SPECTRAL ENHANCEMENT BY TRACKING SPEECH PRESENCE PROBABILITY IN SUBBANDS

Israel Cohen and Baruch Berdugo

Lamar Signal Processing Ltd., P.O.Box 273, Yokneam Ilit 20692, Israel
icohen@lamar.co.il; http://www.AndreaElectronics.com

ABSTRACT

In this paper, we introduce a *Minima Controlled Recursive Averaging* (MCRA) noise estimation approach for robust speech enhancement. The noise spectrum is estimated by recursively 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 noise estimate is unbiased, computationally efficient, robust with respect to the input signal-to-noise ratio and type of underlying additive noise, and characterized by the ability to quickly follow abrupt changes in the noise spectrum. Incorporated in the *Optimally-Modified Log-Spectral Amplitude* estimator, excellent noise suppression is achieved, while retaining weak speech components and avoiding the musical residual noise phenomena.

1. INTRODUCTION

A crucial component of a practicable speech enhancement system is the estimation of the noise power spectrum. The noise can be clearly estimated based on histograms in the power spectral domain [10, 6, 12]. However, such methods are computationally expensive. An alternative commonly used approach 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 [7] or estimated independently for individual subbands using *a posteriori* signal-to-noise ratio (SNR) [8, 6]. Unfortunately, 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 case of varying noise spectrum.

Martin [9] 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 [12], and generally biased. The factor merely compensates the bias for stationary white Gaussian noise and *independent* power estimates, which is obviously not applicable since successive values are correlated. Moreover, this method occasionally attenuates low energy phonemes [9]. In [4], a computationally more efficient minimum tracking scheme is presented. Its main drawback is the very slow update rate of the noise estimate in case of a sudden rise in noise energy level.

In this paper, we introduce a *Minima Controlled Recursive Averaging* (MCRA) noise estimation approach for robust speech enhancement. The noise spectrum is estimated by recursively 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, the speech/non-speech segmentation is "softened" via a temporal smoothing, which exploits the strong correlation of speech presence in neighboring frames. The resultant 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.

The MCRA noise estimate is incorporated in a speech enhancement system [1, 2] and compared to alternative conventional noise estimates. Objective and subjective evaluation show that the proposed approach is superior under all tested environmental conditions.

2. 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,

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

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 sampling 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

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

where $X(k,\ell)$ and $D(k,\ell)$ represent the STFT of the clean and noise signals, respectively. Let $\lambda_d(k,\ell) = E[|D(k,\ell)|^2]$

Proc. HSC'01, Kyoto, Japan, 9–11 April 2001, pp. 95–98.

denote the variance of the noise in the kth subband. Then a common technique to obtain its estimate is to apply a temporal recursive smoothing to the noisy measurement during periods of speech absence. In particular,

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

where α_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 Eqs. (2), used for estimating the clean speech, and the hypotheses in Eqs. (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.* $P(H_1|Y) \geq P(H_1'|Y)$.

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

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

where

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

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.

Tracking the conditional signal presence probability is based on the local statistics in the time-frequency plane of the noisy speech energy. Accordingly, speech absence 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. The ratio is compared to a certain threshold value, where a smaller ratio indicates absence of speech. To reduce fluctuations between speech and non-speech segments, a recursive temporal averaging is carried out, thereby taking into account the strong correlation of speech presence in neighboring frames.

The local energy of the noisy speech 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 $2w+1$:

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

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

$$S(k,\ell) = \alpha_s S(k,\ell-1) + (1-\alpha_s)S_f(k,\ell)\,, \qquad (7)$$

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 [9]. 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

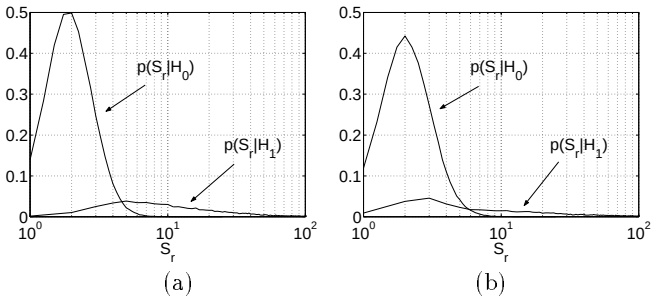

Fig. 1. Hypothetical probability density functions, $p(S_r|H_0)$ and $p(S_r|H_1)$, for: (a) White Gaussian noise; (b) Car interior noise.

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

$$S_{\min}(k,\ell) = \min\{S_{\min}(k,\ell-1), S(k,\ell)\} \qquad (8)$$
$$S_{tmp}(k,\ell) = \min\{S_{tmp}(k,\ell-1), S(k,\ell)\}\,. \qquad (9)$$

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

$$S_{\min}(k,\ell) = \min\{S_{tmp}(k,\ell-1), S(k,\ell)\} \qquad (10)$$
$$S_{tmp}(k,\ell) = S(k,\ell)\,, \qquad (11)$$

and the search for the minimum continues with Eqs. (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 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 noise level increases. According to [9] 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 speech and its derived minimum. A Bayes minimum-cost decision rule is given by

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

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 car interior noise, at -5dB 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

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

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

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

where α_p $(0 < \alpha_p < 1)$ is a smoothing parameter, and $I(k,\ell)$ denotes an indicator function for the result in (13),

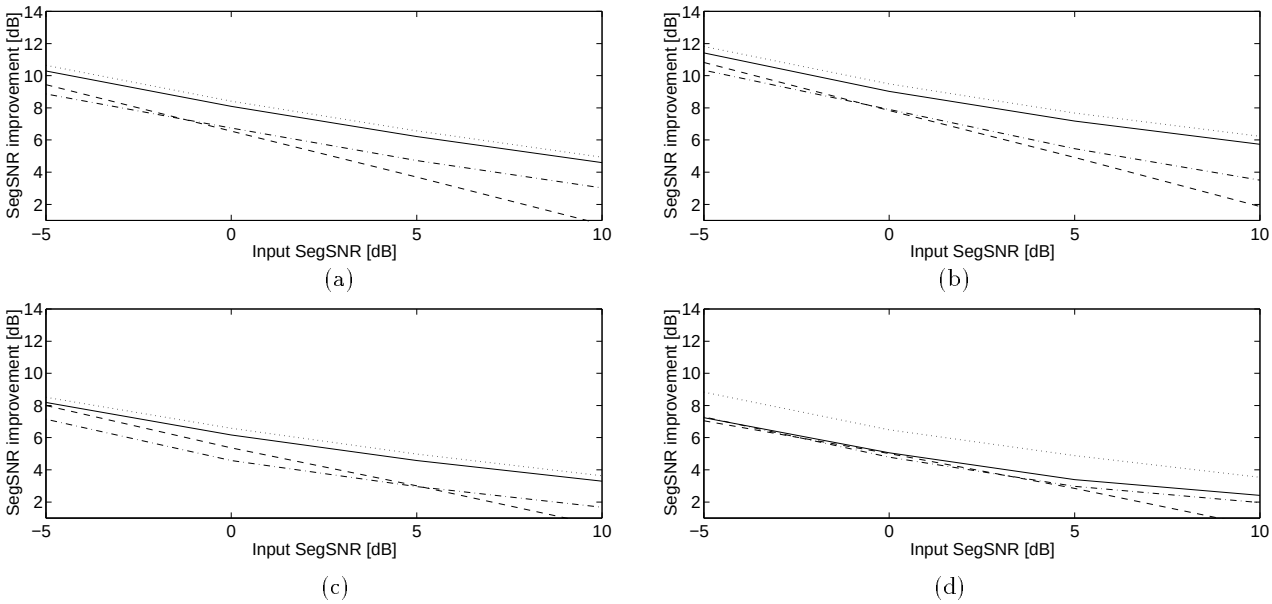

Fig. 2. Average Segmental SNR improvement for various noise types and levels: (a) White Gaussian noise; (b) Car interior noise; (c) F16 cockpit noise; (d) Speech babble noise. The noise spectrum is estimated by the Minimum Statistics method (dashed), Weighted Average (dashdot) and the MCRA approach (solid). A theoretical limit is obtained by calculating the noise spectrum from the noise itself (dotted).

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. 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 not significant. Thirdly, the strong correlation of speech presence in consecutive frames is utilized (via α_p).

3. SPEECH ENHANCEMENT

To estimate the clean speech spectrum, we use the OM-LSA estimator [1, 2]. This estimator minimizes the mean-square error of the log-spectra under signal presence uncertainty. Its superior performance was demonstrated [1, 2] even in the most adverse noise conditions.

Let $G(k,\ell)$ denote a real spectral gain function. Then an estimate of the clean speech STFT is given by

$$\hat{X}(k,\ell) = G(k,\ell)Y(k,\ell)\,. \tag{15}$$

Applying the inverse STFT, with a synthesis window $\tilde{h}$ that is biorthogonal to the analysis window h [13], yields an estimate for the clean speech signal:

$$\hat{x}(n) = \sum_{\ell}\sum_{k=0}^{N-1} \hat{X}(k,\ell)\tilde{h}(n-\ell M)\, e^{j\frac{2\pi}{N}k(n-\ell M)}\,. \tag{16}$$

In practice, (16) is efficiently implemented using the weighted overlap-add method [3].

The spectral gain function in (15) is given by [2]

$$G(k,\ell) = \{G_{H_1}(k,\ell)\}^{p(k,\ell)} \cdot G_{min}^{1-p(k,\ell)}\,, \tag{17}$$

where $p(k,\ell) \triangleq P(H_1(k,\ell)|Y(k,\ell))$ designates the conditional signal presence probability, $G_{H_1}(k,\ell)$ the conditional gain when speech is present, and G_{min} a lower bound constraint for the gain when speech is absent. We have derived [1, 2] an efficient estimator for $p(k,\ell)$, based on the time-frequency distribution of the *a priori* SNR and three parameters that quantify the speech likelihood in subbands. In contrast to other methods, where high *a posteriori* SNR produces high spectral gain resulting in a random appearance of tone-like noise (musical-noise phenomena) [8, 7, 4], the OM-LSA estimator attenuates noise by identifying noise-only regions and reduces the gain correspondingly to G_{min}. Yet, it avoids the attenuation of weak speech components by letting $p(k,\ell)$ increase to one in speech regions.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The MCRA noise estimate is compared to the Minimum Statistics [9] and conventional Weighted Average [6] noise estimates. The comparison is accomplished by evaluating their performance when incorporated in the *Optimally-Modified Log-Spectral Amplitude* (OM-LSA) estimator [1, 2]. A theoretical limit, achievable by calculating the noise spectrum from the noise itself, is also considered.

Four different noise types, taken from Noisex92 database, are used in our evaluation: white Gaussian noise, car noise, F16 cockpit noise, and speech babble noise. Since noise signals have different impacts on different speech signals, the speech enhancement performance is evaluated using six different utterances, taken from the TIMIT database. Half of the utterances are from male speakers, and half are from female speakers. The evaluation consists of an objective segmental SNR measure, a subjective study of speech spectrograms and informal listening tests.

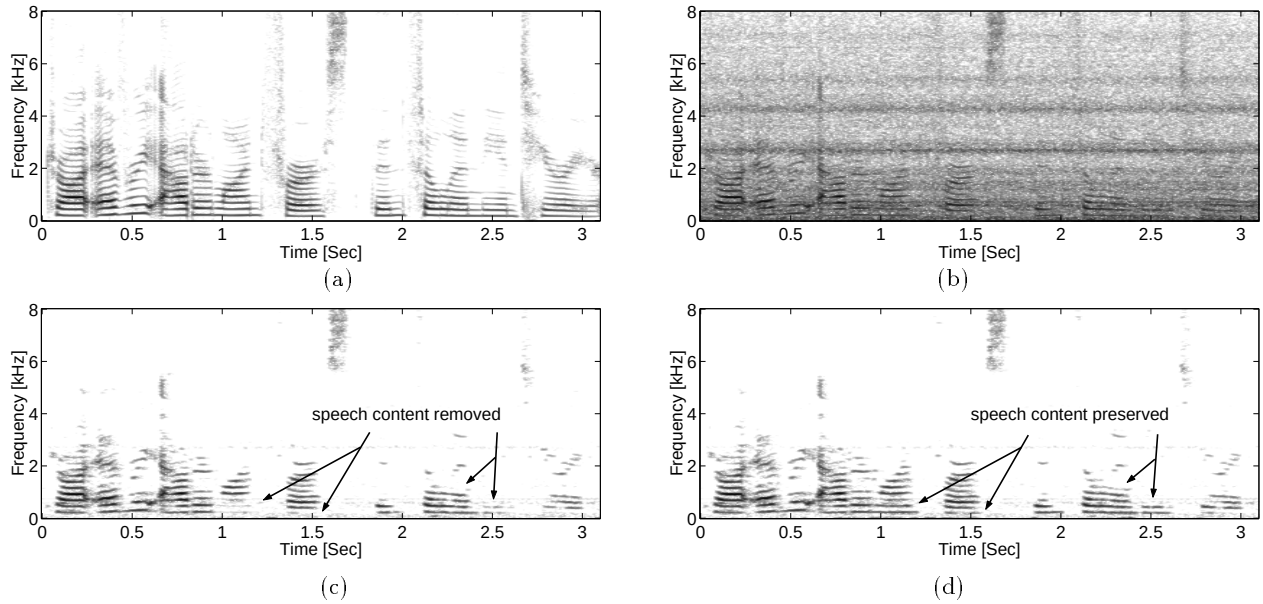

Fig. 3. Speech spectrograms. (a) Original clean speech signal: "Draw every outer line first, then fill in the interior."; (b) Noisy signal (additive F16 cockpit noise at a SegSNR = 0 dB); (c) Speech enhanced with the Minimum Statistics noise estimate (SegSNR = 5.7 dB); (d) Speech enhanced with the MCRA noise estimate (SegSNR = 7.0 dB).

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 [11]

$$SegSNR = \frac{10}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \log \frac{\sum_{k=0}^{N/2} |X(k,\ell)|^2}{\sum_{k=0}^{N/2} |D(k,\ell)|^2} \qquad (18)$$

where $\mathcal{L}$ represents the set of frames that contain speech. The sampling frequency is 16 kHz. Accordingly, the following parameters have been chosen: frame size $N = 512$ (32 ms); time sampling step $M = 128$ (75% overlapping windows); $\alpha_d = 0.95$; $w = 1$; $\alpha_s = 0.8$; $L = 125$ ($1s$ minima search window); $\delta = 5$; $\alpha_p = 0.2$; $G_{min} = -25$ dB. We used biorthogonal Hanning windows [13], and estimated the *a priori* SNR by the decision-directed approach with a smoothing parameter set to 0.92 [5].

Fig. 2 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. A subjective comparison was also conducted using speech spectrograms and validated by informal listening tests. Example of speech spectrograms obtained with the MCRA noise estimate and the Minimum Statistics approach are shown in Fig. 3. Particularly, compare low frequency formants having low input SNR. The proposed method demonstrates excellent noise suppression, while retaining weak speech components and avoiding the musical residual noise phenomena.

5. REFERENCES

[1] I. Cohen, "On Speech Enhancement Under Signal Presence Uncertainty," *ICASSP-01*, Salt Lake City, Utah, May 2001.

[2] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," submitted.

[3] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Prentice-Hall, 1983.

[4] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands," *EUROSPEECH'95*, September 1995, pp. 1513–1516.

[5] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP*, vol. ASSP-32, no. 6, December 1984, pp. 1109–1121.

[6] H. G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," *ICASSP-95*, Detroit, Michigan, May 1995, pp. 153–156.

[7] N. S Kim and J.-H. Chang, "Spectral Enhancement Based on Global Soft Decision," *IEEE Sig. Process. Lett.*, vol. 7, no. 5, May 2000, pp. 108–110.

[8] D. Malah, R. V. Cox and A. J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," *ICASSP-99*, Phoenix, Arizona, March 1999, pp. 789–792.

[9] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *Proc. Euro. Signal Processing Conf., EUSIPCO-94*, September 1994, pp. 1182–1185.

[10] R. J. McAulay and M. L. Malpass "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. ASSP*, vol. ASSP-28, April 1980, pp. 137–145.

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

[12] V. Stahl, A. Fischer and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," *ICASSP-00*, Istanbul, Turkey, June 2000, pp. 1875–1878.

[13] J. Wexler and S. Raz, "Discrete Gabor Expansions," *Signal Processing*, vol. 21, no. 3, November 1990, pp. 207–220.