A LOW-COMPLEXITY NOISE ESTIMATION ALGORITHM BASED ON SMOOTHING OF NOISE POWER ESTIMATION AND ESTIMATION BIAS CORRECTION

Rongshan Yu

Dolby Laboratories, Inc., 100 Potrero Ave, CA 94103, USA
Email: rzyu@dolby.com

ABSTRACT

This paper presents a low-complexity algorithm for tracking the noise spectral variance of speech contaminated by non-stationary noise sources. The proposed algorithm is based upon a recursive refinement process in which each step of the algorithm expectation of the instantaneous noise power is calculated based on information from the incoming signal and the current estimated distribution parameters, and estimation of the distribution parameter is refined accordingly to incorporate the expectation results. A bias estimation correction method is also introduced in the algorithm to avoid estimation errors that may occur when there is a significant mismatch between the statistics of the input signal and the current estimated distribution parameters. The proposed algorithm is compared to the Minimum Statistics method and it is found that the proposed algorithm achieves similar or better performances for various noise conditions and SNR settings.

Index Terms— noise estimation, noise tracking, noise suppression, speech enhancement, expectation-maximization

1. INTRODUCTION

We are living in a noisy world where the noise generated from either natural sources or human activities can be found almost everywhere: car, train, street, restaurant, etc. During voice communication those noises are captured by the microphone and adversely affect the quality of voice communication. To address this problem, noise suppression technology is developed to remove those noise components, and produce an enhanced speech signal that sounds more pleasant to human ears.

In principle, most noise suppression systems rely on some sort of spectral domain process that attenuates the time/frequency (T/F) regions of the input noisy speech that have low Signal-to-Noise-Ratios (SNR), and preserves those with high SNR. The essential parts of the speech signal are thus preserved while the noise level is greatly reduced, leading to an enhanced speech signal. Since many noises are non-stationary in nature, a noise estimator is used in noise suppression systems to trace time-varying statistics of the noise signals. One popular choice of the noise estimator is the VAD (Voice Activity Detection) based approach, where the noise estimation is updated only when speech is not present in the input. The performance of the VAD approach strongly depends on the accuracy of the voice detection, which is a difficult task in particular for signals with low SNR. In addition, this method precludes the possibility to update the noise estimation when the speech signal is present, which is inefficient since there may still be spectral bands where the speech level is weak and noise estimation can still be reliably updated. Another widely quoted method is the Minimum Statistics (MS) noise estimator [1]. In principle, the MS method keeps a record of historical samples for each spectral location, and the noise level is estimated based on the minimum signal level from the record. It is reported that the MS method achieves good noise tracing performance for non-stationary noises; however, it has a high memory demand and is not applicable to devices with limited memory resources.

In this paper we present a noise estimation algorithm that is built upon an expectation-maximization (EM) [2] principle. In the proposed algorithm, the instantaneous noise power is estimated each frame based on information from the incoming signal and the current estimated distribution parameters and the distribution parameters, which include the noise variance we are interested in, are refined from the expectation results. Instead of using a trained gain function for noise power estimation as proposed in [3], naïve minimum mean-square-error (MMSE) noise power expectation is used and the potential estimation bias problem is addressed by using a simple bias estimation correction method. The proposed algorithm has very low computational power and memory requirements and hence is suitable for embedded applications with limited computational resources.

2. PRINCIPLE OF THE ALGORITHM

2.1. Gaussian Model for Speech and Noise Signals

We consider the following additive signal model for the noisy speech signal:

$$\mathbf{Y}_k(m) = \mathbf{X}_k(m) + \mathbf{D}_k(m), \tag{1}$$

where $\mathbf{Y}_k(m)$, $\mathbf{X}_k(m)$, and $\mathbf{D}_k(m)$ are complex-valued short-time DFT coefficients of the noisy speech signal $y(n)$, clean speech signal $x(n)$ and noise signal $d(n)$ respectively. Here k is the frequency bin index and m is the frame index of the DFT analysis window. The complex-valued DFT coefficients can be further decomposed into amplitude and phase representations as follows:

$$\mathbf{Y}_k(m) = R_k(m)\exp\left(j\vartheta_k(m)\right), \tag{2}$$

$$\mathbf{X}_k(m) = A_k(m)\exp\left(j\alpha_k(m)\right), \text{ and} \tag{3}$$

$$\mathbf{D}_k(m) = N_k(m)\exp\left(j\phi_k(m)\right), \tag{4}$$

where $R_k(m)$, $A_k(m)$ and $N_k(m)$ are the amplitudes of the noisy speech, clean speech and noise signals, respectively, and $\vartheta_k(m)$, $\alpha_k(m)$ and $\phi_k(m)$ are their phases.

In most noise suppression systems the speech and the noise signals are usually modeled as independent, zero-mean, complex Gaussian variables with variances $\lambda_x(k)$, and $\lambda_d(k)$ due to the simplicity of this model and its relative good results. This assumption leads to the following marginal and conditional distributions:

$$p(N)=\begin{cases}\dfrac{2N}{\lambda_d}\exp\left(-\dfrac{-N^2}{\lambda_d}\right) & N\in[0,\infty);\\ 0 & \text{otherwise.}\end{cases} \tag{5}$$

$$p(\phi)=\begin{cases}\dfrac{1}{2\pi} & \phi\in[0,2\pi);\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

$$p(\mathbf{Y}\mid N,\phi)=\frac{1}{\pi\lambda_x}\exp\left(\frac{-\left|\mathbf{Y}-Ne^{j\phi}\right|^2}{\lambda_x}\right). \tag{7}$$

For conciseness, the frame index *m* and frequency index *k* are omitted in the above distribution functions as well as in the subsequent discussion whenever it won't cause any confusion.

2.2 Estimation of Noise Power

As a first step of the proposed noise estimation algorithm we try to estimate the instantaneous power of the noise signal from the noisy signal that we observed. To this end, we define:

$$\frac{1}{\lambda}\triangleq\frac{1}{\lambda_x}+\frac{1}{\lambda_d} \tag{8}$$

$$v\triangleq\frac{\gamma}{\xi(1+\xi)};\ \xi\triangleq\frac{\lambda_x}{\lambda_d};\ \gamma\triangleq\frac{R^2}{\lambda_d}. \tag{9}$$

Here ξ and γ are usually referred to as *a prior* SNR and *a posterior* SNR respectively. It can be shown that under the assumed statistical model the posterior distribution of the noise amplitude given the observed signal $\mathbf{Y}$ is Rician [4] with parameters (σ^2,s^2):

$$p(N\mid\mathbf{Y})=\frac{N}{\sigma^2}\exp\left(-\frac{N^2+s^2}{2\sigma^2}\right)I_0\left(\frac{Ns}{\sigma^2}\right), \tag{10}$$

$$\sigma^2\triangleq\frac{\lambda}{2},\ s^2\triangleq v\lambda \tag{11}$$

where $I_i(\bullet)$ denotes the modified Bessel function of order *i*. The MMSE estimation of the power of the noise signal is simply the second moment of this distribution, which is given by:

$$\begin{aligned}\hat{N}^2&=E\left[N^2\mid\mathbf{Y}\right]\\&=\frac{\xi}{1+\xi}\left(\frac{1+v}{\gamma}\right)R^2.\end{aligned} \tag{12}$$

In practical implementation, the *a priori* SNR ξ in the calculation of $\hat{N}^2(m)$ of frame *m* can be obtained by using the decision-directed estimation method proposed in [5]:

$$\hat{\xi}=\alpha\frac{\hat{A}^2(m-1)}{\lambda_d}+(1-\alpha)\max\left(\gamma^2(m)-1,0\right). \tag{13}$$

Here $0\ll\alpha<1$ is a pre-selected constant, and $\hat{A}(m-1)$ is the amplitude estimation of the clean speech of previous analysis frame, i.e.,

$$\hat{A}(m-1)=E\left[A(m-1)\mid\mathbf{Y}(m-1)\right]. \tag{14}$$

In addition, the noise variance estimation updated in the previous frame *m*-1, $\hat{\lambda}_d^{(m)}$, will be used in the calculation of $\hat{N}^2(m)$. The details of the noise variance estimation will be described in Section 2.3.

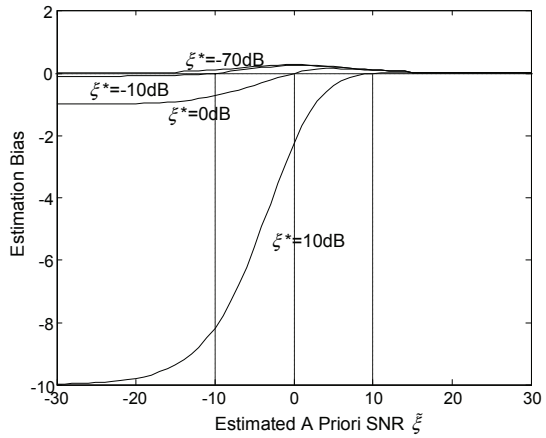
2.3 Estimation of the Noise Variance

The variance of the noise signal can be traced by using a recursive averaging process on the instantaneous noise power $N^2(m)$, which is unfortunately not accessible since in most cases the noise signal is "corrupted" by the speech signal in the input signal. Therefore in the proposed algorithm its MMSE estimation $\hat{N}^2(m)$, which represents the best-effort estimation of $N^2(m)$ under the assumed statistical model and the knowledge that we learn from the input noisy signal, is used instead. This idea leads to the following noise variance estimation algorithm:

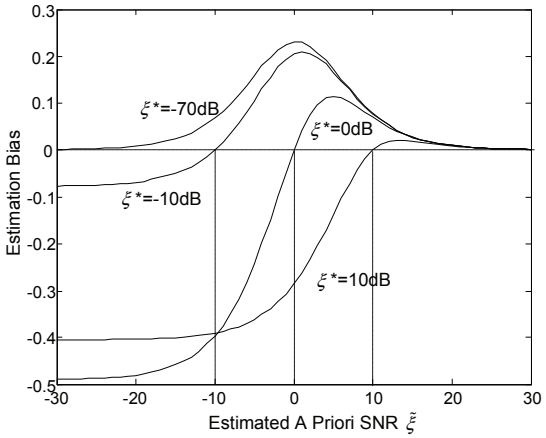
$$\hat{\lambda}_d^{(m+1)}=(1-\beta)\hat{\lambda}_d^{(m)}+\beta\hat{N}^2(m), \tag{15}$$

where $0<\beta<1$ is a constant, $\hat{\lambda}_d^{(m)}$ is the noise variance estimation from the previous frame and $\hat{\lambda}_d^{(m+1)}$ is the updated estimation after incorporating the noise power estimation $\hat{N}^2(m)$. The initial value $\hat{\lambda}_d^{(0)}$ can be simply set to any pre-determined value, or to the noise variance measured at the initialization stage of the noise variance estimator

It can be seen that the algorithm in fact follows closely to the principle of a generalized EM (GEM) algorithm [2]. In the expectation stage (12) the mean value of the noise power, or the hidden variable, is calculated with respect to the *a posterior* probability density function jointly decided by the observed signal $\mathbf{Y}$ and the current estimated noise variance $\hat{\lambda}_d^{(m)}$; in the maximization stage (15) the noise variance estimation is updated as a smoothed version of the maximum likelihood (ML) estimations performing on results from the expectation stage. Moreover, it can be shown that the observation likelihood of the noise variance



a) without the bias estimation correction



b) with the bias estimation correction (ψ =4.5)

Fig. 1 Estimation bias of the MMSE noise estimator.

estimation is increased in each step of the algorithm as what we expect from a GEM algorithm:

$$P\left(\mathbf{Y}(m)\,|\,\lambda_d^{(m+1)}\right) \geq P\left(\mathbf{Y}(m)\,|\,\lambda_d^{(m)}\right). \tag{16}$$

2.4 Bias Estimation Correction

The noise power estimation in (12) is an unbiased estimator for λ_d only when we have perfect knowledge about the *a prior* SNR, i.e., when the estimated SNR $\hat{\xi} = \xi^*$ where ξ^* is the true SNR of the input signal. When $\hat{\xi} \neq \xi^*$ it becomes a bias estimator for λ_d where the estimation bias is given by:

$$\begin{aligned} B &\triangleq E\left\{N^2(m) - \hat{N}^2(m)\right\} \Big/ E\left\{N^2(m)\right\} \\ &= \frac{\hat{\xi} - \xi^*}{\left(1+\hat{\xi}\right)^2}. \end{aligned} \tag{17}$$

which may affect the accuracy of the noise variance estimation. As can be seen from Fig. 1-a) the estimation bias is very unsymmetrical with respect to the error in the SNR estimation $\hat{\xi}$. Large negative bias (or over-estimation of noise power) occurs when the estimated $\hat{\xi} \ll 1$ and $\xi^* \gg \hat{\xi}$. This usually happens during speech onset where the estimated SNR from the decision-directed method (13) lags behind the true SNR, resulting in leakage of speech signal of large amplitudes into the noise variance estimation. To address this problem, we can simply skip samples with large amplitudes that don't fit into the assumed signal model in the noise estimation algorithm. Specifically, we preclude samples where the amplitudes satisfy

$$R^2(n) > \psi\left(1+\hat{\xi}\right)\lambda_d, \tag{18}$$

where ψ is a pre-defined constant. This is equivalent to replace (12) with the following noise power estimator:

$$\hat{N}^2(m) = \begin{cases} E\left[N^2(m)\,|\,\mathbf{Y}\right], & R^2(n) \leq \psi\left(1+\hat{\xi}\right)\lambda_d^{(m)}; \\ \lambda_d^{(m)}, & \text{otherwise.} \end{cases} \tag{19}$$

This estimator successfully avoids the over-estimation problem. Unfortunately, it will introduce an (under) estimation bias even when $\hat{\xi}_k = \xi_k^*$:

$$B = \frac{1}{1+\hat{\xi}_k}\frac{\psi e^{-\psi}}{1-e^{-\psi}} > 0, \tag{20}$$

which, however, can be easily compensated or simply neglected when ψ is sufficiently large.

As evident in Fig.2-b), the estimation bias of the updated noise estimator is now well-bounded even for very inaccurate SNR estimations.

3. EXPERIMENT RESULTS

We evaluated the performance of the proposed noise variance estimator by measuring its estimation error for speech signal contaminated by various noise sources. The speech files used in the experiments contain concatenated sentences of eight short sentences from both male and female speakers. An approximate 0.5 second period of silence is inserted between each sentence pair in the speech files. The sentences are simple meaningful sentences in English extracted from the TIMIT database [6]. The speech levels as measured with the P.56 algorithm [7] are adjusted to -26 dBov. Four different noise files were used in the tests and the noise levels were adjusted using the Root Mean Square (RMS) measure to the level dictated by the SNR settings of the test conditions. The noise files were digitally mixed with the speech files to produce the testing files used in the tests. The sampling rate of the testing files is 8 kHz.

The proposed noise estimator was performed on spectral coefficients derived from a short-time DFT with 50% overlapping windows of 8 ms. The parameters used in the experiments are

Table 1 Mean and variance (in parenthesis) of the estimation error of the proposed algorithm and MS.

NOISE TYPE	6 dB		15 dB	
	Proposed	MS	Proposed	MS
Street	1.14(1.46)	1.52(1.64)	1.47(1.69)	1.71(1.91)
Car	1.07(0.96)	1.36(1.22)	1.25(1.08)	1.64(1.68)
Babble	1.69(1.53)	3.33(2.52)	1.91(1.65)	3.24(2.49)
Train	0.94(1.09)	1.37(1.41)	1.21(1.25)	1.55(1.77)

listed as follows: $\alpha = 0.02$, $\beta = 0.04$; $\psi = 4.5$. For comparison, the MS noise estimator was included in our experiments, for which the original settings from [1] were used. In our experiments the estimation error is defined as the absolute value of the difference between the estimated noise variance and the true variance in the logarithmic domain as follows:

$$LogErr = \left| 10\log_{10}\left(\frac{\hat{\lambda}_d(k,m)}{\tilde{\lambda}_d(k,m)}\right)\right| \text{ (dB)}, \quad (21)$$

where $\hat{\lambda}_d(k,m)$ is the estimated noise variance for frequency bin k and frame index m. The true noise variance $\tilde{\lambda}_d(k,m)$ is calculated separately from the "clean" noise files as

$$\tilde{\lambda}_d(k,m) = \kappa\tilde{\lambda}_d(k,m-1) + (1-\kappa)N_k^2(m), \quad (22)$$

where $\kappa = 0.02$.

Table 1 gives the mean and variance of *LogErr* that are measured over all the frames and frequency bins of the test files for each experiment condition. It can be seen that the proposed algorithm achieves a smaller mean estimation error for all the noise conditions. In addition, its performance is more consistent as suggested by the smaller variances of *LogErr* , which is also important to the noise suppression quality since a consistent estimation error can be easily compensated by tuning other parts of the noise suppression system.

Instantaneous behavior of the proposed algorithm and MS algorithm are illustrated in Fig. 2 with two examples. It can be seen that in both cases the two algorithms are able to track the changing noise variance over time; however, none of them did a perfect job. The proposed algorithm tends to produce smoother estimation of the noise variance compared to the MS algorithm while the latter does a better job when the noise level is fluctuated significantly.

4. CONCLUSION

This paper proposes a low-complexity noise estimation algorithm that relies on recursive smoothing of MMSE estimation of noise power, and a simple estimation bias correction method. The proposed algorithm provides a low-complexity alternative to the popular MS algorithm while it achieves similar or better performance for non-stationary noise sources.

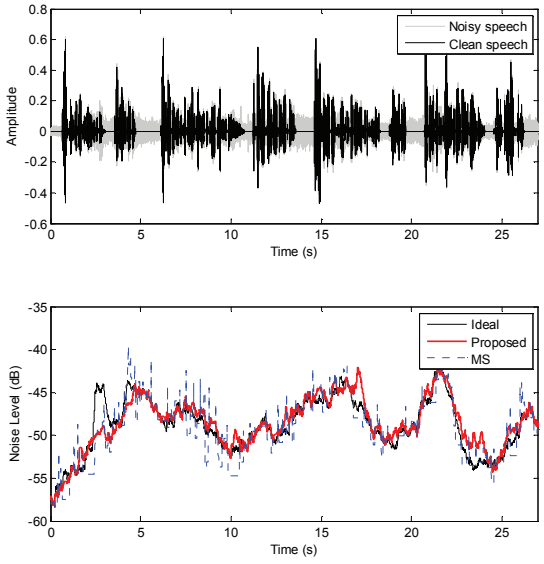
5. REFERENCES

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

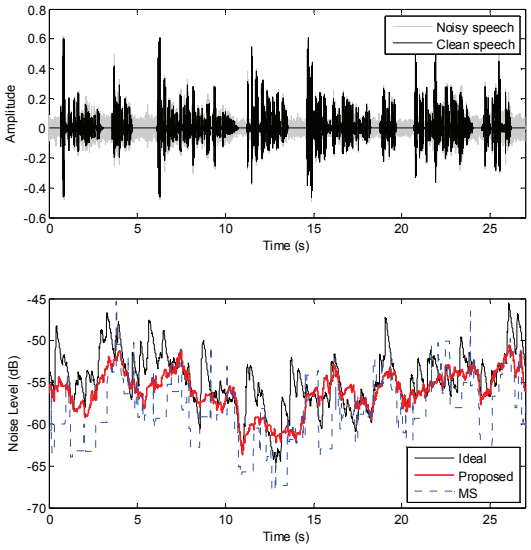
[2] A. Dempster, N. Laird, and D. Rubin. "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39(1):1–38, 1977

[3] J. S. Erkelens and R. Heusdens, "Fast noise tracking based on recursive smoothing of MMSE noise power estimates," in *Proc. ICASSP 2008*.

[4] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell System Technical Journal*, vol. 27, pp. 109-157, 1948.



a) Street noise, SNR = 6 dB



b) Babble noise, SNR = 6 dB

Fig. 2 Output of the noise variance estimation algorithms as a function of time (of frequency bin *k*=10).

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109-1121, Dec. 1984.

[6] Fisher, William M.; Doddington, George R. and Goudie-Marshall, Kathleen M., "The DARPA Speech Recognition Research Database: Specifications and Status". *Proceedings of DARPA Workshop on Speech Recognition: 93-99*

[7] ITU-T Rec. P.56, *Objective Measurement of Active Speech Level - Telephone Transmission Quality Objective Measuring Apparatus*.