Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 48 (2006) 220–231

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A noise-estimation algorithm for highly non-stationary environments

Sundarrajan Rangachari, Philipos C. Loizou *

Department of Electrical Engineering, University of Texas at Dallas, P.O. Box 830688, EC 33 Richardson, TX 75083-0688, USA

Received 17 December 2004; received in revised form 11 August 2005; accepted 24 August 2005

Abstract

A noise-estimation algorithm is proposed for highly non-stationary noise environments. The noise estimate is updated by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors, which are adjusted based on signal-presence probability in individual frequency bins. Signal presence is determined by computing the ratio of the noisy speech power spectrum to its local minimum, which is updated continuously by averaging past values of the noisy speech power spectra with a look-ahead factor. The local minimum estimation algorithm adapts very quickly to highly non-stationary noise environments. This was confirmed with formal listening tests which indicated that the proposed noise-estimation algorithm when integrated in speech enhancement was preferred over other noise-estimation algorithms.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Speech enhancement; Noise estimation; Non-stationary noise

1. Introduction

In most speech-enhancement algorithms, it is assumed that an estimate of the noise spectrum is available. Such an estimate is critical for the performance of speech-enhancement algorithms as it is needed, for instance, to evaluate the Wiener filter in the Wiener algorithms (Lim and Oppenheim, 1978) or to estimate the a priori SNR in the MMSE algorithms (Ephraim and Malah, 1984) or to estimate the noise covariance matrix in the subspace algorithms (Ephraim and Van Trees, 1993). The noise estimate can have a major impact on the quality of the enhanced signal. If the noise estimate is too low, annoying residual noise will be audible, while if the noise estimate is too high, speech will be distorted resulting possibly in intelligibility loss. The simplest approach is to estimate and update the noise spectrum during the silent (e.g., during pauses) segments of the signal using a voice-activity detection (VAD) algorithm (e.g., Sohn and Kim, 1999). Although such an approach might work satisfactorily in stationary noise (e.g., white noise), it will not work well in more realistic environments (e.g., in a restaurant) where the spectral characteristics of the noise might be changing constantly. Hence there is a need to update the noise spectrum continuously over time and this can be done using noise-estimation algorithms.

* Corresponding author. Tel.: +1 972 883 4617; fax: +1 972 883 2710.
E-mail address: loizou@utdallas.edu (P.C. Loizou).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.08.005

Several noise-estimation algorithms have been proposed for speech enhancement applications (Malah et al., 1999; Martin, 2001; Cohen, 2002; Cohen, 2003; Doblinger, 1995; Hirsch and Ehrlicher, 1995; Lin et al., 2003; Stahl et al., 2000; Rangachari et al., 2004; Ris and Dupont, 2001). Martin (2001) proposed a method for estimating the noise spectrum based on tracking the minimum of the noisy speech over a finite window. As the minimum is typically smaller than the mean, unbiased estimates of noise spectrum were computed by introducing a bias factor based on the statistics of the minimum estimates. The main drawback of this method is that it takes slightly more than the duration of the minimum-search window to update the noise spectrum when the noise floor increases abruptly.

Cohen (2002) proposed a minima controlled recursive algorithm (MCRA) which updates the noise estimate by tracking the noise-only regions of the noisy speech spectrum. These regions are found by comparing the ratio of the noisy speech to the local minimum against a threshold. The noise estimate, however, lags by at most twice that window length when the noise spectrum increases abruptly. In the improved MCRA approach (Cohen, 2003), a different method was used to track the noise-only regions of the spectrum based on the estimated speech-presence probability. This probability, however, is also controlled by the minima, and therefore the algorithm incurs roughly the same delay as the MCRA algorithm for increasing noise levels.

Doblinger (1995) updated the noise estimate by continuously tracking the minimum of the noisy speech in each frequency bin. As such, it is computationally more efficient than the method in (Martin, 2001). However, it fails to differentiate between an increase in noise floor and an increase in speech power.

Hirsch and Ehrlicher (1995) updated the noise estimate by comparing the noisy speech power spectrum to the past noise estimate. Their method is also simple to implement, however it fails to update the noise estimate when the noise floor increases abruptly and stays at that level.

Ris and Dupont (2001) combined the above techniques with narrow-band spectral analysis which allowed estimation of the noise levels in the valleys between harmonics of voiced speech segments. Longer time windows were required to achieve the required spectral resolution. Although their approach refines the spectral resolution of the noise level, it does not adapt faster to increasing noise levels. Lastly, in (Stahl et al., 2000) a quantile-based noise-estimation algorithm was proposed which estimates the noise spectrum based on the qth quantile of the noisy speech power spectrum. This method might fail to estimate the noise floor correctly if the noisy speech contains highly-varying noise.

Several other noise-estimation algorithms were proposed for speech recognition applications (Deng et al., 2003; Deng et al., 2003; Afify and Sioham, 2001; Kim, 1998; Yao and Nakamura, 2002). Unlike the above algorithms, these noise-estimation algorithms were based on statistical principles and operated at the feature level (e.g., MFCC coefficients) in the log-spectral domain. Very briefly, the noisy speech feature vectors were modeled using a mixture of Gaussians, and the noise feature vectors were obtained by maximizing a conditional likelihood function based on a recursive EM algorithm. Stochastic approximations were made to sequentially update the noise feature vectors. Some of those noise updates resembled the time-recursive updates of the noise spectrum used in the above noise-estimation algorithms. In fact, some (Afify and Sioham, 2001) proposed the use of optimum smoothing factors for the noise updates similar to (Martin, 2001). Improvements to the EM-based methods were reported in (Yao and Nakamura, 2002) using sequential Monte–Carlo techniques.

In brief, most of the aforementioned noise-estimation algorithms developed for speech-enhancement algorithms do not adapt quickly to increasing noise levels. Recently we introduced a noise-estimation algorithm (Rangachari et al., 2004) which updates the noise estimate faster than the above methods and also avoids overestimation of the noise level. The noise estimate was updated in each frame based on voice-activity detection. If speech was absent in a specific frame, the noise estimate was updated with a constant smoothing factor. The speech-presence decision made in each speech frame was based on the ratio of noisy speech spectrum to its local minimum. Results indicated that the noise-estimation algorithm (Rangachari et al., 2004) took only 0.5 s to adapt to sudden increases in noise levels compared to 1–1.5 s required by other algorithms.

In this paper, we further improve our noise-estimation algorithm in the following aspects: (1) update of the noise estimate without explicit voice-activity decision, (2) estimate of speech-presence probability exploiting the correlation of power spectral components in neighboring frames. The proposed algorithm updates the noise estimate in each frame using a time–frequency dependent smoothing factor computed based on the speech-presence probability.

This paper is organized as follows. Section 2 describes the proposed noise-estimation algorithm. Section 3 compares the proposed method with some of the existing algorithms. Section 4 presents the subjective and objective evaluation of the proposed algorithm and Section 5 gives our conclusions.

2. Proposed noise-estimation algorithm

Let the noisy speech signal in the time domain be denoted as

$$y(n) = x(n) + d(n) \tag{1}$$

where $x(n)$ is the clean speech and $d(n)$ is the additive noise. The smoothed power spectrum of noisy speech is computed using the following first-order recursive equation:

$$P(\lambda,k) = \eta P(\lambda-1,k) + (1-\eta)|Y(\lambda,k)|^2 \tag{2}$$

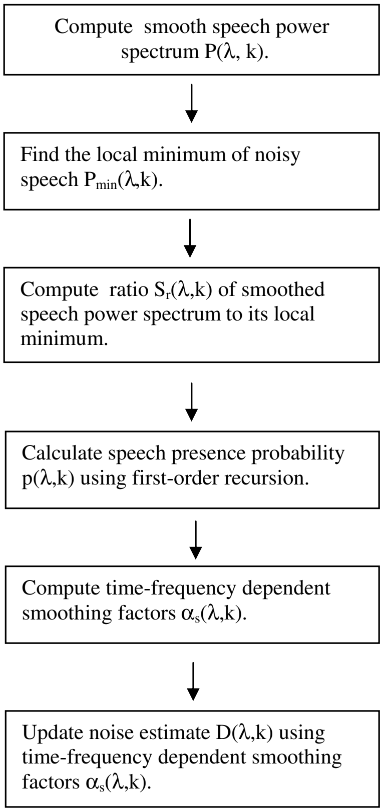
Compute smooth speech power spectrum P(λ, k).
Find the local minimum of noisy speech $P_{min}(\lambda,k)$.
Compute ratio $S_r(\lambda,k)$ of smoothed speech power spectrum to its local minimum.
Calculate speech presence probability p(λ,k) using first-order recursion.
Compute time-frequency dependent smoothing factors $\alpha_s(\lambda,k)$.
Update noise estimate D(λ,k) using time-frequency dependent smoothing factors $\alpha_s(\lambda,k)$.

Fig. 1. Flow diagram of proposed noise-estimation algorithm.

where $P(\lambda,k)$ is the smoothed power spectrum, λ is the frame index, k is the frequency index, $|Y(\lambda,k)|^2$ is the short-time power spectrum of noisy speech and η is a smoothing constant. The proposed algorithm is summarized in the flow chart diagram shown in Fig. 1. Next, we describe each of the individual blocks of the algorithm.

2.1. Tracking the minimum of noisy speech

Various methods (Martin, 2001; Martin, 1994) were proposed for tracking the minimum of the noisy speech power spectrum over a fixed search window length. These methods were sensitive to outliers and also the noise update was dependent on the length of the minimum-search window. A different non-linear rule is used in our method for tracking the minimum of the noisy speech by continuously averaging past spectral values (Doblinger, 1995)

$$\begin{array}{l} \text{If } P_{\min}(\lambda-1,k) < P(\lambda,k) \text{ then} \\ \quad P_{\min}(\lambda,k) = \gamma P_{\min}(\lambda-1,k) \\ \quad + \dfrac{1-\gamma}{1-\beta}(P(\lambda,k) - \beta P(\lambda-1,k)) \\ \text{else} \\ \quad P_{\min}(\lambda,k) = P(\lambda,k) \\ \text{end} \end{array} \tag{3}$$

where $P_{\min}(\lambda,k)$ is the local minimum of the noisy speech power spectrum and β and γ are constants which are determined experimentally. The look-ahead factor β controls the adaptation time of the local minimum. Fig. 2 shows the power spectrum of noisy speech and the local minimum tracked with the above mentioned rule for a sentence degraded by babble noise at 5 dB SNR. The adaptation time for the algorithm is ≈0.5 s for non-stationary noise.

2.2. Speech-presence probability

The approach taken to determine speech presence in each frequency bin is similar to the method used in (Cohen, 2002). Let the ratio of noisy speech power spectrum and its local minimum be defined as

$$S_r(\lambda,k) = P(\lambda,k)/P_{\min}(\lambda,k) \tag{4}$$

This ratio is compared with a frequency dependent threshold, and if the ratio is found to be greater than the threshold, it is taken as a speech-present frequency bin else it is taken as a speech-absent

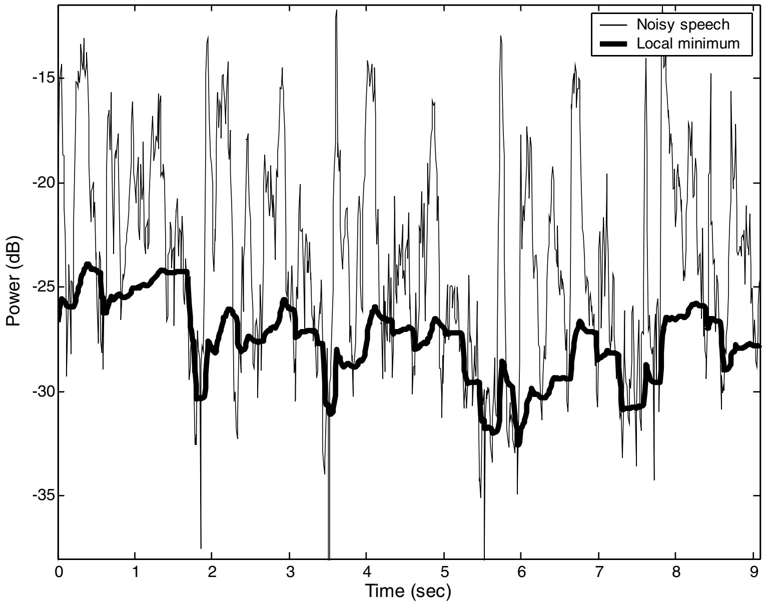

Fig. 2. Plot of noisy speech power spectrum and local minimum using (3) for a speech degraded by babble noise at 5 dB SNR at frequency bin $k = 5$.

frequency bin. This is based on the principle that the power spectrum of noisy speech will be nearly equal to its local minimum when speech is absent. Hence the smaller the ratio is in (4), the higher the probability that it will be a noise-only region and vice versa. The speech-presence decision can be summarized as follows:

```
if $S_r(\lambda,k) > \delta(k)$
    $I(\lambda,k) = 1$   speech present
else                                        (5)
    $I(\lambda,k) = 0$   speech absent
end
```

where $\delta(k)$ is the frequency-dependent threshold determined experimentally. Note that in (Cohen, 2002), a fixed threshold was used in place of $\delta(k)$ for all frequencies. From the above rule, the speech-presence probability, $p(\lambda,k)$, is updated using the following first-order recursion:

$$p(\lambda,k) = \alpha_p p(\lambda-1,k) + (1-\alpha_p)I(\lambda,k) \tag{6}$$

where α_p is a smoothing constant. Note that the above recursion implicitly exploits the correlation for speech presence in adjacent frames. Fig. 3 illustrates the speech presence or absence decision made using the above rule. In this figure, we show the speech present/absent detection for a sentence degraded by babble noise at 5 dB SNR.

The dark regions (top panel in Fig. 3) indicate speech-present regions and the white regions indicate speech-absent regions as identified by Eq. (5). As can be seen, our detection process had detected nearly almost all of the speech-present regions correctly. Also, note that by choosing a smaller threshold value, we can detect speech presence with higher confidence thus avoiding potential speech distortion. This can also be observed from Fig. 3 where we can see that some of the noise-only regions were detected as speech. Only few of the low-energy speech regions were detected as noise-only regions. This may result in slight overestimate of the noise spectrum but will not likely have much effect on the enhanced speech.

2.3. Computing frequency-dependent smoothing constants

Using the above speech-presence probability estimate, we compute the time–frequency dependent smoothing factor as follows (Cohen, 2002):

$$\alpha_s(\lambda,k) \overset{\Delta}{=} = \alpha_d + (1-\alpha_d)p(\lambda,k) \tag{7}$$

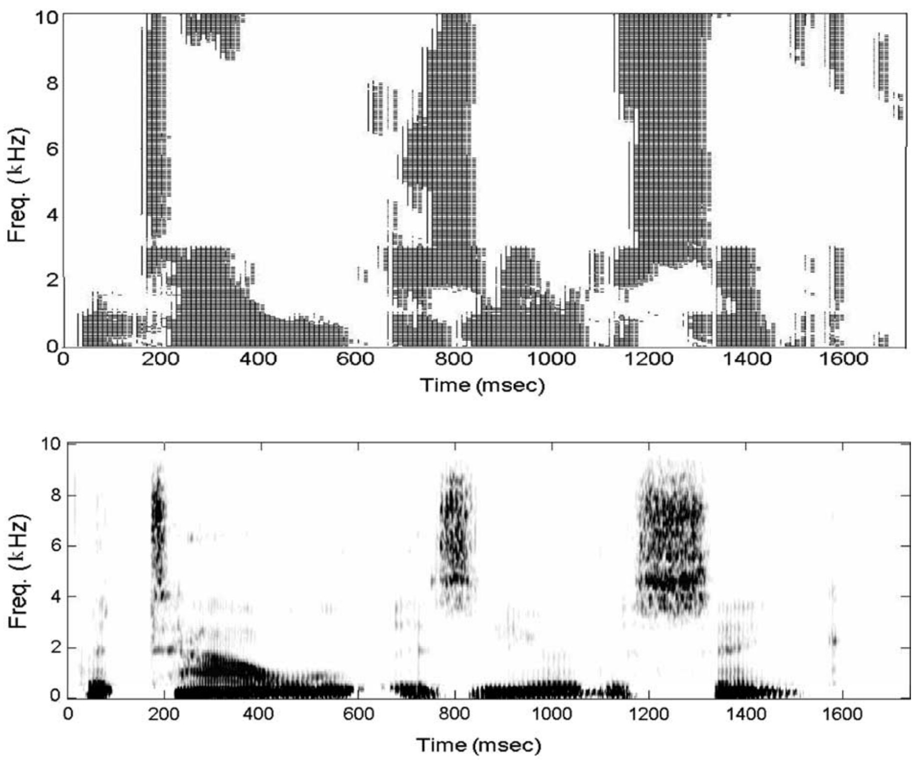

Fig. 3. Top panel: Plot of estimated speech-presence probability based on the ratio $\mathrm{Sr}(\lambda,k)$. Bottom panel: spectrogram of the clean signal.

where α_{d} is a constant. Note that $\alpha_{\mathrm{s}}(\lambda,k)$ takes values in the range of $\alpha_{\mathrm{d}} \leqslant \alpha_{\mathrm{s}}(\lambda,k) \leqslant 1$.

2.4. Update of noise spectrum estimate

Finally, after computing the frequency-dependent smoothing factor $\alpha_{\mathrm{s}}(\lambda,k)$ using Eq. (7), the noise spectrum estimate is updated as

$$D(\lambda,k) = \alpha_{\mathrm{s}}(\lambda,k)D(\lambda-1,k) + (1-\alpha_{\mathrm{s}}(\lambda,k))|Y(\lambda,k)|^2 \tag{8}$$

where $D(\lambda,k)$ is the estimate of the noise power spectrum. Hence, the overall algorithm can be summarized as follows. After classifying the frequency bins into speech present/absent using Eq. (5), we update the speech-presence probability using Eq. (6) and then use this probability to update the time–frequency dependent smoothing factor in Eq. (7). Finally the noise spectrum estimate is updated according to Eq. (8) using the time–frequency dependent smoothing factor.

Fig. 4 shows as an example the true noise spectrum and the estimated noise spectrum calculated with our proposed method for a sentence degraded by babble noise at 5 dB SNR.

3. Comparison of proposed method with existing algorithms

In this section, we provide qualitative comparisons between our proposed algorithm and other existing noise-estimation algorithms.

3.1. Comparison with MS (Martin, 2001)

The minimum statistics (MS) algorithm (Martin, 2001) updates the noise estimate based on tracking the minimum of the noisy speech spectrum. Hence the adaptation time of the noise estimate depends on the adaptation time of the local minimum. For non-stationary noise conditions where the noise power varies slowly over time, our method and the minimum statistics method have the same adaptation time. But for increasing noise levels, the adaptation time might be slightly more than 1.5 s for the MS method whereas for our method it is only 0.5 s. Fig. 5 shows a comparison between the

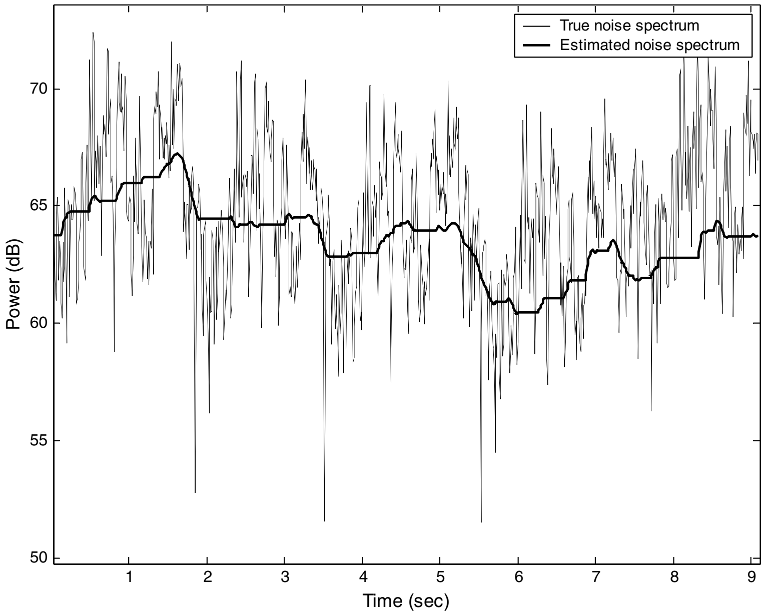

Fig. 4. Plot of true noise spectrum and the estimated noise spectrum using our proposed method for a speech degraded by babble noise at 5 dB SNR and single frequency $f = 250$ Hz.

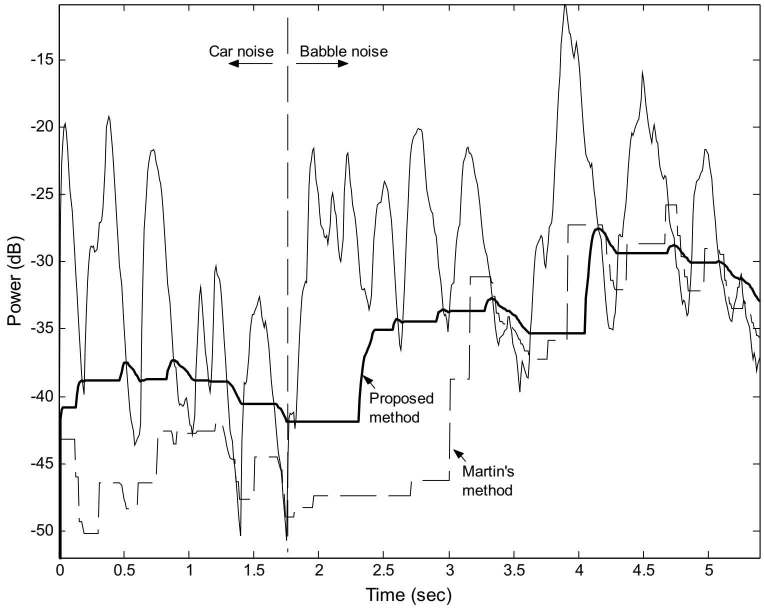

Fig. 5. Comparison between the noise spectrum (for $f = 1$ kHz) estimated using the proposed algorithm (thick line) and Martin's (Martin, 2001) (dashed line) algorithm for a sentence corrupted by car noise ($t < 1.8$ s) followed by a sentence corrupted by multi-talker babble ($t > 1.8$ s).

MS method and our proposed method in a situation where there is a sudden increase in noise power level. From the figure we can see that MS takes more than 1.5 s to update the noise spectrum, whereas our proposed method takes only 0.5 s to update to the higher noise floor.

3.2. Comparison with continuous minima tracking (Doblinger, 1995)

In the method proposed in (Doblinger, 1995), the noise estimate increases whenever the noisy speech power increases. This problem is avoided in our proposed method by using the ratio of the noisy speech to the local minimum. Whenever the noisy speech power increases, the ratio between the noisy speech and local minimum exceeds the threshold, and the noise estimate is not updated. This can be seen from Fig. 6 which compares the noise estimate obtained using the method in (Doblinger, 1995) and our method.

3.3. Comparison with weighted average technique (Hirsch and Ehrlicher, 1995)

Two methods were presented in (Hirsch and Ehrlicher, 1995) for noise estimation, one based on weighted averaging and one based on histograms of past speech segments. In the weighted averaging method, the noise estimate was updated whenever the noisy speech was less than a threshold, which was proportional to the previous noise estimate. Although this approach works satisfactorily in most cases, it fails in the following scenario. Consider an example where there is a sudden increase in noise level. This will result in a situation where the noisy speech spectrum will never be smaller than the threshold, since the threshold is based on the past noise estimates already very low. Thus, the noise estimate will not be updated if the noise power remains at that high level. Fig. 7 shows the comparison of the noise estimate using our proposed method with Hirsch and Ehrlicher (1995) for a sentence corrupted by babble noise initially at a high SNR (20 dB) level followed by a low SNR (5 dB) level. Our proposed method tracked the higher noise power within ≈0.5 s.

3.4. Comparison with MCRA (Cohen, 2002) and IMCRA (Cohen, 2003) methods

The local minimum in (Cohen, 2002) was found by tracking the minimum of noisy speech over a search window spanning L frames. This has some drawbacks. First, the minimum is sensitive to outliers.

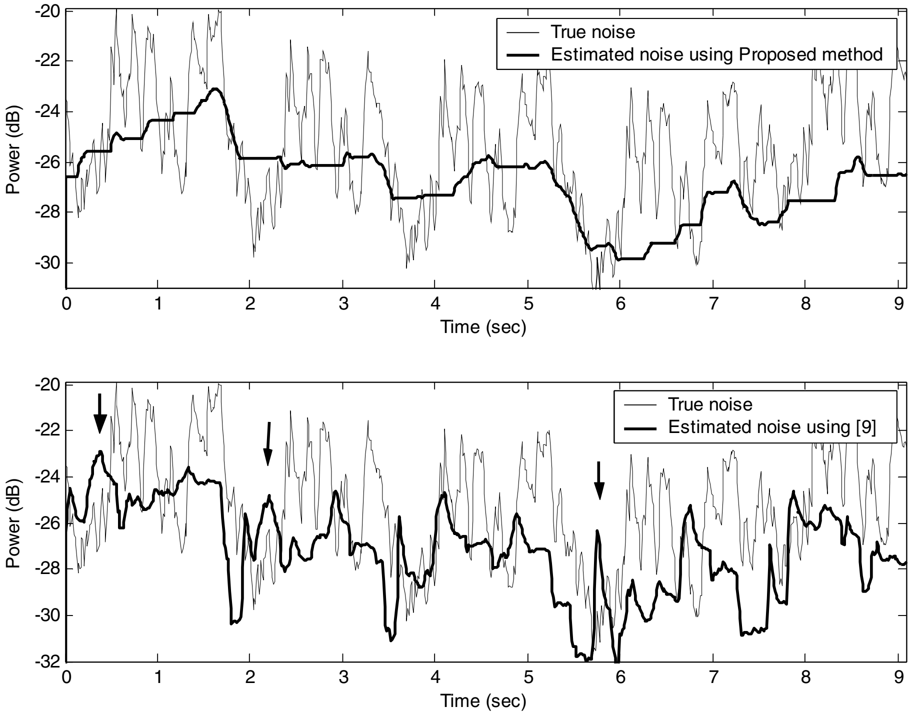

Fig. 6. Top panel: Plot of true noise spectrum and estimated noise spectrum using the proposed method for a noisy speech signal (5 dB SNR) at $f = 250$ Hz. Bottom panel: Plot of true noise spectrum and estimated noise spectrum using (Doblinger, 1995). Arrows indicate regions where noise is overestimated.

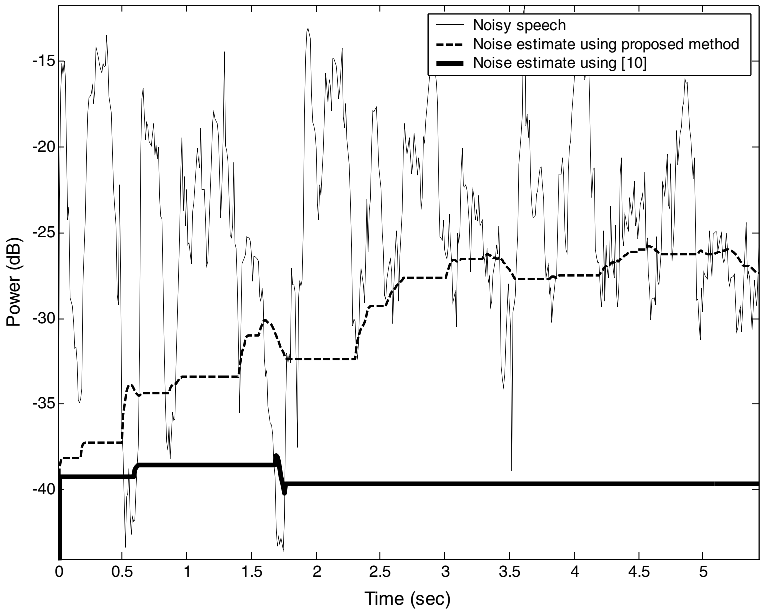

Fig. 7. Comparison of estimated noise spectrum ($f = 500$ Hz) of proposed method (dashed line) with that of Hirsch and Ehrlicher (1995) (solid line) for a noisy speech of SNR 20 dB ($t < 1.8$ s) followed by a noisy speech of SNR 5 dB ($t > 1.8$ s).

Second, the update of minimum can take at most 2L frames for increasing noise levels. Improvements to the method in (Cohen, 2002) were reported in (Cohen, 2003) with the IMCRA method. In the IMCRA algorithm a different formula was used to estimate the speech-presence probability $p(\lambda,k)$, now a function of the a priori speech-absence probability $q(\lambda,k)$. The computation of $q(\lambda,k)$, however, was controlled by the minima values of a smoothed power spectrum of the noisy signal. Hence, the computation of $p(\lambda,k)$ was influenced by the minima tracking. Consequently, the update of the noise estimate is influenced in both MCRA and IMCRA methods by the minima tracking, which may lag by as many as 2L frames.

Unlike the methods in (Cohen, 2002; Cohen, 2003), the estimate of the noise spectrum in the proposed method is not influenced by the minimum-search window. Also, the threshold used in our method for identifying speech presence/absence regions is frequency dependent while that of Cohen (2002) is fixed for all frequencies.

4. Experimental results

The proposed noise-estimation algorithm was combined with a Wiener-type speech-enhancement algorithm (Hu and Loizou, 2004) with the following spectral gain function:

$$G(\lambda,k) = \frac{C(\lambda,k)}{C(\lambda,k) + \mu_k D(\lambda,k)} \tag{9}$$

where $C(\lambda,k)$ is the estimated clean speech spectrum computed from the noisy speech and noise estimates as follows:

$$C(\lambda,k) = \max\{|Y(\lambda,k)|^2 - D(\lambda,k), vD(\lambda,k)\} \tag{10}$$

where $v = 0.001$ is a small positive number. The max(·) operation is used to ensure positive values for the estimated clean speech spectra. The over subtraction factor μ_k in Eq. (9) is determined from the a posteriori segmental SNR as per Hu and Loizou (2004).

The performance of the proposed method was evaluated using both subjective and objective measures. The following values were used in the implementation (assuming a sampling frequency of 20.1 kHz): $\alpha_d = 0.85$, $\alpha_p = 0.2$, $\beta = 0.8$, $\gamma = 0.998$, $\eta = 0.7$ and

$$\delta(k) = \begin{cases} 2 & 1 \leqslant k \leqslant LF \\ 2 & LF < k \leqslant MF \\ 5 & MF < k \leqslant Fs/2 \end{cases}$$

where LF and MF are the bins corresponding to 1 and 3 kHz respectively, and Fs is the sampling frequency.

4.1. Subjective evaluation

The performance of the proposed method was compared with that of the methods in (Martin, 2001; Cohen, 2003; Doblinger, 1995; Hirsch and Ehrlicher, 1995) using formal listening tests. The listening test included two different noise types, namely single noise and triplet noise. In the single noise case, sentences were degraded by either multi-talker babble noise (two male and two female speakers) or factory noise. In the triplet noise case, three different noise signals were concatenated to evaluate the adaptation of the algorithm for different noise types. The three different noise types included multi-talker babble, factory noise and white noise. Thus a noisy (triplet) set of stimuli consisted of a sentence degraded by babble noise followed by a sentence degraded by factory noise and a sentence degraded by white noise without any pauses in the middle. The overall SNR of the noisy speech was 5 dB for both cases. The sentences were taken from the HINT (Nilsson et al., 1994) database.

The quality of speech enhanced by the proposed noise-estimation algorithm was compared against the quality of speech produced by four other noise-estimation algorithms. The same speech-enhancement algorithm (Hu and Loizou, 2004) was used with all noise-estimation algorithms. For the single noise case, 40 sentences were used (20 sentences corrupted by babble noise and 20 sentences corrupted by factory noise) and for the triplet noise case, 20 sets of triplet sentences were used and degraded by the triplet noise for each comparison. The listeners were presented with pairs of sentences, one processed with our proposed method and the one processed with one of the other methods (Martin, 2001; Cohen, 2003; Doblinger, 1995; Hirsch and Ehrlicher, 1995). The order of the sentences was randomized. The listeners were asked to select from the pair of stimuli presented the sentence which was more natural, easier to listen and free of artifacts. The overall preference was assessed for speech enhanced by the proposed method compared to the other methods. The preference score (relative number of times—out of 20 pairs of sentences presented—that listeners preferred the proposed method over the other methods) was averaged over six normal-hearing listeners. A preference score of 100%, for instance, would indicate that listeners preferred the proposed method over the other methods all the time.

Table 1 shows the preference results. From the results, it can be seen that our proposed method had equal preference compared with the other methods in (Martin, 2001; Cohen, 2003; Doblinger, 1995; Hirsch and Ehrlicher, 1995) for the single noise case. But, for the triplet noise case, the proposed method had higher preference scores compared to all the other methods. We suspect that this was due to the fact that our noise-estimation algorithm adapts quickly to the highly non-stationary environments.

Table 1
Percent preference for the proposed method compared to other methods for single and mixed type noise

Method	Single noise Preference (%)	Mixed noise Preference (%)
Cohen (2003)	48.8	81.7
Doblinger (1995)	53.8	81.3
Hirsch and Ehrlicher (1995)	50.0	78.8
Martin (2001)	50.8	63.8

4.2. Objective evaluation

We computed the relative mean squared error between the true noise spectrum and the estimated noise spectrum as follows:

$$\mathrm{MSE} = \frac{1}{M}\sum_{\lambda=0}^{M-1}\frac{\sum_k[D(\lambda,k)-\sigma_D^2(\lambda,k)]^2}{\sum_k\sigma_D^2(\lambda,k)} \tag{11}$$

where $D(\lambda,k)$ is the estimated noise power spectrum (as per Eq. (8)), $\sigma_D^2(\lambda,k)$ is the true noise power spectrum, and M is the total number of frames in the noisy speech.

Two additional objective measures were used to evaluate and compare the performance of the proposed noise-estimation algorithm: the segmental SNR and the log-likelihood ratio (LLR) measure (Quackenbush et al., 1988). The LLR measure for each 20-ms speech frame was computed as follows:

$$d_{\mathrm{LLR}} = \log_{10}\left(\frac{a_yR_xa_y^T}{a_xR_xa_x^T}\right) \tag{12}$$

where a_x and R_x are the linear prediction coefficient vector and autocorrelation matrix of the original (clean) speech frame respectively, and a_y is the linear prediction coefficient vector of the enhanced speech frame. The LLR is a spectral distance measure which mainly models the mismatch between the formants of the original and enhanced signals (Quackenbush et al., 1988). The mean LLR value was obtained by averaging the individual frame LLR values across the sentence.

The MSE results are tabulated in Table 2 and the segmental SNR and LLR results are tabulated in Table 3. Overall, the MSE results are not consistent with the preference outcomes, in that lower MSE values did not suggest better preference. This indicates that the MSE measure might not be a reliable measure for assessing performance of noise-estimation algorithms. For one, this measure is sensitive to outlier values. Secondly, it treats noise overestimation and noise underestimation errors the same. Unlike the MSE values, the segmental SNR values and the LLR values shown in Table 3 were found to be more consistent with the subjective evaluation results. Relatively larger segmental SNR values were obtained with Martin's (2001) method and our proposed method compared to the other methods. Smaller spectral distance values (LLR) were also obtained by our proposed method and that of

Table 2
The normalized mean squared error (MSE) between the estimated and true noise spectra for various methods

Methods	MSE Single noise	MSE Mixed noise
Cohen (2003)	0.40	0.86
Doblinger (1995)	0.52	1.08
Hirsch and Ehrlicher (1995)	0.52	0.87
Martin (2001)	0.53	0.94
Quantile ($q = 0.25$) (Stahl et al., 2000)	0.50	4.95
Proposed method	0.43	0.87

Table 3
Objective evaluation and comparison of the proposed noise-estimation algorithm in terms of segmental SNR values (dB) and LLR values

Method	Single noise		Mixed noise	
	SNRseg	LLR	SNRseg	LLR
Cohen (2003)	7.17	2.90	7.05	5.13
Doblinger (1995)	6.98	1.89	7.48	2.57
Hirsch and Ehrlicher (1995)	7.04	2.10	7.37	3.35
Martin (2001)	8.59	1.94	8.12	2.25
Quantile (Stahl et al., 2000)	7.84	1.96	8.06	2.49
Proposed method	7.43	1.97	8.13	2.15

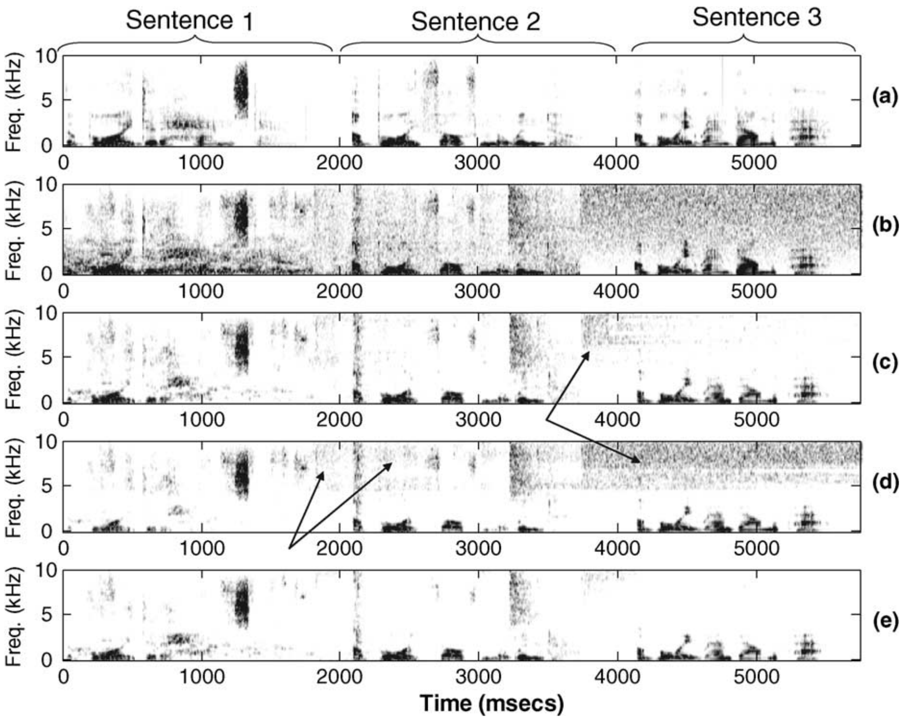

Fig. 8. Spectrograms of speech enhanced using Martin's (2001) noise-estimation method (panel c), Cohen's method (Cohen (2003)) (panel d) and the proposed noise-estimation method (panel e). Spectrograms of the clean and noisy speech signals are given in panels (a) and (b) respectively. Arrows in panels (c) and (d) at $t > 3.8$ s show the presence of residual noise due partly to the inability of the noise-estimation algorithms to track the sudden appearance of high-frequency noise in the last sentence (sentence 3). In contrast, as shown in panel (e), the residual noise is greatly reduced with the proposed noise-estimation algorithm.

Martin (2001) compared to the other methods. Large MSE values were obtained with the quantile method ($q = 0.25$[1]) for the triplet sentences. This is attributed to the fact that three types of noise with different characteristics, and possibly three different quantile values of the noisy speech power spectrum, were used in the triplet sentences Subsequent simulations confirmed that if the q-th quantile of the noisy speech spectrum was estimated separately for each type of noise and each sentence, the MSE value reduces significantly. The objective measures shown in Table 3 suggest that Martin's noise-estimation method (Martin, 2001) performed better than Cohen's method (Cohen, 2003). This outcome is consistent with our listening tests (see Table 1) and is also confirmed by visual inspection of spectrograms of speech enhanced by the various methods (see Fig. 8). A different outcome was observed in (Cohen, 2003), and this could be attributed to several reasons: (a) difference in speech materials and type of noise used and (b) difference in the way non-stationary noise was modeled. In (Cohen, 2003), non-stationary noise was modeled by increasing the level of WGN by 2 dB/s. The change in noise level in our triplet sentences was more abrupt than 2 dB/s (see Fig. 8(b)). Furthermore, only objective measures were reported in (Cohen, 2003) and no subjective listening tests were performed.

5. Summary and conclusions

In this paper we have addressed the issue of noise estimation for enhancement of noisy speech. The noise estimate was updated continuously in every frame using time–frequency smoothing factors calculated based on speech-presence probability in each frequency bin of the noisy speech spectrum. The speech-presence probability was estimated using the ratio of noisy speech power spectrum to its local minimum. Unlike other methods, the update of local minimum was continuous over time and did not depend on some fixed window length. Hence the update of noise estimate was faster for very rapidly varying non-stationary noise environments. This was confirmed by formal listening tests that indicated significantly higher preference for our proposed algorithm compared to the other existing noise-estimation algorithms.

Acknowledgements

The authors would like to thank the anonymous reviewers for all their comments and suggestions. This research was supported by Grant No. R01 DC03421 from NIDCD/NIH.

References

Afify, M., Sioham, O., 2001. Sequential noise estimation with optimal forgetting for robust speech recognition. Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process. 1, 229–232.

Cohen, I., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. 9 (1), 12–15.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11 (5), 466–475.

Deng, L., Droppo, J., Acero, A., 2003. Recursive estimation of non-stationary noise using iterative stochastic approximation for robust speech recognition. IEEE Trans. Speech Audio Process. 11 (6), 568–580.

Deng, L., Droppo, J., Acero, A., 2003. Incremental Bayes learning with prior evolution for tracking nonstationary noise statistics from noisy speech data. Proc. IEEE Internat. on Conf. Acoust. Speech, Signal Process. I, 672–675.

Doblinger, G., 1995. Computationally efficient speech enhancement by spectral minima tracking in subbands. Proc. Eurospeech 2, 1513–1516.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP 32 (6), 1109–1121.

Ephraim, Y., Van Trees, H.L., 1993. A signal subspace approach for speech enhancement. Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Process. II, 355–358.

Hirsch, H., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., 153–156.

Hu, Y., Loizou, P., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Process. 12 (1), 59–67.

Kim, N., 1998. Nonstationary environment compensation based on sequential estimation. IEEE Signal Process. Lett. 5 (3), 57–59.

Lim, J., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. ASSP 26 (3), 197–210.

Lin, L., Holmes, W., Ambikairajah, E., 2003. Subband noise estimation for speech enhancement using a perceptual Wiener filter. Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process. I, 80–83.

Malah, D., Cox, R., Accardi, A., 1999. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary environments. Proc. IEEE Internat. on Conf. Acoust. Speech Signal Process., 789–792.

[1] Better performance, in terms of smaller MSE value, was obtained using $q = 0.25$ than with $q = 0.5$ for the triplet sentences. The qth quantile value was estimated over the whole duration of the sentences as per Stahl et al. (2000).

Martin, R., 1994. Spectral subtraction based on minimum statistics. Proc. Eur. Signal Process., 1182–1185.

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

Nilsson, M., Soli, S., Sullivan, J., 1994. Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Amer. 95 (2), 1085–1099.

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

Rangachari, S., Loizou, P., Hu, Y., 2004. A noise estimation algorithm with rapid adaptation for highly nonstationary environments. Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process. I, 305–308.

Ris, C., Dupont, S., 2001. Assessing local noise level estimation methods: application to noise robust ASR. Speech Comm. 34, 141–158.

Sohn, J., Kim, N., 1999. Statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1), 1–3.

Stahl, V., Fischer, A., Bippus, R., 2000. Quantile based noise estimation for spectral subtraction and Wiener filtering. Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process., 1873–1875.

Yao, K., Nakamura, S., 2002. Sequential noise compensation by sequential Monte Carlo method. Adv. Neural Inform. Process. Systems 14, 1213–1220.