MMSE Estimation of Complex-Valued Discrete Fourier Coefficients with Generalized Gamma Priors

J. Jensen, R. C. Hendriks, J. S. Erkelens, and R. Heusdens

Department of Mediamatics
Delft University of Technology
The Netherlands
{J.Jensen, R.C.Hendriks, J.S.Erkelens, R.Heusdens}@tudelft.nl

10.21437/Interspeech.2006-71

Abstract

We consider DFT based techniques for single-channel speech enhancement. Specifically, we derive minimum mean-square error estimators of clean speech DFT coefficients based on generalized gamma prior probability density functions. Our estimators contain as special cases the well-known Wiener estimator and the more recently derived estimators based on Laplacian and two-sided gamma priors. Simulation experiments with speech signals degraded by various additive noise sources verify that the estimator based on the two-sided gamma prior is close to optimal amongst all the estimators considered in this paper.

Index Terms: DFT based speech enhancement, minimum mean-square error estimation, generalized gamma priors.

1. Introduction

Single-channel speech enhancement methods based on the discrete Fourier transform (DFT) have received significant interest due to their low complexity and relatively good performance, e.g. [1, 2, 3, 4, 5]. Assuming that the noise process is additive and that noise and speech signals are independent, these methods generally estimate either the noise-free complex-valued DFT coefficients, e.g. [4], or the magnitudes of the DFT coefficients [2, 3]. The DFT based methods differ in their statistical assumptions regarding the speech and noise DFT coefficients; speech has traditionally been assumed Gaussian, e.g. [2], but more recently estimators based on supergaussian speech assumptions have been derived, see e.g. [4, 3]. Similarly, the noise is most often assumed Gaussian, but estimators exist which assume the noise to be supergaussian distributed [4]. Finally, existing methods differ in their objective; most methods rely on the minimum mean-square error (MMSE) criterion [2, 4], but sometimes simpler estimators can be found with the maximum a posteriori (MAP) criterion, e.g. [3].

We focus on MMSE estimators of complex-valued speech DFT coefficients and generalize the results of Martin [4]. We assume that noise DFT coefficients are Gaussian distributed, and that the real and imaginary parts of the speech DFT coefficients are statistically independent and distributed according to a two-sided generalized gamma prior density of the following form

$$f_{S_R}(s_R) = \frac{\gamma\beta^\nu}{2\Gamma(\nu)}|s_R|^{\gamma\nu-1}\exp(-\beta|s_R|^\gamma), \qquad (1)$$

where $\beta > 0, \gamma > 0, \nu > 0, -\infty < s_R < \infty$, and where the random variable S_R represents the real part of a complex-valued DFT coefficient; a similar equation holds for the imaginary part.

We derive MMSE estimators for the cases where $\gamma = 1$ and $\gamma = 2$. Since the prior $f_{S_R}(s_R)$ (and $f_{S_I}(s_I)$) in this case is parameterized by β and ν, the resulting estimators are also functions of these parameters. Certain parameter choices lead to priors for which MMSE estimators are already known. Specifically, with $\gamma = 1$, the prior in Eq. (1) has as special cases both the Laplace and Gamma densities for which MMSE estimators are presented in [4]. Further, for $\gamma = 2$, the Gaussian density occurs as a special case, and the well-known Wiener estimator [6] is MMSE optimal.

2. MMSE Estimation of DFT Coefficients

We consider a signal model of the form

$$X(k,m) = S(k,m) + W(k,m),$$

where $X(k,m), S(k,m), W(k,m)$ are complex random variables representing the DFT coefficients in signal frame m at frequency index k of the noisy, clean, and noise signal, respectively. Assuming that $S(k,m)$ and $W(k,m)$ are statistically independent across time and frequency and from each other, the resulting estimators are also time/frequency independent. Thus, we drop the time/frequency indices and introduce the following notation of the real and imaginary parts of the random variables in question

$$X = S + W,$$

with $X = X_R + jX_I$, $S = S_R + jS_I$, and $W = W_R + jW_I$. It is well-known that the MMSE estimator of the clean speech DFT coefficient S is identical to the conditional mean $E\{S|x\}$[1][7]. As in [4] we assume that the real and imaginary parts of S, S_R and S_I, are statistically independent, from which it follows that

$$E\{S|x\} = E\{S_R|x_R\} + jE\{S_I|x_I\}.$$

We now consider estimation of S_R; a similar procedure applies for S_I. Using Bayes' formula we find

$$E\{S_R|x_R\} = \frac{\int_{s_R} s_R f_{X_R|S_R}(x_R|s_R) f_{S_R}(s_R) ds_R}{\int_{s_R} f_{X_R|S_R}(x_R|s_R) f_{S_R}(s_R) ds_R}. \qquad (2)$$

From the Gaussian noise assumption it follows that

$$f_{X_R|S_R}(x_R|s_R) = \frac{1}{\sqrt{2}\pi\sigma^2_{W_R}}\exp(-\frac{1}{2\sigma^2_{W_R}}(x_R - s_R)^2). \qquad (3)$$

where $\sigma^2_{W_R}$ is the variance of W_R. From the assumption that W_R and W_I are independent it follows that $\sigma^2_{W_R} = \sigma^2_{W_I} = \sigma^2_W/2$. Similar results hold for the speech DFT coefficient S.

[1]Lower-case x represents a realization of the random variable X.

2.1. The Case $\gamma = 1$

With $\gamma = 1$ the prior density is of the form

$$f_{S_R}(s_R) = \frac{\beta^\nu}{2\Gamma(\nu)}|s_R|^{\nu-1}\exp(-\beta|s_R|), \tag{4}$$

Choosing $\nu = 1/2$ leads to the two-sided gamma density, while $\nu = 1$ results in a Laplacian density. MMSE estimators for these two special case were presented in [4].

Inserting Eqs. (3) and (4) in Eq. (2) it can be shown (see [8] for details) that the numerator in Eq. (2) is given by

$$\begin{aligned}&\int_{s_R} s_R f_{X_R|S_R}(x_R|s_R) f_{S_R}(s_R) ds_R = \\ &k \times \Big(\int_0^\infty s_R^\nu \exp(-\frac{s_R^2}{2\sigma_{W_R}^2} - s_R(\beta - \frac{x_R}{\sigma_{W_R}^2}))ds_R - \\ &\int_0^\infty s_R^\nu \exp(-\frac{s_R^2}{2\sigma_{W_R}^2} - s_R(\beta + \frac{x_R}{\sigma_{W_R}^2}))ds_R\Big),\end{aligned} \tag{5}$$

and the denominator is given by

$$\begin{aligned}&\int_{s_R} f_{X_R|S_R}(x_R|s_R) f_{S_R}(s_R) ds_R = \\ &k \times \Big(\int_0^\infty s_R^{\nu-1} \exp(-\frac{s_R^2}{2\sigma_{W_R}^2} - s(\beta - \frac{x_R}{\sigma_{W_R}^2}))ds_R + \\ &\int_0^\infty s_R^{\nu-1} \exp(-\frac{s_R^2}{2\sigma_{W_R}^2} - s_R(\beta + \frac{x_R}{\sigma_{W_R}^2}))ds_R\Big),\end{aligned} \tag{6}$$

where we introduced $k = (2\pi\sigma_{W_R}^2)^{-\frac{1}{2}} \frac{\beta^\nu}{2\Gamma(\nu)} \exp(-\frac{x_R^2}{2\sigma_{W_R}^2})$.

In order to find analytical expressions for the integrals in Eqs. (5) and (6) we use [9, Thm. 3.462.1]

$$\begin{aligned}&\int_0^\infty y^{\nu'-1}\exp(-\beta' y^2 - \gamma' y)dy = \\ &(2\beta')^{-\nu'/2}\Gamma(\nu')\exp(\frac{\gamma'^2}{8\beta'})D_{-\nu'}\left(\frac{\gamma'}{\sqrt{2\beta'}}\right),\end{aligned} \tag{7}$$

where $\beta' > 0, \nu' > 0$, and $D_{\nu'}(\cdot)$ is a parabolic cylinder function of order ν'. Applying this theorem to Eqs. (5) and (6), and using that β is related to $\sigma_{S_R}^2$, the variance of S_R, as $\beta^2 = \sigma_{S_R}^{-2}(\nu+1)\nu$, we can write the conditional mean $E\{S_R|x_R\}$ as

$$\begin{aligned}&E\{S_R|x_R\} = \\ &\sigma_{W_R}\nu\frac{\exp(\frac{1}{4}x_-^2)D_{-(\nu+1)}(x_-) - \exp(\frac{1}{4}x_+^2)D_{-(\nu+1)}(x_+)}{\exp(\frac{1}{4}x_-^2)D_{-\nu}(x_-) + \exp(\frac{1}{4}x_+^2)D_{-\nu}(x_+)},\end{aligned}$$

where x_- and x_+ are given by

$$x_\pm = \frac{\sigma_{W_R}}{\sigma_{S_R}}\sqrt{\nu(\nu+1)} \pm \frac{x_R}{\sigma_{W_R}}.$$

We note that $\frac{\sigma_{W_R}}{\sigma_{S_R}} = \xi^{-\frac{1}{2}}$, where $\xi \triangleq \frac{\sigma_S^2}{\sigma_W^2}$ is the a priori SNR [2].

2.2. The Case $\gamma = 2$

When $\gamma = 2$ in Eq. (1) we get

$$f_{S_R}(s_R) = \frac{\beta^\nu}{\Gamma(\nu)}|s_R|^{2\nu-1}\exp(-\beta|s_R|^2), \tag{8}$$

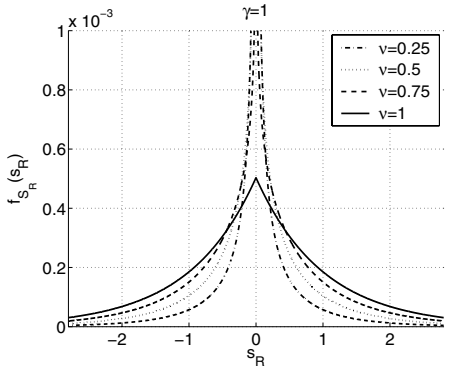

Figure 1: *Prior densities $f_{S_R}(s_R)$ for $\gamma = 1$ with $\nu = \{0.25, 0.50, 0.75, 1.0\}$ (normalized to unit variance).*

with $\beta > 0, \nu > 0$, and $-\infty < s_R < \infty$. We follow a similar strategy as before: Eqs. (8) and (3) are inserted in Eq. (2) and the resulting integrals are solved using the expression in (7). This leads to the following analytical expression for the MMSE estimator (again, we refer to [8] for details):

$$E\{S_R|x_R\} = 2\nu\sigma_{W_R}L_R\frac{D_{-(2\nu+1)}(x_-) - D_{-(2\nu+1)}(-x_-)}{D_{-2\nu}(x_-) + D_{-2\nu}(-x_-)},$$

where x_- can be written as

$$x_- = -\frac{x_R}{\sigma_{W_R}}L_R, \text{ and } L_R = \left(1 + 2\nu\xi^{-1}\right)^{-\frac{1}{2}}.$$

2.3. Input-Output Characteristics of Estimators

In this section we study the input-output characteristics of the derived estimators. For the case of $\gamma = 1$ we consider the following ν values: $\nu = \{0.25, 0.50, 0.75, 1\}$. The resulting prior densities $f_{S_R}(s_R)$ are shown in Fig. 1 (β is chosen such that the variance of S_R equals one). For $\nu = 1.0$ we get a Laplacian (two-sided exponential) prior and for $\nu = 0.5$ the two-sided gamma distribution occurs. Fig. 3A shows examples of input-output characteristics for the corresponding MMSE estimators. For high a priori SNRs, the relation between x_R and the estimator $E\{S_R|x_R\}$ is almost linear. At low a priori SNRs, the relation is non-linear, especially for small values of ν, i.e., more peaked priors.

For the $\gamma = 2$ case we consider $\nu = \{0.1, 0.2, 0.3, 0.5\}$. Fig. 2 shows the corresponding normalized prior densities. Choosing $\nu = 0.5$ gives a Gaussian prior, while lower values of ν give more peaked distributions[2]. Fig. 3B shows input-output characteristics for the resulting MMSE estimators. For $\nu = 0.5$ the Wiener estimator occurs (solid line in Fig. 3B). For all other choices of ν, the estimators are non-linear in the noisy observation x_R.

3. Simulation Results

We study the performance of the derived estimators in simulation experiments with noisy speech signals sampled at 8 kHz. The signals are taken from the Noizeus speech corpus [10] which consists

[2] In principle, the derived estimators remain valid for $\nu > 1.0$ for $\gamma = 1$ and $\nu > 0.5$ for $\gamma = 2$. In this case, however, the priors become bimodal. We have therefore chosen to restrict ν to the range $0 < \nu \leq 1.0$ for $\gamma = 1$ and $0 < \nu \leq 0.5$ for $\gamma = 2$.

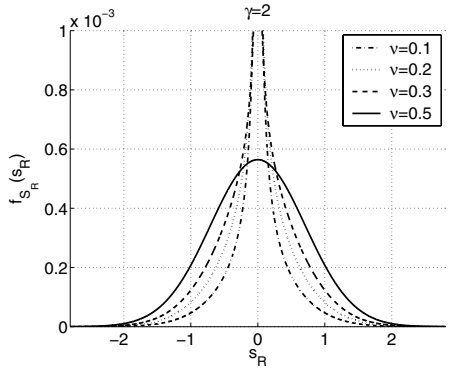

Figure 2: *Prior densities $f_{S_R}(s_R)$ for $\gamma = 2$ with $\nu = \{0.1, 0.2, 0.3, 0.5\}$ (normalized to unit variance).*

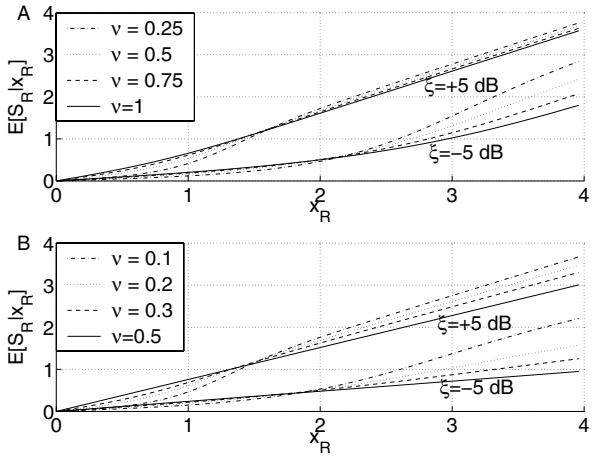

Figure 3: *Input-output characteristics for $\xi = -5$ dB and $\xi = 5$ with $\sigma_S^2 + \sigma_W^2 = 2$. A) $\gamma = 1$, B) $\gamma = 2$.*

of 30 speech signals, of roughly 3 seconds each, contaminated by various additive noise sources. We included signals contaminated by additive white Gaussian noise, since this noise condition was not present in the data base. The noisy speech signals were divided into segments of 256 samples with an overlap of 50% and transformed to the spectral domain using an FFT. After applying the derived gain functions to the noisy FFT coefficients, the enhanced signal segments were generated using an inverse FFT and overlap-added to form an enhanced waveform. To track the noise power spectral density we used the minimum statistics estimator [11]. The a priori SNR ξ was estimated using the decision-directed approach [2] with a fixed smoothing factor of $\alpha = 0.98$, and we limited the maximum suppression to 0.1.

We adopt the procedure of [3] to quantify the performance of the estimators in terms of speech distortion and noise reduction (although, to the authors knowledge, it has not been established to which extent this procedure correlates with subjective evaluations). Define the segmental speech SNR, as

$$\text{SNR-S} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} 10 \cdot \log_{10} \left(\frac{\|\mathbf{s}_p\|_2^2}{\|\mathbf{s}_p - \tilde{\mathbf{s}}_p\|_2^2} \right) \quad [\text{dB}],$$

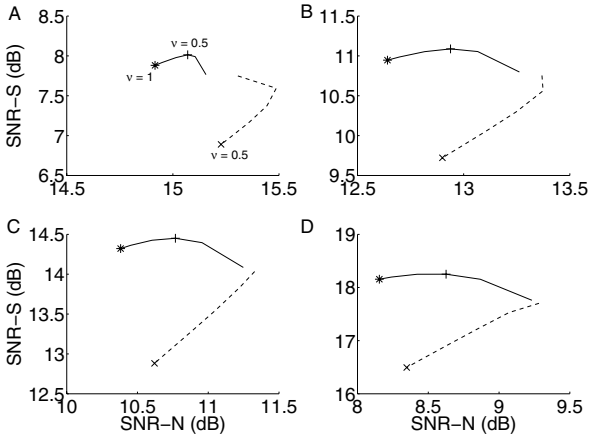

Figure 4: *SNR-S vs. SNR-N for $\gamma = 1$ (solid line) and $\gamma = 2$ (dashed line) for white noise. The special cases that correspond to the Gamma, Laplace and Gaussian priors are indicated by $+$, $*$ and $\times$, respectively. A) Input SNR = 0 dB, B) SNR = 5 dB, C) SNR = 10 dB, D) SNR = 15 dB.*

where the vector $\mathbf{s}_p$ represents a clean speech (time-domain) segment and $\tilde{\mathbf{s}}_p$ is the result of applying the gain functions to the *clean* speech segment[3]. To discard non-speech segments, let $\mathcal{P}$ be an index set of clean signal segments with energy larger than a threshold. More specifically, $\mathcal{P}$ is given by $\mathcal{P} = \{p : 10 \log_{10} \|\mathbf{s}_p\|_2^2 + 30 \geq \max_p 10 \cdot \log_{10}(\|\mathbf{s}_p\|_2^2)\}$, i.e., segments with energy within 30 dB of the maximum segment energy in a particular speech signal. Similarly, we measure the segmental noise reduction using

$$\text{SNR-N} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}}^{P} 10 \cdot \log_{10} \left(\frac{\|\mathbf{w}_p\|_2^2}{\|\tilde{\mathbf{w}}_p\|_2^2} \right) \quad [\text{dB}],$$

where $\mathbf{w}_p$ is the p'th noise segment, and $\tilde{\mathbf{w}}_p$ is the residual noise segment resulting from applying the noise suppression filter to $\mathbf{w}_p$.

Fig. 4 plots SNR-S vs. SNR-N for the derived estimators for different values of ν for white noise. Clearly, the estimator based on the two-sided gamma prior ($+$) gives relatively low speech distortions (high SNR-S) for a given residual noise level. Further, the Wiener estimator ($\times$) provides the weakest SNR-S vs. SNR-N tradeoff in the $\gamma = 2$ class of estimators. However, choosing low ν values in the $\gamma = 2$ class leads to estimators with performance close to that of estimators in the $\gamma = 1$ class[4].

Define the segmental SNR (SNR_{seg}) as

$$\text{SNR}_{\text{seg}} = \frac{1}{P} \sum_{p=1}^{P} T \left[10 \cdot \log_{10} \left(\frac{\|\mathbf{s}_p\|_2^2}{\|\mathbf{s}_p - \hat{\mathbf{s}}_p\|_2^2} \right) \right] \text{[dB]},$$

where $\hat{\mathbf{s}}_p$ is an enhanced signal segment, P is the total number of segments in the speech corpus, and the function $T[y] = \max(\min(y, 35), -10)$ clips per-segment SNRs to the range -10 – 35 dB. Fig. 5 shows the segmental SNR of the enhanced signals as a function of ν for different input SNRs for street noise (Figs.

[3]Clearly, this is only possible since the noisy signals are mixed synthetically, i.e., we have the clean signals available.

[4]When ν/ξ is small and x_R/σ_{W_R} is not, the estimators for $\gamma = 1$ and $\gamma = 2$ are approximately equal when $\gamma\nu$ is the same.

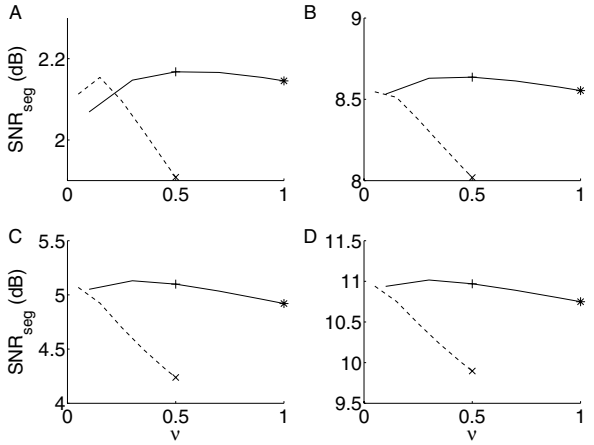

Figure 5: *Performance in terms of SNR$_{seg}$ vs. ν for $\gamma = 1$ (solid line) and $\gamma = 2$ (dashed line). A) Street noise at input SNR=5 dB. B) Street noise, SNR=15 dB. C) White noise, SNR=5 dB. D) White noise, SNR=15 dB.*

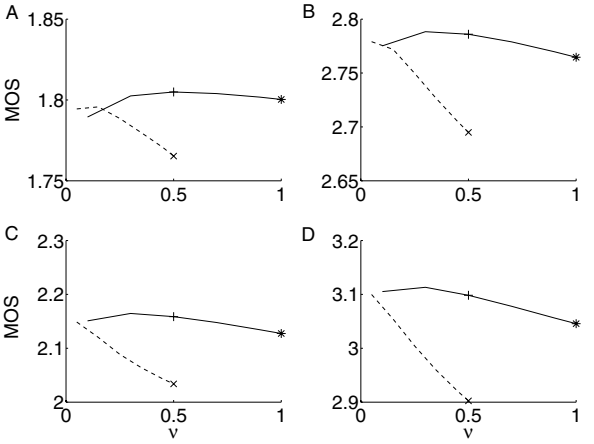

Figure 6: *Performance in terms of MOS (as estimated by PESQ [12]) vs. ν for $\gamma = 1$ (solid line) and $\gamma = 2$ (dashed line). A) Street noise at input SNR=5 dB. B) Street noise, SNR=15 dB. C) White noise, SNR=5 dB. D) White noise, SNR=15 dB.*

5A–B) and white noise (Figs. 5C–D). We see that the estimators based on a Laplacian ($*$) or Gamma ($+$) prior both perform well and that the performance of estimators with $\gamma = 1$ is relatively insensitive to the choice of ν. For $\gamma = 2$, choosing $\nu \approx 0.1 - 0.2$ leads to good performance, while $\nu = 0.5$, i.e. the Wiener estimator ($\times$), leads to the poorest performance in the $\gamma = 2$ class of estimators.

Finally, we evaluate the quality of the enhanced signals using PESQ [12] for different estimators, SNRs and noise sources, see Fig. 6. Whereas the $\gamma = 1$ based estimators are rather insensitive to the choice of ν, we see, as before, that lower values of ν lead to better performance when $\gamma = 2$. Interestingly, the PESQ curves in Fig. 6 are very similar in shape to the SNR$_{seg}$ curves in Fig. 5.

4. Concluding Remarks

This paper considered DFT based techniques for single channel speech enhancement. Specifically, we extended existing MMSE estimators by deriving two classes of estimators based on generalized gamma prior pdfs. Estimators from the class where $\gamma = 1$ typically perform better than the $\gamma = 2$ class, except for very small values of the parameter ν, where the estimators are very similar. A complex Gaussian model assumption for the complex speech DFT coefficients clearly does not perform well.

5. References

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Speech, Audio Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[2] 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. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[3] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[4] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," *IEEE Trans. Speech, Audio Processing*, vol. 13, no. 5, pp. 845–856, September 2005.

[5] P. C. Loizou, "Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum," *IEEE Trans. Speech, Audio Processing*, vol. 13, no. 5, pp. 857–869, September 2005.

[6] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*, Principles of Electrical Engineering Series. MIT Press, 1949.

[7] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall International, Inc., 1992.

[8] J. Jensen, R. C. Hendriks, and J. S. Erkelens, "MMSE Estimation of Discrete Fourier Coefficients with a Generalized Gamma Prior," Tech. Rep., Delft University of Technology, April 2006.

[9] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, Inc., 6 edition, 2000.

[10] *NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms. http://www.utdallas.edu/∼loizou/speech/noizeus/.*

[11] 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, July 2001.

[12] J. G. Beerends, "Extending P.862 PESQ for assessing speech intelligibility," White contribution COM 12-C2 to ITU-T Study Group 12 (equivalent to TNO Information and Communication Technology report 33392), October 2004.