Improvement of Minimum Tracking in Minimum Statistics Noise Estimation Method

Hassan Farsi hfarsi@birjand.ac.ir
Department of Electronics and Communications Engineering,
University of Birjand,
Birjand, IRAN.

Abstract

Noise spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. In this paper we propose a new method for minimum tracking in Minimum Statistics (MS) noise estimation method. This noise estimation algorithm is proposed for 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.

Keywords: Speech enhancement, Statistics noise, noise cancellation, Short time Fourier transform

1. INTRODUCTION

Noise spectrum estimation is a fundamental component of speech enhancement and speech recognition systems. The robustness of such systems, particularly under low signal-to-noise ratio (SNR) conditions and non-stationary noise environments, is greatly affected by the capability to reliably track fast variations in the statistics of the noise. Traditional noise estimation methods, which are based on voice activity detectors (VAD's), restrict the update of the estimate to periods of speech absence.

Additionally, VAD's are generally difficult to tune and their reliability severely deteriorates for weak speech components and low input SNR [1], [2], [3]. Alternative techniques, based on histograms in the power spectral domain [4], [5], [6], are computationally expensive, require much memory resources, and do not perform well in low SNR conditions. Furthermore, the signal segments used for building the histograms are typically of several hundred milliseconds, and thus the update rate of the noise estimate is essentially moderate.

Martin (2001)[7] 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. Moreover, this method may occasionally attenuate low energy phonemes, particularly if the minimum search window is too short [8]. These limitations can be overcome, at the price of significantly higher complexity, by adapting the smoothing parameter and the bias compensation factor in time and frequency [9]. A computationally more efficient minimum tracking scheme is presented in [10]. Its main drawbacks are the very slow update rate of the noise estimate in case of a sudden rise in the noise energy level, and its tendency to cancel the signal [1].In this paper we propose a new approach for minimum tracking , resulted improving the performance of MS method.

The paper is organized as follows. In Section II, we present the MS noise estimator. In Section III, we introduce an method for minimum tracking, and in section IV, evaluate the proposed method, and discuss experimental results, which validate its effectiveness.

2. MINIMUM STATISTICS NOISE ESTIMATOR

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,

$$Y(k,l) = \sum_{n=0}^{N-1} y(n+lM)h(n)e^{-j\left(\frac{2\pi}{N}\right)nk} \quad (1)$$

Where k is the frequency bin index, l is the time frame index, h is an analysis window of size N (e.g., Hamming window), and M is the framing step (number of samples separating two successive frames). Let $X(k,l)$ and $D(k,l)$ denote the STFT of the clean speech and noise, respectively.
For noise estimation in MS method, first compute the short time subband signal power $\lambda_y(k,l)$ using recursively smoothed periodograms. The update recursion is given by eq.(2). The smoothing constant is typically set to values between $\alpha = 0.9, \ldots, 0.95$.

$$\lambda_y(k,l) = \alpha\lambda_y(k,l-1) + (1-\alpha)|Y(k,l)|^2 \tag{2}$$

The noise power estimate $\lambda_d(k,l)$ is obtained as a weighted minimum of the short time power estimate $\lambda_y(k,l)$ within window of D subband power samples [11], i.e.

$$\bar{\lambda}_d(k,l) = B_{min}\lambda_{min}(k,l) \tag{3}$$

$\lambda_{min}(k,l)$ is the estimated minimum power and B_{min} is a factor to compensate the bias of the minimum estimate. The bias compensation factor depends only on known algorithmic parameters [7].
For reasons of computational complexity and delay the data window of length D is decomposed into U sub-windows of length V such that For a sampling rate of fs=8 kHz and a framing step M=64 typical window parameters are V=25 and U=4,thus D=100 corresponding to a time window of ((D-1).M+N)/f_s=0.824s. Whenever V samples are read, the minimum of the current sub-window is determined and stored for later use. The overall minimum is obtained as the minimum of past samples within the current sub-window and the U previous sub-window minima.
In [7] shown that the bias of the minimum subband power estimate is proportional to the noise power $\sigma^2(k)$ and that the bias can be compensated by multiplying the minimum estimate with the inverse of the mean computed for $\sigma^2(k) = 1$.

$$B_{min} = \frac{1}{E\{\lambda_{min}(k,l)\}|_{\sigma^2(k)=1}} \tag{4}$$

Therefore to obtain B_{min} We must generate data of variance $\sigma^2(k) = 1$, compute the smoothed periodogram (eq. (2)), and evaluate the mean and the variance of the minimum estimate.
As discussed earlier, minimum of the smoothed periodograms, obtained within window of D subband power samples. In next section we propose a method to improve this minimum tracking.

3. PROPOSED METHOD FOR MINIMUM TRACKING

The local minimum in MS method was found by tracking the minimum of noisy speech over a search window spanning D frames. Therefore, the noise update was dependent on the length of the minimum-search window. The update of minimum can take at most 2D frames for increasing noise levels. A different non-linear rule is used in our method for tracking the minimum of the noisy speech by continuously averaging past spectral values [12]

$$\begin{aligned}
&if \quad S_{min}(k,l-1) < S(k,l)\\
&S_{min}(k,l) = \alpha\ S_{min}(k,l-1)\\
&\qquad + \frac{1-\alpha}{1-\beta}(S(k,l) - \beta S(k,l-1))\\
&else\\
&S_{min}(k,l) = \alpha S_{min}(k,l-1)\\
&\qquad - \gamma(S_{min}(k,l-1) - \lambda S(k,l))\\
&end
\end{aligned} \tag{5}$$

where $S_{min}(k,l)$ is the local minimum of the noisy speech power spectrum and α, β, γ and λ are constants which are determined experimentally. The lookahead factor β controls the adaptation time of the local minimum. Typically, we use $\alpha = 0.998$, $\beta = 0.8$, $\gamma = 0.01$ and $\lambda = 0.9$. Because Improve the minimum tracking in this method, the bias compensation factor decreases, as in MS method it is obtained $B_{min} = 1.5$ and in this method it is obtained $B_{min} = 1.2$.

4. PERFORMANCE EVALUATION

The performance evaluation of the proposed method (PM), and a comparison to the MS method, consists of three parts. First, we test the tracking capability of the noise estimators for non-stationary noise. Second, we measure the segmental relative estimation error for various noise types and levels. Third, we integrate the noise estimators into a speech enhancement system, and determine the improvement in the segmental SNR. The results are conformed by a subjective study of speech spectrograms and informal listening tests.
The noise signals used in our evaluation are taken from the Noisex92 database [13]. They include white Gaussian noise (WGN), F16 cockpit noise, and babble noise. The speech signal is sampled at 8 kHz and degraded by the various noise types with segmental SNR's in the range [-5, 10] dB. The segmental SNR is defined by [14]

$$SegSNR = \frac{10}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} log\frac{\sum_k |X(k,l)|^2}{\sum_k |D(k,l)|^2} \quad (6)$$

where $\mathcal{L}$ represents the set of frames that contain speech,
and $|\mathcal{L}|$ its cardinality. The spectral analysis is implemented with Hamming windows of 256 samples length (32ms) and 64 samples frame update step.
Fig. 1 plots the ideal (True), PM, and MS noise estimates for a babble noise at 0 dB segmental SNR, and a single frequency bin k = 5 (the ideal noise estimate is taken as the recursively smoothed periodogram of the noise $|D(k,l)|^2$, with a smoothing parameter set to 0.95). Clearly, the PM noise estimate follows the noise power more closely than the MS noise estimate. The update rate of the MS noise estimate is inherently restricted by the size of the minimum search window (D). By contrast, the PM noise estimate is continuously updated even during speech activity.
Fig. 2 shows another example of the improved tracking capability of the PM estimator. In this case, the speech signal is degraded by babble noise at 5 dB segmental SNR. The ideal, PM, and MS noise estimates, averaged out over the frequency, are depicted in this figure.
A quantitative comparison between the PM and MS estimation methods is obtained by evaluating the segmental relative estimation error in various environmental conditions. The segmental relative estimation error is defined by [15]

$$SegEr = \frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_k [\hat{\lambda}_d(k,l) - \lambda_d(k,l)]^2}{\sum_k \lambda_d^2(k,l)} \quad (7)$$

where $\lambda_d(k,l)$ is the ideal noise estimate, $\hat{\lambda}_d(k,l)$ is the noise estimated by the tested method, and L is the number of frames in the analyzed signal. Table 1 presents the results of the segmental relative estimation error achieved by the PM and MS estimators for various noise types and levels. It shows that the PM method obtains significantly lower estimation error than the MS method.
The segmental relative estimation error is a measure that weighs all frames in a uniform manner, without a distinction between speech presence and absence. In practice, the estimation error is more consequential in frames that contain speech, particularly weak speech components, than in frames that contain only noise. We therefore examine the performance of our estimation method when integrated into a speech enhancement system. Specifically, the PM and MS noise estimators are combined with the Optimally-Modified Log-Spectral Amplitude (OM-LSA) estimator, and evaluated both objectively using an improvement in segmental SNR measure, and subjectively by informal listening tests. The OM-LSA estimator [16], [17] is a modified version of the conventional LSA estimator [18-19], based on a binary hypothesis model. The modification includes a lower bound for the gain, which is determined by a subjective criterion for the noise naturalness, and exponential weights, which are given by the conditional speech presence probability [20, 21].

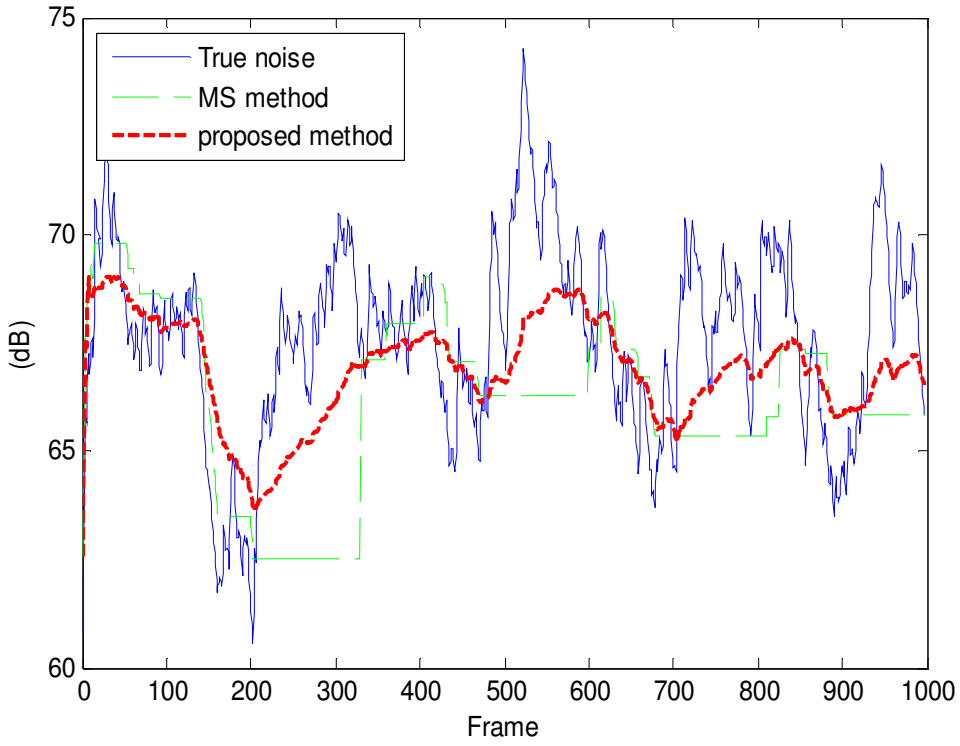

FIGURE 1. Plot of true noise spectrum and estimated noise spectrum using proposed method and MS method for a noisy speech signal degraded by babble noise at 0 dB segmental SNR, and a single frequency bin k = 5.

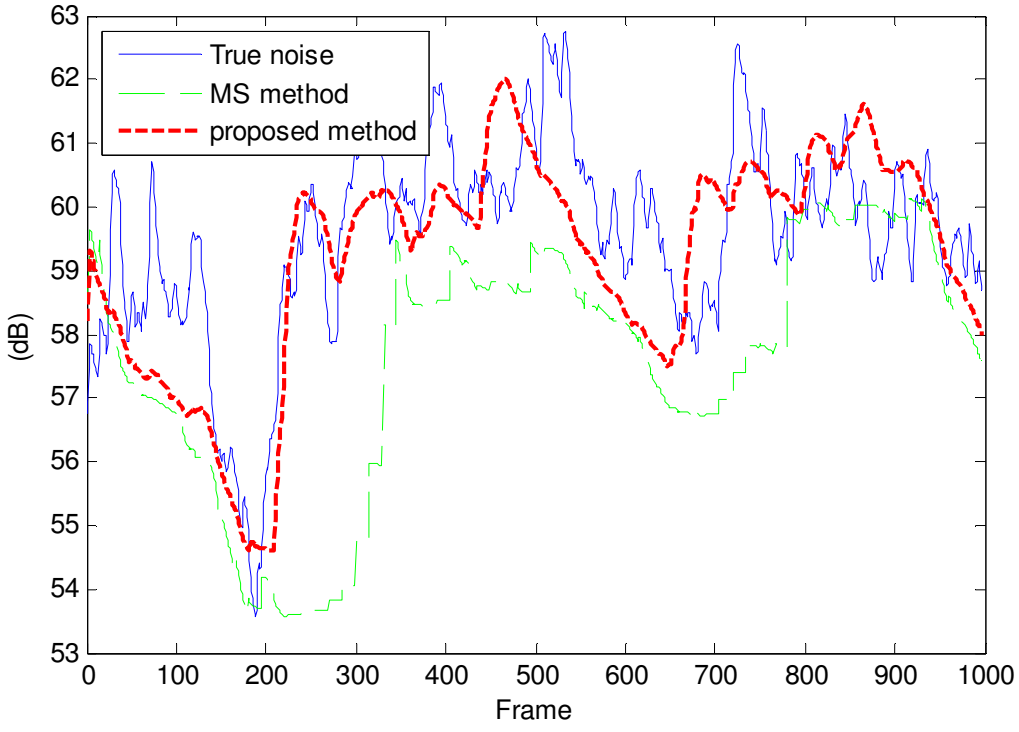

FIGURE 2. Ideal, proposed and MS average noise estimates for babble noise at 5 dB segmental SNR.

Input SegSNR (dB)	WGN Noise		F16 Noise		Babble Noise	
	MS	PM	MS	PM	MS	PM
-5	0.147	0.139	0.192	0.189	0.401	0.397
0	0.170	0.163	0.197	0.193	0.398	0.395
5	0.181	0.173	0.231	0.228	0.427	0.422
10	0.241	0.231	0.519	0.512	0.743	0.736

TABLE 1. Segmental Relative Estimation Error for Various Noise Types and Levels, Obtained Using the MS and proposed method (PM) Estimators.

Input SegSNR (dB)	WGN Noise		F16 Noise		Babble Noise	
	MS	PM	MS	PM	MS	PM
-5	8.213	8.285	6.879	6.924	3.254	3.310
0	7.231	7.312	6.025	6.165	2.581	2.612
5	6.215	6.279	5.214	5.298	2.648	2.697
10	5.114	5.216	3.964	4.034	1.943	1.998

TABLE 2. Segmental SNR Improvement for Various Noise Types and Levels, Obtained Using the MS and proposed method (PM) Estimators.

Table 2 summarizes the results of the segmental SNR improvement for various noise types and levels. The PM estimator consistently yields a higher improvement in the segmental SNR, than the MS estimator, under all tested environmental conditions.

5. SUMMARY AND CONCLUSION

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 minimum of the smoothed noisy speech spectrum. Unlike the MS method, 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 MS noise estimation algorithm.

6. REFERENCES

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

2. J. Sohn, N. S Kim and W. Sung, "*A statistical model-based voice activity detector,*" IEEE Signal Processing Letters, 6(1): 1-3, January 1999.

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

4. R. J. McAulay and M. L. Malpass "*Speech enhancement using a soft-decision noise suppression filter,*" IEEE Trans. Acoustics, Speech and Signal Processing, 28(2): 137-145, April 1980.

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

6. C. Ris and S. Dupont, "*Assessing local noise level estimation methods: application to noise robust ASR,*" Speech Communication, 34(1): 141-158, April 2001.

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

8. I. Cohen and B. Berdugo, "*Speech Enhancement for Non-Stationary Noise Environments,*" Signal Processing, 81(11): 2403-2418, November 2001.

9. R. Martin, "*Noise power spectral density estimation based on optimal smoothing and minimum statistics,*" IEEE Trans. Speech and Audio Processing, 9(5): 504-512, July 2001.

10. G. Doblinger, "*Computationally efficient speech enhancement by spectral minima tracking in subbands,*" Proc. 4th EUROSPEECH'95, Madrid, Spain, 18-21 September 1995, pp. 1513-1516.

11. R. Martin: "*An Efficient Algorithm to Estimate the instantaneous SNR of Speech Signals*," Proc. EUROSPEECH '93, pp. 1093-1096, Berlin, September 21-23, 1993.

12. Doblinger, G., 1995. "*Computationally efficient speech enhancement by spectral minima tracking in subbands,*" in Proc. Eurospeech' 2002, 1513–1516.

13. A. Varga and H. J. M. Steeneken, "*Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effiect of additive noise on speech recognition systems,*" Speech Communication, 12(3): 247-251, July 1993.

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

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

16. I. Cohen, "*On speech enhancement under signal presence uncertainty,*" Proc. 26th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2001, 7-11 May 2001, pp. 167-170.

17 I. Cohen and B. Berdugo, *"Speech Enhancement for Non-Stationary Noise Environments,"* Signal Processing, 81(11): 2403-2418, November 2001.

18 J. Ghasemi, K. Mollaei, *"A new approach for speech enhancement based on eigenvalue spectral subtraction,"* in Signal Processing: An International Journal (SPIJ), 3(4): 34-41, Sep. 2009.

19 Y. Ephraim and D. Malah, *"Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,"* IEEE Trans. Acoustics, Speech and Signal Processing, 33(2): 443-455, April 1985.

20. M. Satya Sai Ram, P. Siddaiah, M. M. Latha, *" Usefullness of speech coding in voice banking,"* in Signal Processing: An International Journal (SPIJ), 3(4): 42-54, Sep. 2009.

21 M.S. Salam, D. Mohammad, S-H Salleh, *" Segmentation of Malay Syllables in connected digit speech using statistical approach,"* in Signal Processing: An International Journal (SPIJ), 2(1): 23-33, February 2008.