Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging

Israel Cohen

Abstract— **Noise spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. In this paper, we present an *Improved Minima Controlled Recursive Averaging* (IMCRA) approach, for noise estimation in adverse environments involving non-stationary noise, weak speech components, and low input signal-to-noise ratio (SNR). The noise estimate is obtained by averaging past spectral power values, using a time-varying frequency-dependent smoothing parameter that is adjusted by the signal presence probability. The speech presence probability is controlled by the minima values of a smoothed periodogram. The proposed procedure comprises two iterations of smoothing and minimum tracking. The first iteration provides a rough voice activity detection in each frequency band. Then, smoothing in the second iteration excludes relatively strong speech components, which makes the minimum tracking during speech activity robust. We show that in non-stationary noise environments and under low SNR conditions, the IMCRA approach is very effective. In particular, compared to a competitive method, it obtains a lower estimation error, and when integrated into a speech enhancement system achieves improved speech quality and lower residual noise.**

I. Introduction

NOISE power spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. The robustness of such systems, particularly under low signal-to-noise ratio (SNR) conditions and non-stationary noise environments, is greatly affected by the capability to reliably track fast variations in the statistics of the noise. Traditional noise estimation methods, which are based on voice activity detectors (VAD's), restrict the update of the estimate to periods of speech absence. Additionally, VAD's are generally difficult to tune and their reliability severely deteriorates for weak speech components and low input SNR [16], [20], [15]. Alternative techniques, based on histograms in the power spectral domain [14], [10], [19], are computationally expensive, require much memory resources, and do not perform well in low SNR conditions. Furthermore, the signal segments used for building the histograms are typically of several hundred milliseconds, and thus the update rate of the noise estimate is essentially moderate.

A useful noise estimation approach, known as the Minimum Statistics (MS) [12], is to track the minima values of a smoothed power estimate of the noisy signal, and multiply the result by a factor that compensates the bias. However, the variance of this noise estimate is about twice as large as the variance of a conventional noise estimator [12]. Moreover, this method may occasionally attenuate low energy phonemes, particularly if the minimum search window is too short [4]. These limitations can be overcome, at the price of significantly higher complexity, by adapting the smoothing parameter and the bias compensation factor in time and frequency [13]. A computationally more efficient minimum tracking scheme is presented in [5]. 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 [16]. Other closely related techniques are the *lower-energy envelope tracking* [19] and the *quantile based* [21] estimation methods. Rather than picking the minima values of a smoothed periodogram, the noise is estimated based on a temporal quantile of a non-smoothed periodogram of the noisy signal. Unfortunately, these methods suffer from the high computational complexity associated with the sorting operation, and the extra memory required for keeping past spectral power values.

Recently, we introduced a noise estimation approach, namely *Minima Controlled Recursive Averaging* (MCRA) [3], [4], that combines the robustness of the minimum tracking with the simplicity of the recursive averaging. The noise estimate is obtained by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands. The speech presence probability is controlled by the minima values of a smoothed periodogram. In contrast to the MS and related methods, the minimum tracking is not crucial, since it only controls the recursive averaging as a secondary procedure. The recursive averaging is carried out without a hard distinction between speech absence and presence, thus continuously updating the noise estimate even during weak speech activity. Additionally, the smoothing of the noisy periodogram is carried out in both time and frequency, which takes into account the strong correlation of speech presence in neighboring frequency bins of consecutive frames. We have shown that the MCRA noise estimate is computationally efficient, and characterized by the ability to quickly follow abrupt changes in the noise spectrum.

In this paper, we further improve the MCRA estimator with regard to the following aspects: Minimum tracking during speech activity, speech presence probability estimation, and derivation of a bias compensation factor. The

The author is with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel (email: icohen@ee.technion.ac.il; tel.: +972-4-8294731; fax: +972-4-8323041).

This work was partly carried out at Lamar Signal Processing Ltd., Andrea Electronics Corp. - Israel, P.O.Box 573, Yokneam Ilit 20692, Israel.

proposed procedure comprises two iterations of smoothing and minimum tracking. The first iteration provides a rough voice activity detection in each frequency band. Then, the smoothing in the second iteration excludes relatively strong speech components, which makes the minimum tracking during speech activity robust. This facilitates larger smoothing windows, and thus a decreased variance of the minima values. The estimation of the speech presence probability is based on a Gaussian statistical model [6]. However, the *a priori* speech absence probability is controlled by the result of the minimum tracking. We show that this prevents the estimated noise from increasing during weak speech activity, especially when the input SNR is low. The speech presence probability is biased toward higher values to avoid speech distortions in speech enhancement applications. Accordingly, we include in the noise estimator a factor to compensate its bias. We show that the value of the bias compensation factor is determined by the *a priori* speech absence probability estimator, and an explicit expression is derived.

Objective and subjective evaluation of the *Improved Minima Controlled Recursive Averaging* (IMCRA) estimator is performed under various environmental conditions. We examine the tracking capability for non-stationary noise, the segmental relative estimation error for various noise types and levels, and the improvement in the segmental SNR when integrated into a speech enhancement system. We show that compared to the MS method, the proposed noise estimate is superior. Specifically, it responses more quickly to noise variations, it obtains significantly lower estimation error, and yields a higher improvement in the segmental SNR. The advantages of the IMCRA method are particularly notable in adverse environments involving non-stationary noise, weak speech components, and low input SNR.

The paper is organized as follows. In Section II, we present the IMCRA noise estimator. The recursive averaging is accomplished through a time-varying frequency-dependent smoothing parameter, which is adapted under the speech presence uncertainty. In Section III, we introduce an estimator for the *a priori* speech absence probability. The estimator is controlled by the minima values of a smoothed periodogram of the noisy signal. In Section IV, we combine the time-varying recursive averaging with the minima-controlled estimation of the *a priori* speech absence probability, and present the IMCRA algorithm. Finally, in Section V, we evaluate the proposed method, and discuss experimental results, which validate its effectiveness.

II. Time-Varying Recursive Averaging

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

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively. The observed signal $y(n)$ is divided into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). In the time-frequency domain we have $Y(k,\ell) = X(k,\ell) + D(k,\ell)$, where k represents the frequency bin index, and ℓ the frame index. Given two hypotheses, $H_0(k,\ell)$ and $H_1(k,\ell)$, which indicate respectively speech absence and presence in the kth frequency bin of the ℓth frame, and assuming a complex Gaussian distribution of the STFT coefficients for both speech and noise [6], the conditional probability density functions (PDF's) of the observed signal are given by

$$f\left(Y(k,\ell) \mid H_0(k,\ell)\right) = \frac{1}{\pi\lambda_d(k,\ell)} \exp\left\{-\frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)}\right\} \quad (1)$$

$$f\left(Y(k,\ell) \mid H_1(k,\ell)\right) = \frac{1}{\pi(\lambda_x(k,\ell)+\lambda_d(k,\ell))} \cdot \exp\left\{-\frac{|Y(k,\ell)|^2}{\lambda_x(k,\ell)+\lambda_d(k,\ell)}\right\} \quad (2)$$

where $\lambda_x(k,\ell) \triangleq E\left\{|X(k,\ell)|^2 \mid H_1(k,\ell)\right\}$ and $\lambda_d(k,\ell) \triangleq E\left\{|D(k,\ell)|^2\right\}$ denote respectively the short-term spectrum of the speech and noise signals.

Let the *a posteriori* and *a priori* SNR's be defined by [14], [6]

$$\gamma(k,\ell) \triangleq \frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)}, \quad (3)$$

$$\xi(k,\ell) \triangleq \frac{\lambda_x(k,\ell)}{\lambda_d(k,\ell)}. \quad (4)$$

Then, the conditional PDF's of the *a posteriori* SNR can be written as

$$f\left(\gamma(k,\ell) \mid H_0(k,\ell)\right) = e^{-\gamma(k,\ell)} u(\gamma(k,\ell)) \quad (5)$$

$$f\left(\gamma(k,\ell) \mid H_1(k,\ell)\right) = \frac{1}{1+\xi(k,\ell)} \cdot \exp\left\{-\frac{\gamma(k,\ell)}{1+\xi(k,\ell)}\right\} u(\gamma(k,\ell)) \quad (6)$$

where $u(\cdot)$ is the unit step function (*i.e.*, $u(\gamma) = 1$ for $\gamma \geq 0$ and $u(\gamma) = 0$ otherwise). Applying Bayes rule for the conditional speech presence probability $p(k,\ell) \triangleq \mathcal{P}\left(H_1(k,\ell) \mid \gamma(k,\ell)\right)$, one obtains

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell)) \exp(-\upsilon(k,\ell))\right\}^{-1} \quad (7)$$

where $q(k,\ell) \triangleq \mathcal{P}\left(H_0(k,\ell)\right)$ is the *a priori* probability for speech absence, and $\upsilon \triangleq \gamma\,\xi/(1+\xi)$.

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

$$\begin{aligned} H_0(k,\ell):\ & \bar{\lambda}_d(k,\ell+1) = \alpha_d \bar{\lambda}_d(k,\ell) + (1-\alpha_d)|Y(k,\ell)|^2 \\ H_1(k,\ell):\ & \bar{\lambda}_d(k,\ell+1) = \bar{\lambda}_d(k,\ell) \end{aligned} \quad (8)$$

where α_d $(0 < \alpha_d < 1)$ denotes a smoothing parameter. Under speech presence uncertainty, we can employ the conditional speech presence probability, and carry out the recursive averaging by

$$\begin{aligned}\bar{\lambda}_d(k,\ell+1) &= \left[\alpha_d\bar{\lambda}_d(k,\ell)+(1-\alpha_d)|Y(k,\ell)|^2\right] \\ &\quad\cdot(1-p(k,\ell))+\bar{\lambda}_d(k,\ell)p(k,\ell)\,. \end{aligned} \quad (9)$$

Equivalently, the recursive averaging can be obtained by

$$\bar{\lambda}_d(k,\ell+1)=\tilde{\alpha}_d(k,\ell)\bar{\lambda}_d(k,\ell)+[1-\tilde{\alpha}_d(k,\ell)]|Y(k,\ell)|^2 \quad (10)$$

where

$$\tilde{\alpha}_d(k,\ell)\triangleq\alpha_d+(1-\alpha_d)\,p(k,\ell) \quad (11)$$

is a time-varying frequency-dependent smoothing parameter. The smoothing parameter $\tilde{\alpha}_d$ is adjusted by the speech presence probability, which is estimated based on the noisy measurement. The speech presence probability also modifies the spectral estimate of the clean speech, and therefore is generally biased toward higher values to avoid speech distortions in speech enhancement applications[1] [4]. Accordingly, estimating the noise spectrum using (10) and (11) would be biased toward lower values. We propose to include a bias compensation factor in the noise estimator

$$\hat{\lambda}_d(k,\ell+1)=\beta\cdot\bar{\lambda}_d(k,\ell+1) \quad (12)$$

such that the factor β compensates the bias when speech is absent:

$$\beta\triangleq\left.\frac{\lambda_d(k,\ell)}{E\left\{\bar{\lambda}_d(k,\ell)\right\}}\right|_{\xi(k,\ell)=0}. \quad (13)$$

In Appendix A, we show that the value of β is completely determined by the particular estimator for the *a priori* speech absence probability. An explicit expression for β is derived in the case of estimating the *a priori* speech absence probability by the method proposed in the next section.

We note that the MS and *lower-energy envelope tracking* methods [12], [13], [19], also entail a multiplicative bias compensation factor. However, its value has to be determined by simulations. Furthermore, these methods estimate the noise at a given frame by processing a fixed time segment, *i.e.*, a fixed number of past frames. Whereas, our noise estimator is based on a variable time segment in each subband, which takes into account the probability of speech presence. The time segment is longer in subbands that contain frequent *speech* portions, and shorter in subbands that contain frequent *silence* portions. This feature has been considered [19] a desirable characteristic of the noise estimator, which improves its robustness and tracking capability.

[1]The spectral gain is minimal when speech is absent. Hence, deciding speech is absent when speech is present results ultimately in the attenuation of speech components. Whereas, the alternative false decision, up to a certain extent, merely introduces some level of residual noise.

III. Minima-Controlled Estimation

In this section, we introduce an estimator $\hat{q}(k,\ell)$ for the *a priori* speech absence probability. The estimator is controlled by the minima values of a smoothed power spectrum of the noisy signal.

In contrast to the MS and related methods [13], [5], the smoothing of the noisy power spectrum is carried out in both time and frequency. This takes into account the strong correlation of speech presence in neighboring frequency bins of consecutive frames [4]. Furthermore, the proposed procedure comprises two iterations of smoothing and minimum tracking. The first iteration provides a rough voice activity detection in each frequency band. Then, the smoothing in the second iteration excludes relatively strong speech components, which makes the minimum tracking during speech activity robust, even when using a relatively large smoothing window[2].

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

$$S_f(k,\ell)=\sum_{i=-w}^{w} b(i)\;|Y(k-i,\ell)|^2\,. \quad (14)$$

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

$$S(k,\ell)=\alpha_s S(k,\ell-1)+(1-\alpha_s)S_f(k,\ell)\,. \quad (15)$$

In accordance with the MS method, the minima values of $S(k,\ell)$ are picked within a finite window of length D, for each frequency bin:

$$S_{min}(k,\ell)\triangleq\min\left\{S(k,\ell')\mid \ell-D+1\le\ell'\le\ell\right\}. \quad (16)$$

It follows [13] that there exists a constant factor B_{min}, independent of the noise power spectrum, such that

$$E\left\{S_{min}(k,\ell)\mid\xi(k,\ell)=0\right\}=B_{min}^{-1}\cdot\lambda_d(k,\ell)\,. \quad (17)$$

The factor B_{min} represents the bias of a minimum noise estimate, and generally depends on the values of D, α_s, w and the spectral analysis parameters (type, length and overlap of the analysis windows)[3].

Let $\gamma_{min}(k,\ell)$ and $\zeta(k,\ell)$ be defined by

$$\begin{aligned}\gamma_{min}(k,\ell)&\triangleq\frac{|Y(k,\ell)|^2}{B_{min}S_{min}(k,\ell)}\\ \zeta(k,\ell)&\triangleq\frac{S(k,\ell)}{B_{min}S_{min}(k,\ell)}\,.\end{aligned} \quad (18)$$

[2]A larger smoothing window decreases the variance of the minima values, but also widens the peaks of the speech activity power. An alternative, computationally expensive, solution is to modify the smoothing in time and frequency based on a smoothed *a posteriori* SNR [13].

[3]The value of B_{min} can be estimated by generating a white Gaussian noise, and computing the inverse of the mean of $S_{min}(k,\ell)$. This takes into account also the time-frequency correlation of the noisy periodogram $|Y(k,\ell)|^2$. Notice that the value of B_{min} is fixed, whereas in [13], it is estimated for each frequency band and each frame.

Under the assumed statistical model, the PDF's of $\gamma_{min}(k,\ell)$ and $\zeta(k,\ell)$, in the absence speech, can respectively be approximated by exponential and chi-square distributions (Appendix B):

$$f\left(\gamma_{min}(k,\ell) \mid H_0(k,\ell)\right) \approx e^{-\gamma_{min}(k,\ell)} u(\gamma_{min}(k,\ell)) \quad (19)$$

$$f\left(\zeta(k,\ell) \mid H_0(k,\ell)\right) \approx \frac{1}{\left(\frac{2}{\mu}\right)^{\mu/2}\Gamma\left(\frac{\mu}{2}\right)} \zeta(k,\ell)^{\mu/2-1} \cdot \exp\left\{-\frac{\mu\,\zeta(k,\ell)}{2}\right\} u(\zeta(k,\ell)) \quad (20)$$

where $\Gamma(\cdot)$ is the gamma function, and μ is the equivalent degrees of freedom. Based on the first iteration smoothing and minimum tracking, we propose the following rough decision about speech presence:

$$I(k,\ell) = \begin{cases} 1, & \text{if } \gamma_{\min}(k,\ell) < \gamma_0 \text{ and } \zeta(k,\ell) < \zeta_0 \\ & \qquad \text{(speech is absent)} \\ 0, & \text{otherwise} \qquad \text{(speech is present)}\,. \end{cases} \quad (21)$$

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

$$\mathcal{P}\left(\gamma_{\min}(k,\ell) \geq \gamma_0 \mid H_0(k,\ell)\right) < \epsilon\,, \quad (22)$$

$$\mathcal{P}\left(\zeta(k,\ell) \geq \zeta_0 \mid H_0(k,\ell)\right) < \epsilon\,. \quad (23)$$

From (19) and (20) we have

$$\gamma_0 = -\log(\epsilon) \quad (24)$$

$$\zeta_0 = \frac{1}{\mu} F^{-1}_{\chi^2;\mu}(1-\epsilon) \quad (25)$$

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

The second iteration of smoothing includes only the power spectral components, which have been identified as containing primarily noise. We set the initial condition for the first frame by $\tilde{S}(k,0) = S_f(k,0)$. Then, for $\ell > 0$ the smoothing in frequency, employing the above voice activity detector, is obtained by

$$\tilde{S}_f(k,\ell) = \begin{cases} \dfrac{\sum_{i=-w}^{w} b(i)\, I(k-i,\ell)|Y(k-i,\ell)|^2}{\sum_{i=-w}^{w} b(i)\, I(k-i,\ell)}, & \text{if } \sum_{i=-w}^{w} I(k-i,\ell) \neq 0 \\ \tilde{S}(k,\ell-1), & \text{otherwise}\,. \end{cases} \quad (26)$$

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

$$\tilde{S}(k,\ell) = \alpha_s \tilde{S}(k,\ell-1) + (1-\alpha_s)\tilde{S}_f(k,\ell)\,. \quad (27)$$

We note that keeping the strong speech components out of the smoothing process enables improved minimum tracking. In particular, a larger smoothing parameter (α_s) and smaller minima search window (D) can be used. This reduces the variance of the minima values [13], and shortens the delay when responding to a rising noise power, which eventually improves the tracking capability of the noise estimator.

Let $\tilde{S}_{min}(k,\ell)$ be the result of the second iteration minimum tracking,

$$\tilde{S}_{min}(k,\ell) \triangleq \min\left\{\tilde{S}(k,\ell') \mid \ell - D + 1 \leq \ell' \leq \ell\right\},$$

and let $\tilde{\gamma}_{\min}(k,\ell)$ and $\tilde{\zeta}(k,\ell)$ be defined by

$$\tilde{\gamma}_{\min}(k,\ell) \triangleq \frac{|Y(k,\ell)|^2}{B_{min}\tilde{S}_{min}(k,\ell)}$$

$$\tilde{\zeta}(k,\ell) \triangleq \frac{S(k,\ell)}{B_{min}\tilde{S}_{min}(k,\ell)}\,. \quad (28)$$

Since we use a relatively small significance level in the first iteration ($\epsilon = 0.01$), the influence of the voice activity detector in noise-only periods can be neglected. That is, the effect of excluding strong *noise* components from the smoothing process is negligible. Accordingly, the conditional PDF's of $\tilde{\gamma}_{\min}(k,\ell)$ and $\tilde{\zeta}(k,\ell)$, in the absence of speech, are approximately the same as those of $\gamma_{\min}(k,\ell)$ and $\zeta(k,\ell)$ (Eqs. (19) and (20)).

We propose the following estimator for the *a priori* speech absence probability:

$$\hat{q}(k,\ell) = \begin{cases} 1, & \text{if } \tilde{\gamma}_{\min}(k,\ell) \leq 1 \\ & \quad \text{and } \tilde{\zeta}(k,\ell) < \zeta_0 \\ \left(\gamma_1 - \tilde{\gamma}_{\min}(k,\ell)\right)/(\gamma_1 - 1), & \text{if } 1 < \tilde{\gamma}_{\min}(k,\ell) < \gamma_1 \\ & \quad \text{and } \tilde{\zeta}(k,\ell) < \zeta_0 \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$

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

$$\mathcal{P}\left(\tilde{\gamma}_{\min}(k,\ell) > \gamma_1 \mid H_0(k,\ell)\right) < \epsilon_1 \quad \Rightarrow \quad \gamma_1 \approx -\log(\epsilon_1)\,. \quad (30)$$

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

The *a priori* speech absence probability estimator assumes speech is present ($\hat{q}(k,\ell) = 0$) whenever $\tilde{\zeta}(k,\ell) \geq \zeta_0$ or $\tilde{\gamma}_{\min}(k,\ell) \geq \gamma_1$. That is, whenever the local measured power, $S(k,\ell)$, or the instantaneous measured power, $|Y(k,\ell)|^2$, are relatively high compared to the noise power $B_{min}\tilde{S}_{min}(k,\ell) \approx \lambda_d(k,\ell)$. The estimator assumes speech is absent ($\hat{q}(k,\ell) = 1$) whenever both the local and instantaneous measured powers are relatively low compared to the noise power ($\tilde{\gamma}_{\min}(k,\ell) \leq 1$ and $\tilde{\zeta}(k,\ell) < \zeta_0$). In between, the estimator provides a soft transition between speech absence and speech presence, based on the value of $\tilde{\gamma}_{\min}(k,\ell)$.

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

Initialize variables at the first frame for all frequency bins k:

$\hat{\lambda}_d(k,0) = |Y(k,0)|^2$; $\bar{\lambda}_d(k,0) = |Y(k,0)|^2$; $\gamma(k,0) = 1$; $S_{min}(k,0) = S_f(k,0)$;
$S(k,0) = S_f(k,0)$; $\tilde{S}(k,0) = S_f(k,0)$; $G_{H_1}(k,0) = 1$; $\tilde{S}_{min}(k,0) = S_f(k,0)$;
$S_{min\_sw}(k) = S_f(k,0)$; $\tilde{S}_{min\_sw}(k) = S_f(k,0)$.

Let $j = 0$. % j is a counter for frames within a sub-window.

For all time frames ℓ

For all frequency bins k

Compute the *a posteriori* SNR $\gamma(k,\ell)$ using Eq. (3), the *a priori* SNR $\hat{\xi}(k,\ell)$ using Eq. (32), and the conditional gain $G_{H_1}(k,\ell)$ using Eq. (33).

Compute the first iteration of smoothed power spectrum $S(k,\ell)$ using Eqs. (14) and (15), and update its running minimum: $S_{min}(k,\ell) = \min\{S_{min}(k,\ell-1),\, S(k,\ell)\}$; $S_{min\_sw}(k) = \min\{S_{min\_sw}(k),\, S(k,\ell)\}$.

Compute the indicator function $I(k,\ell)$ for the voice activity detection using Eqs. (18) and (21).

Compute the second iteration of smoothed power spectrum $\tilde{S}(k,\ell)$ using Eqs. (26) and (27), and update its running minimum: $\tilde{S}_{min}(k,\ell) = \min\left\{\tilde{S}_{min}(k,\ell-1),\, \tilde{S}(k,\ell)\right\}$; $\tilde{S}_{min\_sw}(k) = \min\left\{\tilde{S}_{min\_sw}(k),\, \tilde{S}(k,\ell)\right\}$.

Compute the *a priori* speech absence probability $\hat{q}(k,\ell)$ using Eqs. (28) and (29), the speech presence probability $p(k,\ell)$ using Eq. (7), and the time-varying smoothing parameter $\tilde{\alpha}_d(k,\ell)$ using Eq. (11).

Update the noise spectrum estimate $\hat{\lambda}_d(k,\ell+1)$ using Eqs. (10) and (12).

Let $j = j + 1$.

If $j == V$

For all frequency bins k

Store $S_{min\_sw}(k)$, set $S_{min}(k,\ell)$ to the minimum of the last U stored values of $S_{min\_sw}(k)$, and let $S_{min\_sw}(k) = S(k,\ell)$.

Store $\tilde{S}_{min\_sw}(k)$, set $\tilde{S}_{min}(k,\ell)$ to the minimum of the last U stored values of $\tilde{S}_{min\_sw}(k)$, and let $\tilde{S}_{min\_sw}(k) = \tilde{S}(k,\ell)$.

Let $j = 0$.

Fig. 1. The IMCRA noise estimation algorithm.

TABLE I
VALUES OF PARAMETERS USED IN THE IMPLEMENTATION OF THE IMCRA NOISE ESTIMATOR, FOR A SAMPLING RATE OF 16 KHZ.

$w = 1$	$\alpha_s = 0.9$	$U = 8$	$V = 15$
$D = 120$	$B_{min} = 1.66$	$\gamma_0 = 4.6$	$\gamma_1 = 3$
$\zeta_0 = 1.67$	$\alpha = 0.92$	$\alpha_d = 0.85$	$\beta = 1.47$
b: Hanning window			

$\tilde{\gamma}_{\min}(k,\ell)$. The remaining speech components can hardly affect the noise estimator, since their power is relatively low compared to that of the noise.

IV. Implementation of the Algorithm

In this section, we combine the time-varying recursive averaging with the minima-controlled estimation of the *a priori* speech absence probability, and present the IMCRA noise estimation algorithm.

The noise spectrum estimate, $\hat{\lambda}_d(k,\ell)$, is initialized at the first frame by $\hat{\lambda}_d(k,0) = |Y(k,0)|^2$. Then, at each frame ℓ ($\ell \geq 0$), it is used, jointly with the current observation $Y(k,\ell)$, for estimating the noise power spectrum at the next frame, $\ell+1$. According to Eq. (12), we need to find the bias compensation factor β, and the time-varying smoothing parameter $\tilde{\alpha}_d(k,\ell)$. Appendix A shows that the value of β is given by

$$\beta = \frac{\gamma_1 - 1 - e^{-1} + e^{-\gamma_1}}{\gamma_1 - 1 - 3\,e^{-1} + (\gamma_1 + 2)\,e^{-\gamma_1}}\,. \tag{31}$$

In particular, for $\gamma_1 = 3$, we have $\beta = 1.47$. The value of $\tilde{\alpha}_d(k,\ell)$ is updated for each frequency bin and time frame, using the speech presence probability $p(k,\ell)$, and expression (11).

It follows from Eq. (7), that the computation of the speech presence probability requires an estimate for the *a priori* SNR $\xi(k,\ell)$. The "decision-directed" approach of Ephraim and Malah [6] is commonly used for that purpose. However, we obtained better performance with a modified version proposed in [4]. Specifically, the *a priori* SNR is estimated by

$$\hat{\xi}(k,\ell) = \alpha G^2_{H_1}(k,\ell-1)\gamma(k,\ell-1) + (1-\alpha)\max\{\gamma(k,\ell) - 1, 0\}\,, \tag{32}$$

where α is a weighting factor that controls the trade-off between noise reduction and speech distortion [6], [1], and

$$G_{H_1}(k,\ell) \triangleq \frac{\xi(k,\ell)}{1+\xi(k,\ell)} exp\left(\frac{1}{2}\int_{v(k,\ell)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{33}$$

is the spectral gain function of the *Log-Spectral Amplitude* (LSA) estimator when speech is surely present [7]. We note

that the original "decision-directed" *a priori* SNR estimator of Ephraim and Malah [6], [11] is given by

$$\frac{\alpha G^2(k,\ell-1)\gamma(k,\ell-1)+(1-\alpha)\max\{\gamma(k,\ell)-1,0\}}{1-\hat{q}(k,\ell)}, \tag{34}$$

where $G(k,\ell)$ is the spectral gain function of the LSA estimator under speech presence uncertainty. The advantage of $\hat{\xi}(k,\ell)$ over the original estimator, particularly for weak speech components and low input SNR, is discussed in some detail in [4].

The estimator for the *a priori* speech absence probability, $\hat{q}(k,\ell)$, (29), requires two iterations of time-frequency smoothing ($S(k,\ell)$, $\tilde{S}(k,\ell)$) and minimum tracking ($S_{min}(k,\ell)$, $\tilde{S}_{min}(k,\ell)$). The minimum tracking is implemented by the method proposed in [12], [13], which provides a flexible balance between the computational complexity and the update rate of the minima values. Accordingly, we divide the window of D samples into U sub-windows of V samples ($UV = D$). Whenever V samples are read, the minimum of the current sub-window is determined and stored for later use. The overall minimum is obtained as the minimum of past samples within the current sub-window and the U previous sub-window minima.

The implementation of the IMCRA algorithm is summarized in Fig. 1. Typical values of the respective parameters, for a sampling rate of 16 kHz, are given in Table I.

V. Performance Evaluation

The performance evaluation of the IMCRA method, and a comparison to the MS method, consists of three parts. First, we test the tracking capability of the noise estimators for non-stationary noise. Second, we measure the segmental relative estimation error for various noise types and levels. Third, we integrate the noise estimators into a speech enhancement system, and determine the improvement in the segmental SNR. The results are confirmed by a subjective study of speech spectrograms and informal listening tests.

The noise signals used in our evaluation are taken from the Noisex92 database [22]. They include white Gaussian noise (WGN), car noise, and F16 cockpit noise. A non-stationary WGN was simulated by increasing the level of the stationary WGN at a rate of 2 dB/s for a period of three seconds, and some time afterwards decreasing it back to the original level at the same rate. The speech signal is constructed from six different utterances, without intervening pauses. The utterances, half from male speakers and half from female speakers, are taken from the TIMIT database [8]. The speech signal is sampled at 16 kHz and degraded by the various noise types with segmental SNR's in the range $[-5, 10]$ dB. The segmental SNR is defined by [18]

$$SegSNR = \frac{10}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\log\frac{\sum_k |X(k,\ell)|^2}{\sum_k |D(k,\ell)|^2} \tag{35}$$

where $\mathcal{L}$ represents the set of frames that contain speech, and $|\mathcal{L}|$ its cardinality. The spectral analysis is implemented with Hamming windows of 512 samples length (32 ms) and 128 samples frame update step.

Fig. 2(a) shows the periodogram $|Y(k,\ell)|^2$, a recursively smoothed periodogram with a smoothing parameter set to 0.95, and the noise power $\hat{\lambda}_d(k,\ell)$ estimated by the IMCRA method, for a F16 cockpit noise at 0 dB segmental SNR, and a single frequency bin $k = 40$ (center frequency 1219 Hz). Fig. 2(b) plots the ideal, IMCRA, and MS noise estimates (the ideal noise estimate is taken as the recursively smoothed periodogram of the noise $|D(k,\ell)|^2$, with a smoothing parameter set to 0.95). Clearly, the IMCRA noise estimate follows the noise power more closely than the MS noise estimate. The update rate of the MS noise estimate is inherently restricted by the size of the minimum search window (D). By contrast, the IMCRA noise estimate is continuously updated even during speech activity, as long as the speech components are not too large compared to the noise power. This is a major advantage of the IMCRA method, particularly in adverse noise environments, which involve non-stationary noise, weak speech components, and low input SNR.

Fig. 3 shows another example of the improved tracking capability of the IMCRA estimator. In this case, the speech signal is degraded by non-stationary WGN at 0 dB segmental SNR. The ideal, IMCRA, and MS noise estimates, averaged out over the frequency, are depicted in Fig. 3(b). The response of the IMCRA estimator to increasing or decreasing noise power is essentially much faster than that of the MS estimator, due to the recursive averaging mechanism. For increasing noise power, the MS estimator lags behind with a delay of $D + V$ frames [13]. For decreasing noise power, the delay of the MS estimator stems from the fact that the minimum search window becomes effectively shorter, and therefore the bias compensation factor is practically too large. On the other hand, the delay of the IMCRA estimator in case of increasing noise power results from the increase in the time-varying smoothing parameter, subsequent to the decrease in the *a priori* speech absence probability. This delay is smaller than $D+V$ frames, since the recursive averaging is carried out instantaneously. For decreasing noise power, the *a priori* speech absence probability gets larger and the time-varying smoothing parameter gets smaller, which further shortens the delay of the IMCRA estimator.

A quantitative comparison between the IMCRA and MS estimation methods is obtained by evaluating the segmental relative estimation error in various environmental conditions. The segmental relative estimation error is defined by

$$SegErr = \frac{1}{L}\sum_{\ell=0}^{L-1}\frac{\sum_k\left[\hat{\lambda}_d(k,\ell)-\lambda_d(k,\ell)\right]^2}{\sum_k \lambda_d^2(k,\ell)} \tag{36}$$

where $\lambda_d(k,\ell)$ is the ideal noise estimate, $\hat{\lambda}_d(k,\ell)$ is the noise estimated by the tested method, and L is the number of frames in the analyzed signal. Table II presents the results of the segmental relative estimation error achieved by the IMCRA and MS estimators for various noise types

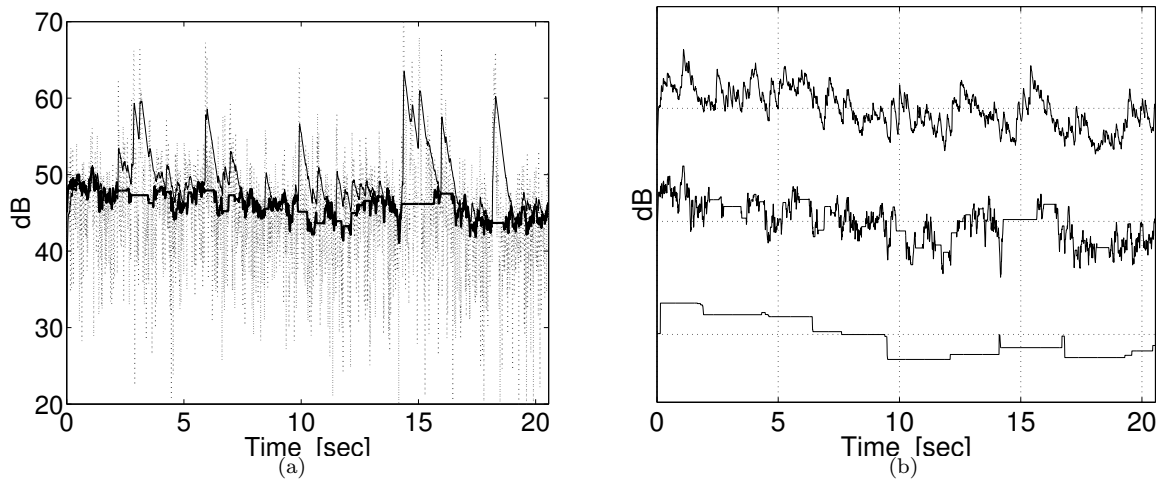

Fig. 2. Noise power estimation for a speech signal, degraded by F16 cockpit noise at 0 dB segmental SNR, and a single frequency bin $k = 40$ (center frequency 1219 Hz): (a) Periodogram (dotted), smoothed periodogram (fine solid), and IMCRA noise estimate (heavy solid); (b) Ideal (top), IMCRA (center), and MS (bottom) noise estimates (top and bottom graphs are displaced by ± 10 dB, for clarity).

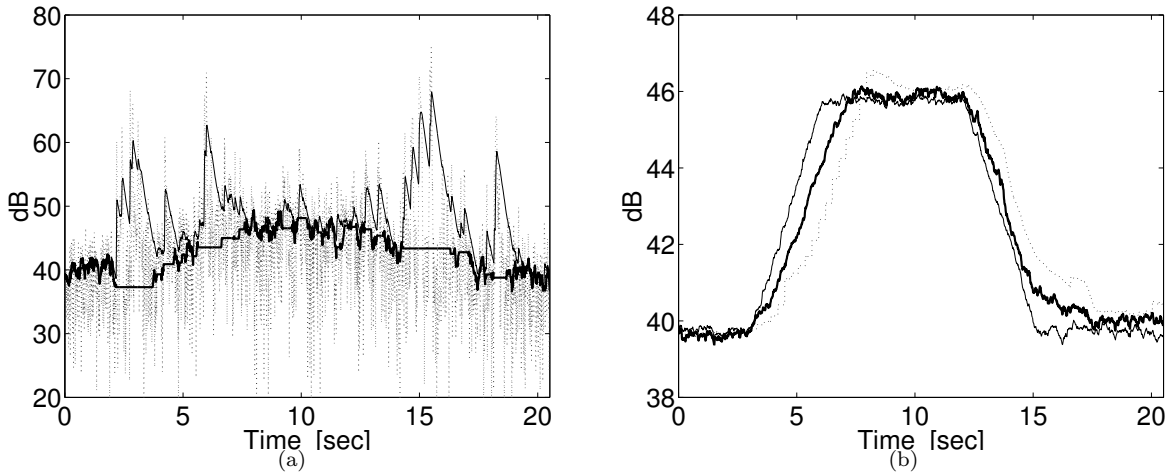

Fig. 3. Noise power estimation for a speech signal, degraded by non-stationary white Gaussian noise at 0 dB segmental SNR: (a) Periodogram (dotted), smoothed periodogram (fine solid), and IMCRA noise estimate (heavy solid) for a single frequency bin $k = 33$ (center frequency 1 kHz); (b) Ideal (fine solid), IMCRA (heavy solid), and MS (dotted) average noise estimates.

TABLE II
SEGMENTAL RELATIVE ESTIMATION ERROR FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING THE MS AND IMCRA ESTIMATORS.

Input SegSNR	Stationary WGN		Non-stationary WGN		Car interior noise		F16 cockpit noise	
[dB]	MS	IMCRA	MS	IMCRA	MS	IMCRA	MS	IMCRA
-5	0.119	0.056	0.321	0.082	0.401	0.129	0.250	0.114
0	0.149	0.065	0.350	0.093	0.404	0.131	0.265	0.119
5	0.177	0.085	0.365	0.114	0.356	0.135	0.255	0.124
10	0.216	0.118	0.353	0.151	0.288	0.131	0.234	0.143

and levels. It shows that the IMCRA method obtains significantly lower estimation error than the MS method.

The segmental relative estimation error is a measure that weighs all frames in a uniform manner, without a distinction between speech presence and absence. In practice, the estimation error is more consequential in frames that contain speech, particularly weak speech components, than in frames that contain only noise. We therefore examine the performance of our estimation method when integrated into a speech enhancement system. Specifically, the IMCRA and MS noise estimators are combined with the *Optimally-Modified Log-Spectral Amplitude* (OM-LSA) estimator, and evaluated both objectively using an improvement in segmental SNR measure, and subjectively by informal listening tests. The OM-LSA estimator [2], [4] is a modified version of the conventional LSA estimator [7], based on a binary hypothesis model. The modification includes a lower bound for the gain, which is determined by a

TABLE III
SEGMENTAL SNR IMPROVEMENT FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING THE MS AND IMCRA ESTIMATORS.

Input SegSNR	Stationary WGN		Non-stationary WGN		Car interior noise		F16 cockpit noise	
[dB]	MS	IMCRA	MS	IMCRA	MS	IMCRA	MS	IMCRA
-5	9.91	10.45	9.11	10.06	9.67	10.76	8.08	8.49
0	7.93	8.39	7.33	8.07	8.26	8.91	6.45	6.60
5	6.15	6.43	5.67	6.14	6.78	7.21	4.84	4.92
10	4.53	4.62	4.14	4.35	5.37	5.83	3.44	3.44

subjective criteria for the noise naturalness, and exponential weights, which are given by the conditional speech presence probability. Moreover, the *a priori* SNR is estimated using (32), rather than the standard "decision-directed" estimator (34).

Table III summarizes the results of the segmental SNR improvement for various noise types and levels. The IMCRA estimator consistently yields a higher improvement in the segmental SNR, than the MS estimator, under all tested environmental conditions. The fact that the benefit is greater for low input SNR implies that weak speech components are better preserved when the noise is estimated by the IMCRA method. This is confirmed by a subjective study of speech spectrograms and informal listening tests.

Another major advantage of the IMCRA noise estimation method, as discussed earlier, is its tracking capability under non-stationary noise environments. In speech enhancement applications, this quality is often not fully appreciated when considering the *average* improvement in the segmental SNR, since variations in the statistics of the noise are usually sparse. However, a *frame-by-frame* trace of the improvement in the segmental SNR, as illustrated in Fig. 4, revels that the effectiveness of the IMCRA method is particularly notable during alteration in noise characteristics. Figs. 4(a) and (b) are plots of the speech waveform in noise-free and noisy conditions (additive non-stationary WGN at -5 dB segmental SNR). Figs. 4(c) and (d) are, respectively, plots of the enhanced speech waveforms using the IMCRA and MS noise estimates. While the increase in the segmental SNR, gained by the IMCRA method over the MS method, is on average less than 1 dB in this example, it surpasses 5 dB in some instances (Fig. 4(e)).

VI. Conclusion

Recursive averaging is a commonly used procedure for estimating the noise power spectrum during sections which do not contain speech. However, rather than employing a voice activity detector and restricting the update of the noise estimator to periods of speech absence, we adapt the smoothing parameter in time and frequency according to the speech presence probability. The noise estimate is thereby continuously updated even during weak speech activity. We have proposed an estimator for the *a priori* speech absence probability that is controlled by the minima values of a smoothed periodogram of the noisy measurement. It combines conditions on both the instantaneous and local measured power, and provides a soft transition between speech absence and presence. This prevents an occasional increase in the noise estimate during speech activity. Furthermore, carrying out the smoothing and minimum tracking in two iterations allows larger smoothing windows and smaller minimum search windows, while reliably tracking the minima even during strong speech activity. This yields a reduced variance of the minima values and shorter delay when responding to a rising noise power, which eventually improves the tracking capability of the noise estimator. We have shown that in non-stationary noise environments and under low SNR conditions, the IMCRA approach is extremely effective. In particular, it obtains a lower estimation error, and when integrated into a speech enhancement system achieves improved speech quality and lower residual noise.

Appendix

I. Derivation of the Bias Compensation Factor

The factor β in (12), by definition, compensates the bias of the noise spectrum estimator when speech is absent. It stems from Eqs. (10) and (13) and the definition of the *a posteriori* SNR that

$$\beta = \left. \frac{E\left\{1-\tilde{\alpha}(k,\ell)\right\}}{E\left\{(1-\tilde{\alpha}(k,\ell))\gamma(k,\ell)\right\}} \right|_{\xi(k,\ell)=0} . \tag{37}$$

By Eq. (7), the conditional speech presence probability $p(k,\ell)$ degenerates, in the absence of speech ($\xi(k,\ell)=0$), to the *a priori* speech presence probability $1-q(k,\ell)$. Hence, Eq. (11) implies that the value of β is completely determined by the particular estimator for the *a priori* speech absence probability:

$$\beta = \left. \frac{E\left\{\hat{q}(k,\ell)\right\}}{E\left\{\hat{q}(k,\ell)\gamma(k,\ell)\right\}} \right|_{\xi(k,\ell)=0} . \tag{38}$$

In our case, the estimate for the *a priori* speech absence probability, $\hat{q}(k,\ell)$, is given by (29). Since we are using a relatively low significance level in the first iteration ($\epsilon = 0.01$), the conditional PDF of $\tilde{\gamma}_{min}(k,\ell)$ in the absence of speech is approximately the same as that of $\gamma_{min}(k,\ell)$:

$$f\left(\tilde{\gamma}_{min}(k,\ell) \mid H_0(k,\ell)\right) \approx e^{-\tilde{\gamma}_{min}(k,\ell)} u(\tilde{\gamma}_{min}(k,\ell)) . \tag{39}$$

Similarly, the conditional PDF of $\tilde{\zeta}(k,\ell)$ in the absence of speech is approximately the same as that of $\zeta(k,\ell)$. Then

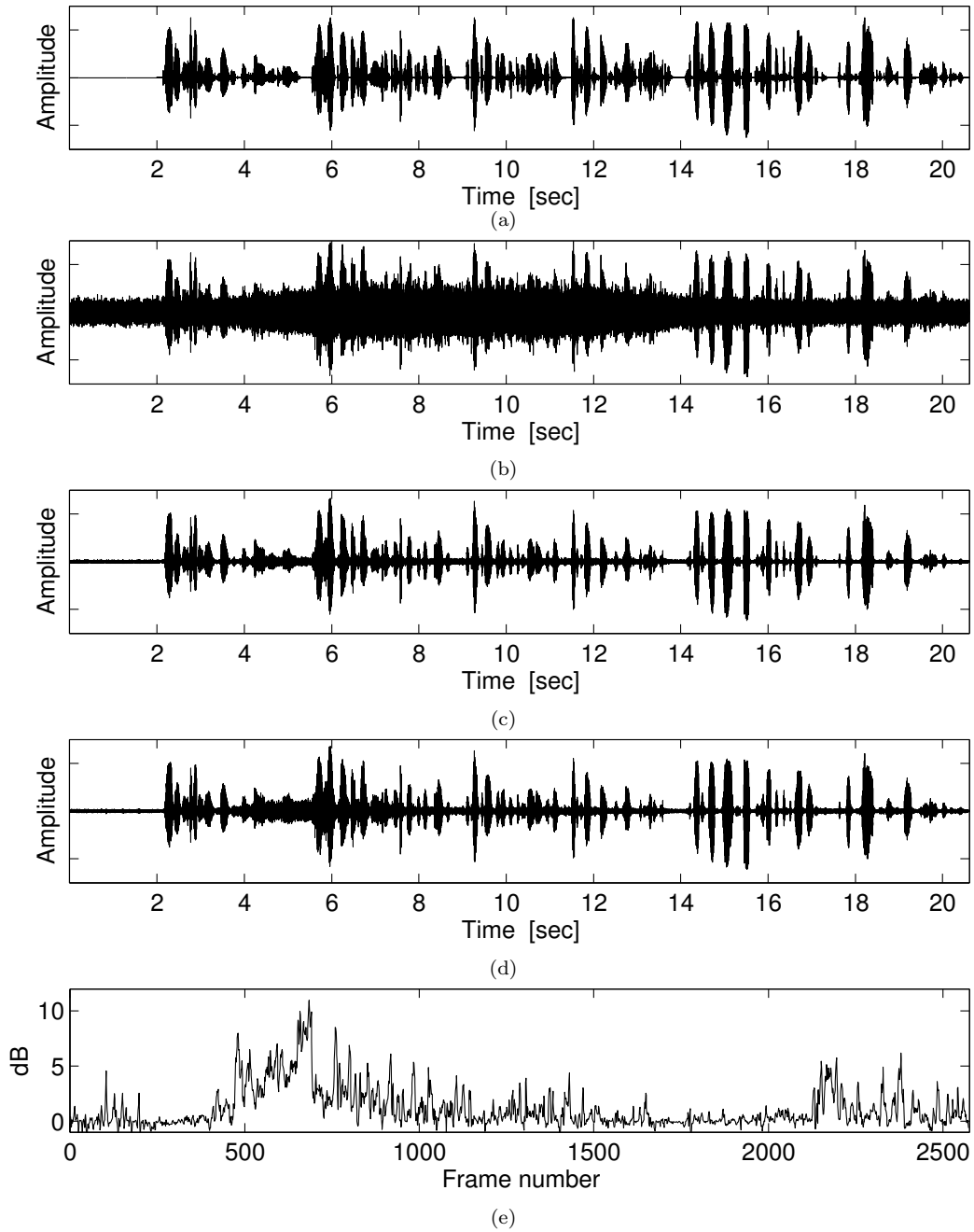

Fig. 4. Example of speech enhancement using the IMCRA and MS noise estimators: (a) Original speech waveform; (b) Noisy speech waveform (additive non-stationary white Gaussian noise at −5 dB segmental SNR); (c) Enhanced speech waveform using the IMCRA noise estimate (SegSNR= 5.05 dB); (d) Enhanced speech waveform using the MS noise estimate (SegSNR= 4.11 dB); (e) Trace of the increase in segmental SNR, gained by the IMCRA method over the MS method.

by Eq. (23), the probability of $\tilde{\zeta}(k,\ell) \geq \zeta_0$ is relatively low ($\approx \epsilon$). Hence, in the absence of speech we can assume that $\tilde{\zeta}(k,\ell) < \zeta_0$ for all k and ℓ. Accordingly,

$$\begin{aligned} E\left\{\hat{q}(k,\ell) \mid \xi(k,\ell)=0\right\} &\cong \int_0^1 e^{-z}\,dz + \int_1^{\gamma_1} \frac{\gamma_1 - z}{\gamma_1 - 1} e^{-z}\,dz \\ &= 1 - \frac{1}{\gamma_1 - 1}\left(e^{-1} - e^{-\gamma_1}\right) \end{aligned} \quad (40)$$

and

$$\begin{aligned} &E\left\{\hat{q}(k,\ell)\gamma(k,\ell) \mid \xi(k,\ell)=0\right\} \\ &\cong \int_0^1 z\,e^{-z}\,dz + \int_1^{\gamma_1} \frac{\gamma_1 z - z^2}{\gamma_1 - 1} e^{-z}\,dz \\ &= 1 - \frac{3}{\gamma_1 - 1}\,e^{-1} + \frac{\gamma_1 + 2}{\gamma_1 - 1}\,e^{-\gamma_1}. \end{aligned} \quad (41)$$

Substituting (40) and (41) into (38), we have

$$\beta = \frac{\gamma_1 - 1 - e^{-1} + e^{-\gamma_1}}{\gamma_1 - 1 - 3\,e^{-1} + (\gamma_1 + 2)\,e^{-\gamma_1}}. \quad (42)$$

II. Statistics of γ_{min} and ζ

Generally, successive values of $|Y(k,\ell)|^2$ are correlated, and there is no closed form solution for the probability density functions of $\gamma_{min}(k,\ell)$ and $\zeta(k,\ell)$. However, based on certain assumptions and results from [12], [13], we can obtain an approximate solution. To simplify notation, speech absence is implicitly assumed throughout this appendix.

Let the spectral power values of the noisy measurement $|Y(k,\ell)|^2$ be independent, exponentially and identically distributed. Substituting (14) into (15), the recursively averaged periodogram can be written as

$$S(k,\ell) = (1-\alpha_s) \sum_{i=-w}^{w} \sum_{j=0}^{\infty} b(i)\, \alpha^j\, |Y(k-i,\ell-j)|^2 \,. \quad (43)$$

If we approximate $S(k,\ell)$ as the sum of μ squared mutually independent normal variables, then its density and distribution functions can be obtained by

$$f_{S(k,\ell)}(x) \approx \frac{\mu}{\lambda_d(k,\ell)}\, f_{\chi^2;\mu}\left(\frac{\mu\, x}{\lambda_d(k,\ell)}\right) \quad (44)$$

$$F_{S(k,\ell)}(x) \approx F_{\chi^2;\mu}\left(\frac{\mu\, x}{\lambda_d(k,\ell)}\right) \quad (45)$$

where $f_{\chi^2;\mu}(x)$ and $F_{\chi^2;\mu}(x)$ denote, respectively, the standard chi-square density and distribution functions, with μ degrees of freedom. Specifically,

$$f_{\chi^2;\mu}(x) = \frac{x^{\mu/2-1}\, e^{-\frac{x}{2}}\, u(x)}{2^{\mu/2}\, \Gamma\left(\frac{\mu}{2}\right)} \quad (46)$$

$$F_{\chi^2;\mu}(x) = \frac{\Gamma\left(\frac{\mu}{2},\frac{x}{2}\right)\, u(x)}{\Gamma\left(\frac{\mu}{2}\right)} \quad (47)$$

where $\Gamma(\cdot)$ is the gamma function, and $\Gamma(a,x) \triangleq \int_0^\infty e^{-t} t^{a-1}\, dt$ is the incomplete gamma function. We note that μ, the equivalent degrees of freedom, is determined by the smoothing parameter α_s and the window function b. For a normalized Hanning window function of size $2w+1$, it was found experimentally that $\mu \approx \frac{1+\alpha_s}{1-\alpha_s}\ (1+0.7\,w)$.

The value of $S_{min}(k,\ell)$ (Eq. (16)) is based on D successive values of $S(k,\ell)$, which are clearly correlated. However, to approximate the statistics of $S_{min}(k,\ell)$, we assume that $S_{min}(k,\ell)$ is based on equivalent $\mathcal{D}$ i.i.d random variables. Hence, the probability density function of $S_{min}(k,\ell)$ is given by [13], [9]

$$f_{S_{min}(k,\ell)}(x) \approx \mathcal{D}\left(1-F_{S(k,\ell)}(x)\right)^{\mathcal{D}-1} f_{S(k,\ell)}(x)\,. \quad (48)$$

Since $\gamma_{min}(k,\ell)$ is defined as the ratio of two random variables, $|Y(k,\ell)|^2$ and $S_{min}(k,\ell)$ scaled by B_{min}, its density function is given by [17]

$$f_{\gamma_{min}(k,\ell)}(x) = \int_0^\infty B_{min}\, y\, f_{|Y(k,\ell)|^2, S_{min}(k,\ell)}(B_{min}\, y\, x, y)\, dy\,. \quad (49)$$

Similarly, the density function of $\zeta(k,\ell)$ is given by

$$f_{\zeta(k,\ell)}(x) = \int_0^\infty B_{min}\, y\, f_{S(k,\ell), S_{min}(k,\ell)}(B_{min}\, y\, x, y)\, dy\,. \quad (50)$$

For large $\mathcal{D}$ and μ ($\mathcal{D}, \mu > 10$), we can assume that $S_{min}(k,\ell)$ is independent of either $|Y(k,\ell)|^2$ or $S(k,\ell)$. Furthermore, the variance of $S_{min}(k,\ell)$ is significantly smaller than its squared mean value. Hence, Eqs. (49) and (50) can be simplified to

$$f_{\gamma_{min}(k,\ell)}(x) \approx \int_0^\infty f_{S_{min}(k,\ell)}(y)\, B_{min}\, E\left\{S_{min}(k,\ell)\right\} \cdot f_{|Y(k,\ell)|^2}\left(x\, B_{min} E\left\{S_{min}(k,\ell)\right\}\right)\, dy \quad (51)$$

$$f_{\zeta(k,\ell)}(x) \approx \int_0^\infty f_{S_{min}(k,\ell)}(y)\, B_{min}\, E\left\{S_{min}(k,\ell)\right\} \cdot f_{S(k,\ell)}\left(x\, B_{min} E\left\{S_{min}(k,\ell)\right\}\right)\, dy\,. \quad (52)$$

Substituting Eq. (17) into (51) and (52) we have

$$f_{\gamma_{min}(k,\ell)}(x) \approx \lambda_d(k,\ell)\, f_{|Y(k,\ell)|^2}\left(x\, \lambda_d(k,\ell)\right) = f_{\gamma(k,\ell)}(x) \quad (53)$$

$$f_{\zeta(k,\ell)}(x) \approx \lambda_d(k,\ell)\, f_{S(k,\ell)}\left(x\, \lambda_d(k,\ell)\right) = \mu\, f_{\chi^2;\mu}(\mu\, x)\,. \quad (54)$$

Acknowledgement

The author thanks Dr. Baruch Berdugo for helpful discussions, Dr. Rainer Martin for making his Minimum Statistics code available, and the anonymous reviewers for proofreading the manuscript.

References

[1] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, April 1994, pp. 345–349.

[2] I. Cohen, "On speech enhancement under signal presence uncertainty," *Proc. 26th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2001*, Salt Lake City, Utah, 7–11 May 2001, pp. 167–170.

[3] I. Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands," *Proc. IEEE Workshop on Hands Free Speech Communication, HSC'01*, Kyoto, Japan, 9–11 April 2001, pp. 95–98.

[4] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403–2418, November 2001.

[5] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Proc. 4th European Conf. Speech, Communication and Technology, EUROSPEECH'95*, Madrid, Spain, 18–21 September 1995, pp. 1513–1516.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No. 6, December 1984, pp. 1109–1121.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 2, April 1985, pp. 443–445.

[8] 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).

[9] E. J. Gumbel, *Statistics of Extremes*. New York, NY: Columbia University Press, 1958.

[10] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *Proc. 20th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-95*, Detroit, Michigan, 8–12 May 1995, pp. 153–156.

[11] D. Malah, R. V. Cox and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proc. 24th IEEE Internat.*

Conf. Acoust. Speech Signal Process., ICASSP-99, Phoenix, Arizona, 15–19 March 1999, pp. 789–792.

[12] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. 7th European Signal Processing Conf., EUSIPCO-94*, Edinburgh, Scotland, 13–16 September 1994, pp. 1182–1185.

[13] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 5, July 2001, pp. 504–512.

[14] R. J. McAulay and M. L. Malpass "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 2, April 1980, pp. 137–145.

[15] B. L. McKinley and G. H. Whipple, "Model based speech pause detection," *Proc. 22th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-97*, Munich, Germany, 20–24 April 1997, pp. 1179–1182.

[16] J. Meyer, K. U. Simmer and K. D. Kammeyer "Comparison of one- and two-channel noise-estimation techniques," *Proc. 5th International Workshop on Acoustic Echo and Noise Control, IWAENC-97*, London, UK, 11–12 September 1997, pp. 137–145.

[17] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. third edition, New York, NY: McGraw-Hill Inc., 1991.

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

[19] C. Ris and S. Dupont, "Assessing local noise level estimation methods: application to noise robust ASR," *Speech Communication*, Vol. 34, No. 1-2, April 2001, pp. 141–158.

[20] J. Sohn, N. S Kim and W. Sung, "A statistical model-based voice activity detector," *IEEE Signal Processing Letters*, Vol. 6, No. 1, January 1999, pp. 1–3.

[21] V. Stahl, A. Fischer and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," *Proc. 25th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2000*, Istanbul, Turkey, 5–9 June 2000, pp. 1875–1878.

[22] A. Varga and 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 Communication, Vol. 12, No. 3, July 1993, pp. 247–251.