OPTIMAL RECURSIVE SMOOTHING OF NON-STATIONARY PERIODOGRAMS

Rainer Martin and Thomas Lotter

Institute of Communication Systems and Data Processing
Aachen University of Technology, D-52056 Aachen, Germany
Phone: +49 241 8026984, Fax: +49 241 8022186
E-mail: {martin | thomas}@ind.rwth-aachen.de

ABSTRACT

This contribution presents a time varying optimal smoothing parameter for periodograms which are smoothed over frequency and time. While the spectral smoothing is implemented as an average over adjacent Discrete Fourier Transform (DFT) bins the temporal smoothing is achieved by means of a first order recursive system. The smoothing parameter of this recursive system minimizes a conditional mean square error criterion and is optimal for chi-square distributed frequency domain data. The optimal smoothing is especially useful for tracking the power spectral density of noisy speech signals.

1. INTRODUCTION

Most single-microphone speech enhancement algorithms require an explicit estimate of the noise power spectral density (psd). E.g., the MMSE estimator of the clean speech Discrete Fourier Transform (DFT) coefficients or their magnitude is often defined in terms of *a priori* and *a posteriori* SNR values [1, 2, 3]. The computation of these SNR values necessitates the estimation of the noise psd. This noise psd estimate is usually computed by smoothing magnitude squared DFT coefficients $|Y(\lambda,k)|^2$ of the noisy signal $y(i)$ and by employing a voice activity detector or the Minimum Statistics [4, 5] noise tracking approach. The DFT coefficients $Y(\lambda,k)$ are obtained from

$$Y(\lambda,k)=\sum_{\mu=0}^{L-1} y(\lambda R+\mu)h(\mu)\exp(-j2\pi k\mu/L) \qquad (1)$$

where $h(\mu)$ denotes a window function, λ is the subsampled time index, $\lambda\in\mathbb{Z}$, and k denotes the frequency bin index, $k\in\{0,1,...,L-1\}$, which is related to the normalized center frequency Ω_k by $\Omega_k=2\pi k/L$. R denotes the shift between successive signal frames. To simplify our notation we assume $\sum_{\mu=0}^{L-1}h^2(\mu)=1$. Because the human ear has less frequency resolution at higher frequencies than at lower frequencies speech (and audio) signals are often processed in Bark scaled frequency bands. Bark scale processing can also lead to significant computational savings and is therefore attractive for real time implementations of speech enhancement systems. A simple method to obtain approximately Bark scaled noise psd estimates is to average magnitude squared Fourier coefficients over adjacent DFT bins, i.e.,

$$B(\lambda,k)=\frac{1}{2N_k+1}\sum_{\ell=k-N_k}^{k+N_k}|Y(\lambda,\ell)|^2 \qquad (2)$$

where k is the index of the center frequency bin and $2N_k+1$ is the number of averaged bins of this band. Obviously, another consequence of smoothing over frequency is the reduction of the variance of the spectral data which is most pronounced when adjacent frequency bins are statistically independent.

Given the averaged periodogram $B(\lambda,k)$, the final noise psd estimate is obtained by temporally smoothing over successive DFT frames using a first order recursive system

$$P(\lambda,k)=\alpha(\lambda,k)P(\lambda-1,k)+(1-\alpha(\lambda,k))B(\lambda,k) \qquad (3)$$

where $\alpha(\lambda,k)$ is a time and frequency dependent smoothing parameter.

The smoothing parameter $\alpha(\lambda,k)$ has to satisfy conflicting requirements. When the noise is stationary $\alpha(\lambda,k)$ should be close to one to achieve a small variance of the psd estimate. When the noise is not stationary, the smoothing parameter must be small enough to enable fast tracking. In this contribution we derive a solution for a time and frequency dependent $\alpha(\lambda,k)$ which minimizes a mean square error criterion. As an extension of a result given in [5], the new optimal smoothing parameter also accounts for spectral smoothing.

2. STATISTICAL MODELLING OF AVERAGED DFT COEFFICIENTS

When $y(i)$ is stationary with a relatively small span of correlation and the frame size L is large the real and imaginary part of a DFT coefficient $Y(\lambda,k)$, $k\notin\{0,L/2\}$, can be considered to be independent and can be modelled as zero mean Gaussian random variables [6]. Under these assumptions each periodogram bin $|Y(\lambda,k)|^2$ is an exponentially distributed random variable with probability density function (pdf)

$$f_{|Y(\lambda,k)|^2}(x)=\frac{U(x)}{\sigma_Y^2(\lambda,k)}\exp(-x/\sigma_Y^2(\lambda,k)) \qquad (4)$$

where $\sigma_Y^2(\lambda,k)=E\{|Y(\lambda,k)|^2\}$ is the power spectral densities of the noisy signal, $\sigma_Y^2(\lambda,k)=\sigma_S^2(\lambda,k)+\sigma_N^2(\lambda,k)$. $\sigma_S^2(\lambda,k)$ and $\sigma_N^2(\lambda,k)$ denote the power spectral density of the speech and noise signal, respectively. The speech and the noise signal are considered to be statistically independent. $U(x)$ is the unit step function, i.e., $U(x)=1$ for $x\geq 0$ and $U(x)=0$ otherwise. Obviously, during speech pause, $\sigma_S^2(\lambda,k)\equiv 0$, the mean and the variance of $|Y(\lambda,k)|^2$ are equal to $\sigma_N^2(\lambda,k)$ and $\sigma_N^4(\lambda,k)$, respectively. To simplify

the discussion we will now assume that no speech is present and will discuss the case of speech activity later on.

The exponential density can be also interpreted as a special case of the more general χ^2 density with $K = 2$ degrees of freedom

$$f_{\chi^2}(x) = \frac{KU(x)}{2\sigma_N^2\Gamma(K/2)}(\frac{xK}{2\sigma_N^2})^{K/2-1}\exp(\frac{-xK}{2\sigma_N^2}) \quad (5)$$

which arises as the distribution of the sum of squared i.i.d. Gaussian random variables. $\Gamma(\cdot)$ is the complete Gamma function [7]. K is determined by the number of independent random variables and is related to the variance of the χ^2-distributed variable by $var\{\chi^2\} = 2\sigma_N^4/K$.

Neighboring periodogram bins are asymptotically independent. For $k > N_k$ and $k < L/2 - N_k$ the probability density of the spectrally averaged periodogram $B(\lambda, k)$ can be therefore approximated by a χ^2-distribution with $K = 4 * N_k + 2$ degrees of freedom [6, Theorem 5.4.3]. When $h(\mu)$ is a tapered analysis window the variance of the averaged periodogram is larger than for the untapered (rectangular) case. The difference between these two cases is approximately ($L \to \infty$, $N_k \to \infty$, $N_k/L \to 0$) given by the factor [6, Theorem 5.6.4]

$$F_W = L\sum_{\mu=0}^{L-1} h^4(\mu) \Big/ \left[\sum_{\mu=0}^{L-1} h^2(\mu)\right]^2 . \quad (6)$$

Hence, in the case of a tapered analysis window we might approximate the pdf of $B(\lambda, k)$ by a χ^2-distribution with

$$\widetilde{K}(k) = K(k)/F_W = (4N_k + 2)/F_W \quad (7)$$

"equivalent" degrees of freedom. Generally, N_k varies with k. Therefore, the equivalent degrees of freedom $\widetilde{K}(k)$ are also a function of k. Equation (7) will be used below for estimating the degrees of freedom of the Bark scaled averaged periodograms when a large number of DFT bins is averaged. E.g., for a 256 point Hann window we obtain $\widetilde{K}(k) = 0.516K(k)$. When a relatively small number of bands is averaged the theoretical limit is not accurate and the equivalent degrees of freedom $\widetilde{K}(k)$ must be determined numerically. For the 256 point Hann window, Figure 1 plots the equivalent degrees of freedom $\widetilde{K}$ normalized on the degrees of freedom $K = 4 * N_k + 2$ which are obtained in the untapered case for $L \to \infty$ and an input signal $y(i)$ which is white Gaussian noise.

To conclude this Section, we note that approximations for the variance of the averaged periodogram $B(\lambda, k)$ are also available when neighboring DFT bins are not equally weighted [6]. This modified weighting can be used to emphasize those bins which are close to the center bin of a given frequency band. Thus, the bias with respect to the true power spectral density at the frequency of the center bin can be reduced when the power spectral density varies in the neighborhood of the center bin.

3. OPTIMAL FIRST ORDER RECURSIVE SMOOTHING

To derive the optimal smoothing parameter we assume that no speech is present. In this case we want $P(\lambda, k)$ in (3) to be as close as possible to the true noise psd $\sigma_N^2(\lambda, k)$. Therefore, our objective is to minimize the conditional mean square error

$$E\{(P(\lambda, k) - \sigma_N^2(\lambda, k))^2 \mid P(\lambda - 1, k)\} \quad (8)$$

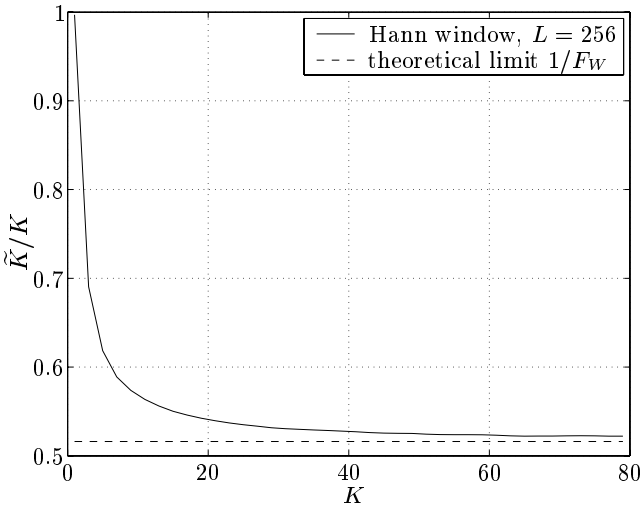

Figure 1: Equivalent degrees of freedom $\widetilde{K}$ normalized on $K = 4 * N_k + 2$ for an $L = 256$ point Hann window (solid) and the theoretical limit (dashed) according to (7). The number of averaged DFT bins equals $K/2$.

from one iteration to the next. After substituting (3) into (8) and using $E\{B(\lambda, k)\} \approx \sigma_N^2(\lambda, k)$ and $E\{B^2(\lambda, k)\} \approx (\widetilde{K}(k)+2)\sigma_N^4(\lambda, k)/\widetilde{K}(k)$ for χ^2-distributed data with $\widetilde{K}(k)$ degrees of freedom, the optimal smoothing parameter is given by (see Appendix A)

$$\alpha_{opt}(\lambda, k) = \frac{2}{2 + \widetilde{K}(k)(P(\lambda - 1, k)/\sigma_N^2(\lambda, k) - 1)^2} \quad (9)$$

and for $\widetilde{K}(k) = 2$ (no frequency averaging) by

$$\alpha_{opt}(\lambda, k) = \frac{1}{1 + (P(\lambda - 1, k)/\sigma_N^2(\lambda, k) - 1)^2} . \quad (10)$$

The quotient $P(\lambda-1, k)/\sigma_N^2(\lambda, k) = \overline{\gamma}(\lambda, k)$ in (9) and (10) is a smoothed version of the *a posteriori* SNR [1],

$$\gamma(\lambda, k) = \frac{|Y(\lambda - 1, k)|^2}{\sigma_N^2(\lambda, k)} . \quad (11)$$

Figure 2 plots the optimal smoothing parameter α_{opt} for $0 \le \overline{\gamma} \le 10$ and $\widetilde{K} \in \{2, 8, 32\}$. Since the optimal smoothing parameter α_{opt} is between zero and one a stable and non-negative noise power estimate $P(\lambda, k)$ is guaranteed. Finally, after substituting $\alpha_{opt}(\lambda, k)$ into (8) we find the minimum mean square error

$$E\{(P(\lambda, k) - \sigma_N^2(\lambda, k))^2 \mid P(\lambda - 1, k)\}_{|\alpha=\alpha_{opt}} = \frac{2}{\widetilde{K}(k)}\sigma_N^4(\lambda, k)(1 - \alpha_{opt}(\lambda, k)) . \quad (12)$$

For stationary noise $\overline{\gamma}(\lambda, k)$ will be close to one since the smoothed periodogram converges to the true noise psd. In this case $\alpha_{opt}(\lambda, k)$ is near unity and a low variance of the smoothed periodogram is achieved. When the noise is non-stationary the smoothing parameter will be reduced to enable rapid tracking. For $\widetilde{K} > 2$ the periodogram data is smoothed over frequency and has therefore a smaller variance. Thus, for $\widetilde{K} > 2$ deviations of the smoothed psd estimate $P(\lambda, k)$ from the target psd $\sigma_N^2(\lambda, k)$ lead to a more

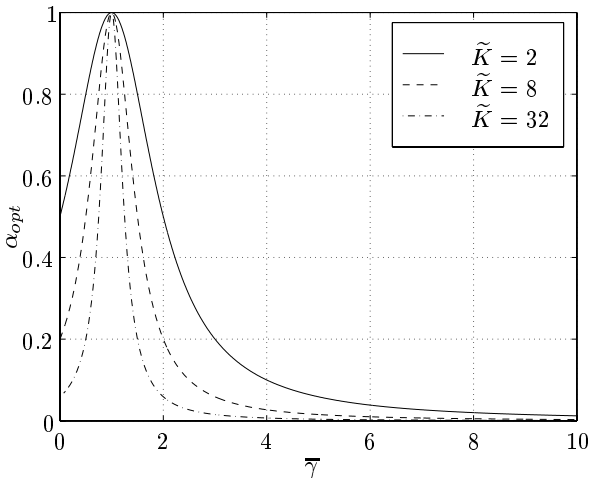

Figure 2: Optimal smoothing parameter α_{opt} as a function of the smoothed *a posteriori* SNR $\overline{\gamma}(\lambda, k)$ for $\widetilde{K} \in \{2, 8, 32\}$.

rapid decay of the smoothing parameter. When speech is present in the noisy signal $y(i)$, $P(\lambda, k)$ is not an estimate of the noise psd anymore. However, the influence of speech on the optimal smoothing parameter is similar to highly non-stationary noise. In any case, the smoothing parameter will be reduced to small values and thus enables fast tracking of the time varying signal power. This behavior is actually of great advantage when this smoothing method is combined with a Minimum Statistics noise estimator [5].

Note, that the smoothing parameter $\alpha_{opt}(\lambda, k)$ is quite different to the soft-decision smoothing law proposed in [8]. The update control in [8] shifts the balance in (3) towards a lower proportion of the smoothed variable $P(\lambda - 1, k)$ whenever the *a posteriori* SNR is close to one. This results in a high update rate for the noise estimate during speech pause but also a higher variance of the noise psd estimate. The non-stationarity of the noise is therefore not explicitly taken into account. In our smoothing approach we utilize the statistics of the *a posteriori* SNR (or averaged versions thereof) to track non-stationary noise and to minimize the variance of the noise psd estimate.

3.1. Error Monitoring

Since the application of the optimal smoothing parameter requires a noise psd estimate the proposed smoothing method is used in conjunction with, e.g., the Minimum Statistics noise psd estimator [5]. In a practical implementation of the optimal smoothing parameter (9) we replace the true noise psd $\sigma_N^2(\lambda, k)$ by its latest estimated value $\widehat{\sigma}_N^2(\lambda - 1, k)$ and limit the smoothing parameter to a maximum value α_{max}, e.g. $\alpha_{max} = 0.96$, to avoid dead lock for $\overline{\gamma}(\lambda, k) = 1$.

In general, the time evolution of the estimated noise psd $\widehat{\sigma}_N^2(\lambda, k)$ lags behind the time evolution of the true noise psd (tracking delay). As a consequence, the estimated noise psd might be smaller or larger than the true noise psd and thus, the estimated smoothing parameter might be too small or too large. Problems may arise when the smoothing parameter is close to one since then the smoothed psd estimate $P(\lambda, k)$ cannot react quickly to changes in the true noise psd. Given this uncertainty in the noise psd estimate the tracking error in the smoothed short term psd $P(\lambda, k)$ must be monitored. When tracking errors are detected the optimal smoothing parameter must be decreased to guarantee reliable operation under all circumstances.

Tracking errors in the short term estimate $P(\lambda, k)$ can be monitored by comparing $P(\lambda, k)$ to a reference quantity, for instance the frequency averaged periodogram. Our monitoring algorithm therefore compares the average short term psd estimate of the previous frame $\frac{1}{L}\sum_{k=0}^{L-1} P(\lambda - 1, k)$ to the average periodogram $\frac{1}{L}\sum_{k=0}^{L-1} |Y(\lambda, k)|^2$ and thus detects deviations of the short term psd estimate from the actual averaged periodogram. The result of this comparison can be used to modify the smoothing parameter in case of large deviations.

The comparison between the average smoothed psd estimate and the average actual periodogram is implemented by means of a "soft" $1/(1 + x^2)$ characteristic

$$\widetilde{\alpha}_c(\lambda) = \frac{1}{1 + (\sum_{k=0}^{L-1} P(\lambda - 1, k) / \sum_{k=0}^{L-1} |Y(\lambda, k)|^2 - 1)^2} \tag{13}$$

and the resulting correction factor is limited to values larger than 0.7 and smoothed over time

$$\alpha_c(\lambda) = 0.7\ \alpha_c(\lambda - 1) + 0.3\ \max(\widetilde{\alpha}_c(\lambda), 0.7)\ . \tag{14}$$

The smoothing parameter in recursion (14) was chosen empirically. The multiplication of the correction factor with the optimal smoothing parameter then yields the final time and frequency dependent smoothing parameter

$$\widehat{\alpha}(\lambda, k) = \frac{2\alpha_{max}\ \alpha_c(\lambda)}{2 + \widetilde{K}(k)(P(\lambda - 1, k)/\widehat{\sigma}_N^2(\lambda - 1, k) - 1)^2}\ . \tag{15}$$

4. EXPERIMENTAL RESULTS

We verify the benefits of the proposed smoothing method by processing a time varying synthetic white Gaussian noise signal with the above algorithm (sampling rate 16 kHz, $L = 512$). The first 3 seconds of this signal have constant power. For the next 5 seconds we add a white Gaussian noise which is amplitude modulated by a sine wave with a modulation frequency of 2 Hz. The composite signal is therefore representative of a stationary noise signal with an additive non-stationary component such as speech. We compare the estimated power for our time varying smoothing method with the estimated power for a constant smoothing parameter of $\alpha = 0.96$. The noise psd estimate required for the computation of $\widehat{\alpha}(\lambda, k)$ was obtained from a Minimum Statistics noise psd estimator [5]. Subplot A of Figure 3 plots the resulting time varying smoothing parameter (dashed) and the constant smoothing parameter (dotted). Subplot B plots the smoothed power for the proposed smoothing algorithm, the true signal power, and the noise power estimate as obtained from the Minimum Statistics method. Subplot C depicts the smoothed power, the true signal power and the Minimum Statistics noise power estimate for the smoothing with a constant smoothing parameter $\alpha = 0.96$. Finally, in subplot D we plot the mean square error (MSE) between the smoothed power and the true noise power for both smoothing methods. Note, that all quantities in Fig. 3 are averaged over all frequency bands. We find that for stationary noise the time varying smoothing parameter achieves approximately the same MSE as the constant smoothing parameter and that the Minimum Statistics noise power estimator delivers a noise power estimate which is representative of the

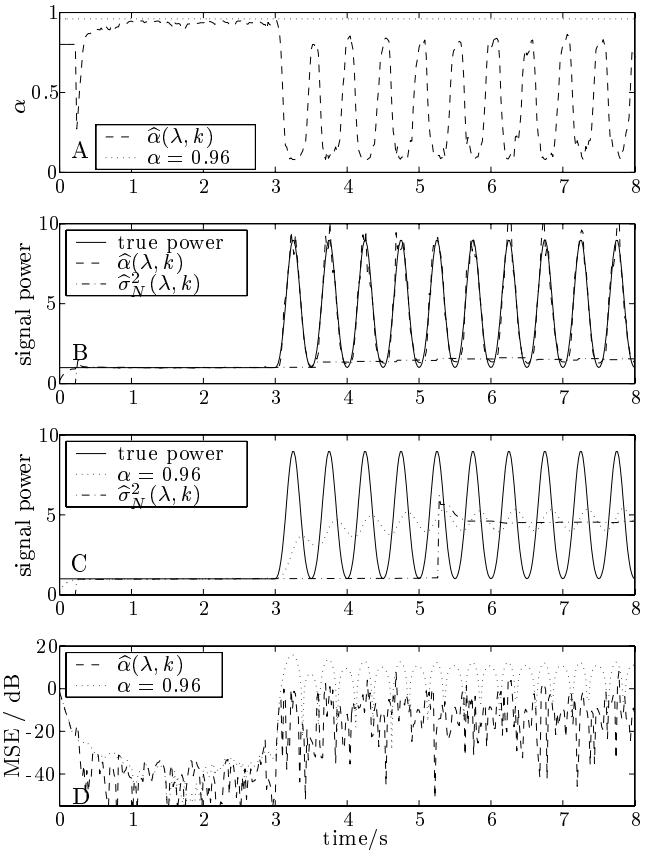

Figure 3: Results of processing a modulated white Gaussian noise with proposed (subplot B) and with a constant ($\alpha = 0.96$, subplot C) smoothing parameter. A: Smoothing parameters; B,C: Estimated power, true power, and estimated noise floor; D: Mean square errors.

stationary component within the signal (subplot B). However, the constant parameter does not allow tracking of the signal power when the signal is modulated. Therefore, after a transient period the noise power estimate delivered by the Minimum Statistics estimator is significantly too large for $\alpha = 0.96$ (subplot C). On the other hand, a small constant smoothing parameter such as $\alpha = 0.6$ results in tracking properties similar to the proposed method. However, the error variance for the stationary part of the signal is significantly larger for $\alpha = 0.6$ than for the time varying smoothing parameter (not shown in Figure 3).

The proposed smoothing algorithm was successfully employed in a wideband speech enhancement system (sampling rate 16 kHz, $L = 512$). For the computation of the noise estimate 257 DFT bins were grouped into 22 Bark scaled frequency bands with significant computational savings.

APPENDIX A: DERIVATION OF THE OPTIMAL SMOOTHING PARAMETER

We minimize $E\{(P(\lambda,k) - \sigma_N^2(\lambda,k))^2 | P(\lambda-1,k)\}$ where $P(\lambda,k)$ is given by (3). After substituting (3) into the error criterion we obtain

$$E\{(\alpha(\lambda,k)(P(\lambda-1,k) - B(\lambda,k)) + B(\lambda,k) - \sigma_N^2(\lambda,k))^2\}. \tag{16}$$

Differentiating (16) with respect to $\alpha(\lambda,k)$ and equating the result to zero leads to

$$E\{(\alpha(\lambda,k)(P(\lambda-1,k) - B(\lambda,k)) + B(\lambda,k) - \sigma_N^2(\lambda,k)) \cdot (P(\lambda-1,k) - B(\lambda,k))\} = 0 \tag{17}$$

By solving for $\alpha(\lambda,k)$ we obtain

$$\alpha_{opt}(\lambda,k) = \frac{E\{(P(\lambda-1,k) - B(\lambda,k))(\sigma_N^2(\lambda,k) - B(\lambda,k))\}}{E\{(P(\lambda-1,k) - B(\lambda,k))^2\}}. \tag{18}$$

Substituting $E\{B(\lambda,k)\} = \sigma_N^2(\lambda,k)$ and $E\{B^2(\lambda,k)\} = (1 + 2/\widetilde{K}(k))\sigma_N^4(\lambda,k)$, where $\widetilde{K}(k)$ denotes the equivalent degrees of freedom we obtain for the numerator of (18)

$$E\{(P(\lambda-1,k) - B(\lambda,k))(\sigma_N^2(\lambda,k) - B(\lambda,k))\} = \frac{2\sigma_N^4(\lambda,k)}{\widetilde{K}(k)} \tag{19}$$

and the denominator of (18)

$$E\{(P(\lambda-1,k) - B(\lambda,k))^2\} = \frac{2\sigma_N^4(\lambda,k)}{\widetilde{K}(k)} + (P(\lambda-1,k) - \sigma_N^2(\lambda,k))^2 . \tag{20}$$

The optimal smoothing parameter is therefore given by (9). Since the second derivative of the mean square error, $2E\{(P(\lambda-1,k) - B(\lambda,k))^2\}$, is non-negative a minimum is obtained.

REFERENCES

[1] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, December 1980.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, December 1984.

[3] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on a Priori Signal to Noise Estimation," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 629–632, 1996.

[4] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, pp. 1182–1185, 1994.

[5] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, July 2001.

[6] D. Brillinger, *Time Series: Data Analysis and Theory.* Holden-Day, 1981.

[7] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products.* Academic Press, 5th ed., 1994.

[8] J. Sohn and W. Sung, "A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, pp. 365–368, 1998.