expansions, Speech Process. **21**(3), 207–220 (1990)

44.37 R.E. Crochiere, L.R. Rabiner: *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs 1983)

44.38 J.S. Garofolo: *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database* (NIST, Gaithersburg 1988)

44.39 A. Stuart, J.K. Ord: *Kendall's Advanced Theory of Statistics*, Vol. 1, 6th edn. (Arnold, London 1994)

44.40 R. Martin: Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors, Proc. 27th ICASSP **02**, 253–256 (2002)

44.41 I. Cohen: Modeling speech signals in the time-frequency domain using GARCH, Signal Process. **84**(12), 2453–2459 (2004)

44.42 I. Cohen, B. Berdugo: Speech enhancement for non-stationary noise environments, Signal Process. **81**(11), 2403–2418 (2001)

44.43 I. Cohen: Speech enhancement using supergaussian speech models and noncausal a priori SNR estimation, Speech Commun. **47**(3), 336–350 (2005)

44.44 I.S. Gradshteyn, I.M. Ryzhik: *Table of Integrals, Series, and Products*, 4th edn. (Academic Press, New York 1980)

44.45 R. Martin, C. Breithaupt: Speech enhancement in the DFT domain using Laplacian speech priors. In: *Proc. 8th Int. Workshop on Acoustic Echo and Noise Control* (Kyoto, Japan 2003) pp. 87–90

44.46 J. Porter, S. Boll: Optimal estimators for spectral restoration of noisy speech, Proc. ICASSP **84**, 18A.2.1–18A.2.4 (1984)

44.47 O. Cappé: Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Trans. Acoust. Speech Signal Process. **2**(2), 345–349 (1994)

44.48 P. Scalart, J. Vieira-Filho: Speech enhancement based on a priori signal to noise estimation, Proc. 21th ICASSP **96**, 629–632 (1996)

44.49 D. Malah, R.V. Cox, A.J. Accardi: Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments, Proc. 24th ICASSP **99**, 789–792 (1999)

44.50 I. Cohen: On speech enhancement under signal presence uncertainty, Proc. 26th ICASSP **2001**, 167–170 (2001)

44.51 I.Y. Soon, S.N. Koh, C.K. Yeo: Improved noise suppression filter using self-adaptive estimator of probability of speech absence, Signal Process. **75**(2), 151–159 (1999)

44.52 M. Marzinzik: *Noise reduction schemes for digital hearing aids and their use for the hearing impaired*, Ph.D. Thesis (Oldenburg University, Oldenburg 2000)

44.53 I. Cohen: Speech enhancement using a noncausal a priori SNR estimator, IEEE Signal Process. Lett. **11**(9), 725–728 (2004)

44.54 I. Cohen: Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging, IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)

44.55 C. Ris, S. Dupont: Assessing local noise level estimation methods: Application to noise robust ASR, Speech Commun. **34**(1–2), 141–158 (2001)

44.56 R. Martin: Noise power spectral density estimation based on optimal smoothing and minimum statis-

tics, IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
44.57 G. Doblinger: Computationally efficient speech enhancement by spectral minima tracking in subbands, Proc. 4th Eurospeech **95**, 1513–1516 (1995)
44.58 S. Qian, D. Chen: Discrete Gabor transform, IEEE Trans. Signal Process. **41**(7), 2429–2438 (1993)
44.59 R. Martin: Spectral subtraction based on minimum statistics, Proc. 7th EUSIPCO **94**, 1182–1185 (1994)
44.60 A. Varga, H.J.M. Steeneken: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. **12**(3), 247–251 (1993)
44.61 S.R. Quackenbush, T.P. Barnwell, M.A. Clements: *Objective Measures of Speech Quality* (Prentice-Hall, Englewood Cliffs 1988)
44.62 J.R. Deller, J.H.L. Hansen, J.G. Proakis: *Discrete-Time Processing of Speech Signals*, 2nd edn. (IEEE, New York 2000)
44.63 P.E. Papamichalis: *Practical Approaches to Speech Coding* (Prentice-Hall, Englewood Cliffs 1987)
44.64 A.J. Accardi, R.V. Cox: A modular approach to speech enhancement with an application to speech coding, Proc. 24th ICASSP **99**, 201–204 (1999)
44.65 J. Sohn, N.S. Kim, W. Sung: A statistical model-based voice activity detector, IEEE Signal Process. Lett. **6**(1), 1–3 (1999)
44.66 T. Lotter, C. Benien, P. Vary: Multichannel speech enhancement using bayesian spectral amplitude estimation, Proc. 28th ICASSP **03**, 832–835 (2003)
44.67 J.W.B. Davenport: *Probability and Random Processes: An Introduction for Applied Scientists and Engineers* (McGraw-Hill, New York 1970)
44.68 C. Breithaupt, R. Martin: MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors, Proc. 28th ICASSP **03**, 896–899 (2003)
44.69 T. Lotter, P. Vary: Noise reduction by maximum a posteriori spectral amplitude estimation with supergaussian speech modeling. In: *Proc. 8th Internat. Workshop on Acoustic Echo and Noise Control* (2003) pp. 83–86
44.70 Y. Ephraim, D. Malah: *Signal to Noise Ratio Estimation for Enhancing Speech Using the Viterbi Algorithm*, Tech. Rep. EE PUB 489 (Technion – Israel Institute of Technology, Haifa 1984)
44.71 J. Meyer, K.U. Simmer, K.D. Kammeyer: Comparison of one- and two-channel noise-estimation techniques, Proc. 5th IWAENC **97**, 137–145 (1997)
44.72 B.L. McKinley, G.H. Whipple: Model based speech pause detection, Proc. 22th ICASSP **97**, 1179–1182 (1997)
44.73 R.J. McAulay, M.L. Malpass: Speech enhancement using a soft-decision noise suppression filter, IEEE Trans. Acoust. Speech Signal Process. **ASSP-28**(2), 137–145 (1980)
44.74 H.G. Hirsch, C. Ehrlicher: Noise estimation techniques for robust speech recognition, Proc. 20th ICASSP **95**, 153–156 (1995)
44.75 I. Cohen, B. Berdugo: Speech enhancement for non-stationary noise environments, Signal Process. **81**(11), 2403–2418 (2001)
44.76 V. Stahl, A. Fischer, R. Bippus: Quantile based noise estimation for spectral subtraction and Wiener filtering, Proc. 25th ICASSP **2000**, 1875–1878 (2000)