Chapter 5
Simultaneous Detection and Estimation Approach for Speech Enhancement and Interference Suppression

Ari Abramson and Israel Cohen

Abstract [1]In this chapter, we present a simultaneous detection and estimation approach for speech enhancement. A detector for speech presence in the short-time Fourier transform domain is combined with an estimator, which jointly minimizes a cost function that takes into account both detection and estimation errors. Cost parameters control the trade-off between speech distortion, caused by missed detection of speech components, and residual musical noise resulting from false-detection. Furthermore, a modified decision-directed *a priori* signal-to-noise ratio (SNR) estimation is proposed for transient-noise environments. Experimental results demonstrate the advantage of using the proposed simultaneous detection and estimation approach with the proposed *a priori* SNR estimator, which facilitate suppression of transient noise with a controlled level of speech distortion.

5.1 Introduction

In many signal processing applications as well as communication applications, the signal to be estimated is not surely present in the available noisy observation. Therefore, algorithms often try to estimate the signal under uncertainty by using some *a priori* probability for the existence of the signal, e.g., [1, 2, 3, 4], or alternatively, apply an independent detector for signal presence, e.g., [5, 6, 7, 8, 9, 10]. The detector may be designed based on the noisy observation, or, on the estimated signal. Considering speech sig-

Ari Abramson
Technion–Israel Institute of Technology, Israel, e-mail: aari@tx.technion.ac.il

Israel Cohen
Technion–Israel Institute of Technology, Israel e-mail: icohen@ee.technion.ac.il

[1] This work was supported by the Israel Science Foundation under Grant 1085/05 and by the European Commission under project Memories FP6-IST-035300.

I. Cohen et al. (Eds.): Speech Processing in Modern Communication, STSP 3, pp. 127–150.
springerlink.com

nals, the spectral coefficients are generally sparse in the short-time Fourier transform (STFT) domain in the sense that speech is present only in some of the frames, and in each frame only some of the frequency-bins contain the significant part of the signal energy. Therefore, both signal estimation and detection are generally required while processing noisy speech signals. The well-known spectral-subtraction algorithm [11, 12] contains an elementary detector for speech activity in the time-frequency domain, but it generates musical noise caused by falsely detecting noise peaks as bins that contain speech, which are randomly scattered in the STFT domain. Subspace approaches for speech enhancement [13, 14, 15, 16] decompose the vector of the noisy signal into a signal-plus-noise subspace and a noise subspace, and the speech spectral coefficients are estimated after removing the noise subspace. Accordingly, these algorithms are aimed at detecting the speech coefficients and subsequently estimating their values. McAulay and Malpass [2] were the first to propose a speech spectral estimator under a two-state model. They derived a maximum likelihood (ML) estimator for the speech spectral amplitude under speech-presence uncertainty. Ephraim and Malah followed this approach of signal estimation under speech presence uncertainty and derived an estimator which minimizes the mean-squared error (MSE) of the short-term spectral amplitude (STSA) [3]. In [17], speech presence probability is evaluated to improve the minimum MSE (MMSE) of the LSA estimator, and in [4] a further improvement of the MMSE-LSA estimator is achieved based on a two-state model.

Middleton *et al.* [18, 19] were the first to propose simultaneous signal detection and estimation within the framework of statistical decision theory. This approach was recently generalized to speech enhancement, as well as single sensor audio source separation [20, 21]. The speech enhancement problem is formulated by incorporating simultaneous operations of detection and estimation. A detector for the speech coefficients is combined with an estimator, which jointly minimizes a cost function that takes into account both estimation and detection errors. Under speech-presence, the cost is proportional to some distortion between the desired and estimated signals, while under speech-absence, the distortion depends on a certain attenuation factor [12, 4, 22]. A combined detector and estimator enables to control the trade-off between speech distortion, caused by missed detection of speech components, and residual musical noise resulting from false-detection. The combined solutions generalize standard algorithms, which involve merely estimation under signal presence uncertainty.

In some speech processing applications, an indicator for the transient noise activity may be available, *e.g.*, a siren noise in an emergency car, lens-motor noise of a digital video camera or a keyboard typing noise in a computer-based communication system. The transient spectral variances can be estimated in such cases from training signals. However, applying a standard estimator to the spectral coefficients may result in removal of critical speech components in case of falsely detecting the speech components, or under-suppression of

transient noise in case of miss detecting the noise transients. For cases where some indicator (or detector) for the presence of noise transients in the STFT domain is available, the speech enhancement problem is reformulated using two hypotheses. Cost parameters control the trade-off between speech distortion and residual transient noise. The optimal signal estimator is derived which employs the available detector. The resulting estimator generalizes the optimally-modified log-spectral amplitude (OM-LSA) estimator [4].

This chapter is organized as follows. In Section 5.2, we briefly review classical speech enhancement under signal presence uncertainty. In Section 5.3, the speech enhancement problem is reformulated as a simultaneous detection and estimation problem in the STFT domain. A detector for the speech coefficients is combined with an estimator, which jointly minimizes a cost function that takes into account both estimation and detection errors. The combined solution is derived for the quadratic distortion measure as well as the quadratic spectral amplitude distortion measure. In Section 5.4, we consider the integration of a spectral estimator with a *given* detector for noise transients and derive an optimal estimator which minimizes the mean-square error of the log-spectral amplitude. In Section 5.5, a modification of the decision-directed *a priori* signal-to-noise ratio (SNR) estimator is presented which better suites transient-noise environments. Experimental results are given in Section 5.6. It shows that the proposed approaches facilitate improved noise reduction with a controlled level of speech distortion.

5.2 Classical Speech Enhancement in Nonstationary Noise Environments

Let us start with a short presentation of classical approach for spectral speech enhancement while considering nonstationary noise environments. Specifically, we may assume that some indicator for transient noise activity is available.

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, and let $y(n) = x(n) + d(n)$ be the observed signal. Applying the STFT to the observed signal, we have

$$Y_{\ell k} = X_{\ell k} + D_{\ell k}\,, \tag{5.1}$$

where $\ell = 0, 1, ...$ is the time frame index and $k = 0, 1, ..., K-1$ is the frequency-bin index. Let $H_1^{\ell k}$ and $H_0^{\ell k}$ denote, respectively, speech presence and absence hypotheses in the time-frequency bin (ℓ, k), *i.e.*,

$$\begin{aligned} H_1^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k}\,, \\ H_0^{\ell k} &: Y_{\ell k} = D_{\ell k}\,. \end{aligned} \tag{5.2}$$

The noise expansion coefficients can be represented as the sum of two uncorrelated noise components $D_{\ell k} = D^s_{\ell k} + D^t_{\ell k}$, where $D^s_{\ell k}$ denotes a quasi-stationary noise component and $D^t_{\ell k}$ denotes a highly nonstationary transient component. The transient components are generally rare, but they may be of high energy and thus cause significant degradation to speech quality and intelligibility. However, in many applications, a reliable indicator for the transient noise activity may be available in the system. For example, in an emergency car (*e.g.*, police or ambulance) the engine noise may be considered as quasi-stationary, but activating a siren results in a highly nonstationary noise which is perceptually very annoying. Since the sound generation in the siren is nonlinear, linear echo cancelers may be inappropriate. In a computer-based communication system, a transient noise such as a keyboard typing noise may be present in addition to quasi-stationary background office noise. Another example is a digital camera, where activating the lens-motor (zooming in/out) may result in high-energy transient noise components, which degrade the recorded audio. In the above examples, an indicator for the transient noise activity may be available, *i.e.*, siren source signal, keyboard output signal and the lens-motor controller output. Furthermore, given that a transient noise source is active, a detector for the transient noise in the STFT domain may be designed and its spectrum can be estimated based on training data.

The objective of a speech enhancement system is to reconstruct the spectral coefficients of the speech signal such that under speech-presence a certain distortion measure between the spectral coefficient and its estimate, $d\left(X_{\ell k}, \hat{X}_{\ell k}\right)$, is minimized. Under speech-absence a constant attenuation of the noisy coefficient would be desired to maintain a natural background noise [22, 4]. Although the speech expansion coefficients are not necessarily present, most classical speech enhancement algorithms try to estimate the spectral coefficients rather than detecting their existence, or try to independently design detectors and estimators. The well-known spectral subtraction algorithm estimates the speech spectrum by subtracting the estimated noise spectrum from the noisy squared absolute coefficients [11, 12], and thresholding the result by some desired residual noise level. Thresholding the spectral coefficients is in fact a detection operation in the time-frequency domain, in the sense that speech coefficients are assumed to be absent in the low-energy time-frequency bins and present in noisy coefficients whose energy is above the threshold.

McAulay and Malpass were the first to propose a two-state model for the speech signal in the time-frequency domain [2]. Accordingly, the MMSE estimator follows

$$\begin{aligned} \hat{X}_{\ell k} &= E\left\{X_{\ell k} \,|\, Y_{\ell k}\right\} \\ &= E\left\{X_{\ell k} \,|\, Y_{\ell k}, H_1^{\ell k}\right\} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) . \end{aligned} \tag{5.3}$$

The resulting estimator does not detect speech components, but rather, a soft-decision is performed to further attenuate the signal estimate by the *a posteriori* speech presence probability. Ephraim and Malah followed the same approach and derived an estimator which minimizes the MSE of the STSA under signal presence uncertainty [3]. Accordingly,

$$\left|\hat{X}_{\ell k}\right| = E\left\{|X_{\ell k}| \mid Y_{\ell k}, H_1^{\ell k}\right\} p\left(H_1^{\ell k} \mid Y_{\ell k}\right) . \tag{5.4}$$

Both in [2] and [3], under $H_0^{\ell k}$ the speech components are assumed zero and the *a priori* probability of speech presence is both time and frequency invariant, *i.e.*, $p\left(H_1^{\ell k}\right) = p\left(H_1\right)$. In [17, 4], the speech presence probability is evaluated for each frequency-bin and time-frame to improve the performance of the MMSE-LSA estimator [23]. Further improvement of the MMSE-LSA suppression rule can be achieved by considering under $H_0^{\ell k}$ a constant attenuation factor $G_f << 1$, which is determined by subjective criteria for residual noise naturalness, see also [22]. The OM-LSA estimator [4] is given by

$$\left|\hat{X}_{\ell k}\right| = \left(\exp\left[E\left\{\log|X_{\ell k}| \mid Y_{\ell k}, H_1^{\ell k}\right\}\right]\right)^{p\left(H_1^{\ell k} \mid Y_{\ell k}\right)} \left(G_f \ |Y_{\ell k}|\right)^{1-p\left(H_1^{\ell k} \mid Y_{\ell k}\right)} . \tag{5.5}$$

Suppose that an indicator for the presence of transient noise components is available in a highly nonstationary noise environment, then high-energy transients may be attenuated by using one of the above-mentioned estimators (5.3)–(5.5) and heuristically setting the *a priori* speech presence probability $p\left(H_1^{\ell k}\right)$ to a sufficiently small value. Unfortunately, this also results in suppression of desired speech components and intolerable degradation of speech quality. In general, an estimation-only approach under signal presence uncertainty produces larger speech degradation for small $p\left(H_1^{\ell k}\right)$, since the optimal estimate is attenuated by the *a posteriori* speech presence probability. On the other hand, increasing $p\left(H_1^{\ell k}\right)$ prevents the estimator from sufficiently attenuating noise components. Integrating a jointly optimal detector and estimator into the speech enhancement system may significantly improve the speech enhancement performance under highly non-stationary noise conditions and may allow further reduction of transient components without much degradation of the desired signal.

5.3 Simultaneous Detection and Estimation for Speech Enhancement

Middleton and Esposito [18] were the first to propose simultaneous signal detection and estimation within the framework of statistical decision theory. This approach was generalized to speech enhancement formalism in [20]. A decision space, $\left\{\eta_0^{\ell k}, \eta_1^{\ell k}\right\}$, is assumed for the detection operation where under

the decision $\eta_j^{\ell k}$, signal hypothesis $H_j^{\ell k}$ is accepted and a corresponding estimate $\hat{X}_{\ell k} = \hat{X}_{\ell k,j}$ is considered. The detection and estimation are strongly coupled so that the detector is optimized with the knowledge of the specific structure of the estimator, and the estimator is optimized in the sense of minimizing a Bayesian risk associated with the combined operations. For notation simplification, we omit the time-frequency indices (ℓ, k). Let

$$C_j\left(X, \hat{X}\right) \geq 0 \tag{5.6}$$

denote the cost of making a decision η_j (and choosing an estimator $\hat{X}_j$) where X is the desired signal. Then, the Bayes risk of the two operations associated with simultaneous detection and estimation is defined by [18, 19, 20]

$$R = \sum_{j=0}^{1} \int_{\Omega_y} \int_{\Omega_x} C_j\left(X, \hat{X}\right) p\left(\eta_j \,|\, Y\right) p\left(Y \,|\, X\right) p\left(X\right) dX dY\,, \tag{5.7}$$

where Ω_x and Ω_y are the spaces of the speech and noisy signals, respectively. The simultaneous detection and estimation approach is aimed at jointly minimizing the Bayes risk over both the decision rule and the corresponding signal estimate. Let $q \triangleq p\left(H_1\right)$ denote the *a priori* speech presence probability. Then, the *a priori* distribution of the speech expansion coefficient follows

$$p\left(X\right) = q\, p\left(X \,|\, H_1\right) + \left(1 - q\right) p\left(X \,|\, H_0\right)\,, \tag{5.8}$$

where $p\left(X \,|\, H_0\right) = \delta\left(X\right)$ denotes the Dirac-delta function. The cost function $C_j\left(X, \hat{X}\right)$ may be defined differently whether H_1 or H_0 is true. Therefore, we let

$$C_{ij}\left(X, \hat{X}\right) \triangleq C_j\left(X, \hat{X} \,|\, H_i\right) \tag{5.9}$$

denote the cost which is conditioned on the true hypothesis[2]. The cost function $C_{ij}\left(X, \hat{X}\right)$ depends on both the true signal value and its estimate under the decision η_j and therefore couples the operations of detection and estimation. By substituting (5.8) and (5.9) into (5.7) we obtain [20]

$$\begin{aligned} R = & \int_{\Omega_y} p\left(\eta_0 \,|\, Y\right) \left[\, q\, r_{10}\left(Y\right) + \left(1 - q\right) r_{00}\left(Y\right)\right] dY \\ & + \int_{\Omega_y} p\left(\eta_1 \,|\, Y\right) \left[\, q\, r_{11}\left(Y\right) + \left(1 - q\right) r_{01}\left(Y\right)\right] dY\,, \end{aligned} \tag{5.10}$$

where

$$r_{ij}\left(Y\right) = \int_{\Omega_x} C_{ij}\left(X, \hat{X}\right) p\left(X \,|\, H_i\right) p\left(Y \,|\, X\right) dX \tag{5.11}$$

[2] Note that $X = 0$ implies that H_0 is true and $X \neq 0$ implies H_1 so the sub-index i may seem to be redundant. However, this notation simplifies the subsequent formulations.

denotes the risk associated with the pair $\{H_i, \eta_j\}$ and the observation Y.

Since the detector's decision under a given observation is binary, *i.e.,* $p(\eta_j \,|\, Y) \in \{0, 1\}$, for minimizing the combined risk we first evaluate the optimal estimator under each of the decisions, then the optimal decision rule is derived based on the optimal estimators $\hat{X}_0$, $\hat{X}_1$ to further minimize the combined risk. The two-stage minimization guaranties minimum combined risk [19]. The optimal *nonrandom* decision rule which minimizes the combined risk (5.10) is given by

Decide η_1 (*i.e.,* $p(\eta_1 \,|\, Y) = 1$) if

$$q\left[r_{10}(Y) - r_{11}(Y)\right] \geq (1-q)\left[r_{01}(Y) - r_{00}(Y)\right] \ , \tag{5.12}$$

otherwise, decide η_0.

The optimal estimator under a decision η_j is obtained from (5.10) by

$$\arg\min_{\hat{X}_j} \left\{ q\, r_{1j}(Y) + (1-q)\, r_{0j}(Y) \right\} . \tag{5.13}$$

Note that $r_{ij}(Y)$ depends on the estimate $\hat{X}_j$ through the cost function.

The cost function associated with the pair $\{H_i,\, \eta_j\}$ is generally defined by

$$C_{ij}\left(X, \hat{X}\right) = b_{ij}\, d_{ij}\left(X, \hat{X}\right) \ , \tag{5.14}$$

where $d_{ij}\left(X, \hat{X}\right)$ is an appropriate distortion measure and the cost parameters b_{ij} control the trade-off between the costs associated with the pairs $\{H_i, \eta_j\}$. That is, a high valued b_{01} raises the cost of a false alarm (*i.e.,* decision of speech presence when speech is actually absent), which may result in residual musical noise. Similarly, b_{10} is associated with the cost of missed detection of a signal component, which may cause perceptual signal distortion. Under a correct classification, normalized cost parameters are generally used, $b_{00} = b_{11} = 1$. However, $d_{ii}(\cdot, \cdot)$ is not necessarily zero since estimation errors are still possible even when there is no detection error.

Contrary to the approach in [18, 19, 24], in speech enhancement application the signal is not rejected when a decision η_0 is made. Instead, the estimator $\hat{X}_0 \neq 0$ compensates for any detection errors to reduce potential musical noise and audible distortions. Furthermore, when speech is indeed absent the distortion function is defined to allow some natural background noise level such that under H_0 the attenuation factor will be lower bounded by a constant gain floor $G_f << 1$ as proposed in [12, 4, 25, 22].

In the following subsections we specify two distortions measures which are of interest for spectral speech enhancement and derive their combined solution of detection and estimation.

5.3.1 Quadratic Distortion Measure

The distortion measure associated with the MMSE estimation is the quadratic distortion measure which follows:

$$d_{ij}\left(X,\hat{X}\right)=\begin{cases}\left|X-\hat{X}_j\right|^2, & i=1\\ \left|G_f Y-\hat{X}_j\right|^2, & i=0\end{cases}. \tag{5.15}$$

We assume that both X and D are statistically independent, zero-mean, complex-valued Gaussian random variables with variances λ_x and λ_d, respectively. Let $\xi \triangleq \lambda_x/\lambda_d$ denote the *a priori* SNR under hypothesis H_1, let $\gamma \triangleq |Y|^2/\lambda_d$ denote the *a posteriori* SNR and let $\upsilon \triangleq \gamma\,\xi/(1+\xi)$. Then, the optimal estimation under the quadratic cost function is obtained by substituting (5.15) and (5.14) into (5.13) and set its derivation to zero [19, 21]

$$\begin{aligned}\hat{X}_j &= \left[b_{1j}\,\Lambda(\xi,\gamma)\,G_{MSE}(\xi)+b_{0j}\,G_f\right]\phi_j(\xi,\gamma)^{-1}\,Y\\ &\triangleq G_j\,Y\,, \qquad j=0,1\,,\end{aligned} \tag{5.16}$$

where $G_{MSE}(\xi)=\left(1+\xi^{-1}\right)^{-1}$ is the MSE gain function under signal presence, $\Lambda(\xi,\gamma)$ is the generalized likelihood ratio

$$\begin{aligned}\Lambda(\xi,\gamma) &= \frac{q}{(1-q)}\frac{p(Y\,|\,H_1)}{p(Y\,|\,H_0)}\\ &= \frac{q}{(1-q)}\frac{e^{\upsilon}}{1+\xi},\end{aligned} \tag{5.17}$$

and the normalization factor $\phi_j(\xi,\gamma)$ is given by

$$\phi_j(\xi,\gamma)=b_{0j}+b_{1j}\,\Lambda(\xi,\gamma)\,. \tag{5.18}$$

The risk $r_{ij}(Y)$, required for the detector decision is obtained by substituting the quadratic cost function (5.15) into (5.11). Let X_R and X_I denote the real and imaginary parts of the expansion coefficient X. Then, for the case where H_1 is true we have

$$\begin{aligned}&r_{1j}(Y)=\\ &\quad b_{1j}\int_{-\infty}^{\infty}\int\Big[X_R^2+X_I^2+G_j^2(\xi,\gamma)\,|Y|^2-2G_j(\xi,\gamma)\,(X_RY_R+X_IY_I)\Big]\\ &\quad\times p(X_R,X_I\,|\,H_1)\,p(Y\,|\,X_R,X_I)\,dX_R\,dX_I\,,\end{aligned} \tag{5.19}$$

where

$$p\left(X_R, X_I \mid H_1\right) = \frac{1}{\pi\lambda_x}\exp\left(-\frac{|X|^2}{\lambda_x}\right), \tag{5.20}$$

$$p\left(Y \mid X_R, X_I\right) = \frac{1}{\pi\lambda_d}\exp\left(-\frac{|Y-X|^2}{\lambda_d}\right). \tag{5.21}$$

By using [26, eq. 3.323.2] we have

$$\int\int_{-\infty}^{\infty} p\left(X_R, X_I \mid H_1\right) p\left(Y \mid X_R, X_I\right) dX_R\, dX_I = \frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{\pi\lambda_d\left(1+\xi\right)}, \tag{5.22}$$

and using [26, eq. 3.462.2] we obtain

$$\int\int_{-\infty}^{\infty}\left(X_R Y_R + X_I Y_I\right) p\left(X_R, X_I \mid H_1\right) p\left(Y \mid X_R, X_I\right) dX_R\, dX_I = \frac{\upsilon}{\pi(1+\xi)}\exp\left(-\frac{\gamma}{1+\xi}\right), \tag{5.23}$$

and

$$\int\int_{-\infty}^{\infty}\left(X_R^2 + X_I^2\right) p\left(X_R, X_I \mid H_1\right) p\left(Y \mid X_R, X_I\right) dX_R\, dX_I = \frac{\xi(1+\upsilon)}{\pi(1+\xi)^2}\exp\left(-\frac{\gamma}{1+\xi}\right). \tag{5.24}$$

Substituting (5.22)–(5.24) into (5.19) we obtain

$$r_{1j}\left(Y\right) = \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi}\left[G_{MSE}\left(\xi\right) + \left(G_j\left(\xi,\gamma\right) - G_{MSE}\left(\xi\right)\right)^2\gamma\right], \tag{5.25}$$

where $G_j\left(\xi,\gamma\right)$ is defined by (5.16).

Under hypothesis H_0 speech is absent and $p\left(X_R, X_I \mid H_0\right) = \delta\left(X_R, X_I\right)$. Consequently,

$$\begin{aligned} r_{0j}\left(Y\right) &= b_{0j}\int\int_{-\infty}^{\infty}\left[\left(G_j\left(\xi,\gamma\right) - G_f\right)^2|Y|^2\right] \\ &\quad \times p\left(X_R, X_I \mid H_i\right) p\left(Y \mid X_R, X_I\right) dX_R\, dX_I \\ &= \frac{b_{0j}}{\pi}\left[G_j\left(\xi,\gamma\right) - G_f\right]^2\gamma\, e^{-\gamma}. \end{aligned} \tag{5.26}$$

By substituting (5.25) and (5.26) into (5.12) we obtain the optimal decision rule for the detection operation under the quadratic distortion measure:

$$\frac{(1-q)(1+\xi)}{q\, e^{\upsilon}}\left[\left(G_0 - G_f\right)^2 - b_{01}\left(G_1 - G_f\right)^2\right] \underset{\eta_0}{\overset{\eta_1}{\gtrless}} \frac{1-b_{10}}{\gamma}G_{MSE} + \left(G_1 - G_{MSE}\right)^2 - b_{10}\left(G_0 - G_{MSE}\right)^2. \tag{5.27}$$

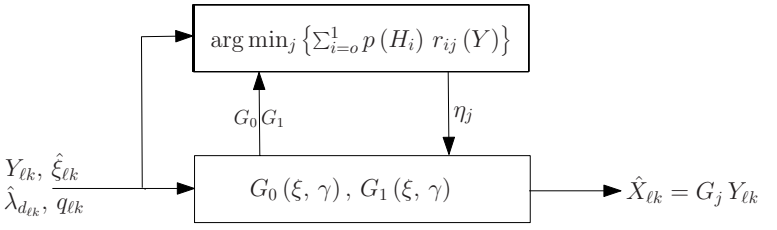

Fig. 5.1 Block diagram of the simultaneous detection and estimation.

Here, G_j, $j = 0, 1$ and G_{MSE} hold for $G_j\left(\xi,\gamma\right)$ and $G_{MSE}\left(\xi\right)$, respectively.

To conclude, a block diagram of the signal simultaneous detection and estimation from a noisy observation is shown in Fig. 5.1. The gain factor G_j is calculated under any of the decisions using (5.16). Then, the optimal decision η_j is calculated using (5.27) and accordingly, the estimated signal is given by applying selected gain to the noisy signal.

It is of interest to examine the asymptotic behavior of the estimator (5.16) under each of the decisions made by the detector. When the cost parameter associated with false alarm is much smaller than the generalized likelihood ratio, *i.e.*, $b_{01} << \Lambda\left(\xi,\gamma\right)$, the spectral gain under the decision η_1 implies $G_1\left(\xi,\gamma\right) \cong G_{MSE}\left(\xi\right)$ which is the optimal estimation under no uncertainty. However, if $b_{01} >> \Lambda\left(\xi,\gamma\right)$, the spectral gain under η_1 needs to compensate the possibility of a high-cost false-decision made by the detector and thus $G_1\left(\xi,\gamma\right) \cong G_f$. On the other hand, if the cost parameter associated with missed detection is small and we have $b_{10} << \Lambda\left(\xi,\gamma\right)^{-1}$, than $G_0\left(\xi,\gamma\right) \cong G_f$ (*i.e.*, estimation where speech is surely absence) but under $b_{10} >> \Lambda\left(\xi,\gamma\right)^{-1}$, in order to overcome the high cost related to missed detection, we have $G_0\left(\xi,\gamma\right) \cong G_{MSE}\left(\xi\right)$.

Recall that

$$\frac{\Lambda\left(\xi,\gamma\right)}{1+\Lambda\left(\xi,\gamma\right)} = p\left(H_1 \,|\, Y\right) \tag{5.28}$$

is the *a posteriori* probability of speech presence, it can be seen that the proposed estimator (5.16) generalizes existing estimators. For the case of equal parameters $b_{ij} = 1\ \forall i, j$ and $G_f = 0$ we get the estimation under signal presence uncertainty (5.3). In that case the detection operation is not needed since the estimation is independent of the detection rule.

Figure 5.2 shows attenuation curves under quadratic cost function as a function of the *a priori* SNR, ξ, with *a posteriori* SNR of $\gamma = 5$ dB, $q = 0.8$, $G_f = -15$ dB and cost parameters $b_{01} = 4$, $b_{10} = 2$. In (a), the gains G_1 (dash line), G_0 (dotted line) and the total detection and estimation system gain G (solid line) are shown and compared with (b) the MSE gain function under no uncertainty G_{MSE} (dashed line) and the MMSE estimation under signal presence uncertainty which is defined by (5.3) (dashed line). It can be seen that for *a priori* SNRs higher than about -10 dB the detector decision is η_1

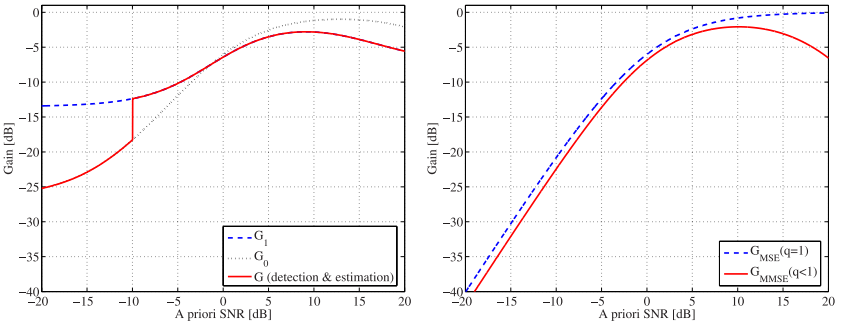

Fig. 5.2 Gain curves under quadratic cost function with $\gamma = 5$ [dB], $q = 0.8$, and $G_f = -15$, [dB]; (a) G_1, G_0 and the detection and estimation gain G with $b_{01} = 4$, $b_{10} = 2$, and (b) G_{MSE} gain curve with $q = 1$ and the MMSE gain curve under uncertainty ($q = 0.8$).

and therefore the total gain is $G = G_1$. For lower *a priori* SNRs, η_0 is decided and consequently the total gain is G_0. Note that if an ideal detector for the speech coefficients would be available, a more significantly non-continuous gain would be desired to block the noise-only coefficients. However, in the simultaneous detection and estimation approach the detector is not ideal but optimized to minimize the combined risk and the non-continuity of the system gain depends on the chosen cost parameters as well as on the gain floor.

5.3.2 Quadratic Spectral Amplitude Distortion Measure

The distortion measure of the quadratic spectral amplitude (QSA) is defined by

$$d_{ij}\left(X,\hat{X}\right) = \begin{cases} \left(|X| - \left|\hat{X}_j\right|\right)^2, & i = 1 \\ \left(G_f\,|Y| - \left|\hat{X}_j\right|\right)^2, & i = 0 \end{cases}, \tag{5.29}$$

and is related to the STSA suppression rule of Ephraim and Malah [3]. For evaluating the optimal detector and estimator under the QSA distortion measure we denote by $X \triangleq A\,e^{j\alpha}$ and $Y \triangleq R\,e^{j\theta}$ the clean and noisy spectral coefficients, respectively, where $A = |X|$ and $R = |Y|$. Accordingly, the pdf of the speech expansion coefficient under H_1 satisfies

$$p\left(a,\alpha \mid H_1\right) = \frac{a}{\pi\lambda_x}\exp\left(-\frac{a^2}{\lambda_x}\right). \tag{5.30}$$

Since the combined risk under the QSA distortion measure is independent of the signal phase nor the estimation phase, the estimated amplitude under η_j

is given by

$$\hat{A}_j = \arg\min_{\hat{a}} \left\{ q\, b_{1j} \int_0^{\infty} \int_0^{2\pi} (a-\hat{a})^2\, p\,(a,\alpha \mid H_1)\, p\,(Y \mid a,\alpha)\, d\alpha\, da + (1-q)\, b_{0j}\, (G_f\, R - \hat{a})^2\, p\,(Y \mid H_0) \right\}. \tag{5.31}$$

By using the phase of the noisy signal, the optimal estimation under the decision η_j, $j \in \{0,1\}$ is given by [20]

$$\begin{aligned} \hat{X}_j &= [b_{1j}\, \Lambda\,(\xi,\gamma)\, G_{STSA}\,(\xi,\gamma) + b_{0j}\, G_f]\, \phi_j\,(\xi,\gamma)^{-1}\, Y \\ &\triangleq G_j\,(\xi,\gamma)\, Y\,, \end{aligned} \tag{5.32}$$

where

$$G_{STSA}\,(\xi,\gamma) = \left[(1+\upsilon)\, I_0\left(\frac{\upsilon}{2}\right) + \upsilon\, I_1\left(\frac{\upsilon}{2}\right)\right] \frac{\sqrt{\pi\,\upsilon}}{2\gamma} \exp\left(-\frac{\upsilon}{2}\right) \tag{5.33}$$

denotes the STSA gain function of Ephraim and Malah [3], and $I_\nu\,(\cdot)$ denotes the modified Bessel function of order ν. For evaluating the optimal decision rule under the QSA distortion measure we need to compute the risk $r_{ij}(Y)$. Under H_1 we have [20]

$$r_{1j}\,(Y) = \frac{b_{1j}}{\pi} \frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi} \left\{ G_j^2\,\gamma + \frac{\xi}{1+\xi}\,(1+\upsilon) - 2\gamma\, G_j\, G_{STSA} \right\}, \tag{5.34}$$

and under H_0 we have

$$r_{0j}\,(Y) = \frac{b_{0j}}{\pi}\, [G_j\,(\xi,\gamma) - G_f]^2\, \gamma\, e^{-\gamma}\,. \tag{5.35}$$

Substituting (5.34) and (5.35) into (5.12), we obtain the optimal decision rule under the QSA distortion measure:

$$\begin{aligned} &\frac{b_{01}\,(G_1 - G_f)^2 - (G_0 - G_f)^2}{\Lambda\,(\xi,\gamma)} \underset{\eta_0}{\overset{\eta_1}{\lessgtr}} \\ &\qquad b_{10}\, G_0^2 - G_1^2 + \frac{\xi\,(1+\upsilon)\,(b_{10}-1)}{(1+\xi)\,\gamma} + 2\,(G_1 - b_{10}\, G_0)\, G_{STSA}\,. \end{aligned} \tag{5.36}$$

Figure 5.3 demonstrates attenuation curves under QSA cost function as a function of the *instantaneous* SNR defined by $\gamma - 1$, for several *a priori* SNRs, using the parameters $q = 0.8$, $G_f = -25$ dB and cost parameters $b_{01} = 5$ and $b_{10} = 1.1$. The gains G_1 (dashed line), G_0 (dotted line) and the total detection and estimation system gain (solid line) are compared to the STSA gain under signal presence uncertainty of Ephraim and Malah [3] (dashed-dotted line). The *a priori* SNRs range from -15 dB to 15 dB. Not only

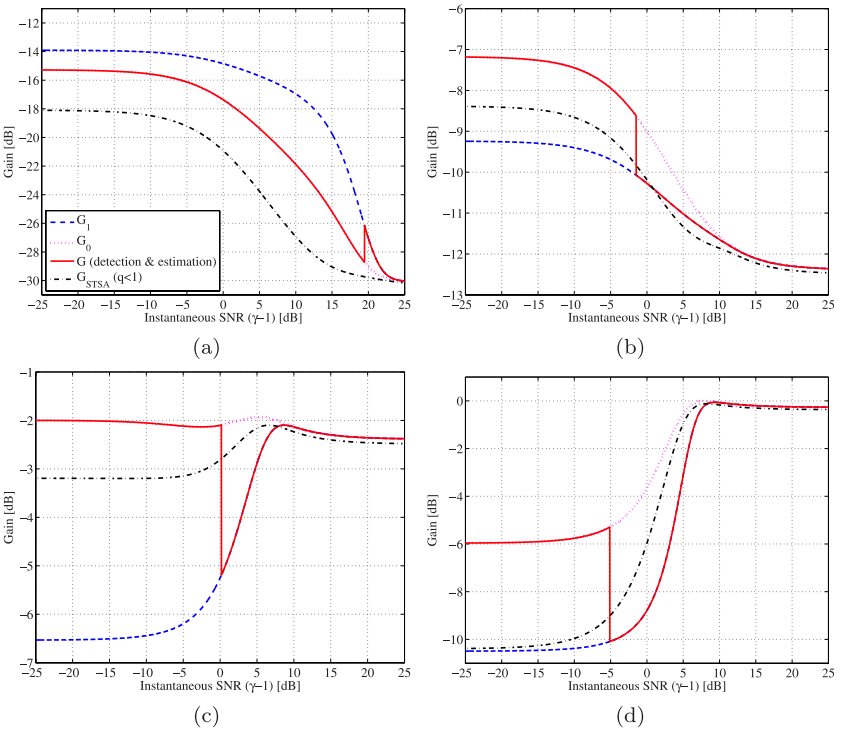

Fig. 5.3 Gain curves of G_1 (dashed line), G_0 (dotted line) and the total detection and estimation system gain curve (solid line), compared with the STSA gain under signal presence uncertainty (dashed-dotted line). The *a priori* SNRs are (a) $\xi = -15$ dB, (b) $\xi = -5$ dB, (c) $\xi = 5$ dB, and (d) $\xi = 15$ dB.

that the cost parameters shape the STSA gain curve, when combined with the detector the simultaneous detection and estimation provides a significant non-continuous modification of the standard STSA estimator. For example, for *a priori* SNRs of $\xi = -5$ and $\xi = 15$ dB, as shown in Fig. 5.3(b) and (d) respectively, as long as the instantaneous SNR is higher than about -2 dB (for $\xi = -5$ dB) or -5 dB (for $\xi = 15$ dB), the detector decision is η_1, while for lower instantaneous SNRs, the detector decision is η_0. The resulting non-continues gain function may yield greater noise reduction with slightly higher level of musicality, while not degrading speech quality.

Similarly to the case of a quadratic distortion measure, when the false alarm parameter is much smaller than the generalized likelihood ratio, $b_{01} << \Lambda(\xi, \gamma)$, the spectral gain of the estimator under the decision η_1 is $G_1(\xi, \gamma) \cong G_{STSA}(\xi, \gamma)$, which is optimal when the signal is surely present. When $b_{01} >> \Lambda(\xi, \gamma)$, the spectral gain under η_1 is $G_1(\xi, \gamma) \cong G_f$ to compensate for false decision made by the detector. If the cost parameter associated with missed detection is small and we have $b_{10} << \Lambda(\xi, \gamma)^{-1}$, then

$G_0(\xi,\gamma) \cong G_f$, and under $b_{10} >> \Lambda(\xi,\gamma)^{-1}$ we have $G_0(\xi,\gamma) \cong G_{STSA}(\xi)$ in order to overcome the high cost related to missed detection.

If one chooses constant cost parameters $b_{ij} = 1\ \forall i,j$, than the detection operation is not required (the estimation is independent of the decision rule), and we have

$$\begin{aligned} \hat{X}_0 &= [\,p(H_1 \mid Y)\, G_{STSA}(\xi,\gamma) + (1 - p(H_1 \mid Y))\, G_f]\, Y \\ &= \hat{X}_1\,. \end{aligned} \tag{5.37}$$

If we also set G_f to zero, the estimation reduces to the STSA suppression rule under signal presence uncertainty [3].

5.4 Spectral Estimation Under a Transient Noise Indication

In the previous section we introduced a method for optimal integration of a detector and an estimator for speech spectral components. In this section, we consider the integration of a spectral estimator with a *given* detector for noise transients. In many speech enhancement applications, an indicator for the transient source may be available, *e.g.*, siren noise in an emergency car, keyboard typing in computer-based communication system and a lens-motor noise in a digital video camera. In such cases, *a priori* information based on a training phase may yield a reliable detector for the transient noise. However, false detection of transient noise components when signal components are present may significantly degrade the speech quality and intelligibility. Furthermore, missed detection of transient noise components may result in a residual transient noise, which is perceptually annoying. The transient spectral variances can be estimated in such cases from training signals. However, applying a standard estimator to the spectral coefficients may result in removal of critical speech components in case of falsely detecting the speech components, or under-suppression of transient noise in case of miss detecting the noise transients. Consider a reliable detector for transient noise, we can define a cost (5.14) by using the quadratic log-spectral amplitude (QLSA) distortion measure is given by

$$d_{ij}\left(X,\hat{X}\right) = \begin{cases} \left(\log A - \log \hat{A}_j\right)^2\,, & i=1 \\ \left(\log\left(G_f R\right) - \log \hat{A}_j\right)^2\,, & i=0 \end{cases}\,, \tag{5.38}$$

and is related with the LSA estimation [23].

Similarly to the case of a QSA distortion measure, the average risk is independent of the signal phase nor on the estimation phase. Thus, by substituting the cost function into (5.13) we have

$$\hat{A}_j = \arg\min_{\hat{a}_j} \left\{ q\, b_{1j} \int_0^\infty \int_0^{2\pi} \left(\log a - \log \hat{a}_j\right)^2 p\left(a, \alpha \mid H_1\right) p\left(Y \mid a, \alpha\right) d\alpha da + (1-q)\, b_{0j} \left(\log\left(G_f R\right) - \log \hat{a}_j\right)^2 p\left(Y \mid H_0\right) \right\} . \tag{5.39}$$

By setting the derivative of (5.39) according to $\hat{a}_j$ equal to zero, we obtain[3]

$$q\, b_{1j} \int_0^\infty \int_0^{2\pi} \left(\log a - \log \hat{A}_j\right) p\left(a, \alpha \mid H_1\right) p\left(Y \mid a, \alpha\right) d\alpha da + (1-q)\, b_{0j} \left(\log\left(G_f R\right) - \log \hat{A}_j\right) p\left(Y \mid H_0\right) = 0. \tag{5.40}$$

The integration over $\log a$ yields

$$\begin{aligned} &\int_0^\infty \int_0^{2\pi} \log a\, p\left(a, \alpha \mid H_1\right) p\left(Y \mid a, \alpha\right) d\alpha da \\ &= \int_0^\infty \int_0^{2\pi} \log a\, p\left(a, \alpha \mid Y, H_1\right) p\left(Y \mid H_1\right) d\alpha da \\ &= E\left\{\log a \mid Y, H_1\right\} p\left(Y \mid H_1\right) , \end{aligned} \tag{5.41}$$

and

$$\begin{aligned} &\int_0^\infty \int_0^{2\pi} p\left(a, \alpha \mid H_1\right) p\left(Y \mid a, \alpha\right) d\alpha da \\ &= \int_0^\infty \int_0^{2\pi} p\left(a, \alpha, Y \mid H_1\right) d\alpha da = p\left(Y \mid H_1\right) . \end{aligned} \tag{5.42}$$

Substituting (5.41) and (5.42) into (5.40) we obtain

$$\hat{A}_j = \exp\left\{ b_{1j} \Lambda\left(Y\right) E\left\{\log a \mid Y\right\} \phi_j\left(Y\right)^{-1} + b_{0j} \log\left(G_f R\right) \phi_j\left(Y\right)^{-1} \right\} , \tag{5.43}$$

where

$$\begin{aligned} \exp\left[E\left\{\log a \mid Y\right\}\right] &= \frac{\xi}{1+\xi} \exp\left(\frac{1}{2} \int_v^\infty \frac{e^{-t}}{t} dt\right) R \\ &\triangleq G_{LSA}\left(\xi, \gamma\right) R , \end{aligned} \tag{5.44}$$

is the LSA suppression rule [23].

Substituting (5.44) into (5.43) and applying the noisy phase, we obtain the optimal estimation under a decision η_j, $j \in \{0, 1\}$:

$$\hat{X}_j = \left[G_f^{b_{0j}} G_{LSA}\left(\xi, \gamma\right)^{b_{1j} \Lambda(\xi, \gamma)} \right]^{\phi_j(\xi,\gamma)^{-1}} Y \triangleq G_j\left(\xi, \gamma\right) Y . \tag{5.45}$$

[3] Note that this solution is not dependent on the basis of the log and in addition, the optimal solution is strictly positive.

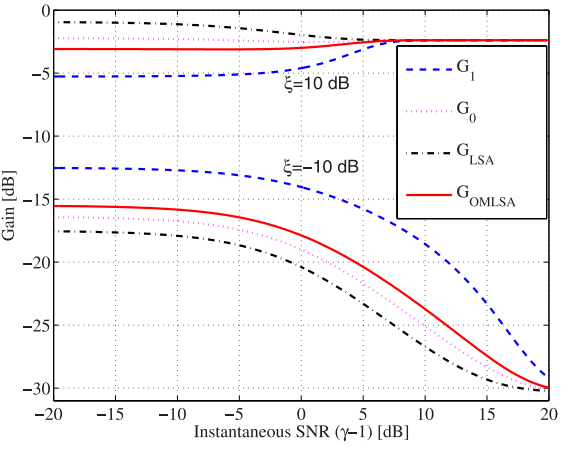

Fig. 5.4 Gain curves under QLSA distortion measure with $q = 0.8$, $b_{01} = 4$, $b_{10} = 2$, and $G_f = -15$ dB.

If we consider the simultaneous detection and estimation which was formulated in Section 5.3, the derivation of the optimal decision rule for the QLSA distortion measure is mathematically intractable. However, the estimation (5.45) under *any given detector* is still optimal in the sense of minimizing the average cost under the given decision. Therefore, the estimation (5.45) can be incorporated with any reliable detector to yield a sub-optimal detection and estimation system. Even where a non optimal detector is considered, the use of the cost parameters enables better control on the spectral gain under any decision made by the detector. In [27], a further generalization is considered by incorporating the probabilities for the detector decisions.

Figure 5.4 shows attenuation curves under QLSA distortion measure as a function of the instantaneous SNR with different *a priori* SNRs and with $q = 0.8$, $G_f = -15$ dB and cost parameters $b_{01} = 4$ and $b_{10} = 2$.

The signal estimate (5.45) generalizes existing suppression rules. For equal cost parameters (5.45) reduces to the OM-LSA estimator [4] which is given by (5.5). If we also let $G_f = 0$ and $q = 1$ we get the LSA suppression rule [23]. Both these estimators are shown in Fig. 5.4 with comparison to the spectral gains under any potential decision made by the detector.

5.5 A Priori SNR Estimation

In spectral speech enhancement applications, the *a priori* SNR is often estimated by using the decision-directed approach [3]. Accordingly, in each time-frequency bin we compute

$$\hat{\xi}_{\ell k} = \max\left\{\alpha\, G^2\left(\hat{\xi}_{\ell-1,k}, \gamma_{\ell-1,k}\right)\gamma_{\ell-1,k}\,(1-\alpha)\,(\gamma_{\ell k}-1)\,,\, \xi_{\min}\right\}, \quad (5.46)$$

where α $(0 \le \alpha \le 1)$ is a weighting factor that controls the trade-off between noise reduction and transient distortion introduced into the signal, and $\xi_{\min}$ is a lower bound for the *a priori* SNR which is necessary for reducing the residual musical noise in the enhanced signal [3, 22]. Since the *a priori* SNR is defined under the assumption that $H_1^{\ell k}$ is true, it is proposed in [4] to replace the gain G in (5.46) by G_{H_1} which represents the spectral gain when the signal is surely present (*i.e.*, $q = 1$). Increasing the value of α results in a greater reduction of the musical noise phenomena, at the expense of further attenuation of transient speech components (*e.g.*, speech onsets) [22]. By using the proposed approach with high cost for false speech detection, the musical noise can be reduced without increasing the value of α, which enables rapid changes in the *a priori* SNR estimate. The lower bound for the *a priori* SNR is related to the spectral gain floor G_f since both imply a lower bound on the spectral gain. The latter parameter is used to evaluate both the optimal detector and estimator while taking into account the desired residual noise level.

The decision-directed estimator is widely used, but is not suitable for transient noise environments, since a high-energy noise burst may yield an instantaneous increase in the *a posteriori* SNR and a corresponding increase in $\hat{\xi}_{\ell k}$ as can be seen from (5.46). The spectral gain would then be higher than the desired value, and the transient noise component would not be sufficiently attenuated. Let $\hat{\lambda}^s_{d_{\ell k}}$ denote the estimated spectral variance of the stationary noise component and let $\hat{\lambda}^t_{d_{\ell k}}$ denote the estimated spectral variance of the transient noise component. The former may be practically estimated by using the improved minima-controlled recursive averaging (IMCRA) algorithm [4, 28] or by using the minimum-statistics approach [29], while $\lambda^t_{d_{\ell k}}$ may be evaluated based on a training signals as assumed in [27]. The total variance of the noise component is $\hat{\lambda}_{d_{\ell k}} = \hat{\lambda}^s_{d_{\ell k}} + \hat{\lambda}^t_{d_{\ell k}}$. Note that $\lambda^t_{d_{\ell k}} = 0$ in time-frequency bins where the transient noise source is inactive. Since the *a priori* SNR is highly dependent on the noise variance, we first estimate the speech spectral variance by

$$\hat{\lambda}_{x_{\ell k}} = \max\left\{\alpha\, G^2_{H_1}\left(\hat{\xi}_{\ell-1,k}, \gamma_{\ell-1,k}\right)|Y_{\ell-1,k}|^2\,(1-\alpha)\left(|Y_{\ell k}|^2 - \hat{\lambda}_{d_{\ell k}}\right), \lambda_{\min}\right\}, \quad (5.47)$$

where $\lambda_{\min} = \xi_{\min}\,\hat{\lambda}^s_{d_{\ell k}}$. Then, the *a priori* SNR is evaluated by $\hat{\xi}_{\ell k} = \hat{\lambda}_{x_{\ell k}}/\hat{\lambda}_{d_{\ell k}}$. In a stationary noise environment this estimator reduces to the decision-directed estimator (5.46), with G_{H_1} substituting G. However, under the presence of a transient noise component, this method yields a lower *a priori* SNR estimate, which enables higher attenuation of the high-energy transient noise component. Furthermore, to allow further reduction of the transient noise component to the level of the residual stationary noise, the gain floor is modified by $\tilde{G}_f = G_f\,\hat{\lambda}^s_{d_{\ell k}}/\hat{\lambda}_{d_{\ell k}}$ as proposed in [30].

The different behaviors under transient noise conditions of this modified decision-directed *a priori* SNR estimator and the decision-directed estimator as proposed in [4] are illustrated in Figs 5.5 and 5.6. Figure 5.5 shows the signals in the time domain: the analyzed signal contains a sinusoidal wave which is active in only two specific segments. The noisy signal contains both additive white Gaussian noise with 5 dB SNR and high-energy transient noise components. The signal enhanced by using the decision-directed estimator and the STSA suppression rule is shown in Fig. 5.5(c). The signal enhanced by using the modified *a priori* SNR estimator and the STSA suppression rule is shown in Fig. 5.5(d), and the result obtained by using the proposed modified *a priori* SNR estimation with the detection and estimation approach is shown in Fig. 5.5(d) (using the same parameters as in the previous section). Both the decision-directed estimator and the modified *a priori* SNR estimator are applied with $\alpha = 0.98$ and $\xi_{\min} = -20$ dB. Clearly, in stationary noise intervals, and where the SNR is high, similar results are obtained by both *a priori* SNR estimators. However, the proposed modified *a priori* SNR estimator obtain higher attenuation of the transient noise, whether it is incorporated with the STSA or the simultaneous detection and estimation approach. Figure 5.6 shows the amplitudes of the STFT coefficients of the noisy and enhanced signals at the frequency band which contains the desired sinusoidal component. Accordingly, the modified *a priori* SNR estimator enables a greater reduction of the background noise, particularly transient noise components. Moreover, it can be seen that using the simultaneous detection and estimation yields better attenuation of both the stationary and background noise compared to the STSA estimator, even while using the same *a priori* SNR estimator.

5.6 Experimental Results

For the experimental study, speech signals from the TIMIT database [31] were sampled at 16 kHz and degraded by additive noise. The noisy signals are transformed into the STFT domain using half-overlapping Hamming windows of 32 msec length, and the background-noise spectrum is estimated by using the IMCRA algorithm (for all the considered enhancement algorithms) [28, 4]. The performance evaluation includes objective quality measures, a subjective study of spectrograms and informal listening tests. The first quality measure is the segmental SNR defined by [32]

$$SegSNR = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathcal{T} \left\{ 10 \log_{10} \frac{\sum_{n=0}^{K-1} x^2 \left(n + \ell\, K/2\right)}{\sum_{n=0}^{K-1} \left[x \left(n + \ell\, K/2\right) - \hat{x} \left(n + \ell\, K/2\right)\right]^2} \right\}, \tag{5.48}$$

where $\mathcal{L}$ represents the set of frames which contain speech, $|\mathcal{L}|$ denotes the number of elements in $\mathcal{L}$, $K = 512$ is the number of samples per frame and

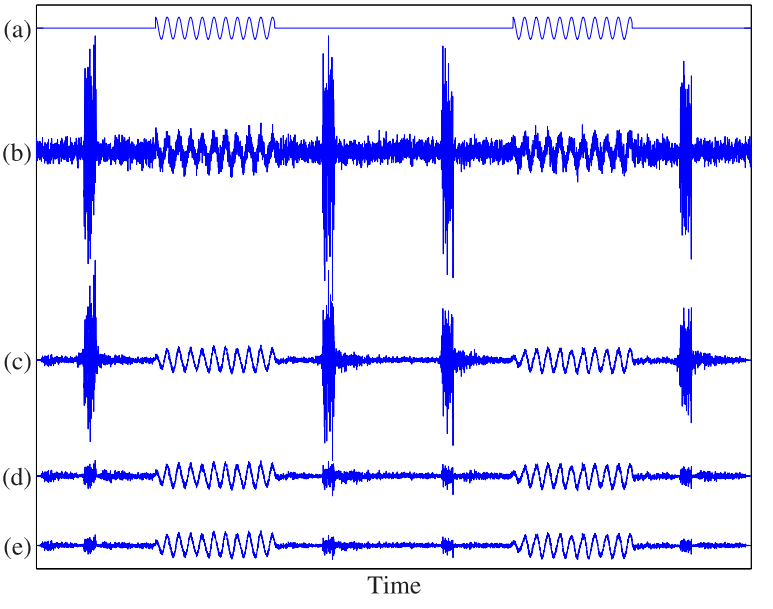

Fig. 5.5 Signals in the time domain. (a) Clean sinusoidal signal; (b) noisy signal with both stationary and transient components; (c) enhanced signal obtained by using the STSA and the decision-directed estimators; (d) enhanced signal obtained by using the STSA and the modified *a priori* SNR estimators; (e) enhanced signal obtained by using the detection and estimation approach and the modified *a priori* SNR estimator.

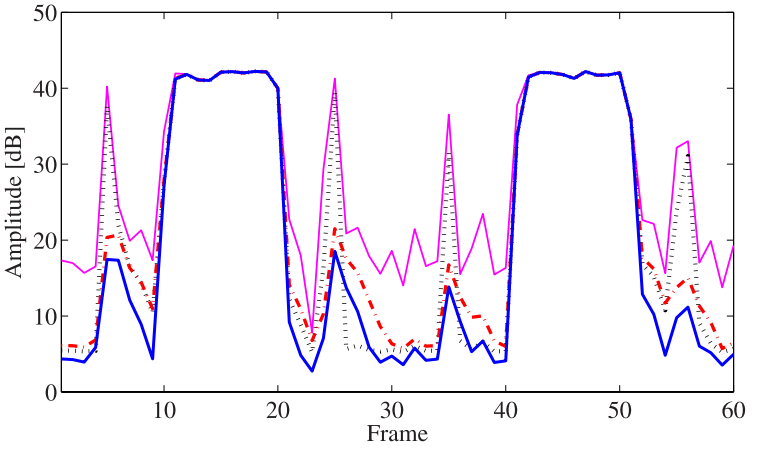

Fig. 5.6 Amplitudes of the STFT coefficients along time-trajectory corresponding to the frequency of the sinusoidal signal: noisy signal (light solid line), STSA with decision-directed estimation (dotted line), STSA with the modified *a priori* SNR estimator (dashed-dotted line) and simultaneous detection and estimation with the modified *a priori* SNR estimator (dark solid line).

Table 5.1 Segmental SNR and log spectral distortion obtained by using either the simultaneous detection and estimation approach or the STSA estimator in stationary noise environment.

Input SNR	Input Signal		Detection & Estimation		STSA ($\alpha = 0.98$)		STSA ($\alpha = 0.92$)	
dB	SegSNR	LSD	SegSNR	LSD	SegSNR	LSD	SegSNR	LSD
-5	-6.801	20.897	1.255	7.462	0.085	9.556	-0.684	10.875
0	-3.797	16.405	4.136	5.242	3.169	6.386	2.692	7.391
5	0.013	12.130	5.98	3.887	5.266	4.238	5.110	4.747
10	4.380	8.194	6.27	3.143	5.93	3.167	6.014	3.157

the operator $\mathcal{T}$ confines the SNR at each frame to a perceptually meaningful range between -10 dB and 35 dB. The second quality measure is log-spectral distortion (LSD) which is defined by

$$LSD = \frac{1}{L}\sum_{\ell=0}^{L-1}\left\{\frac{1}{K/2+1}\sum_{k=0}^{K/2}\left[10\log_{10}\mathcal{C}X_{\ell k} - 10\log_{10}\mathcal{C}\hat{X}_{\ell k}\right]^2\right\}^{\frac{1}{2}}, \quad (5.49)$$

where $\mathcal{C}X \triangleq \max\left\{|X|^2, \epsilon\right\}$ is a spectral power clipped such that the log-spectrum dynamic range is confined to about 50 dB, that is, $\epsilon = 10^{-50/10} \cdot \max_{\ell,k}\left\{|X_{\ell k}|^2\right\}$. The third quality measure (used in Section 5.6-B) is the perceptual evaluation of speech quality (PESQ) score [33].

5.6.1 Simultaneous Detection and Estimation

The suppression rule results from the proposed simultaneous detection and estimation approach with the QSA distortion measure is compared to the STSA estimation [3] for stationary white Gaussian noise with SNRs in the range $[-5, 10]$ dB. For both algorithms the *a priori* SNR is estimated by the decision-directed approach (5.46) with $\xi_{\min} = -15$ dB, and the *a priori* speech presence probability is $q = 0.8$. For the STSA estimator a decision-directed estimation [4] with $\alpha = 0.98$ reduces the residual musical noise but generally implies transient distortion of the speech signal [3, 22]. However, the inherent detector obtained by the simultaneous detection and estimation approach may improve the residual noise reduction and therefore a lower weighting factor α may be used to allow lower speech distortion. Indeed, for the simultaneous detection and estimation approach $\alpha = 0.92$ implies better results, while for the STSA algorithm, better results are achieved with $\alpha = 0.98$. The cost parameters for the simultaneous detection and estimation should be chosen according to the system specification, *i.e.*, whether the quality of the speech signal or the amount of noise reduction is of higher importance. Table 5.1 summarizes the average segmental SNR and LSD for these two enhancement algorithms, with cost parameters $b_{01} = 10$ and $b_{10} = 2$,

Table 5.2 Objective quality measures.

Method	SegSNR [dB]	LSD [dB]	PESQ
Noisy speech	−2.23	7.69	1.07
OM-LSA	−1.31	6.77	0.97
Proposed Alg.	5.41	1.67	2.87

and $G_f = -15$ dB for the simultaneous detection and estimation algorithm. The results for the STSA algorithm are presented for $\alpha = 0.98$ as well as for $\alpha = 0.92$ (note that for the STSA estimator $G_f = 0$ is considered as originally proposed). It shows that the simultaneous detection and estimation yields improved segmental SNR and LSD, while a greater improvement is achieved for lower input SNR. Informal subjective listening tests and inspection of spectrograms demonstrate improved speech quality with higher attenuation of the background noise. However, since the weighting factor used for the *a priori* SNR estimate is lower, and the gain function is discontinuous, the residual noise resulting from the simultaneous detection and estimation algorithm is slightly more musical than that resulting from the STSA algorithm.

5.6.2 Spectral Estimation Under a Transient Noise Indication

The application of the spectral estimation under an indicator for the transient noise presented in Section 5.4, with the a priori SNR estimation for nonstationary environment of Section 5.5, is demonstrated in a computer-based communication system. The background office noise is slowly-varying while possible keyboard typing interference may exist. Since the keyboard signal is available to the computer, a reliable detector for the transient-like keyboard noise is assumed to be available based on a training phase but still, erroneous detections are reasonable. The speech signals degraded by a stationary background noise with 15 dB SNR and a keyboard typing noise such that the total SNR is 0.8 dB. The transient noise detector is assumed to have an error probability of 10% and the missed detection and false detection costs are set to 1.2.

Figure 5.7 demonstrates the spectrograms and waveforms of a signal enhanced by using the proposed algorithm, compared to using the OM-LSA algorithm. It can be seen that using the proposed approach, the transient noise is significantly attenuated, while the OM-LSA is unable to eliminate the keyboard transients.

The results of the objective measures are summarized in Table 5.2. It can be seen that the proposed detection and estimation approach significantly

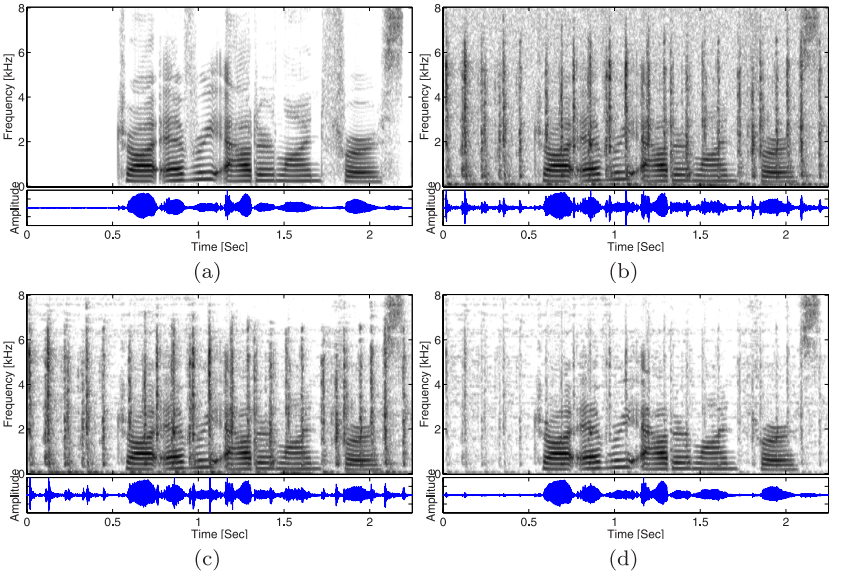

Fig. 5.7 Speech spectrograms and waveforms. (a) Clean signal ("Draw any outer line first"); (b) noisy signal (office noise including keyboard typing noise, SNR=0.8 dB); (c) speech enhanced by using the OM-LSA estimator; (d) speech enhanced by using the proposed algorithm.

improves speech quality compared to using the OM-LSA algorithm. Informal listening tests confirm that the annoying keyboard typing noise is dramatically reduced and the speech quality is significantly improved.

5.7 Conclusions

We have presented a novel formulation of the single-channel speech enhancement problem in the time-frequency domain. The formulation relies on coupled operations of detection and estimation in the STFT domain, and a cost function that combines both the estimation and detection errors. A detector for the speech coefficients and a corresponding estimator for their values are jointly designed to minimize a combined Bayes risk. In addition, cost parameters enable to control the trade-off between speech quality, noise reduction and residual musical noise. The proposed method generalizes the traditional spectral enhancement approach which considers estimation-only under signal presence uncertainty. In addition we propose a modified decision-directed *a priori* SNR estimator which is adapted to transient noise environment. Experimental results show greater noise reduction with improved speech quality when compared with the STSA suppression rules under stationary noise. Fur-

thermore, it is demonstrated that under transient noise environment, greater reduction of transient noise components may be achieved by exploiting a reliable detector for interfering transients.

References

1. I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2007, ch. 45.
2. R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
3. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
4. I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Processing*, vol. 81, pp. 2403–2418, Nov. 2001.
5. J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. 23rd IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-98*, vol. 1, Seattle, Washington, May 1998, pp. 365–368.
6. J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
7. Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Lett.*, vol. 8, no. 10, pp. 276–278, Oct. 2001.
8. S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505, Sep. 2003.
9. A. Davis, S. Nordholm, and R. Tongneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–423, Mar. 2006.
10. J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.
11. S. F. Boll, "Suppression of acousting noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
12. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-79*, vol. 4, Apr. 1979, pp. 208–211.
13. Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
14. H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.
15. Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 4, pp. 334–341, Jul. 2003.
16. F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. 27th IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-02*, Orlando, Florida, May 2002, pp. 569–572.
17. D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. 24th*

IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-99, Phoenix, Arizona, Mar. 1999, pp. 789–792.
18. D. Middleton and F. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 434–444, May 1968.
19. A. Fredriksen, D. Middleton, and D. Vandelinde, "Simultaneous signal detection and estimation under multiple hypotheses," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 5, pp. 607–614, 1972.
20. A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2348–2359, Nov. 2007.
21. ——, "Single-sensor blind source separation using classification and estimation approach and GARCH modeling," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, Nov. 2008.
22. O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
23. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
24. A. G. Jaffer and S. C. Gupta, "Coupled detection-estimation of gaussian processes in gaussian noise," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 1, pp. 106–110, Jan. 1972.
25. I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Processing*, vol. 86, no. 4, pp. 698–709, Apr. 2006.
26. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., A. Jefferey and D. Zwillinger, Eds. Academic Press, 2000.
27. A. Abramson and I. Cohen, "Enhancement of speech signals under multiple hypotheses using an indicator for transient noise presence," in *Proc. 32nd IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP-07*, Honolulu, Hawaii, Apr. 2007, pp. 553–556.
28. I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.
29. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, Jul. 2001.
30. E. Habets, I. Cohen, and S. Gannot, "MMSE log-spectral amplitude estimator for multiple interferences," in *Proc. Int. Workshop on Acoust. Echo and Noise Control., IWAENC-06*, Paris, France, Sept. 2006.
31. J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database," *Technical report, National Institute of Standards and Technology (NIST)*, Gaithersburg, Maryland (prototype as of December 1988).
32. S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Meaasures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
33. ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *International Telecommunication Union, Geneva, Switzerland*, Feb. 2001.