Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement

Israel Cohen, *Member, IEEE,* and Baruch Berdugo

***Abstract*—In this letter, we introduce a *minima controlled recursive averaging* (MCRA) approach for noise estimation. The noise estimate is given by averaging past spectral power values and using a smoothing parameter that is adjusted by the signal presence probability in subbands. Presence of speech in subbands is determined by the ratio between the local energy of the noisy speech and its minimum within a specified time window. The noise estimate is computationally efficient, robust with respect to the input signal-to-noise ratio (SNR) and type of underlying additive noise, and characterized by the ability to quickly follow abrupt changes in the noise spectrum.**

***Index Terms*—Acoustic noise, signal detection, spectral analysis, speech enhancement.**

I. Introduction

A CRUCIAL component of a practical speech enhancement system is the estimation of the noise power spectrum. A common approach is to average the noisy signal over nonspeech sections. A speech pause detection is either implemented on a frame-by-frame basis [1] or estimated independently for individual subbands using *a posteriori* signal-to-noise ratio (SNR) [2], [3]. However, the detection reliability severely deteriorates for week speech components and low-input SNR. Additionally, the amount of presumable nonspeech sections in the signal may not be sufficient, which restricts the tracking capability of the noise estimator in case of varying noise spectrum. Alternatively, the noise can be estimated from histograms in the power spectral domain [3]–[5]. Unfortunately, such methods are computationally expensive.

Martin [6] 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. However, this noise estimate is sensitive to outliers [5] and its variance is about twice as large as the variance of a conventional noise estimator [6]. Moreover, this method may occasionally attenuate low energy phonemes, particularly if the minimum search window is too short [7]. A computationally more efficient minimum tracking scheme is presented in [8]. Its main drawback is the very slow update rate of the noise estimate in case of a sudden rise in noise energy level and its tendency to cancel the signal [9].

In this letter, we introduce a *minima controlled recursive averaging* (MCRA) approach for noise estimation. The noise estimate is given by averaging past spectral power values, using a smoothing parameter that is adjusted by the signal presence probability in subbands. We show that presence of speech in a given frame of a subband can be determined by the ratio between the local energy of the noisy speech 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 temporal smoothing is carried out to reduce fluctuations between speech and nonspeech segments, thereby exploiting the strong correlation of speech presence in neighboring frames. The resultant noise estimate is 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.

The letter is organized as follows. In Section II, we present the noise spectrum estimation approach. In Section III, we introduce a minima controlled estimator for the speech presence probability. In Section IV, we evaluate the proposed method and discuss experimental results, which validate its usefulness.

II. Noise Spectrum Estimation

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,

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

where k is the frequency bin index, ℓ is the time frame index, h is an analysis window of size N, and M is the frame update step in time. Given two hypotheses, $H_0(k,\ell)$ and $H_1(k,\ell)$, which indicate, respectively, speech absence and presence in the ℓth frame of the kth subband, we have

$$\begin{aligned} H_0(k,\ell) &: Y(k,\ell) = D(k,\ell) \\ H_1(k,\ell) &: Y(k,\ell) = X(k,\ell) + D(k,\ell) \end{aligned} \tag{2}$$

where $X(k,\ell)$ and $D(k,\ell)$ represent the STFT of the clean and noise signals, respectively. Let $\lambda_d(k,\ell) = E[|D(k,\ell)|^2]$ denote the variance of the noise in the kth subband. Then a common

Manuscript received February 22, 2001; October 26, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. A. S. Spanias.

I. Cohen is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il).

B. Berdugo is with Lamar Signal Processing, Ltd., Yokneam Ilit 20692, Israel (e-mail: bberdugo@lamar.co.il).

Publisher Item Identifier S 1070-9908(02)02410-0.

technique to obtain its estimate is to apply a temporal recursive smoothing to the noisy measurement during periods of speech absence. In particular,

$$\begin{aligned} H_0'(k,\ell) : \hat{\lambda}_d(k,\ell+1) =& \alpha_d \hat{\lambda}_d(k,\ell) + (1-\alpha_d)\,|Y(k,\ell)|^2 \\ H_1'(k,\ell) : \hat{\lambda}_d(k,\ell+1) =& \hat{\lambda}_d(k,\ell) \end{aligned} \tag{3}$$

where $\alpha_d (0 < \alpha_d > 1)$ is a smoothing parameter and H_0' and H_1' designate hypothetical speech absence and presence, respectively. Here, we make a distinction between the hypotheses in (2), used for estimating the clean speech and the hypotheses in (3), 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 signal than when estimating the noise. Hence, different decision rules are employed and generally we tend to decide H_1 with a higher confidence than H_1', i.e., $P(H_1|Y) \geq P(H_1'|Y)$ [7].

Let $p'(k,\ell) \triangleq P(H_1'(k,\ell)|Y(k,\ell))$ denote the conditional signal presence probability. Then (3) implies

$$\begin{aligned} \hat{\lambda}_d(k,\ell+1) =& \hat{\lambda}_d(k,\ell) p'(k,\ell) \\ &+ \left[\alpha_d \hat{\lambda}_d(k,\ell) + (1-\alpha_d)\,|Y(k,\ell)|^2\right] \\ &\times (1 - p'(k,\ell)) \\ =& \tilde{\alpha}_d(k,\ell)\hat{\lambda}_d(k,\ell) \\ &+ [1 - \tilde{\alpha}_d(k,\ell)]\,|Y(k,\ell)|^2 \end{aligned} \tag{4}$$

where

$$\tilde{\alpha}_d(k,\ell) \triangleq \alpha_d + (1-\alpha_d)\,p'(k,\ell) \tag{5}$$

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 signal presence probability.

III. Signal Presence Probability

Speech presence in a given frame of a subband is determined by the ratio between the local energy of the noisy speech and its minimum within a specified time window. Let the local energy of the noisy speech be obtained by smoothing the magnitude squared of its STFT in time and frequency. In frequency, we use a window function b whose length is $2w+1$

$$S_f(k,\ell) = \sum_{i=-w}^{w} b(i)\,|Y(k-i,\ell)|^2. \tag{6}$$

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

$$S(k,\ell) = \alpha_s S(k,\ell-1) + (1-\alpha_s)\,S_f(k,\ell) \tag{7}$$

where $\alpha_s (0 < \alpha_s < 1)$ is a parameter. The minimum of the local energy, $S_{\min}(k,\ell)$, is searched using a simplified form of the procedure proposed in [6]. First, the minimum and a temporary variable $S_{tmp}(k,\ell)$ are initialized by $S_{\min}(k,0) = S(k,0)$ and $S_{tmp}(k,0) = S(k,0)$. Then, a samplewise comparison of

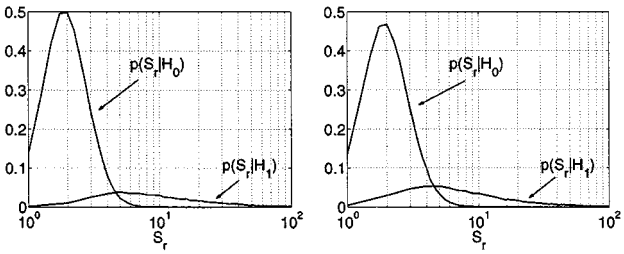

Fig. 1. Hypothetical probability density functions, $p\,(S_r|H_0)$ and $p\,(S_r|H_1)$, for: (a) White Gaussian noise and (b) F16 cockpit noise.

the local energy and the minimum value of the previous frame yields the minimum value for the current frame

$$S_{\min}(k,\ell) = \min\{S_{\min}(k,\ell-1), S(k,\ell)\} \tag{8}$$

$$S_{tmp}(k,\ell) = \min\{S_{tmp}(k,\ell-1), S(k,\ell)\}. \tag{9}$$

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

$$S_{\min}(k,\ell) = \min\{S_{tmp}(k,\ell-1), S(k,\ell)\} \tag{10}$$

$$S_{tmp}(k,\ell) = S(k,\ell) \tag{11}$$

and the search for the minimum continues with (8) and (9). 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 length of the window controls the bias upwards during "continuous" speech and the bias downwards when noise level increases. According to [6] and our own experiments with different speakers and environmental conditions, a suitable window is typically 0.5–1.5 s.

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

$$\frac{p\,(S_r|H_1)}{p\,(S_r|H_0)} \underset{H_0'}{\overset{H_1'}{\gtrless}} \frac{c_{10}P(H_0)}{c_{01}P(H_1)} \tag{12}$$

where $P(H_0)$ and $P(H_1)$ 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. 1 shows representative examples of conditional probability density functions, $p(S_r|H_0)$ and $p(S_r|H_1)$, obtained experimentally for white Gaussian noise and F16 cockpit noise, at -5 dB segmental SNR. Since the likelihood ratio $p(S_r|H_1)/p(S_r|H_0)$ is a monotonic function, the decision rule of (12) can be expressed as

$$S_r(k,\ell) \underset{H_0'}{\overset{H_1'}{\gtrless}} \delta. \tag{13}$$

We propose the following estimator for $p'(k,\ell)$:

$$\hat{p'}(k,\ell) = \alpha_p \hat{p'}(k,\ell-1) + (1-\alpha_p)\,I(k,\ell) \tag{14}$$

where $\alpha_p (0 < \alpha_p < 1)$ is a smoothing parameter and $I(k,\ell)$ denotes an indicator function for the result in (13), 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. First, δ is not sensitive to the type and intensity of environmental noise. Second, the probability of

TABLE I
SEGMENTAL SNR IMPROVEMENT FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING THE WEIGHTED AVERAGE (WA) AND MCRA NOISE ESTIMATORS

Input SegSNR	White Gaussian noise		Car interior noise		F16 cockpit noise	
[dB]	WA	MCRA	WA	MCRA	WA	MCRA
-5	9.82	9.95	9.32	10.07	6.50	7.36
0	7.73	7.94	7.11	7.93	4.36	5.43
5	5.85	6.12	5.01	6.18	2.73	3.83
10	4.10	4.47	3.41	4.81	1.34	2.50

$|Y|^2 \gg \lambda_d$ is very small when $S_r < \delta$. Hence, an increase in the estimated noise, consequent upon falsely deciding H_0' when H_1', is not significant. Third, the strong correlation of speech presence in consecutive frames is utilized (via α_p).

IV. EXPERIMENTAL RESULTS

The performance of the MCRA approach is evaluated and compared to that of the weighted average technique [3]. The evaluation is carried out within the framework of speech enhancement. Specifically, the MCRA and weighted average noise estimates are combined with the *optimally modified log-spectral amplitude* (OM-LSA) estimator [7] for obtaining a clean speech estimate in noisy environments. The assessment is based on an objective improvement in the segmental SNR, a subjective study of speech spectrograms and informal listening tests.

Three different noise types, taken from Noisex92 database [10], are used in our evaluation: white Gaussian noise, car interior noise, and F16 cockpit noise. The speech data include six different utterances, taken from the TIMIT database [11]. Half of the utterances are from male speakers and half are from female speakers. Each speech signal is degraded by the various noise types with segmental SNRs in the range $[-5, 10]$ dB. The segmental SNR is defined by [12]

$$SegSNR = \frac{10}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \log \frac{\sum_{k=0}^{N/2} |X(k,\ell)|^2}{\sum_{k=0}^{N/2} |D(k,\ell)|^2} \qquad (15)$$

where $\mathcal{L}$ represents the set of frames that contain speech and $|\mathcal{L}|$ its cardinality. The sampling frequency is 16 kHz. Accordingly, the following parameters have been chosen: frame size $N = 512$ (32 ms); frame update step $M = 128$ (75% overlapping windows); $\alpha_d = 0.95$; $w = 1$; $\alpha_s = 0.8$; $L = 125$ (1 s minima search window); $\delta = 5$; $\alpha_p = 0.2$. The weighted average technique is implemented with a weighting parameter $\alpha = 0.95$ and a threshold $\beta = 2$ (the parameters and method are described in [3]).

Table I shows the average segmental SNR improvement obtained for various noise types and at various noise levels. The MCRA estimator consistently achieves a higher improvement in the segmental SNR, than the weighted average estimator, under all tested environmental conditions. Its advantage is more significant in nonstationary noise environments. This is attributable to the fact that the weighted average noise estimate heavily relies on the instantaneous ratio between the spectral magnitudes of the degraded speech and the estimated noise. Presumably, considerably higher values occur at the onset of speech. However, in nonstationary and low-SNR noise environments, this assumption is occasionally not valid. Under nonstationary noise conditions, strong noise components are falsely considered as speech components. This yields an underestimated noise, resulting in high level of musical residual noise. On the other hand, low SNR may produce an overestimated noise, due to weak speech components that are partially processed as noise components. Consequently, the SNR for weak speech components gets even worse. This was confirmed by a subjective study of speech spectrograms and informal listening tests.

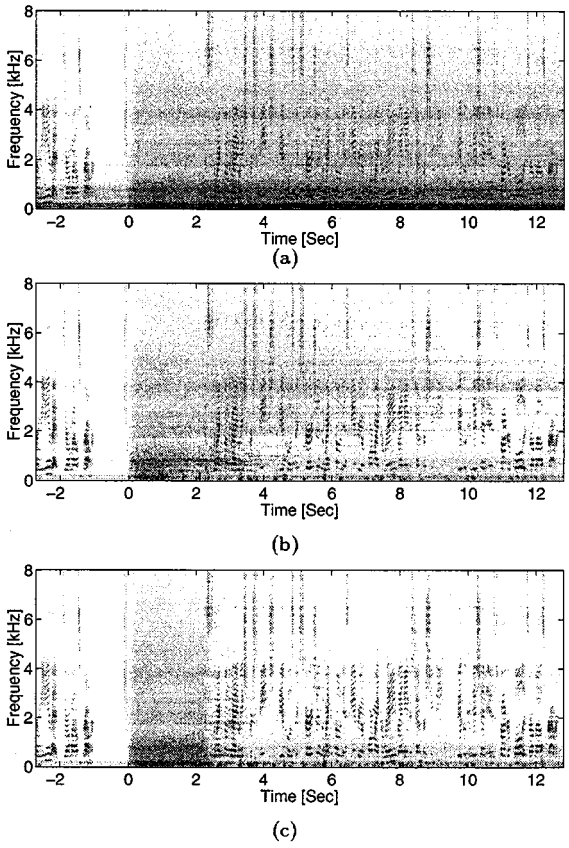

Fig. 2. Speech spectrograms: (a) noisy speech signal (car interior noise, defroster turned on at 0 s on full), (b) speech enhanced with the weighted average noise estimate, and (c) speech enhanced with the MCRA noise estimate.

Fig. 2 demonstrates the effect of a sudden rise in the noise energy level on the enhanced speech. Fig. 2(a) shows a noisy speech signal, recorded in a moving car, which contains a sudden increase in the noise level at 0 s. The increase in the noise was generated by turning on the defroster on full. Clearly, the speech enhanced with the weighted average noise estimate [Fig. 2(b)] is impaired by high level of residual noise, even as much as 12 s after the increase. By contrast, the MCRA noise estimate is built up in less than 3 s, allowing an efficient speech enhancement shortly after the substantial change in the noise statistics [Fig. 2(c)]. Furthermore, since the update of the weighted average estimate is based on relatively weak noise components *compared to the noise estimate*, its response is slower to greater changes in the noise statistics. On the other hand, the update of the MCRA estimate is controlled by the minima values *within a finite time window*. Hence, its response is not sensitive to the extent of noise variations.

V. Conclusion

Recursive averaging is a commonly used procedure for estimating the noise power spectrum. However, rather than employing a voice activity detector and restricting the update of the noise estimator to periods of speech absence and rather than computing a weighted average based on the instantaneous spectral magnitudes of the degraded speech and estimated noise, we adapt the smoothing parameter in time and frequency according to the speech presence probability. The speech presence probability is controlled by the minima values of a smoothed periodogram of the noisy measurement. Compared to a competitive method, the MCRA noise estimate responses more quickly to noise variations and, when integrated into a speech enhancement system, yields higher segmental SNR and a lower level of musical residual noise.

References

[1] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, pp. 108–110, May 2000.

[2] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments," in *Proc. 24th IEEE ICASSP'99*, Phoenix, AZ, Mar. 15–19, 1999, pp. 789–792.

[3] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. 20th IEEE ICASSP'95*, Detroit, MI, May 8–12, 1995, pp. 153–156.

[4] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137–145, Apr. 1980.

[5] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. 25th IEEE ICASSP'2000*, Istanbul, Turkey, June 5–9, 2000, pp. 1875–1878.

[6] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. 7th EUSIPCO'94*, Edinburgh, U.K., Sept. 13–16, 1994, pp. 1182–1185.

[7] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, pp. 2403–2418, Nov. 2001.

[8] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. 4th EUROSPEECH'95*, Madrid, Spain, Sept. 18–21, 1995, pp. 1513–1516.

[9] J. Meyer, K. U. Simmer, and K. D. Kammeyer, "Comparison of one- and two-channel noise-estimation techniques," in *Proc. 5th IWAENC'97*, London, U.K., Sept. 11–12, 1997, pp. 137–145.

[10] 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 Commun.*, vol. 12, no. 3, pp. 247–251, July 1993.

[11] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Nat. Inst. Stand. Technol. (NIST), Gaithersburg, MD, 1988.

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