A NOISE ESTIMATION ALGORITHM WITH RAPID ADAPTATION FOR HIGHLY NON-STATIONARY ENVIRONMENTS

Sundarrajan Rangachari, Philipos C. Loizou and Yi Hu
Dept. of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083
rsrajan@student.utdallas.edu, {loizou,yihuyxy}@utdallas.edu

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 a time and frequency dependent smoothing factor, which is adjusted based on signal presence probability in subbands. Signal presence is determined by computing the ratio of the noisy speech power spectrum to its local minimum, which is computed 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 that indicated that our noise estimation algorithm when integrated in speech enhancement was preferred over other noise estimation algorithms.

1. INTRODUCTION

An important aspect of speech enhancement is noise spectrum estimation. For stationary noise, averaging the spectrum of the noisy signal during the initial silence period can often be sufficient. That is not the case, however, with non-stationary noise since the noise spectrum will be varying rapidly over time. To overcome this problem, the noise spectrum needs to be estimated and updated continuously. This is a challenging task since we only have access to the noisy speech signal. Noise estimation algorithms are therefore needed which can track the noise without explicitly doing speech/silence detection.

Several algorithms have been proposed for estimating the noise spectrum based on the noisy speech signal. Martin [2] proposed a method which is based on finding the minimum of the noisy speech over a window. This method takes slightly more than the window duration to update the noise spectrum when the noise floor increases abruptly [2]. Doblinger's method [1] updates the noise estimate continuously and is computationally more efficient. However, it fails to differentiate between an increase in noise floor and an increase in the speech spectrum level. Hirsch and Ehrlicher [3] update the noise estimate by comparing the noisy speech power spectrum to the past noise estimate. The major drawback of their method is that it fails to update the noise estimate when the noise floor increases abruptly and stays at that level. Similar problems are encountered with the method proposed in [4]. Cohen and Berdugo [5] proposed an algorithm which tracks the noise-only regions by finding the ratio of the noisy speech to the local minimum over a period of 0.5-1.5 sec. The noise estimate, however, lags by at most twice that period when the noise spectrum increases abruptly.

Most of the above noise estimation methods are slow in adapting to increasing levels of noise. In this paper, we propose a method which tracks the noise spectrum quickly, even when the noise levels suddenly increase. It uses a fast method for tracking the minimum of the noisy speech power spectrum, and also makes use of the fact that speech may be absent in some frequencies even in speech-present frames. The advantage of this method over other methods is that it updates the noise spectrum faster since the minimum tracking is not constrained by a specified time window and also does not overestimate the noise spectrum.

This paper is organized as follows. Section 2 gives the analysis of different noise estimation algorithms is some detail. Section 3 presents the proposed method and Section 4 presents our experimental results.

2. ANALYSIS OF EXISTING NOISE ESTIMATION METHODS

Martin's minimum statistics method [2] is based on the observation that the power level of the noisy speech signal often decays to the power level of the noise. Hence by tracking the minimum of the noisy speech spectrum, we can get an estimate of the noise spectrum. The spectral minima in each frequency bin is sought over a window of approximately 1.5 secs, and then compared with the power spectrum of the noisy speech after bias compensation. The local minimum is updated whenever the power spectrum of the noisy speech is smaller than the local minimum with some bias compensation. To make the adaptation faster, the window is subdivided into smaller windows and the noise estimate is updated in every sub-window. In spite of that, the algorithm lags behind particularly when there is sudden rise in the noise power. It takes slightly more than the window length (1.5 secs) to track the new noise floor [2]. Reducing the

window duration within which the minimum is sought may result in speech distortion particularly if the speech extends continuously for more than the window duration.

Cohen and Berdugo [5] introduced a minima controlled recursive averaging approach for noise estimation. The noise estimate is updated continuously by averaging the past spectral values of the noisy speech with time and frequency-dependent smoothing factors. The smoothing factors are controlled by the ratio of the noisy speech power spectrum to its local minimum found over a period of 0.5-1.5 secs. The method is based on the principle that if the estimated ratio is less than some fixed threshold, then it is taken as noise-only region and the noise estimate is updated accordingly. This method also suffers from the similar problem as [2] when there is a sudden increase in noise power. It takes at most 3 secs to update the local minimum.

Doblinger [1] proposed an efficient method for tracking the noise spectrum that was not constrained by any window length for updating the estimate of the noise spectrum. In his algorithm, the noise estimate is updated by averaging the past values of noisy speech power spectrum in a way that it tracks the minimum of the noisy speech in each frequency band. The adaptation period is around 0.2-0.4 secs. The major drawback of this algorithm, however, is that the noise estimate increases whenever the noisy speech power increases. This overestimation of the noise power level can produce speech distortion.

The method proposed by Hirsch and Ehrlicher [3], is based on estimating a histogram of past spectral values, which are compared against a threshold. The histogram is found over the past 400ms of noise segments and the maximum in each band is taken as the noise spectrum. Since the threshold is based on the past noise estimate, the algorithm fails to adapt when the noise estimate suddenly increases. In that case, the threshold may always be less than the current noisy speech power and hence the noise spectrum estimate might never be updated.

In summary, the major drawback of most noise estimation algorithms is that they are either slow in tracking sudden increases of noise power or that they are overestimating the noise energy resulting in speech distortion. To overcome these drawbacks, we propose a method, which updates the noise spectrum quickly without overestimating the noise spectrum.

3. PROPOSED NOISE ESTIMATION ALGORITHM

Let y(n)=x(n)+d(n), where y(n) is the noisy speech signal, x(n) is the clean signal and d(n) is the additive noise. The smoothed power spectrum of the noisy speech signal can be estimated using a first-order recursive formula as follows:

$$P(\lambda,k) = \eta P(\lambda-1,k) \; + \; (1-\eta)\left|Y(\lambda,k)\right|^2 \qquad (1)$$

where $|Y(\lambda,k)|^2$ is an estimate of the short-time power spectrum of y(n) obtained by wavelet-thresholding the multitaper spectrum of y(n) [6], η is a smoothing constant, λ is the frame index and k is the frequency bin index.

Since the noisy speech power spectrum in the speech-absent frames is equal to the power spectrum of the noise, we can update the estimate of the noise spectrum by tracking the speech-absent frames. To do that, we compute the ratio of the energy of the noisy speech power spectrum in three different frequency bands (low: 0-1 kHz, middle: 1-3 kHz, high: 3 kHz and above) to the energy of the corresponding frequency band in the previous noise estimate. The following three ratios are computed:

$$\xi_L(\lambda) = \frac{\sum_{k=1}^{LF} P(\lambda,k)}{\sum_{k=1}^{LF} N(\lambda-1,k)}, \quad \xi_M(\lambda) = \frac{\sum_{k=LF+1}^{MF} P(\lambda,k)}{\sum_{k=LF+1}^{MF} N(\lambda-1,k)}$$
$$\xi_H(\lambda) = \frac{\sum_{k=MF+1}^{Fs/2} P(\lambda,k)}{\sum_{k=MF+1}^{Fs/2} N(\lambda-1,k)} \qquad (2)$$

where N(λ,k) is the estimate of the noise power spectrum at frame λ, and LF, MF, Fs correspond to the frequency bins of 1 kHz, 3 kHz and the sampling frequency respectively. If the above three ratios ($\xi_L(\lambda), \xi_M(\lambda), \xi_H(\lambda)$) are all smaller than a threshold σ, then it is concluded that it is a speech-absent frame and the noise estimate is updated according to:

$$N(\lambda,k) = \varepsilon N(\lambda-1,k) + (1-\varepsilon)\left|Y(\lambda,k)\right|^2 \qquad (3)$$

where ε is a smoothing constant. If any or all of the above three ratios are larger than the threshold σ, then a different algorithm is used for updating and estimating the noise spectrum.

The proposed algorithm used for speech-present segments is based on first finding the minimum of the noisy speech spectrum, and using that minimum to determine signal presence probability in subbands. The signal presence probability is used to determine a frequency-dependent smoothing parameter which replaces the fixed smoothing constant ε in Eq. (3).

The local minimum of the noisy speech is computed by averaging the past spectral values with a look-ahead factor as defined in [1]:

if $P_{min}(\lambda-1,k) < P(\lambda,k)$

then

$$P_{min}(\lambda,k) = \gamma P_{min}(\lambda-1,k) + \frac{1-\gamma}{1-\beta}(P(\lambda,k) - \beta P(\lambda-1,k)) \qquad (4)$$

else $P_{min}(\lambda,k) = P(\lambda,k)$

where $P_{min}(\lambda,k)$ denotes the local minimum of the noisy speech power spectrum and β and γ are constants determined experimentally. This method for computing the local minimum is effective since it adapts to sudden increases in noise power within 0.4-0.6 secs, which is considerably fast compared to the adaptation period of the local minimum in [2] and [5].

The approach taken to determine signal presence probability in subbands is similar to that proposed in [5]. Let $S_r(\lambda,k) \triangleq P(\lambda,k)/P_{min}(\lambda,k)$ denote the ratio between the energy of the noisy speech to its local minimum. This ratio is compared against a frequency-dependent threshold and if it is found to be larger than that threshold, then the corresponding frequency is considered to contain speech. Note that the aforementioned problem with [1] is avoided with the use of this ratio $S_r(\lambda,k)$. Figure 1 shows an example of speech-presence determination. Using the above ratio $S_r(\lambda,k)$, the new frequency-dependent smoothing constant can be estimated as follows:

$$\alpha_s(\lambda,k) = \begin{cases} \alpha_1 & \text{if } S_r(\lambda,k) < \delta(k) \\ \alpha_2 & \text{otherwise} \end{cases} \quad (5)$$

where α_1, α_2 are smoothing constants ($\alpha_2 > \alpha_1$) and $\delta(k)$ is a frequency-dependent threshold given by:

$$\delta(k) = \begin{cases} 1.3 & 1 \le k \le LF \\ 3 & LF < k \le MF \\ 5 & MF < k \le Fs/2 \end{cases} \quad (6)$$

Finally, after computing the frequency-depending smoothing factor $\alpha_s(\lambda,k)$, the noise spectrum estimate is updated according to:

$$N(\lambda,k) = \alpha_s(\lambda,k)N(\lambda-1,k) + (1-\alpha_s(\lambda,k))|Y(\lambda,k)|^2 \quad (7)$$

To summarize, if the ratios defined in Eq. 2 indicate that the current frame is a speech-absent frame, then Eq. 3 is used to update the noise spectrum. Otherwise, Eq. 7 is used to update the noise spectrum.

Note that there are two major differences between our method and that in [5] for estimating the smoothing parameter $\alpha_s(\lambda,k)$. First, in our method the threshold $\delta(k)$ used to determine signal presence is frequency dependent, while in [5], it was fixed. Second, we use a different method for obtaining the minimum of the noisy speech spectrum that adapts faster than the method used in [5].

Figure 2 shows an example noise spectrum estimated with our algorithm and with Martin's algorithm [2] for a scenario in which the noise level suddenly increases. Our algorithm is able to adapt to the new environment within 0.6 secs, while Martin's algorithm required 1.5 secs to adapt.

4. EXPERIMENTAL RESULTS

The proposed noise spectrum estimation method was combined with a Wiener-type speech enhancement algorithm [6], which had a spectral gain function of the form:

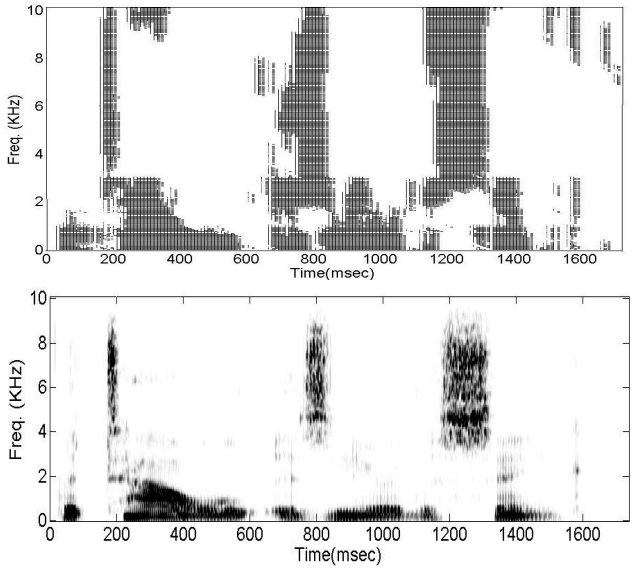

Fig 1: Top panel: Plot of estimated speech presence probability based on the ratio Sr(λ,k). Bottom panel: spectrogram of the clean signal.

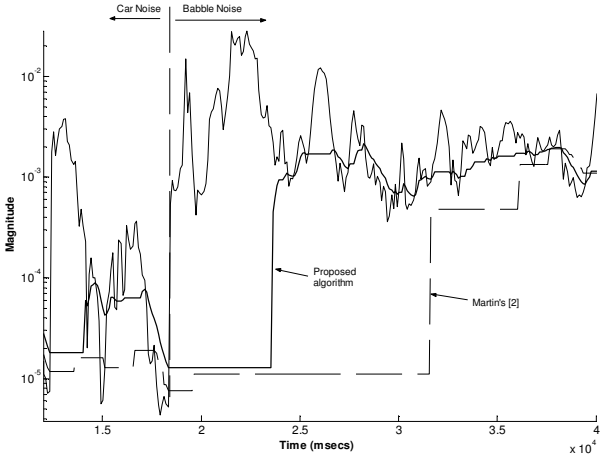

Fig. 2. Comparison between the noise spectrum (for f=1.5 kHz) estimated using the proposed algorithm (thick line) and Martin's [2] (dashed line) algorithm for a sentence corrupted by car noise (t <1.8 s) followed by a sentence corrupted by multi-talker babble (t>1.8 s).

$$G(\lambda,k) = \frac{C(\lambda,k)}{C(\lambda,k) + \mu_k N(\lambda,k)} \quad (8)$$

where $C(\lambda,k)$ is the estimated clean speech spectrum based on wavelet-thresholding the multitaper spectra of the noisy speech signal, and μ_k is a factor which is dependent on the *a posteriori* segmental SNR [6]. Speech was segmented into 20-ms frames using a Hamming window with 50% overlap. The following parameters were used in the implementation of the noise estimation algorithm: $\alpha_1 = 0.8$, $\alpha_2 = 1$, $\beta = 0.8$, $\gamma = 0.998$, $\eta = 0.7$, $\sigma = 1.3$ and $\varepsilon = 0.8$.

The performance of the proposed method was compared with that of the other methods described in [1-3,5]. The proposed noise estimation algorithm was evaluated using formal listening tests and objective measures. To evaluate the effect of adaptation speed on speech quality, we used sets of sentences composed of a sequence of three (concatenated) sentences, each corrupted by a different type of noise. The three types of noise included multi-talker (2 female and 2 male) babble, factory noise and white noise. So, a typical triplet sentence set included a sentence corrupted by multi-talker babble followed by a sentence corrupted by factory noise, and followed by a sentence corrupted by white noise, with no gap between sentences. These sets of sentences reflect a scenario in which the environment is rapidly changing, thereby requiring a noise estimation algorithm with fast adaptation. Twenty sets of triplet sentences taken from the HINT [7] database were used in the listening tests. In addition, 40 single sentences (20 corrupted by multi-talker babble, and 20 corrupted by factory noise) were used for testing. The overall SNR of the corrupted sentences was 5 dB for both conditions.

The quality of the enhanced speech obtained using the proposed noise estimation algorithm was compared against the quality of speech produced by four other noise estimation algorithms using a paired-preference paradigm. The same speech enhancement algorithm [6] was used in all conditions. Six normal-hearing listeners participated in the paired-preference tests (all were native speakers of American English). The listeners were presented via headphones pairs of sentences (single or triplet sets) processed via two different noise estimation algorithms and asked to indicate their preference in terms of having better speech quality and least amount of distortion. The order of the sentences in each comparison was randomized for all listeners. Table 1 shows the percentage of time listeners preferred the proposed noise estimation method over the other methods. Table 1 also shows the normalized mean squared error (MSE) between the estimated and true noise spectra.

Results indicated that for the single sentences corrupted by either babble or factory noise, the performance of the proposed algorithm was comparable to that of other algorithms. But for the triplet sentences in which the noise type was changing, the proposed method was preferred over all the other methods. We attribute this outcome to the quicker adaptation of the proposed method compared to the other methods.

Method	Single Noise		Mixed Noise	
	Preference	MSE	Preference	MSE
Cohen and Berdugo [5]	60.8%	0.95	80.4%	1.12
Doblinger [1]	40.6%	0.52	82.2%	1.08
Hirsch and Ehrlicher [3]	47.8%	0.52	87.1%	0.87
Martin [2]	55.0%	0.53	58.8%	0.94
Proposed	-	0.54	-	0.75

Table 1: Percentage of preference for the proposed method compared to other methods for single and mixed type noise. The normalized mean squared error (MSE) between the estimated and true noise spectra is also given.

5. CONCLUSION

In this paper, we presented a fast noise estimation algorithm, which is well suited for rapidly varying noise environments. The noise estimate was found by averaging past spectral power values using a smoothing parameter that was adjusted by the signal presence probability in subbands. Unlike other methods, the adaptation of this frequency-dependent smoothing parameter did not depend on a specified time window and was therefore fast. This was substantiated by our formal listening tests results which showed preference for our method compared to other methods for estimating the noise spectrum.

6. REFERENCES

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

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

[3] Hirsch, H. G. and Ehrlicher, C., "Noise estimation techniques for robust speech recognition," in *Proc. 20th IEEE Int. Conf. Acoustics, Speech, Signal Proc.,* Detroit, MI, May 8-12, 1995, pp. 153-156.

[4] Lin, L., Holmes, W. H. and Ambikairajah, E., "Subband noise estimation for speech enhancement using a perceptual wiener filter," in *Proc. Int. Conf. Acoustics, Speech, Signal Proc.,* Hong Kong, April 6-10, 2003, pp. I_80-I_83.

[5] Cohen, I. and Berdugo, B., "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Proc. Letters*, vol. 9, no. 1, pp. 12-15, January 2002.

[6] Hu, Y. and Loizou, P., "Speech enhancement based on wavelet thresholding and multitaper spectrum," *IEEE Trans. on Speech and Audio Processing, to appear in*, Jan 2004.

[7] Nilsson, M., Soli, S. and Sullivan, J., "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *Journal of Acoustic Society of America*, vol. 95, pp. 1085-1099, 1994.