MMSE ESTIMATION OF MAGNITUDE-SQUARED DFT COEFFICIENTS WITH SUPERGAUSSIAN PRIORS

Colin Breithaupt and Rainer Martin

Institute of Communications Technology
Technical University of Braunschweig
Schleinitzstr. 22, D-38092 Braunschweig, Germany
{breithaupt, martin}@ifn.ing.tu-bs.de

ABSTRACT

We present two minimum mean square error (MMSE) frequency domain estimators of the squared magnitude of a clean speech signal that is degraded by additive noise. These estimators are derived under the assumption that the DFT (discrete Fourier transform) coefficients of the clean speech are best modelled by the Gamma probability distribution function (pdf) instead of the common Gaussian pdf. The statistics of the perturbing noise is the Gaussian pdf in one case and the Laplacian pdf in the other. The estimators are used as noise reduction filters in the experimental evaluation. We give a comparison with a previously derived estimator which uses the Gaussian pdf as the pdf for speech and noise coefficients.

1. INTRODUCTION

Speech enhancement algorithms have found many applications in mobile communications and human-machine interfaces. Although numerous algorithms are available and significant improvements have been obtained there is no single algorithm which suits all kinds of applications. Even for a single application such as low bit rate speech coding Accardi and Cox [1] proposed to use more than one estimator in order to deliver optimally preprocessed signals to the various parts of the speech coder. In this context they developed the notion of “core estimators”. They used a MMSE-LSA estimator [4] as a core estimator to enhance the prediction residual and an estimator for the magnitude-squared DFT coefficients to enhance the autocorrelation coefficients which are in turn used to compute the LSF coefficients. In this paper we focus on the estimation of the magnitude-squared DFT coefficients, however with a significantly improved statistical model.

It was observed in [8] and [10] that DFT coefficients of speech signals derived from speech frames having a length of about the span of correlation within the signal are not normally distributed. The Gamma pdf is found to be a more appropriate statistical model for the real and imaginary part of the complex DFT coefficients of clean speech. As the analytic and in general non-linear MMSE estimation requires that the pdf of the signal and the disturbing noise is known, the observation in [8] suggests to newly derive estimators based on supergaussian priors. In this contribution we present the Squared Spectral Magnitude (SSM) estimator of clean speech DFT coefficients based on the Gamma pdf. As statistical models of the noise we choose the common Gaussian pdf and the Laplacian pdf. The SSM estimator can be used e.g. in the LPC analysis as in [1].

The structure of this paper is as follows: In the next section we present the MMSE estimation in a general form and define the pdf’s that yield our two new estimators. In section 3 we discuss the estimators. Finally we summarize experimental results.

2. MMSE ESTIMATION IN THE DFT DOMAIN

Our goal is to recover information about a clean speech signal s that is degraded by an additive noise n. The disturbed time signal y is digitized at a sampling rate $f_s = 8000\,\mathrm{Hz}$ yielding $y(k) = s(k) + n(k)$ at times $t = k/f_s$, $k \in \mathbb{Z}$. $s(k)$ and $n(k)$ shall be statistically independent signals both having zero mean.

We obtain frequency domain coefficients via a framewise DFT transform of length M , e.g. $M = 256$. A frame λ of M consecutive samples starts at $k = \lambda \cdot D$, with frame shift $D = M/2 = 128$ and frame index $\lambda \in \mathbb{Z}$. Two frames λ and $\lambda + 1$ overlap by $M - D = 128$ samples (50 per cent overlap). A Hann window h_{hann} is used in the transform. We get the DFT transform of a frame λ by

$$\begin{aligned} Y(\mu,\lambda) &= S(\mu,\lambda) + N(\mu,\lambda) \\ &= \sum_{\kappa=0}^{M-1} y(\lambda D + \kappa) h_{hann}(\kappa) e^{-j2\pi\frac{\mu}{M}\kappa}, \end{aligned}$$

where $\mu \in \{0, \dots, M-1\}$ is the index for the normalized center frequency $\Omega_\mu = 2\pi\ \mu/M$.

2.1. MMSE Estimation

The information we wish to extract from the noisy signal $Y(\mu,\lambda)$ is the squared magnitude of the clean speech spectral coefficient, i.e. $|S(\mu,\lambda)|^2$. The MMSE estimation of $|S(\mu,\lambda)|^2$ is obtained by the conditional expectation $E\left\{\,|S(\mu,\lambda)|^2 \middle|\, Y(\mu,\lambda)\right\}$. We may omit the indices for the frame λ and the frequency bin μ in the sequel, because any estimation is done independently for one specific λ and one specific μ. We assume the real and imaginary parts of a DFT coefficient to be independent, identically distributed (i.i.d.). Thus the estimation can be split into an estimator for the real part $S_R = \mathrm{Real}\,\{S\}$ and the imaginary part $S_I = \mathrm{Imag}\,\{S\}$. The estimator then becomes

$$E\left\{\,|S|^2 \middle|\, Y\right\} = E\left\{\,{S_R}^2 \middle|\, Y_R\right\} + E\left\{\,{S_I}^2 \middle|\, Y_I\right\}. \qquad (1)$$

For an analytic evaluation of this expression we need to describe the statistical behavior of the DFT coefficients. As is illustrated in [8] the Gamma pdf approximates the histogram of the clean speech coefficients S_R and S_I more precisely for the short time DFT (M reasonably small) than the Gaussian distribution. Similarly noise originating from a predominant source does not have to be normally distributed. We found that the Laplacian pdf is a good alternative for car noise (e.g. noise in the interior of a car driving at 90 km per hour which we used in our evaluation).

We derive two analytic solutions of (1). One for Gaussian distributed noise coefficients and one for noise coefficients having a Laplacian pdf. In both cases the Gamma distribution is used as the pdf of S_R and S_I. The corresponding densities are given in (2) – (4) for the real part of the DFT coefficients. N_R and S_R denote the real part of the noise and the speech, respectively.

2.1.1. Gaussian pdf, i.e. Normal distribution

$$p\left(N_R\right) = \frac{1}{\sqrt{\pi}\sigma_n} \exp\left(-\frac{N_R^2}{\sigma_n^2}\right) \tag{2}$$

2.1.2. Laplacian noise model

$$p\left(N_R\right) = \frac{1}{\sigma_n} \exp\left(-2\frac{|N_R|}{\sigma_n}\right) \tag{3}$$

2.1.3. Gamma speech model

$$p\left(S_R\right) = \frac{\sqrt[4]{3}}{2\sqrt[4]{2}\sqrt{\pi\sigma_s}} |S_R|^{-1/2} \exp\left(-\frac{\sqrt{3}}{\sqrt{2}}\frac{|S_R|}{\sigma_s}\right) \tag{4}$$

The signal power is evenly split between the real and the imaginary part, i.e. the variance of the real and imaginary parts of N and S is $\sigma_n^2/2$ for noise and $\sigma_s^2/2$ for speech. As the real and imaginary part of the signals are identically distributed, the pdf's of the imaginary parts are given by substituting S_R by S_I and N_R by N_I .

2.2. Experimental Data

We use the Kullback divergence [6] to measure the ability of the Gamma pdf to better describe the histogram of speech DFT coefficients compared to the Gaussian pdf. It is defined for two discrete pdf's $p_s(\nu)$ and $p_h(\nu)$ with N bins as

$$J(s:h) = \sum_{\nu=1}^{N} \left(p_s(\nu) - p_h(\nu)\right) \log\left(\frac{p_s(\nu)}{p_h(\nu)}\right)$$

with $p_h(\nu)$ the histogram of S_R (obtained from clean speech samples) and $p_s(\nu)$ the discrete pdf derived from (2) and (4), respectively. If we normalize the result by the Kullback divergence of the Gaussian pdf (using $p_s(\nu)$ derived from (2)) we have

$p_s(\nu)$	$J(s:h)\,/\,J(s:h)_{Gaussian}$
Gaussian	1.0
Gamma	0.514

Thus, the Kullback divergence for Gamma pdf is about half of the Kullback divergence of the Gaussian pdf. Therefore, the Gamma pdf delivers a better fit to the data than the Gaussian pdf. The same result applies to the imaginary part S_I . Similarly, one can investigate if for a specific noise another pdf is more appropriate than the Gaussian distribution (2). We found that for some noise types, such as low frequency car noise, also the Laplacian pdf may present a good fit to the data.

In the analytical derivation of the optimal estimators we had to assume that the real part and the imaginary part of DFT coefficients are statistically independent. For a complex supergaussian distributed variable this implies a dependence between the magnitude and phase. It can be shown [9], however, that this dependence is very weak and that with the above assumption improved estimators result.

3. MAGNITUDE-SQUARED ESTIMATORS

Choosing different density functions as a statistical model for the spectral coefficients of the clean speech and the noise signal, we derive the corresponding estimators (1) for the squared magnitude $|S|^2$ of the clean speech DFT coefficients.

3.1. Gaussian Noise and Gaussian Speech Model

In [1] Accardi and Cox presented the estimator of the power spectral density $E\left\{\,|S|^2\middle|\,Y\right\}$ of the speech signal for normally distributed spectral coefficients:

$$E\left\{\,|S|^2\middle|\,Y\right\} = \frac{\xi}{1+\xi}\sigma_n^2 + \left(\frac{\xi}{1+\xi}|Y|\right)^2 . \tag{5}$$

We have $\xi = \sigma_s^2/\sigma_n^2$ the *a priori* SNR, with σ_s^2 the expected power of the speech and σ_n^2 the expected power of the noise signal. In (5) the following relation, also mentioned in [2], is evident:

$$E\left\{\,|S|^2\middle|\,Y\right\} = \mathit{Var}\left\{\,|S|\middle|\,Y\right\} + \left|E\left\{\,S\middle|\,Y\right\}\right|^2 . \tag{6}$$

3.2. Gaussian Noise and Gamma Speech Model

In this section we assume that the real part and the imaginary part of the clean speech spectral coefficients can be modeled by the Gamma density function (4). The DFT coefficients of the noise signal follow a Gaussian pdf (2).

We define the auxiliary variables

$$G_{\pm} = \frac{\sqrt{3}}{2\sqrt{2}\sqrt{\xi}} \pm \frac{|Y_R|}{\sigma_n}. \tag{7}$$

The estimator for the squared real part of S as needed in (1) is

$$E\left\{\,{S_R}^2\middle|\,Y_R\right\} = \frac{3}{16}{\sigma_n}^2 \cdot \frac{\Psi\left(\frac{5}{4},\frac{1}{2};G_+^2\right) + \Psi\left(\frac{5}{4},\frac{1}{2};G_-^2\right)}{\Psi\left(\frac{1}{4},\frac{1}{2};G_+^2\right) + \Psi\left(\frac{1}{4},\frac{1}{2};G_-^2\right)} \tag{8}$$

for $G_- \geq 0$ and

$$\begin{aligned} E\left\{\,{S_R}^2\middle|\,Y_R\right\} = \tfrac{3}{16}{\sigma_n}^2 \cdot \tfrac{1}{\mathrm{Den}} \cdot \Bigg\{ & \tfrac{16\sqrt{\pi}}{\Gamma\left(\frac{1}{4}\right)} G_-\ \Phi\left(-\tfrac{1}{4},\tfrac{3}{2};-G_-^2\right) \\ & - \exp\left(-G_-^2\right)\left[\,\Psi\left(\tfrac{5}{4},\tfrac{1}{2};G_+^2\right) + \Psi\left(\tfrac{5}{4},\tfrac{1}{2};G_-^2\right)\right]\Bigg\} \end{aligned} \tag{9}$$

with

$$\begin{aligned} \mathrm{Den} = & \tfrac{4\sqrt{\pi}}{\Gamma\left(\frac{1}{4}\right)} G_-\ \Phi\left(\tfrac{3}{4},\tfrac{3}{2};-G_-^2\right) \\ & - \exp\left(-G_-^2\right)\left[\,\Psi\left(\tfrac{1}{4},\tfrac{1}{2};G_+^2\right) + \Psi\left(\tfrac{1}{4},\tfrac{1}{2};G_-^2\right)\right] \end{aligned}$$

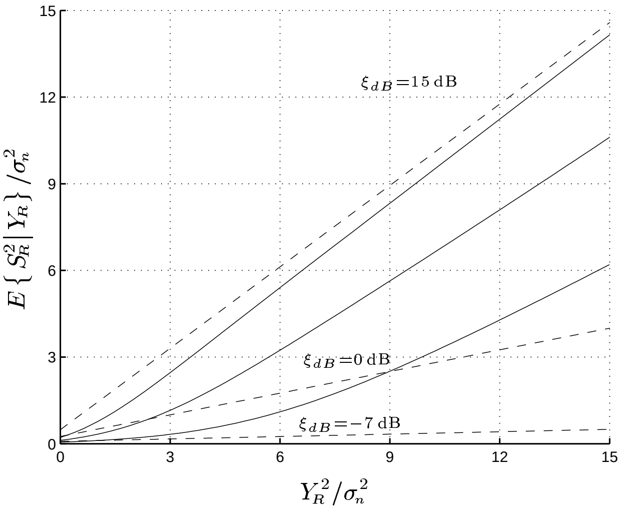

Fig. 1. The Gaussian/Gamma SSM estimate (8) and (9) of the squared real part S_R^2 of the clean speech DFT coefficients (solid line). The estimator output is plotted for values of the *a priori* SNR $\xi_{dB} = 15\,\mathrm{dB}, 0\,\mathrm{dB}, -7\,\mathrm{dB}$. For comparison, the Gaussian/Gaussian solution (5) for the same SNR values is given (dashed line). We have $\xi_{dB} = 10\log\left(\sigma_s^2/\sigma_n^2\right)$ and $\sigma_s^2 + \sigma_n^2 = 2$.

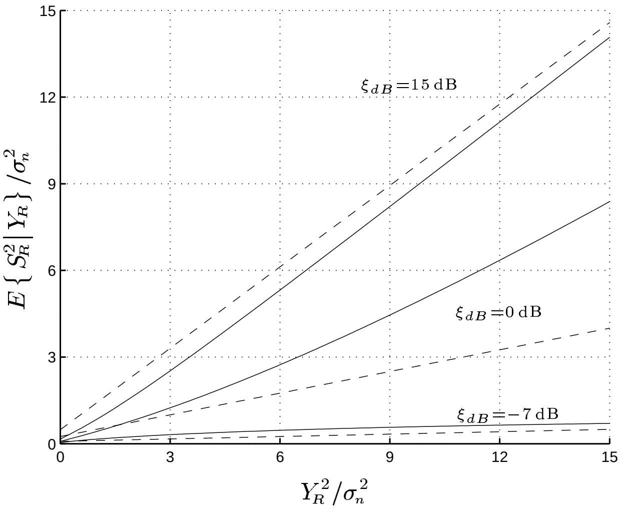

Fig. 2. The Laplace/Gamma SSM estimate (10) of the squared real part S_R^2 of the clean speech DFT coefficients (solid line). The estimator output is plotted for values of the *a priori* SNR $\xi_{dB} = 15\,\mathrm{dB}, 0\,\mathrm{dB}, -7\,\mathrm{dB}$. For comparison, the Gaussian/Gaussian solution (5) for the same SNR values is given (dashed line). We have $\xi_{dB} = 10\log\left(\sigma_s^2/\sigma_n^2\right)$ and $\sigma_s^2 + \sigma_n^2 = 2$.

for $G_- < 0$, whereby $\Gamma(z)$ is the Gamma function [5, (8.310.1)], $\Phi(\alpha,\gamma;x) = {}_1F_1\left(\alpha;\gamma;z\right)$ denotes the confluent hypergeometric function [5, (9.210.1)] and the function $\Psi(\alpha,\gamma;x)$ is defined as in [5, (9.210.1)].

The auxiliary variables $G_\pm$ and hence the optimal estimator use the magnitude $|Y_R|$ only, because estimating the squared real part results in an even function of Y_R. As we assume identically distributed DFT coefficients for the real and imaginary part of a spectral value, the estimator for the squared imaginary part is obtained by substituting Y_R with Y_I in (7) to (9). Equation (1) then gives the estimator for the squared spectral magnitude.

Figure 1 compares the estimator derived in this section (solid line) with (5) (dashed line). The plot shows the estimated squared real part of the clean speech DFT coefficient $E\left\{\left.{S_R}^2\right| Y_R\right\}$ for *a priori* SNR values $\xi_{dB} = 10\log(\xi) = 15\,\mathrm{dB}, 0\,\mathrm{dB}, -7\,\mathrm{dB}$ and a total signal power of $\sigma_s^2 + \sigma_n^2 = 2$ of the disturbed signal $Y = S + N$. Taking Y_R^2/σ_n^2 as the abscissa, (5) results in a straight line.

As the estimators are optimized for a function of the magnitude, i.e. the squared magnitude, they have an offset for $Y_R^2/\sigma_n^2 = 0$. The estimators derived by Ephraim und Malah ([3],[4]) show the same behavior. This offset (non-zero estimate for very small input amplitudes) contributes to the high quality of the residual noise signal as it helps to mask residual noise fluctuations.

For low values of the SNR the estimator using Gamma speech priors and the Gaussian noise model emphasizes coefficients of the noisy signal that are significantly larger than the standard deviation of the noise. This corresponds to the observation made in [8] for the spectral estimators of the complex DFT coefficients.

3.3. Laplacian Noise and Gamma Speech Model

We now change the assumption to that the statistics of the noise is best represented by a Laplacian pdf. We keep the Gamma pdf for the speech coefficients.

Again we define auxiliary variables

$$G_\pm = \frac{\sqrt{3}}{\sqrt{2}}\frac{1}{\sqrt{\xi}} \pm 2.$$

The MMSE estimator of the squared real part of the clean speech DFT coefficient then is

$$\begin{aligned} E\left\{\left.{S_R}^2\right| Y_R\right\} = {} & \tfrac{\sigma_n^2}{4}\tfrac{1}{G_+^2}\cdot\tfrac{1}{\mathrm{Den}}\cdot\Big\{\, 3\sqrt{\pi} \\ & + 4\exp\left(-G_-\tfrac{|Y_R|}{\sigma_n}\right)\,\Psi\left(-\tfrac{3}{2},-\tfrac{3}{2};G_+\tfrac{|Y_R|}{\sigma_n}\right) \\ & + \tfrac{8}{5}\left(G_+\tfrac{|Y_R|}{\sigma_n}\right)^{5/2}\,\Phi\left(\tfrac{8}{5},\tfrac{7}{2};-G_-\tfrac{|Y_R|}{\sigma_n}\right)\Big\} \end{aligned} \qquad (10)$$

with

$$\begin{aligned} \mathrm{Den} = {} & \sqrt{\pi} + \exp\left(-G_-\tfrac{|Y_R|}{\sigma_n}\right)\,\Psi\left(\tfrac{1}{2},\tfrac{1}{2};G_+\tfrac{|Y_R|}{\sigma_n}\right) \\ & + 2\left(G_+\tfrac{|Y_R|}{\sigma_n}\right)^{1/2}\,\Phi\left(\tfrac{1}{2},\tfrac{3}{2};-G_-\tfrac{|Y_R|}{\sigma_n}\right) \end{aligned}$$

Again, we exploit the even symmetry of the result using the magnitude $|Y_R|$. Figure 2 shows the comparison between this estimator (solid line) and (5). The use of the Laplacian noise prior yields a graph with negative curvature for very low *a priori* SNR values (see graph for $\xi_{dB} = -7\,\mathrm{dB}$). Again this shows the relationship to the corresponding estimator in [8] as indicated by (6). As the estimate is almost constant for low *a priori* SNR values, no musical tones are generated.

4. EXPERIMENTAL RESULTS

In the evaluation we use the estimators (5), (8)/(9), and (10) as noise reduction filters. The clean speech samples encompass six different speakers each speaking twelve sentences. The sampling

noise/speech pdf	Gaussian noise: SNR_{seg}		
	0 dB	10 dB	20 dB
Gaussian/Gaussian	3.20 dB	11.89 dB	20.68 dB
Gaussian/Gamma	4.46 dB	13.08 dB	21.71 dB
Laplace/Gamma	4.01 dB	12.84 dB	21.59 dB

(a) Gaussian noise

noise/speech pdf	car noise: SNR_{seg}		
	0 dB	10 dB	20 dB
Gaussian/Gaussian	3.01 dB	11.75 dB	20.52 dB
Gaussian/Gamma	4.12 dB	12.79 dB	21.28 dB
Laplace/Gamma	3.80 dB	12.60 dB	21.24 dB

(b) Car noise

Table 1. In the experimental evaluation the SSM estimators were used as noise reduction filters. The results are given as segmental SNR after filtering. Two different noises, white Gaussian noise (a) and car noise (b), were applied to clean speech samples at three different levels ($\mathrm{SNR}_{seg} = 0\,\mathrm{dB}, 10\,\mathrm{dB}$, and $20\,\mathrm{dB}$ before filtering).

frequency was 8 kHz. Filtering is done for two different noises: a stationary Gaussian noise of known variance σ_n^2 (Gaussian noise) and a recording made in the interior of a driving car (car noise) as described above. The degree of perturbation is measured as segmental SNR (denoted by SNR_{seg}) before and after filtering. Speech pauses are not considered. The segmental SNR is a quantity to measure noise reduction and signal distortion at the same time. SNR_{seg} compares two signals in the time domain considering only speech active sections. The estimated speech signal is constructed on an overlap-add basis, whereby each frame λ is calculated as follows:

$$\hat{s}(\kappa) = \mathrm{IDFT}\left\{\sqrt{E\left\{\,|S(\mu,\lambda)|^2 \middle| Y(\mu,\lambda)\right\}}\cdot e^{j\cdot\varphi_y(\mu,\lambda)}\right\}_M$$

with $\kappa, \mu \in \{0, \ldots, M-1\}$, $\mathrm{IDFT}\{\cdot\}_M$ the inverse DFT of length M , and φ_y the complex angle of Y. We generated a time domain speech signal in order to be also able to listen to the result and to compare the achieved SNR to other estimators. Note that for computing autocorrelation taps we would not draw the square root and would not overlap the time domain frames. The results of the filtering are presented in Table 1. Both Gaussian and car noise were applied at three different degrees: $\mathrm{SNR}_{seg} = 0\,\mathrm{dB}, 10\,\mathrm{dB}, 20\,\mathrm{dB}$ before filtering. For the enhancement processing the *a priori* SNR was estimated using the “decision directed” approach of [3]. In the case of car noise we employed the “Minimum Statistics” method [7] to estimate the noise power σ_n^2.

The previously known SSM estimator (5) using Gaussian pdf's to model both speech and noise coefficients is called Gaussian/ Gaussian in this table. We introduce the names Gaussian/Gamma for (8)/(9) and Laplace/Gamma for (10) accordingly. The use of the more precise statistical models leads to a consistent improvement in terms of the segmental SNR. Modelling the car noise with the Laplace pdf (Laplace/Gamma estimator) does not yield an estimator superior to the Gaussian/Gamma estimator. On the other hand the results for car noise disturbance are relatively close compared to the Gaussian noise case. Moreover, the strong positive curvature of the Gaussian/Gamma estimator for large $|Y_R|^2/\sigma_n^2$ (see Figure 1) results in a tendency to produce *musical tones* – a phenomenon unusual for SSM estimators, because the offset estimate for low input amplitudes mitigates the *musical tones* phenomenon.

5. CONCLUSIONS

MMSE speech enhancement for mobile communications in the frequency domain is based on relatively short framelengths for the DFT transform (M in the range of 64 to 512). In this case, the complex DFT coefficients of speech signals are not Gaussian distributed. They can be approximated more precisely by the Gamma pdf. Similarly, a specific noise like the car noise used in the evaluation might be better modeled with a Laplacian pdf. The results for the two new estimators confirm that more accurate statistical models deliver consistently better results. On the other hand, the *musical tones* arising from the Gaussian/Gamma estimator show that optimizing in the MMSE sense does not necessarily go along with an improved auditory impression. However, in an application where the SSM estimator might serve as a “core estimator” for autocorrelation coefficients or for other features of the speech signal, *musical noise* is not necessarily of primary concern. Moreover, the *musical noise* is completely avoided when the Laplacian noise model is used.

6. REFERENCES

[1] A. J. Accardi and R. V. Cox. A modular approach to speech enhancement with an application to speech coding. *IEEE ICASSP*, 1999.

[2] Y. Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *IEEE Transactions on Signal Processing*, 40(4):725–735, April 1992.

[3] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[4] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(ASSP-2):443–445, April 1985.

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

[6] S. Kullback. *Information Theory and Statistics*. Dover Publication, 1958.

[7] R. Martin. Spectral Subtraction Based on Minimum Statistics. In *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, pages 1182–1185, 1994.

[8] R. Martin. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. *IEEE ICASSP*, 2002.

[9] R. Martin. Speech Enhancement Based on Minimum Mean Square Error Estimation and Supergaussian Priors, 2002 (submitted).

[10] J. Porter and S. Boll. Optimal estimators for spectral restoration of noisy speech. *IEEE ICASSP*, pages pp. 18A.2.1.–18A.2.4., 1984.