ELSEVIER

Signal Processing 81 (2001) 2403–2418

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Speech enhancement for non-stationary noise environments

Israel Cohen *, Baruch Berdugo

Lamar Signal Processing Ltd., P.O.Box 573, Yokneam Ilit 20692, Israel

Received 18 February 2001; received in revised form 26 June 2001

Abstract

In this paper, we present an *optimally-modified log-spectral amplitude* (OM-LSA) speech estimator and a *minima controlled recursive averaging* (MCRA) noise estimation approach for robust speech enhancement. The spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty. The noise estimate is given by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands. We introduce two distinct speech presence probability functions, one for estimating the speech and one for controlling the adaptation of the noise spectrum. The former is based on the time–frequency distribution of the a priori signal-to-noise ratio. The latter is determined by the ratio between the local energy of the noisy signal and its minimum within a specified time window. Objective and subjective evaluation under various environmental conditions confirm the superiority of the OM-LSA and MCRA estimators. Excellent noise suppression is achieved, while retaining weak speech components and avoiding the musical residual noise phenomena.

1. Introduction

A practical speech enhancement system generally consists of two major components: the estimation of noise power spectrum, and the estimation of speech. The estimation of noise, when only one microphone source is provided, is based on the assumption of a slowly varying noise environment. In particular, the noise spectrum remains virtually stationary during speech activity. The estimation of speech is based on the assumed statistical model, distortion measure, and the estimated noise.

A commonly used approach for estimating the noise power spectrum is to average the noisy signal over sections which do not contain speech. A soft-decision speech pause detection is either implemented on a frame-by-frame basis [12,22] or estimated independently for individual subbands using an a posteriori signal-to-noise ratio (SNR) [11,13]. However, the detection reliability severely deteriorates for weak speech components and low input SNR. Additionally, the amount of presumable non-speech sections in the signal may not be sufficient, which restricts the tracking capability of the noise estimator in non-stationary environments. Alternatively, the noise can be estimated from histograms in the power spectral domain [11,18,24]. Unfortunately, such methods are computationally expensive.

Martin [14,15] has proposed an algorithm for noise estimation based on minimum statistics. The noise estimate is obtained as the minima values of a smoothed power estimate of the noisy signal, multiplied by a factor that compensates the bias.

* Corresponding author. Tel.: +972-4-993-7066; fax: +972-4-993-7064.

E-mail address: icohen@lamar.co.il (I. Cohen).

0165-1684/01/$ - see front matter
PII: S0165-1684(01)00128-1

Nomenclature

A	spectral speech amplitude
b	smoothing window for computing S_f
c_{ij}	cost for deciding H'_i when H'_j
D	short-time Fourier transform of the noise signal
d	noise signal
G	spectral gain function
G_{H_1}	conditional gain function
$G_{\min}$	spectral gain floor
H_0	speech absence hypothesis for speech estimation
H_1	speech presence hypothesis for speech estimation
H'_0	speech absence hypothesis for noise estimation
H'_1	speech presence hypothesis for noise estimation
h	analysis window
$\tilde{h}$	synthesis window
h_{local}, h_{global}	local and global smoothing windows
I	indicator function for hypothesis testing
k	frequency bin (subband) index
L	number of frames used for finding S_{tmp}
ℓ	time frame index
$\mathscr{L}$	set of frames that contain speech
M	framing step
N	size of the analysis window
n	discrete time index
P_{local}, P_{global}	local and global likelihood of speech
P_{frame}	frame likelihood of speech
p	speech presence probability for speech estimation
p'	speech presence probability for noise estimation
q	a priori probability for speech absence
$q_{\max}$	upper threshold for q
S	local energy of the noisy signal
S_f	frequency average of the noisy signal's energy
$S_{\min}$	local minimum of S
S_{tmp}	temporary minimum of S
S_r	ratio between the local energy and local minimum
w	length of b is $2w+1$
X	short-time Fourier transform of the speech signal
x	speech signal
Y	short-time Fourier transform of the noisy signal
y	noisy signal
α	weighting factor for the a priori SNR estimation
α_d	smoothing parameter for estimating the noise spectrum
$\tilde{\alpha}_d$	time-varying smoothing parameter
α_p	smoothing parameter for computing p'
α_s	smoothing parameter for computing S
β	smoothing parameter for computing ζ
γ	a posteriori SNR
δ	threshold value of S_r for hypothesis testing
ζ	recursive average of the a priori SNR
ζ_{frame}	frame average of the a priori SNR
ζ_{local}, ζ_{global}	local and global averages of the a priori SNR
$\zeta_{\min}$, $\zeta_{\max}$	empirical constants
$\zeta_{p\,\min}$, $\zeta_{p\,\max}$	empirical constants
ζ_{peak}	confined peak value of ζ_{frame}
Λ	generalized likelihood ratio
λ	designates either "local" or "global"
λ_d	variance of D
λ_x	variance of X given speech is present
μ	transition function from speech to noise
Ξ	a priori SNR estimate assuming speech is present
ξ	a priori SNR
$\hat{\xi}$	a priori SNR estimate under speech presence uncertainty

Abbreviations

LSA	log-spectral amplitude
MCRA	minima controlled recursive averaging
MM-LSA	multiplicatively-modified log-spectral amplitude
OM-LSA	optimally modified log-spectral amplitude
PDF	probability density function
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
STSA	short-time spectral amplitude

However, this noise estimate is sensitive to outliers [24], generally biased [16], and its variance is about twice as large as the variance of a conventional noise estimator [15]. Additionally, this method occasionally attenuates low energy phonemes [15]. To overcome these limitations, the smoothing parameter and the bias compensation factor are turned into time and frequency dependent, and estimated for each spectral component and each time frame [16]. In [6], a computationally more efficient minimum tracking scheme is presented. Its main drawbacks are the very slow update rate of the noise estimate in case of a sudden rise in the noise energy level, and its tendency to cancel the signal [19].

Considering the speech estimation, Ephraim and Malah [8] derived a log-spectral amplitude (LSA) estimator, which minimizes the mean-square error of the log-spectra, based on a Gaussian statistical model. This estimator proved very efficient in reducing musical residual noise phenomena [6,12,17]. However, the speech spectrum is estimated under speech presence hypothesis. In contrast to other estimators, whose performance improves by utilizing the speech presence probability [7,10,18,23,25], it was believed that modification of the LSA estimator under speech presence uncertainty is "unworthy" [8]. Malah et al. [13] have recently proposed a *multiplicatively modified* LSA (MM-LSA) estimator. Accordingly, the spectral gain is multiplied by the conditional speech presence probability, which is estimated for each frequency bin and each frame. Unfortunately, the multiplicative modifier is not optimal [13]. Moreover, their estimate for the a priori SNR interacts with the estimated a priori speech absence probability [17]. This adversely affects the total gain for noise-only bins, and results in an unnaturally structured residual noise. [1]

Kim and Chang [12] proposed to use a small fixed a priori speech absence probability q ($q = 0.0625$) and a multiplicative modifier, which is based on the *global* conditional speech absence probability in each frame. This modifier is applied to the a priori and a posteriori SNRs. Not only such a modification is inconsistent with the statistical model, but also insignificant due to the small value of q and the influence of a few noise-only bins on the global speech absence probability.

In this paper, we present an *optimally modified* LSA (OM-LSA) speech estimator and a *minima controlled recursive averaging* (MCRA) noise estimation approach for robust speech enhancement. The optimal spectral gain function is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty. The exponential weight of each hypothetical gain is its corresponding probability, conditional on the observed signal. The noise spectrum is estimated by recursively averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands.

We introduce two distinct speech presence probability functions, one for estimating the speech and one for controlling the adaptation of the noise spectrum. The former is based on the time–frequency distribution of the a priori SNR. The latter is determined by the ratio between the local energy of the noisy signal and its minimum within a specified time window. The probability functions are estimated for each frame and each subband via a soft-decision approach, which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

Objective and subjective evaluation of the OM-LSA and MCRA estimators is performed under various environmental conditions. We show that these estimators are superior, particularly for low input SNRs and non-stationary noise. The MCRA noise estimate is unbiased, computationally efficient, robust with respect to the input SNR and type of underlying additive noise, and characterized by the ability to quickly follow abrupt changes in the noise spectrum. Its performance is close to the theoretical limit. The OM-LSA estimator demonstrates excellent noise suppression, while retaining weak speech components and avoiding the musical residual noise phenomena.

The paper is organized as follows. In Section 2, we derive the OM-LSA speech estimator and its corresponding speech presence probability function. In Section 3, we discuss the problem of the a priori SNR estimation under speech presence

[1] Applying a uniform attenuation factor to frames that do not contain speech eliminates the noise structuring in such frames [13]. Yet, in speech-plus-noise frames the noise structuring persists.

uncertainty. In Section 4, an expression for the a priori speech absence probability is formulated, based on the time–frequency distribution of the a priori SNR. In Section 5, we present the MCRA noise estimation approach and propose an appropriate speech presence probability function for controlling the adaptation of the noise spectrum. Finally, an objective and subjective evaluation of the OM-LSA and MCRA estimators is performed in Section 6.

2. Optimal gain modification

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 divided into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Specifically,

$$Y(k,\ell)=\sum_{n=0}^{N-1} y(n+\ell M)h(n)\mathrm{e}^{-\mathrm{j}(2\pi/N)nk}, \tag{1}$$

where k is the frequency bin index, ℓ is the time frame index, h is an analysis window of size N (e.g., Hanning window), and M is the framing step (number of samples separating two successive frames). Let $X(k,\ell)$ denote the STFT of the clean speech, then its estimate is obtained by applying a specific gain function to each spectral component of the noisy speech signal:

$$\hat{X}(k,\ell)=G(k,\ell)Y(k,\ell). \tag{2}$$

Using the inverse STFT, with a synthesis window $\tilde{h}$ that is biorthogonal to the analysis window h [28], the estimate for the clean speech signal is given by

$$\hat{x}(n)=\sum_{\ell}\sum_{k=0}^{N-1}\hat{X}(k,\ell)\tilde{h}(n-\ell M)\mathrm{e}^{\mathrm{j}(2\pi/N)k(n-\ell M)}, \tag{3}$$

where the inverse STFT is efficiently implemented using the weighted overlap-add method [5].

Among various existing speech enhancement methods, which can be represented by different spectral gain functions, we choose the LSA estimator [8] due to its superiority in reducing musical noise phenomena. The LSA estimator minimizes

$$E\{(\log A(k,\ell)-\log\hat{A}(k,\ell))^2\},$$

where $A(k,\ell)=|X(k,\ell)|$ denotes the spectral speech amplitude, and $\hat{A}(k,\ell)$ its optimal estimate. Assuming statistically independent spectral components [8], the LSA estimator is defined by

$$\hat{A}(k,\ell)=\exp\{E[\log A(k,\ell)|Y(k,\ell)]\}. \tag{4}$$

Given two hypotheses, $\mathrm{H}_0(k,\ell)$ and $\mathrm{H}_1(k,\ell)$, which indicate, respectively, speech absence and presence in the kth frequency bin of the ℓth frame, we have

$$\begin{aligned}&\mathrm{H}_0(k,\ell):\ Y(k,\ell)=D(k,\ell),\\ &\mathrm{H}_1(k,\ell):\ Y(k,\ell)=X(k,\ell)+D(k,\ell),\end{aligned} \tag{5}$$

where $D(k,\ell)$ represents the STFT of the noise signal. We assume that the STFT coefficients, for both speech and noise, are complex Gaussian variables [7]. Accordingly, the conditional PDFs of the observed signal are given by

$$\begin{aligned}p(Y(k,\ell)|\mathrm{H}_0(k,\ell))&=\frac{1}{\pi\lambda_d(k,\ell)}\exp\left\{-\frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)}\right\},\\ p(Y(k,\ell)|\mathrm{H}_1(k,\ell))&=\frac{1}{\pi(\lambda_x(k,\ell)+\lambda_d(k,\ell))}\\ &\quad\times\exp\left\{-\frac{|Y(k,\ell)|^2}{\lambda_x(k,\ell)+\lambda_d(k,\ell)}\right\},\end{aligned} \tag{6}$$

where $\lambda_x(k,\ell)=E[|X(k,\ell)|^2|\mathrm{H}_1(k,\ell)]$ and $\lambda_d(k,\ell)=E[|D(k,\ell)|^2]$ denote, respectively, the variances of speech and noise. Applying Bayes rule for the conditional speech presence probability, one obtains

$$P(\mathrm{H}_1(k,\ell)|Y(k,\ell))=\frac{\Lambda(k,\ell)}{1+\Lambda(k,\ell)}\triangleq p(k,\ell), \tag{7}$$

where $\Lambda(k,\ell)$ is the generalized likelihood ratio defined by

$$\Lambda(k,\ell)=\frac{1-q(k,\ell)}{q(k,\ell)}\frac{p(Y(k,\ell)|\mathrm{H}_1(k,\ell))}{p(Y(k,\ell)|\mathrm{H}_0(k,\ell))} \tag{8}$$

and $q(k,\ell)\triangleq P(\mathrm{H}_0(k,\ell))$ is the a priori probability for speech absence. Substituting (6) and (8) into

(7), we have

$$p(k,\ell)=\left\{1+\frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\times\exp(-\upsilon(k,\ell))\right\}^{-1}, \quad (9)$$

where

$$\xi(k,\ell)\triangleq\frac{\lambda_x(k,\ell)}{\lambda_d(k,\ell)},\quad \gamma(k,\ell)\triangleq\frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)},$$
$$\upsilon(k,\ell)\triangleq\frac{\gamma(k,\ell)\xi(k,\ell)}{1+\xi(k,\ell)}, \quad (10)$$

$\xi(k,\ell)$ and $\gamma(k,\ell)$ represent the a priori and a posteriori SNRs [7,18].

Based on the binary hypothesis model,

$$E[\log A(k,\ell)|Y(k,\ell)] = E[\log A(k,\ell)|Y(k,\ell),\mathrm{H}_1(k,\ell)]\,p(k,\ell) + E[\log A(k,\ell)|Y(k,\ell),\mathrm{H}_0(k,\ell)]\times(1-p(k,\ell)). \quad (11)$$

Using (4), we get an optimally modified LSA estimator [3] defined by

$$\hat{A}(k,\ell) = (\exp\{E[\log A(k,\ell)|Y(k,\ell),\mathrm{H}_1(k,\ell)]\})^{p(k,\ell)} \times(\exp\{E[\log A(k,\ell)|Y(k,\ell),\mathrm{H}_0(k,\ell)]\})^{(1-p(k,\ell))}. \quad (12)$$

When speech is absent, the gain is constrained to be larger than a threshold $G_{\min}$, which is determined by a subjective criteria for the noise naturalness [1,2,29]. Hence,

$$\exp\{E[\log A(k,\ell)|Y(k,\ell),\mathrm{H}_0(k,\ell)]\} = G_{\min}|Y(k,\ell)|. \quad (13)$$

When speech is present, the conditional gain function, defined by

$$\exp\{E[\log A(k,\ell)|Y(k,\ell),\mathrm{H}_1(k,\ell)]\} = G_{\mathrm{H}_1}(k,\ell)|Y(k,\ell)| \quad (14)$$

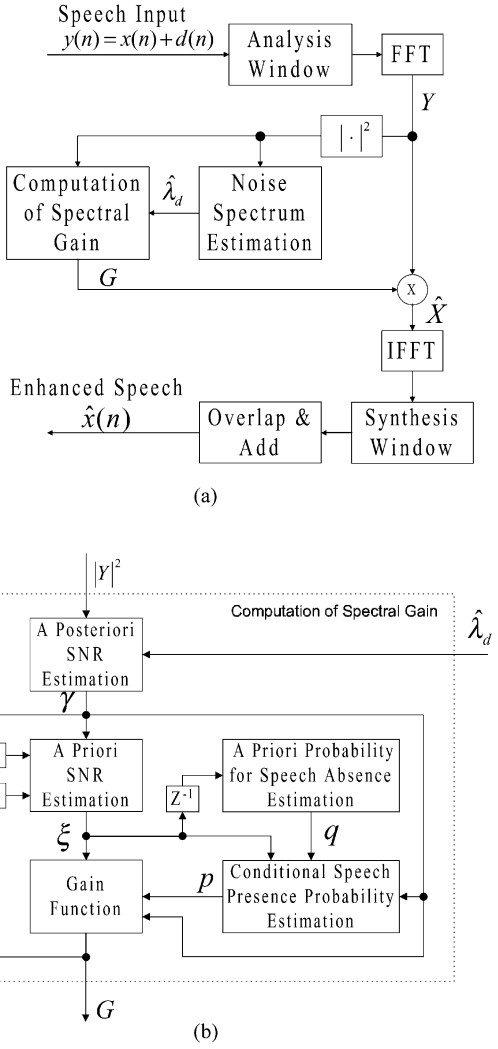

Fig. 1. (a) Block diagram of the speech enhancement configuration; (b) block diagram of the spectral gain computation.

is derived in [8] to be

$$G_{\mathrm{H}_1}(k,\ell)=\frac{\xi(k,\ell)}{1+\xi(k,\ell)}\exp\left(\frac{1}{2}\int_{\upsilon(k,\ell)}^{\infty}\frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t\right). \quad (15)$$

Substituting (13) and (14) into (12), the spectral gain for the OM-LSA is given by

$$G(k,\ell)=\{G_{\mathrm{H}_1}(k,\ell)\}^{p(k,\ell)}G_{\min}^{1-p(k,\ell)}. \quad (16)$$

Fig. 1 describes a block diagram of the speech enhancement configuration. The phase of the noisy

speech is not processed [27], since the optimal phase estimate is the noisy phase itself (in fact, the minimum mean-square error estimator of the complex exponential of the phase, under unity modulus constraint, is the complex exponential of the noisy phase [7]). It is worthwhile mentioning that trying to optimally modify the spectral gain function for the LSA estimator without taking into account a lower bound threshold ($G_{\min}$) results in a non-multiplicative modification, which fails to provide a meaningful improvement over using G_{H_1} alone [8,13].

3. A priori SNR estimation

In this section we address the problem of the a priori SNR estimation under speech presence uncertainty. The a priori SNR $\xi(k,\ell)$ is estimated for each spectral component and each analysis frame due to the non-stationarity of the speech signal. It is used for evaluating both the conditional gain $G_{\mathrm{H}_1}(k,\ell)$ (Eq. (15)) and the speech presence probability $p(k,\ell)$ (Eq. (9)).

Ephraim and Malah [7] have proposed a decision-directed approach, which provides a very useful estimation method for the a priori SNR [2,21]. Accordingly, if speech presence is assumed ($q(k,\ell) \equiv 0$), then the expression

$$\Xi(k,\ell) = \alpha G^2(k,\ell-1)\gamma(k,\ell-1) + (1-\alpha)\max\{\gamma(k,\ell)-1,0\} \tag{17}$$

can be substituted for the a priori SNR. The first term, $G^2(k,\ell-1)\gamma(k,\ell-1)$, represents the a priori SNR resulting from the processing of the previous frame. The second term, $\max\{\gamma(k,\ell)-1,0\}$, is a maximum likelihood estimate for the a priori SNR, based entirely on the current frame. The parameter α is a weighting factor that controls the trade-off between the noise reduction and the transient distortion brought into the signal [2,7].

Under speech presence uncertainty, according to [7,13], the expression in Eq. (17) estimates a *non-conditional a priori* SNR $\eta(k,\ell) \triangleq E[|X(k,\ell)|^2]/\lambda_d(k,\ell)$, and therefore the estimate for the a priori SNR $\xi(k,\ell)$ should be given by $\Xi(k,\ell)/(1-q(k,\ell))$. However, the division by $1-q(k,\ell)$ may deteriorate the performance of the speech enhancement system [3,23]. In some cases, it introduces interaction between the estimated $q(k,\ell)$ and the a priori SNR, that adversely affects the total gain for noise-only bins and results in an unnaturally structured residual noise [17]. We propose the following expression as an estimate for the a priori SNR when speech presence uncertainty is taken into account:

$$\hat{\xi}(k,\ell) = \alpha G^2_{\mathrm{H}_1}(k,\ell-1)\gamma(k,\ell-1) + (1-\alpha)\max\{\gamma(k,\ell)-1,0\}. \tag{18}$$

Notice that for $q(k,\ell-1)\neq 0$, this expression differs from either $\Xi(k,\ell)$ or $\Xi(k,\ell)/(1-q(k,\ell))$. Generally, we could use two different estimators for $\xi(k,\ell)$ (possibly other than the proposed $\hat{\xi}(k,\ell)$), one for the computation of the conditional gain $G_{\mathrm{H}_1}(k,\ell)$, and one for the computation of the speech presence probability $p(k,\ell)$. However, in the scope of this paper we confine ourselves to show that in place of ξ it is more desirable to use $\hat{\xi}$ rather than $\Xi/(1-q)$, either when evaluating G_{H_1} or p.

Let $\hat{q}(k,\ell)$ be an estimate for the a priori speech absence probability, and let $\Xi(k,\ell)$ and $\hat{\xi}(k,\ell)$ be given by Eqs. (17) and (18), respectively. By definition, if $\mathrm{H}_1(k,\ell)$ is true, then the spectral gain $G(k,\ell)$ should degenerate to $G_{\mathrm{H}_1}(k,\ell)$, and the a priori SNR estimate should coincide with $\Xi(k,\ell)$. On the contrary, if $\mathrm{H}_0(k,\ell)$ is true, then $G(k,\ell)$ should decrease to $G_{\min}$, or equivalently the a priori SNR estimate should be as small as possible. Indeed, if $\mathrm{H}_1(k,\ell)$ is true then

$$\hat{\xi}(k,\ell)|_{\mathrm{H}_1(k,\ell)} \approx \Xi(k,\ell)|_{\mathrm{H}_1(k,\ell)} \leqslant \left.\frac{\Xi(k,\ell)}{1-\hat{q}(k,\ell)}\right|_{\mathrm{H}_1(k,\ell)}, \tag{19}$$

where in (18), $G_{\mathrm{H}_1}(k,\ell-1)$ is used instead of $G(k,\ell-1)$, since if $\mathrm{H}_1(k,\ell)$ is true then $\mathrm{H}_1(k,\ell-1)$ is likely to be true as well, due to the strong correlation of speech presence in successive frames. On the other hand, if $\mathrm{H}_0(k,\ell)$ is true, then $\hat{q}(k,\ell)$ is expected to approach one, and $\hat{\xi}(k,\ell)$ is likely to be

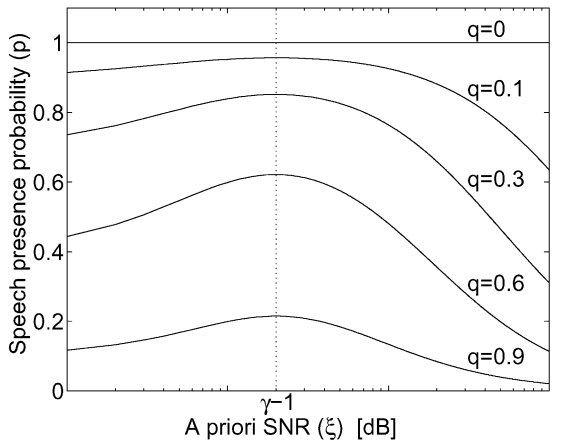

Fig. 2. Conditional speech presence probability $p(k,\ell)$ (Eq. (9)) versus the a priori SNR $\xi(k,\ell)$ for different values of the a priori speech absence probability $q(k,\ell)$ and a fixed a posteriori SNR $\gamma(k,\ell) > 1$.

much smaller than $\Xi(k,\ell)/(1-\hat{q}(k,\ell))$:

$$\begin{aligned}\hat{\xi}(k,\ell)|_{\mathrm{H}_0(k,\ell)} &\approx G_{\min}^2 \ll \left.\frac{\Xi(k,\ell)}{1-\hat{q}(k,\ell)}\right|_{\mathrm{H}_0(k,\ell)} \\ &\approx \frac{\alpha G_{\min}^2 + (1-\alpha)\max\{\gamma(k,\ell)-1,0\}}{1-\hat{q}(k,\ell)}. \end{aligned} \quad (20)$$

Therefore, for the computation of $G_{\mathrm{H}_1}(k,\ell)$, the a priori SNR is more effectively estimated by $\hat{\xi}(k,\ell)$, rather than by $\Xi(k,\ell)/(1-\hat{q}(k,\ell))$.

In case of evaluating the conditional speech presence probability $p(k,\ell)$ (Eq. (9)), the estimator for the a priori SNR should satisfy the following: If speech is present, then $p(k,\ell)$ is maximal; If speech is absent, then $p(k,\ell)$ coincides with the a priori speech presence probability (i.e., $p(k,\ell) \approx 1-\hat{q}(k,\ell)$), since the speech variance cannot be estimated based on frames that do not contain speech. In Fig. 2 several curves for the conditional speech presence probability $p(k,\ell)$ are plotted as a function of the a priori SNR $\xi(k,\ell)$ for various values of the a priori speech absence probability $q(k,\ell)$ and a fixed a posteriori SNR $\gamma(k,\ell) > 1$. The maximal value of $p(k,\ell)$ under speech presence uncertainty ($\hat{q}(k,\ell) > 0$) is obtained for $\xi(k,\ell) = \gamma(k,\ell) - 1$. Accordingly, when speech is present, the estimator for $\xi(k,\ell)$ should on the average resemble $\gamma(k,\ell) - 1$, and when speech is absent that estimator should be kept close to zero (since for $\xi(k,\ell) \approx 0$ we have $p(k,\ell) \approx 1-\hat{q}(k,\ell)$). This is indeed satisfied more favorably by the proposed $\hat{\xi}$ rather than $\Xi/(1-\hat{q})$, since if $\mathrm{H}_1(k,\ell)$ is true then

$$\begin{aligned} &E[\hat{\xi}(k,\ell)|\mathrm{H}_1(k,\ell)] \\ &\quad\approx E[G_{\mathrm{H}_1}^2(k,\ell-1)\gamma(k,\ell-1)|\mathrm{H}_1(k,\ell)] \\ &\quad\approx \frac{E[|X(k,\ell-1)|^2|\mathrm{H}_1(k,\ell-1)]}{\lambda_d(k,\ell-1)} = \xi(k,\ell-1) \\ &\quad\approx E[\gamma(k,\ell)-1|\mathrm{H}_1(k,\ell)], \end{aligned} \quad (21)$$

$$\begin{aligned} &E\left[\left.\frac{\Xi(k,\ell)}{1-\hat{q}(k,\ell)}\right|\mathrm{H}_1(k,\ell)\right] \\ &\quad\approx \frac{1}{1-\hat{q}(k,\ell)}E[G^2(k,\ell-1) \\ &\qquad\gamma(k,\ell-1)|\mathrm{H}_1(k,\ell)] \\ &\quad\approx \frac{\xi(k,\ell)}{1-\hat{q}(k,\ell)} \\ &\quad\geqslant E[\gamma(k,\ell)-1|\mathrm{H}_1(k,\ell)] \end{aligned} \quad (22)$$

and if $\mathrm{H}_0(k,\ell)$ is true, then by (20) $\hat{\xi}(k,\ell) \ll \Xi(k,\ell)/(1-\hat{q}(k,\ell))$. Consequently, under speech presence uncertainty it is preferable to estimate the a priori SNR by $\hat{\xi}(k,\ell)$ rather than $\Xi(k,\ell)/(1-\hat{q}(k,\ell))$, whether evaluating the conditional gain $G_{\mathrm{H}_1}(k,\ell)$ or the conditional speech presence probability $p(k,\ell)$.

4. A priori speech absence probability estimation

In this section we derive an efficient estimator $\hat{q}(k,\ell)$ for the a priori speech absence probability. This estimator uses a soft-decision approach to compute three parameters based on the time–frequency distribution of the estimated a priori SNR, $\hat{\xi}(k,\ell)$. The parameters exploit the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

Let $\zeta(k,\ell)$ be a recursive average of the a priori SNR with a time constant β,

$$\zeta(k,\ell) = \beta\zeta(k,\ell-1) + (1-\beta)\hat{\xi}(k,\ell-1). \quad (23)$$

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

$$\zeta_\lambda(k,\ell) = \sum_{i=-w_\lambda}^{w_\lambda} h_\lambda(i)\zeta(k-i,\ell), \quad (24)$$

where the subscript λ designates either "local" or "global", and h_λ is a normalized window of size $2w_\lambda+1$. We define two parameters, $P_{\text{local}}(k,\ell)$ and $P_{\text{global}}(k,\ell)$, which represent the relation between the above averages and the likelihood of speech in the kth frequency bin of the ℓth frame. These parameters are given by

$$P_\lambda(k,\ell) = \begin{cases} 0 & \text{if } \zeta_\lambda(k,\ell) \leqslant \zeta_{\min}, \\ 1 & \text{if } \zeta_\lambda(k,\ell) \geqslant \zeta_{\max}, \\ \frac{\log(\zeta_\lambda(k,\ell)/\zeta_{\min})}{\log(\zeta_{\max}/\zeta_{\min})} & \text{otherwise,} \end{cases} \quad (25)$$

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

In order to further attenuate noise in noise-only frames, we define a third parameter, $P_{\text{frame}}(\ell)$, which is based on the speech energy in neighboring frames. An averaging of $\zeta(k,\ell)$ in the frequency domain (possibly over a certain frequency band) yields

$$\zeta_{\text{frame}}(\ell) = \underset{1\leqslant k\leqslant M/2+1}{\text{mean}}\{\zeta(k,\ell)\}. \quad (26)$$

To prevent clipping of speech startings or weak components, speech is assumed whenever $\zeta_{\text{frame}}(\cdot)$ increases. Moreover, the transition from H_1 to H_0 is delayed, which reduces the misdetection of weak speech tails, by allowing for a certain decrease in the value of ζ_{frame}. Fig. 3 describes a block diagram for computing $P_{\text{frame}}(\ell)$, where

$$\mu(\ell) \triangleq \begin{cases} 0, & \text{if } \zeta_{\text{frame}}(\ell) \leqslant \zeta_{\text{peak}}(\ell)\zeta_{\min}, \\ 1, & \text{if } \zeta_{\text{frame}}(\ell) \geqslant \zeta_{\text{peak}}(\ell)\zeta_{\max}, \\ \frac{\log(\zeta_{\text{frame}}(\ell)/\zeta_{\text{peak}}(\ell)/\zeta_{\min})}{\log(\zeta_{\max}/\zeta_{\min})} & \text{otherwise} \end{cases} \quad (27)$$

represents a soft transition from "speech" to "noise", ζ_{peak} is a confined peak value of ζ_{frame}, and $\zeta_{p\min}$ and $\zeta_{p\max}$ are empirical constants that determine the delay of the transition.

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

$$\hat{q}(k,\ell) = 1 - P_{\text{local}}(k,\ell)P_{\text{global}}(k,\ell)P_{\text{frame}}(\ell). \quad (28)$$

Accordingly, $\hat{q}(k,\ell)$ is larger if either previous frames, or recent neighboring frequency bins, do not contain speech. When $q(k,\ell) \to 1$, the conditional speech presence probability $p(k,\ell) \to 0$ by (9), and consequently the gain function reduces to $G_{\min}$ by (16). Therefore, to reduce the possibility of speech distortion we restrict $\hat{q}(k,\ell)$ to be smaller than a threshold $q_{\max}$ ($q_{\max} < 1$).

5. Noise spectrum estimation

In this section we address the problem of estimating the time-varying spectrum of the noise, $\lambda_d(k,\ell)$. A common technique is to apply a temporal recursive smoothing to the noisy measurement during periods of speech absence. In particular,

$$\begin{aligned} &H_0'(k,\ell)\text{: } \hat{\lambda}_d(k,\ell+1) \\ &\quad = \alpha_d\hat{\lambda}_d(k,\ell) + (1-\alpha_d)|Y(k,\ell)|^2, \\ &H_1'(k,\ell)\text{: } \hat{\lambda}_d(k,\ell+1) = \hat{\lambda}_d(k,\ell), \end{aligned} \quad (29)$$

where α_d ($0 < \alpha_d < 1$) is a smoothing parameter, and H_0' and H_1' designate hypothetical speech absence and presence, respectively. In the proposed MCRA estimation approach, we make a distinction between the hypotheses in Eqs. (5), used for estimating the clean speech, and the hypotheses in Eqs. (29), which control 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 [4], and generally we tend to decide H_1 with a higher confidence than H_1', i.e. $P(H_1|Y) \geqslant P(H_1'|Y)$.

Let $p'(k,\ell) \triangleq P(H_1'(k,\ell)|Y(k,\ell))$ denote the conditional speech presence probability. Then (29)

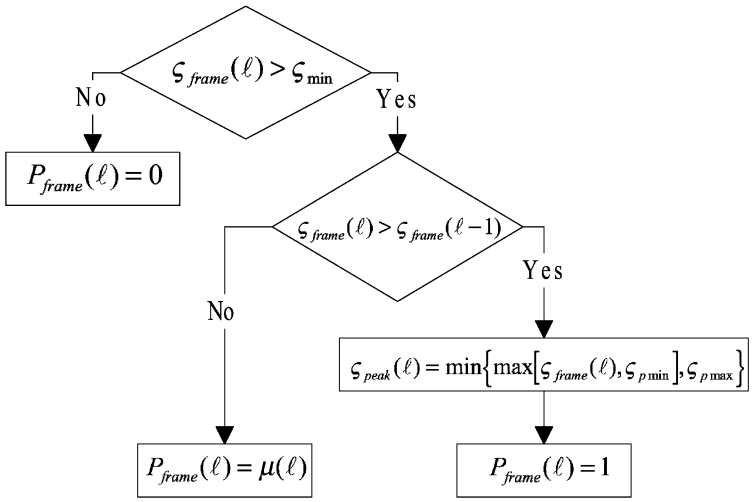

Fig. 3. Block diagram for computing P_{frame} (a parameter representing the likelihood of speech in a given frame).

implies

$$\begin{aligned}\hat{\lambda}_d(k,\ell+1) &= \hat{\lambda}_d(k,\ell)p'(k,\ell) + [\alpha_d\hat{\lambda}_d(k,\ell) \\ &\quad + (1-\alpha_d)|Y(k,\ell)|^2](1-p'(k,\ell)) \\ &= \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) + [1-\tilde{\alpha}_d(k,\ell)]|Y(k,\ell)|^2,\end{aligned} \tag{30}$$

where

$$\tilde{\alpha}_d(k,\ell) \triangleq \alpha_d + (1-\alpha_d)p'(k,\ell) \tag{31}$$

is a time-varying smoothing parameter. Accordingly, the noise spectrum can be estimated by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability.

Tracking the conditional speech presence probability is based on the local statistics in the time–frequency plane of the energy of the noisy signal. We determine speech absence in a given frame of a subband by the ratio between the local energy of the noisy signal and its minimum within a specified time window. The ratio is compared to a certain threshold value, where a smaller ratio indicates absence of speech. Subsequently, a recursive temporal averaging is carried out, to reduce fluctuations between speech and non-speech segments. Thus, the strong correlation of speech presence in neighboring frames is once more taken into account.

The local energy of the noisy signal is obtained by smoothing the magnitude squared of its STFT in time and frequency. In frequency, we use a window function b whose length is $2w+1$:

$$S_f(k,\ell) = \sum_{i=-w}^{w} b(i)\,|Y(k-i,\ell)|^2. \tag{32}$$

In time, the smoothing is performed by a first order recursive averaging, given by

$$S(k,\ell) = \alpha_s S(k,\ell-1) + (1-\alpha_s)S_f(k,\ell), \tag{33}$$

where α_s $(0 < \alpha_s < 1)$ is a parameter. The minimum of the local energy, $S_{\min}(k,\ell)$, is searched using a simplified form of the procedure proposed in [15]. First, the minimum and a temporary variable $S_{\text{tmp}}(k,\ell)$ are initialized by $S_{\min}(k,0) = S(k,0)$ and $S_{\text{tmp}}(k,0) = S(k,0)$. Then, a samplewise comparison of the local energy and the minimum value of the previous frame yields the minimum value for the current frame:

$$S_{\min}(k,\ell) = \min\{S_{\min}(k,\ell-1), S(k,\ell)\}, \tag{34}$$

$$S_{\text{tmp}}(k,\ell) = \min\{S_{\text{tmp}}(k,\ell-1), S(k,\ell)\}. \tag{35}$$

Whenever L frames have been read, i.e. ℓ is divisible by L, the temporary variable is employed and initialized by

$$S_{\min}(k,\ell) = \min\{S_{\text{tmp}}(k,\ell-1), S(k,\ell)\}, \tag{36}$$

$$S_{\text{tmp}}(k,\ell) = S(k,\ell) \tag{37}$$

and the search for the minimum continues with Eqs. (34) and (35). The parameter L determines the resolution of the local minima search. The local minimum is based on a window of at least L frames, but not more than $2L$ frames. The lower limit constraint should guarantee that the local minimum is associated with the noise, and not biased upwards during "continuous" speech. The upper limit, on the other hand, should control the bias downwards when the noise level increases. According to [15] and our own experiments with different speakers and environmental conditions, this can be satisfied with window lengths of approximately $0.5s$–$1.5s$.

Let $S_r(k,\ell) \triangleq S(k,\ell)/S_{\min}(k,\ell)$ denote the ratio between the local energy of the noisy signal and its derived minimum. A Bayes minimum-cost decision rule is given by

$$\frac{p(S_r|\mathrm{H}_1)}{p(S_r|\mathrm{H}_0)} \underset{\mathrm{H}_0'}{\overset{\mathrm{H}_0'}{\gtrless}} \frac{c_{10}P(\mathrm{H}_0)}{c_{01}P(\mathrm{H}_1)}, \tag{38}$$

where $P(\mathrm{H}_0)$ and $P(\mathrm{H}_1)$ are the a priori probabilities for speech absence and presence, respectively, and c_{ij} is the cost for deciding H_i' when H_j'. Fig. 4 shows representative examples of conditional probability density functions, $p(S_r|\mathrm{H}_0)$ and $p(S_r|\mathrm{H}_1)$, obtained experimentally for white Gaussian noise, car interior noise and F16 cockpit noise, at -5 dB segmental SNR. Since the likelihood ratio $p(S_r|\mathrm{H}_1)/p(S_r|\mathrm{H}_0)$ is a monotonic function, the decision rule of (38) can be expressed as

$$S_r(k,\ell) \underset{\mathrm{H}_0'}{\overset{\mathrm{H}_1'}{\gtrless}} \delta. \tag{39}$$

We propose the following estimator for the conditional speech presence probability:

$$\hat{p}'(k,\ell) = \alpha_p \hat{p}'(k,\ell-1) + (1-\alpha_p) I(k,\ell), \tag{40}$$

where α_p $(0 < \alpha_p < 1)$ is a smoothing parameter, and $I(k,\ell)$ denotes an indicator function for the result in (39), i.e. $I(k,\ell) = 1$ if $S_r(k,\ell) > \delta$ and $I(k,\ell) = 0$ otherwise. The merit of this estimate is threefold. Firstly, δ is not sensitive to the type and intensity of environmental noise. Secondly, the probability of $|Y|^2 \gg \lambda_d$ is very small when $S_r < \delta$. Hence, an increase in the estimated noise, consequent upon falsely deciding H_0' when H_1', is

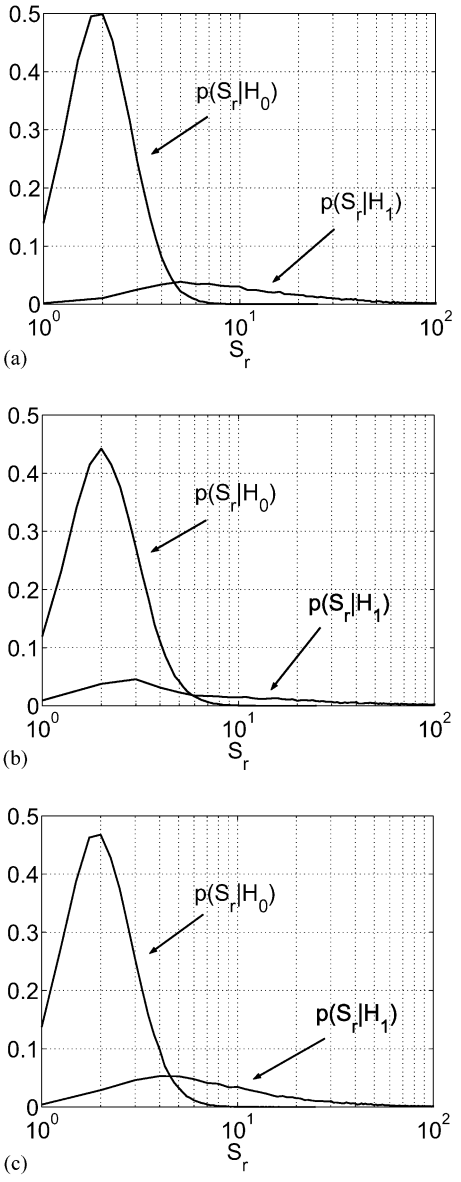

Fig. 4. Hypothetical probability density functions, $p(S_r|\mathrm{H}_0)$ and $p(S_r|\mathrm{H}_1)$, for: (a) white Gaussian noise; (b) car interior noise; (c) F16 cockpit noise.

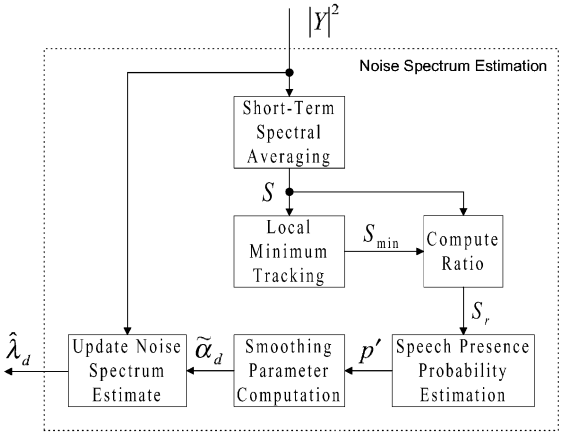

Fig. 5. Block diagram of the MCRA noise spectrum estimation.

not significant.[2] Thirdly, the strong correlation of speech presence in consecutive frames is utilized (via α_p). A block diagram of the MCRA noise spectrum estimation is described in Fig. 5.

6. Performance evaluation

The performance evaluation consists of two parts. First, we compare the MCRA noise estimate to the minimum statistics [15] and conventional weighted average [11] noise estimates. Then, the OM-LSA speech estimator is examined in comparison to the MM-LSA estimator [13] and to the original STSA and LSA estimators [7,8]. The evaluation includes an objective improvement in segmental SNR measure, a subjective study of speech spectrograms and informal listening tests.

Three different noise types, taken from Noisex92 database [26], are used in our evaluation: white Gaussian noise, car noise, and F16 cockpit noise. Since noise signals have different impacts on different speech signals, the performance results are averaged out using six different utterances, taken from the TIMIT database [9]. Half of the utterances are from male speakers, and half are from female speakers. The speech signals, sampled at a frequency of 16 kHz, are degraded by the various noise types with segmental SNR's in the range $[-5, 10]$ dB. The segmental SNR is defined by [20]

$$\text{SegSNR} = \frac{10}{|\mathscr{L}|}\sum_{\ell\in\mathscr{L}} \log\frac{\sum_{k=0}^{N/2}|X(k,\ell)|^2}{\sum_{k=0}^{N/2}|D(k,\ell)|^2}, \qquad (41)$$

where $\mathscr{L}$ represents the set of frames that contain speech, and $|\mathscr{L}|$ its cardinality. Values of

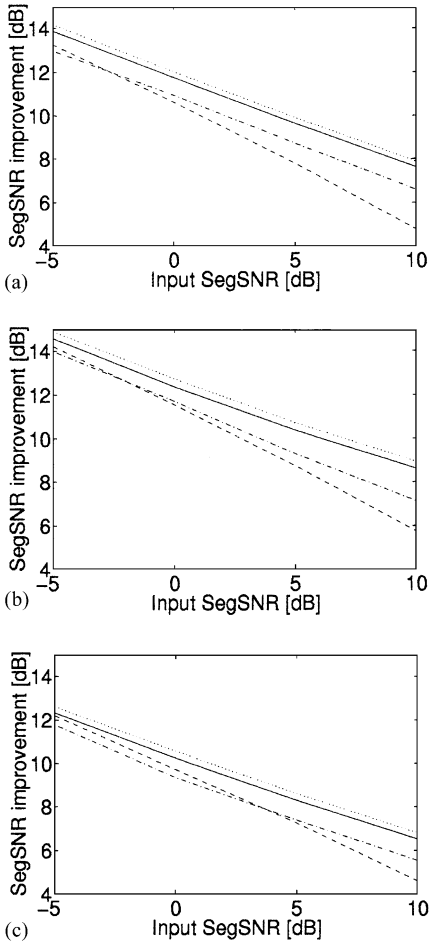

Fig. 6. Comparison of noise estimation methods, minimum statistics (dashed) weighted average (dashdot) and MCRA (solid), for various noise types and levels. Average segmental SNR improvement using the OM-LSA speech estimator for: (a) white Gaussian noise; (b) car interior noise; (c) F16 cockpit noise. A theoretical limit is obtained by calculating the noise spectrum from the noise itself (dotted).

[2] In practice, to completely avoid such an increase in the estimated noise, the decision rule is formulated based also on the ratio between the *instantaneous* energy and the local minimum. Specifically, the indicator function is given by $I(k,\ell)=1$ if $S_r(k,\ell) > \delta$ or $|Y(k,\ell)|^2/S_{\min}(k,\ell) > 1.8\,\delta$, and $I(k,\ell)=0$ otherwise.

Table 1
Values of parameters used in the implementation of the OM-LSA speech estimator and the MCRA noise estimator

Spectral analysis and synthesis:		
$N = 512$	$M = 128$	
h, $\tilde{h}$: Biorthogonal Hanning windows		
Noise spectrum estimation		
$\alpha_d = 0.95$	$\alpha_p = 0.2$	$L = 125$
$\alpha_s = 0.8$	$w = 1$	$\delta = 5$
b: Hanning window		
A priori speech absence probability estimation		
$\beta = 0.7$	$\zeta_{\min} = -10$ dB	$\zeta_{p\min} = 0$ dB
$w_{\text{local}} = 1$	$\zeta_{\max} = -5$ dB	$\zeta_{p\max} = 10$ dB
$w_{\text{global}} = 15$	$q_{\max} = 0.95$	
h_{local}, h_{global}: Hanning windows		
Noise reduction and speech distortion tradeoff		
$\alpha = 0.92$	$G_{\min} = -25$ dB	

parameters used in the implementation of the OM-LSA and MCRA estimators are summarized in Table 1, and those used in the implementation of competitive methods are summarized in Table 2.

Table 2
Values of parameters used in the implementation of competitive speech enhancement and noise estimation methods (the notation is after the referenced publications). The spectral analysis and synthesis parameters are the same as in Table 1

Minimum statistics noise estimation [15]	
$\alpha = 0.95$	Smoothing constant
$D = 100$	Window length
$M = 25$	Sub-window length
$o_{\min} = 1.5$	Bias compensation factor
Weighted average noise estimation [11]	
$\alpha = 0.95$	Weighting parameter
$\beta = 2$	Threshold
Short-time spectral amplitude (STSA) estimator [7] *and log-spectral amplitude (LSA) estimator* [8]	
$\alpha = 0.92$	Weighting factor for the a priori SNR estimation
$\eta_{\min} = -25$ dB	Lower limit for the a priori SNR
Multiplicatively modified LSA (MM-LSA) estimator [13]	
$\alpha = 0.92$	Weighting factor for the a priori SNR estimation
$\eta_{\min} = -25$ dB	Lower limit for the a priori SNR
$\gamma_{\text{TH}} = 0.8$	Threshold for hypothesis testing
$\alpha_q = 0.95$	Smoothing parameter for q estimation

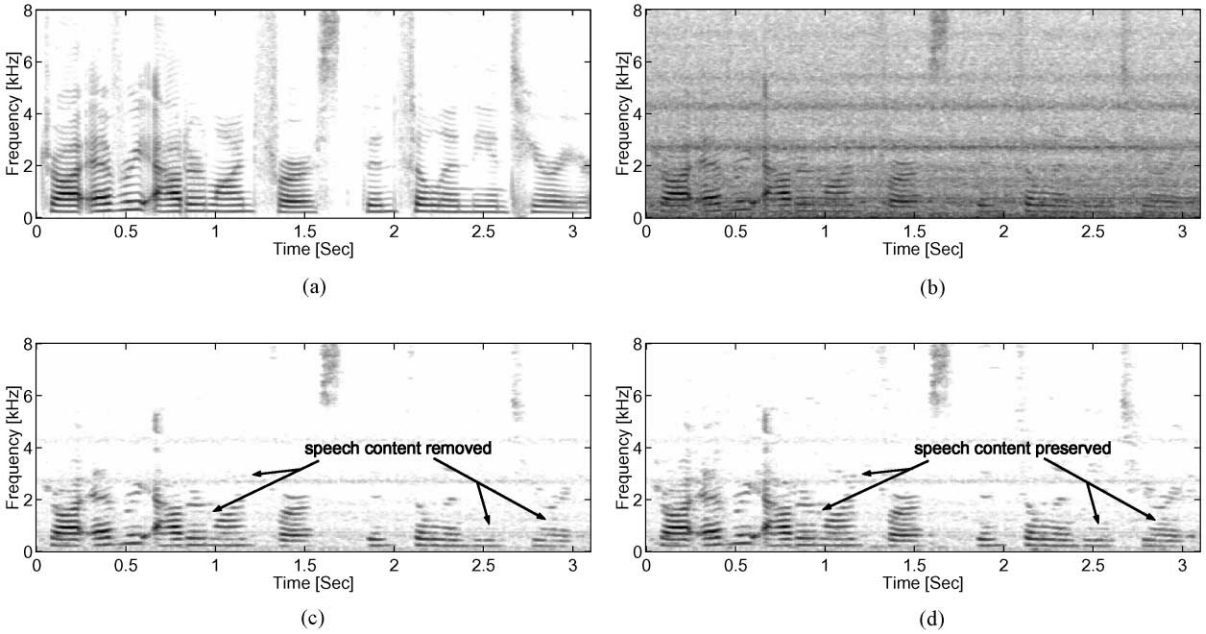

Fig. 7. Speech spectrograms. (a) Original clean speech signal: Draw every outer line first, then fill in the interior; (b) noisy signal (additive F16 cockpit noise at a SegSNR = 0 dB); (c) speech enhanced using the OM-LSA estimator and the minimum statistics noise estimate (SegSNR = 8.8 dB); (d) speech enhanced using the OM-LSA estimator and the MCRA noise estimate (SegSNR = 9.17 dB).

The comparison between the MCRA noise estimate and the minimum statistics and weighted average noise estimates is accomplished by evaluating their performance when incorporated in the OM-LSA estimator. A theoretical limit, achievable by calculating the noise spectrum from the noise itself, is also considered. Fig. 6 shows the average segmental SNR improvement obtained for various noise types and at various noise levels. It can be readily seen that the MCRA approach consistently achieves the best results under all noise conditions, and its performance is close to the theoretical limit.

The segmental SNR measure takes into account both residual noise and speech distortion. Since it lacks indication about the structure of the residual noise, a subjective comparison was conducted using speech spectrograms and validated by informal listening tests. Example of speech spectrograms obtained with the MCRA and minimum statistics noise estimates are shown in Fig. 7. Particularly, compare the low frequency formants, which have been removed with the minimum statistics noise estimate, but are well preserved with the MCRA noise estimate. The minimum statistics noise estimate relies on the assumptions that the power of the noisy speech frequently decays to the power of the noise signal, and its bias can be compensated by a constant correction factor [15,17]. Occasionally, these assumptions are not satisfied and the noise is overestimated, resulting in the attenuation of low SNR phonemes.

The comparison between the OM-LSA estimator and the MM-LSA, STSA and LSA estimators is performed with the noise estimated by the MCRA approach. The average improvements in segmental SNR, obtained by these estimators, are shown in Fig. 8 (notice that the solid lines in Figs. 6 and 8 are the same). The OM-LSA estimate achieves the best results under all noise conditions. Its superiority is more significant for low input SNRs. A comparison using speech spectrograms is shown in Fig. 9. In contrast to the MM-LSA, STSA and LSA estimators, where high a posteriori SNR produces high spectral gains resulting in a random appearance of tone-like noise (musical-noise phenomena), the OM-LSA estimator attenuates noise by identifying noise-only regions ($\hat{q} \to q_{\max}$) and reducing the gain correspondingly to $G_{\min}$. Yet, it avoids the attenuation of weak speech components by letting $\hat{q}$ descend to zero in speech regions.

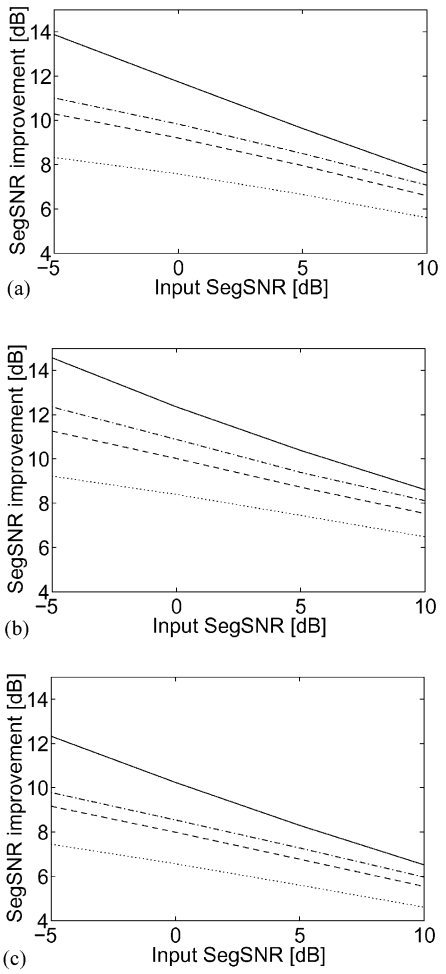

Fig. 8. Comparison of speech estimators, STSA (dotted) LSA (dashed) MM-LSA (dashdot) and OM-LSA (solid), for various noise types and levels. Average segmental SNR improvement using the MCRA noise estimate for: (a) white Gaussian noise; (b) car interior noise; (c) F16 cockpit noise.

7. Conclusion

We have described a speech enhancement system for non-stationary noise environments. The system comprises an OM-LSA speech estimator and a MCRA noise estimate. The spectral gain function

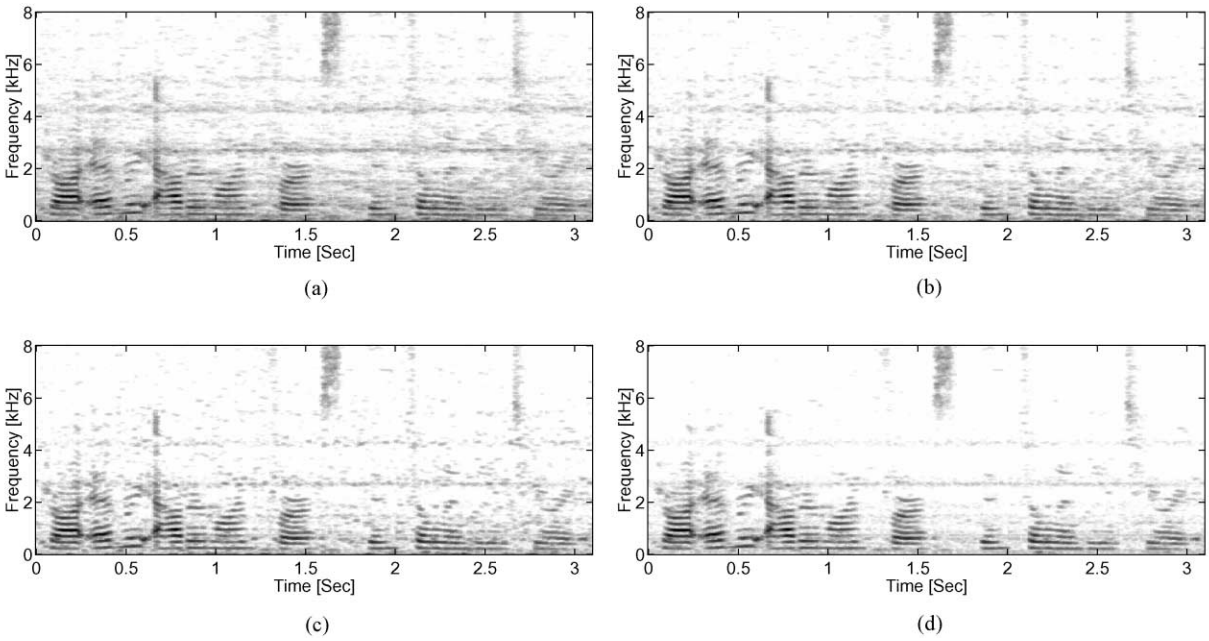

Fig. 9. Comparison of speech estimators. Spectrograms of enhanced speech using: (a) the STSA estimator (SegSNR = 6.5 dB); (b) the LSA estimator (SegSNR = 7.8 dB); (c) the MM-LSA estimator (SegSNR = 8.4 dB); (d) the OM-LSA estimator (SegSNR = 9.7 dB). Noise is estimated by the MCRA method. The original and noisy speech are depicted in Fig. 7.

is obtained by modifying the gain function of the conventional LSA estimator, based on a binary hypothesis model. The modification includes a lower bound for the gain, which is determined by a subjective criteria for the noise naturalness, and exponential weights, which are given by the conditional speech presence probability. Moreover, based on the decision-directed approach of Ephraim and Malah [7], we proposed an estimator $\hat{\xi}(k,\ell)$ (Eq. (18)) for the a priori SNR, which is shown to be preferable to that commonly used in other works (e.g. [7,8,12,13,16]).

The noise spectrum estimate is formulated as a recursive average of the noisy observed signal's STFT, where the smoothing parameter is adjusted by the speech presence probability. We differentiate between the speech presence/absence hypotheses used for estimating the clean speech, and the hypotheses that control the adaptation of the noise spectrum. Presence of speech, in the case of noise estimation, is determined by the ratio between the local energy of the noisy signal and its minimum within a specified time window. In the case of speech estimation, on the other hand, the a priori speech absence probability is related to the time-frequency distribution of the estimated a priori SNR.

The proposed OM-LSA and MCRA estimators have been tested and compared to conventional speech and noise estimators, in various noise types and levels. The evaluation consisted of an objective improvement in segmental SNR measure, study of speech spectrograms and subjective listening tests. Results show that both the OM-LSA and MCRA estimators achieve the best performance under all tested environmental conditions. Combining these estimators, excellent noise suppression is obtained, while retaining weak speech components and avoiding the musical residual noise phenomena.

Acknowledgements

The authors thank Prof. David Malah for enlightening discussions particularly on the estimation of the a priori SNR, Dr. Rainer Martin for making his

Minimum Statistics code available, and the anonymous reviewers for their helpful comments.

References

[1] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, Proceedings of the Fourth IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-79, Washington, DC, 2–4 April 1979, pp. 208–211.

[2] O. Cappé, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Trans. Speech Audio Process. 2 (2) (April 1994) 345–349.

[3] I. Cohen, On speech enhancement under signal presence uncertainty, Proceedings of the 26th IEEE International Conference on Acoustics Speech, and Signal Processing, ICASSP-01, Salt Lake City, Utah, 7–11 May 2001.

[4] I. Cohen, B. Berdugo, spectral enhancement by tracking speech presence probability in subbands, Proceedings of IEEE Workshop on Hands Free Speech Communication, HSC'01, Kyoto, Japan, 9–11 April 2001, pp. 95–98.

[5] R.E. Crochiere, L.R. Rabiner, Multirate Digital Signal Processing., Prentice-Hall, Englewood Cliffs, NJ, 1983.

[6] G. Doblinger, Computationally efficient speech enhancement by spectral minima tracking in subbands, Proceedings of the Fourth European Conference on Speech, Communication and Technology, EUROSPEECH'95, Madrid, Spain, 18–21 September 1995, pp. 1513–1516.

[7] 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) (December 1984) 1109–1121.

[8] 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) (April 1985) 443–445.

[9] J.S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, (prototype as of December 1988).

[10] E.B. George, Single-sensor speech enhancement using a soft-decision/variable attenuation algorithm, Proceedings of the 20th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 816–819.

[11] H.G. Hirsch, C. Ehrlicher, Noise estimation techniques for robust speech recognition, Proceedings of the 20th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 153–156.

[12] N.S. Kim, J.-H. Chang, Spectral enhancement based on global soft decision, IEEE Signal Process. Lett. 7 (5) (May 2000) 108–110.

[13] D. Malah, R.V. Cox, A.J. Accardi, Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments, Proceedings of the 24th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-99, Phoenix, Arizona, 15–19 March 1999, pp. 789–792.

[14] R. Martin, An efficient algorithm to estimate the instantaneous SNR of speech signals, Proceedings of the Second European Conference on Speech, Communication and Technology, EUROSPEECH'93, Berlin, Germany, 21–23 September 1993, pp. 1093–1096.

[15] R. Martin, Spectral subtraction based on minimum statistics, Proceedings of the Seventh European Signal Processing Conference, EUSIPCO-94, Edinburgh, Scotland, 13–16 September 1994, pp. 1182–1185.

[16] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Trans. Speech Audio Process. 9 (5) (July 2001) 504–512.

[17] R. Martin, I. Wittke, P. Jax, Optimized estimation of spectral parameters for the coding of noisy speech, Proceedings of the 25th IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP-00, Istanbul, Turkey, 5–9 June 2000, pp. 1479–1482.

[18] R.J. McAulay, M.L. Malpass, Speech enhancement using a soft-decision noise suppression filter, IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (2) (April 1980) 137–145.

[19] J. Meyer, K.U. Simmer, K.D. Kammeyer, Comparison of one- and two-channel noise-estimation techniques, Proceedings of the Fifth International Workshop on Acoustic Echo and Noise Control, IWAENC-97, London, UK, 11–12 September 1997, pp. 137–145.

[20] S. Quackenbush, T. Barnwell, M. Clements, Objective Measures of Speech Quality, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[21] P. Scalart, J. Vieira-Filho, Speech enhancement based on a priori signal to noise estimation, Proceedings of the 21th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-96, Atlanta, Georgia, 7–10 May 1996, pp. 629–632.

[22] J. Sohn, N.S Kim, W. Sung, A statistical model-based voice activity detector, IEEE Signal Process. Lett. 6 (1) (January 1999) 1–3.

[23] I.Y. Soon, S.N. Koh, C.K. Yeo, Improved noise suppression filter using self-adaptive estimator of probability of speech absence, Signal Processing 75 (1999) 151–159.

[24] V. Stahl, A. Fischer, R. Bippus, Quantile based noise estimation for spectral subtraction and Wiener filtering, Proceedings of the 25th IEEE International Conference on Acoustics, Speech and Signal

Processing, ICASSP-00, Istanbul, Turkey, 5–9 June 2000, pp. 1875–1878.

[25] T.S. Sun, S. Nandkumar, J. Carmody, J. Rothweiler, A. Goldschen, N. Russell, S. Mpasi, P. Green, Speech enhancement using a ternary-decision based filter, Proceedings of the 20th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 820–823.

[26] 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) (July 1993) 247–251.

[27] D.L. Wang, J.S. Lim, The unimportance of phase in speech enhancement, IEEE Trans. Acoust. Speech Signal Process. ASSP-30 (4) (August 1982) 679–681.

[28] J. Wexler, S. Raz, Discrete Gabor expansions, Signal Processing 21 (3) (November 1990) 207–220.

[29] J. Yang, Frequency domain noise suppression approaches in mobile telephone systems, Proceedings of the 18th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-93, Minneapolis, Minnesota, 27–30 April 1993, pp. 363–366.