Log-Spectral Magnitude MMSE Estimators under Super-Gaussian Densities

Richard C. Hendriks[1], *Richard Heusdens*[1] *and Jesper Jensen*[2]

[1]Department of Mediamatics, Delft University of Technology, The Netherlands
[2]Oticon A/S, Smørum, Denmark
R.C.Hendriks@tudelft.nl, R.Heusdens@tudelft.nl and jsj@oticon.dk

Abstract

Despite the fact that histograms of speech DFT coefficients are super-Gaussian, not much attention has been paid to develop estimators under these super-Gaussian distributions in combination with perceptual meaningful distortion measures. In this paper we present log-spectral magnitude MMSE estimators under super-Gaussian densities, resulting in an estimator that is perceptually more meaningful and in line with measured histograms of speech DFT coefficients. Compared to state-of-the-art reference methods, the presented estimator leads to an improvement of the segmental SNR in the order of 0.5 dB up to 1 dB. Moreover, listening tests show that the proposed estimator leads to significant improvement for the presented estimator over state-of-the-art methods.

Index Terms: speech enhancement, log-spectral magnitude MMSE, super-Gaussian

1. Introduction

Commonly, speech communication systems are combined with a noise reduction algorithm in order to increase their applicability in noisy environments. An often used procedure for single-microphone noise reduction is to apply discrete Fourier transform (DFT) domain noise reduction, where clean speech DFT coefficients are estimated on a frame-by-frame basis by applying Bayesian estimators [1]. In [2] it was proposed to estimate the clean speech DFT magnitudes by minimizing the mean-squared error (MSE) under the assumption that the clean speech and noise DFT coefficients are complex Gaussian distributed. Based on the observation that measured histograms of clean speech DFT coefficients show super-Gaussian behavior, e.g., [3][4], spectral magnitude minimum MSE (MMSE) estimators were also derived under super-Gaussian distributions, e.g., [5]. However, it can be argued that other distortion measures than the MSE of magnitude-spectra might be perceptually more meaningful [6]. Based on such arguments it was proposed in [7] to estimate clean speech DFT magnitudes by minimizing the MSE of log-spectral magnitudes.

Despite the development of magnitude MMSE estimators under super-Gaussian distributions on one hand and the fact that minimizing the MSE of log-spectral magnitudes is perceptually more relevant on the other hand, hardly any attention has been paid to the development of log-spectral magnitude MMSE estimators under super-Gaussian distributions. In this paper we derive log-spectral magnitude MMSE estimators under super-Gaussian distributions. This is done by assuming that the clean speech DFT magnitudes are distributed according to a one-sided chi-distribution. The one-sided chi-distribution is a general distribution and can model super-Gaussian data. The one-sided chi-distribution also possesses some special cases, e.g., the one-sided Gaussian distribution and the Rayleigh distribution. The estimator presented in [7] is therefore a special case of the estimator that we present in this paper.

2. Notation and Basic Assumptions

We assume that the noisy microphone signal is windowed and transformed to the DFT domain, leading to the noisy DFT coefficient $Y(k,i)$ with k the frequency bin-index and i the time-frame index. Further, we assume an additive noise model, i.e.,

$$Y(k,i) = X(k,i) + N(k,i) \tag{1}$$

where Y, X and N are the noisy speech, clean speech and noise DFT coefficient. The DFT coefficients Y, X and N are assumed to be complex zero-mean random variables statistically independent across time and frequency, and X and N are assumed to be statistically independent. We will use uppercase letters to denote random variables and the corresponding lowercase letters for their realizations. Although all expressions in this paper are per time-frame i and frequency bin-index k, we will leave out these indices for notational convenience.

For the random variables Y and X we use a polar domain notation, i.e., $Y = Re^{j\Theta}$ and $X = Ae^{j\Phi}$, respectively, where $j = \sqrt{-1}$. We assume that the noise DFT coefficients N have a complex Gaussian distribution. Generally speaking, this assumption holds for noise DFT coefficients as the time-span of dependency [8] for many noise sources is relatively short, see e.g. [4]. Together with the assumption that X and N are independent, this assumption implies that

$$f_{Y|A,\Phi}(y|a,\phi) = \frac{1}{\pi\sigma_N^2}\exp\left(\frac{2ar\cos(\phi-\theta) - r^2 - a^2}{\sigma_N^2}\right), \tag{2}$$

with σ_N^2 the variance of the noise DFT coefficients. The phase Φ of the clean speech DFT coefficients is assumed to be uniformly distributed and independent from the clean speech magnitude-DFT coefficients A according to measured histograms presented in [4][5]. Further, we assume that the speech magnitude-DFT coefficients are distributed according to a chi-distribution, that is

$$f_A(a) = \frac{2\beta^\nu}{\Gamma(\nu)}a^{2\nu-1}\exp(-\beta a^2), \qquad \beta > 0,\ \nu > 0,\ a \geq 0. \tag{3}$$

with $\beta = \nu/\sigma_X^2$ [5] and σ_X^2 the variance of the speech DFT coefficients. By computing the kurtosis of complex random variables for the clean speech DFT coefficient X it can be shown that for a ν-value $\nu < 1$, X has positive kurtosis and is therefore super-Gaussian.

The research is supported by the Oticon foundation and the Dutch Technology Foundation STW.

Further, we use the terms *a priori* SNR and the *a posteriori* SNR, defined as $\xi = \sigma_X^2/\sigma_N^2$ and $\zeta = |y|^2/\sigma_N^2$, respectively.

3. Log-Spectral Magnitude MMSE Estimators

The estimator $\hat{A}$ that minimizes the MSE in the logarithmic domain, i.e.,

$$\min_{\hat{A}} E\left[\left(\log A - \log \hat{A}\right)^2\right], \tag{4}$$

is given by [7]

$$\hat{A} \quad = \quad \exp\left[E\left[\log A|y\right]\right]. \tag{5}$$

For complex-Gaussian distributed clean speech and noise DFT coefficients it was shown in [7] that the estimator according to Eq. (5) can be derived elegantly by exploiting the moment generating function of $\log A|y$. To be more in line with histograms of speech DFT coefficients we extend these results and present estimators according to Eq. (5) under the chi-distribution.

Similar to [7] we exploit the moment generating function Ψ of the random variable $\log A|y$ in order to derive the estimator according to Eq. (5). Let the moment generating function of $\log A|y$ be given by

$$\Psi_{\log A|y}(\mu) = E\left[\exp\left(\mu \log A\right)|y\right] = E\left[A^{\mu}|y\right]. \tag{6}$$

By exploiting the first derivative of $\Psi_{\log A|y}(\mu)$ in Eq. (6) the estimator in Eq. (5) is given by

$$\hat{A} = \exp\left[E\left[\log A|y\right]\right] = \exp\left[\frac{d}{d\mu}\Psi_{\log A|y}(\mu)|_{\mu=0}\right]. \tag{7}$$

Using the assumption that the clean speech DFT phase Φ is uniformly distributed and independent from A, the conditional expectation in Eq. (6) can be expressed as

$$E\{A^{\mu}|y\} = \frac{\int_0^{+\infty}\int_0^{2\pi} a^{\mu} f_{Y|A,\Phi}(y|a,\phi) f_A(a) d\phi da}{\int_0^{+\infty}\int_0^{2\pi} f_{Y|A,\Phi}(y|a,\phi) f_A(a) d\phi da}. \tag{8}$$

By substituting Eqs. (2) and (3) into Eq. (8), followed by using [9, Eqs. 8.431.5, 6.643.2, and 9.220.2] we obtain for $\nu > 0$

$$E\{A^{\mu}|y\} = \frac{\Gamma(\nu+\mu/2)}{\Gamma(\nu)} K^{\mu/2} \frac{\mathcal{M}(\nu+\mu/2;1;P)}{\mathcal{M}(\nu;1;P)}, \tag{9}$$

with $K = \frac{r^2\xi}{(\xi+\nu)\zeta}$, $P = \frac{\zeta\xi}{\nu+\xi}$ and where $\mathcal{M}(\cdot)$ is the confluent hypergeometric function [10, Ch. 13].

The derivative $\frac{d}{d\mu}E\{A^{\mu}|y\}|_{\mu=0}$ in Eq. (7) is then given by

$$\begin{aligned}\frac{d}{d\mu}E\{A^{\mu}|y\}|_{\mu=0} &= \underbrace{\frac{\frac{d}{d\mu}\left\{\Gamma(\nu+\mu/2)\right\}|_{\mu=0}}{\Gamma(\nu)}}_{\text{part 1}} \\ &+ \underbrace{\frac{d}{d\mu}\left\{K^{\mu/2}\right\}|_{\mu=0}}_{\text{part 2}} + \underbrace{\frac{\frac{d}{d\mu}\left\{\mathcal{M}(\nu+\mu/2;1;P)\right\}|_{\mu=0}}{\mathcal{M}(\nu;1;P)}}_{\text{part 3}}\end{aligned} \tag{10}$$

In the following part of this section we further elaborate on part 1, 2 and 3 in Eq. (10).

part 1:

Let $\psi(\nu)$ denote the psi or digamma function [10]. Using [9, Eqs. 8.310.1, 0.410 and 4.352.1] it can then be shown that

$$\frac{\frac{d}{d\mu}\left\{\Gamma(\nu+\mu/2)\right\}|_{\mu=0}}{\Gamma(\nu)} \quad = \quad \frac{1}{2}\psi(\nu). \tag{11}$$

part 2

$$\frac{d}{d\mu}\left\{(K^{1/2})^{\mu}\right\}|_{\mu=0} = \frac{1}{2}\log K \tag{12}$$

part 3

Using [9, Eqs. 9.210.1 and 8.331.1] and the relation in Eq. (11) we obtain

$$\begin{aligned}&\tfrac{d}{d\mu}\left\{\mathcal{M}(\nu+\mu/2;1;P)\right\}|_{\mu=0}\tfrac{1}{\mathcal{M}(\nu;1;P)} \\ &= \textstyle\sum_{l=0}^{\infty} \tfrac{\Gamma(\nu+l)(\psi(\nu+l)-\psi(\nu))}{2\Gamma(\nu)} \tfrac{P^l}{(l!)^2} \tfrac{1}{\mathcal{M}(\nu;1;P)}.\end{aligned} \tag{13}$$

Summing the results of Eqs. (11), (12) and (13) followed by substitution into Eq. (7), the MMSE estimator in Eq. (5) for the distribution in Eq. (3) is obtained as

$$\hat{A} = \left(\frac{\xi}{(\xi+\nu)\zeta}\right)^{1/2} \exp\left[\frac{\psi(\nu)}{2}\right] \exp\left[\frac{T}{\mathcal{M}(\nu;1;P)}\right] r, \tag{14}$$

with

$$T = \sum_{l=0}^{L=\infty} \frac{\Gamma(\nu+l)\left(\psi(\nu+l)-\psi(\nu)\right)}{2\Gamma(\nu)} P^l \frac{1}{(l!)^2}. \tag{15}$$

In general this estimator needs evaluation of the summation in Eq. (15), which can be truncated to L terms. The value of L depends on the maximum value of the range of *a priori* and *a posteriori* SNRs for which the estimator needs to be expressed. From simulation experiments it followed that for *a priori* and *a posteriori* SNRs in the range of $-40\,dB < \xi < 40\,dB$ and $-40\,dB < \zeta < 40\,dB$ and ν in the range of $0 < \nu < 2$ $L = 15 \cdot 10^3$ is sufficient. For slightly lower *a priori* and *a posteriori* SNRs or lower ν-values, a much lower number of terms L is sufficient. For all experimental results in this paper the estimator was tabulated with $L = 15 \cdot 10^3$. Further, notice that when we set ν in Eq. (3) to $\nu = 1$, i.e., speech DFT coefficients are assumed to have a complex Gaussian distribution, we obtain the estimator proposed in [7].

3.1. Comparison of Log-Spectral Magnitude MMSE and spectral Magnitude MMSE Gain Curves

In Fig. 1 we compare the gain curves of the presented log-spectral magnitude MMSE estimator derived under the prior distribution in Eq. (3) and the corresponding spectral magnitude MMSE estimator under the same distributional assumptions, as a function of the *a posteriori* SNR ζ, for several values for ν and several values of the *a priori* SNR. The gain G for the log-magnitude and spectral magnitude MMSE estimator, respectively, are defined as

$$G = \frac{\exp\left[E\left[\log A|y\right]\right]}{r}$$

and

$$G = \frac{E\left[A|y\right]}{r},$$

respectively.

In Fig. 1a and 1b gain curves are shown for *a priori* SNR of $\xi = 0$ dB and 15 dB, respectively, and for ν-values $\nu = 0.4$ and $\nu = 1$. From Fig. 1 we see that for the same ν-value, the log-magnitude MMSE estimators lead to lower gain values than the spectral magnitude MMSE estimators at lower *a posteriori* SNRs. At higher *a posteriori* SNRs the gain values of the log-spectral magnitude and spectral magnitude MMSE estimators

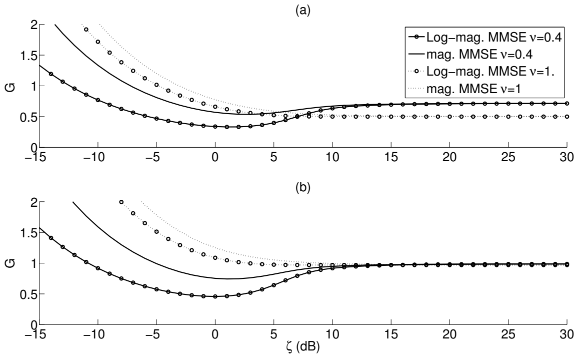

Figure 1: Gain curves at an *a priori* SNR of (a) 0 dB (b) 15 dB.

are very similar. This behavior is not only typical for the ν-values demonstrated in Fig. 1, but is generally valid for all ν-values. That the log-spectral magnitude MMSE estimator has lower gain values than the spectral magnitude MMSE estimator can be proven by using Jensen's inequality as was also shown in [7], that is

$$\hat{A} = \exp\left[E\left[\log A|y\right]\right] \leq \exp\left[\log\left\{E\left[A|y\right]\right\}\right] = E\left[A|y\right].$$

4. Experimental Results

To evaluate the presented estimator we compare its performance with the spectral magnitude MMSE estimator under the same distributional assumptions, i.e. the estimator presented in [5], using objective and subjective experiments. The speech signals originate from the Noizeus [6] database and are degraded with computer generated white noise and babble noise. All signals are filtered at telephone bandwidth and sampled at 8 kHz. The noisy time domain signals are divided in frames of 256 samples with 50 % overlap. For both analysis and synthesis a square-root Hann window is used. For estimation of the *a priori* SNR we use the decision-directed approach [2] and for estimation of the noise variance we use the DFT-subspace based method presented in [11].

4.1. Objective Evaluation

For objective evaluation we use segmental SNR defined as [12]

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{I}\sum_{i=0}^{I-1}\mathcal{T}\left\{10\log_{10}\frac{\|x_n(i)\|_2^2}{\|x_n(i)-\hat{x}_n(i)\|_2^2}\right\},$$

where $x_n(i)$ and $\hat{x}_n(i)$ denote time-frame i of the clean speech signal x_n and the enhanced speech signal $\hat{x}_n$, respectively, I is the number of frames and $\mathcal{T}(x) = \min\{\max(x,-10),35\}$ constrains the estimated SNR per frame to the range between -10 dB and 35 dB [12]. For further comparison of enhancement performance we measure speech segmental SNR as

$$\mathrm{SSNR}_{\mathrm{seg}} = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\mathcal{T}\left\{10\log_{10}\left(\frac{\|\mathbf{x}_n(i)\|_2^2}{\|\mathbf{x}_n(i)-\tilde{\mathbf{x}}_n(i)\|_2^2}\right)\right\},$$

where $\tilde{\mathbf{x}}_n(i)$ is a frame of the time domain signal that is the result of applying the gain functions to the clean speech frame. To discard non-speech frames, an index set $\mathcal{P}$ is used of all clean speech frames with energy within 50 dB of the maximum frame energy in a particular speech signal. $|\mathcal{P}|$ denotes the cardinality of $\mathcal{P}$. Similarly, noise segmental SNR is measured as

$$\mathrm{NSNR}_{\mathrm{seg}} = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\mathcal{T}\left\{10\log_{10}\left(\frac{\|\mathbf{n}_n(i)\|_2^2}{\|\tilde{\mathbf{n}}_n(i)\|_2^2}\right)\right\},$$

where $\mathbf{n}_n(i)$ is a noise frame, and $\tilde{\mathbf{n}}_n(i)$ is the residual noise frame resulting from applying the noise suppression filter to the noise only.

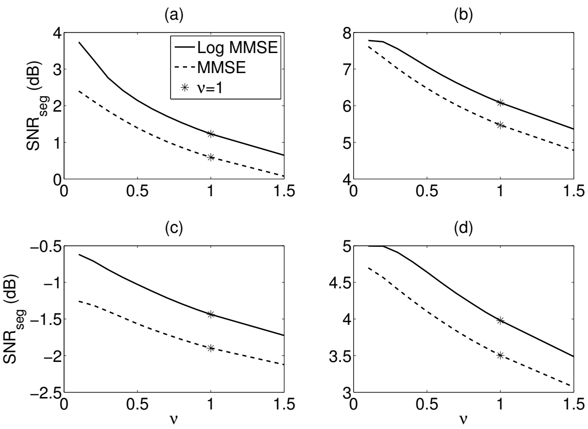

Figure 2: Comparison in terms of $\mathrm{SNR}_{\mathrm{seg}}$ (dB) for speech degraded with (a)-(b) white noise at (a) 0 dB SNR (b) 10 dB SNR, and (c)-(d) babble noise at (c) 0 dB SNR (d) 10 dB SNR.

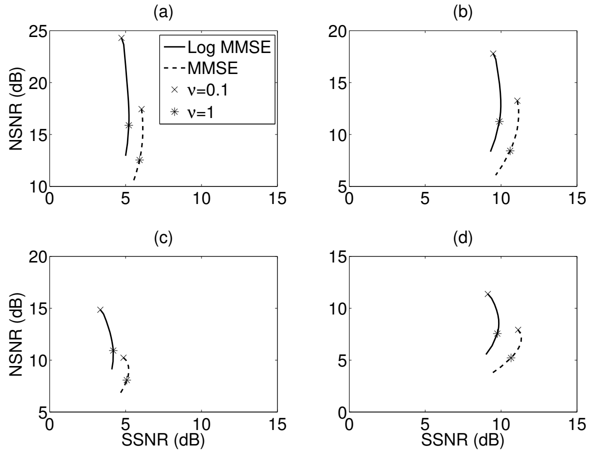

Figure 3: Comparison in term of $\mathrm{SSNR}_{\mathrm{seg}}$ (dB) versus $\mathrm{NSNR}_{\mathrm{seg}}$ (dB) for speech degraded with (a)-(b) white noise at (a) 0 dB SNR (b) 10 dB SNR, and (c)-(d) babble noise at (c) 0 dB SNR (d) 10 dB SNR.

In Fig. 2 the performance of the log-spectral magnitude MMSE and spectral magnitude MMSE estimator is expressed in terms of $\mathrm{SNR}_{\mathrm{seg}}$. For practically all input SNRs and the whole range of ν-parameters the log-spectral magnitude MMSE estimators lead to an improvement in the order of 0.5 dB up to 1 dB. Since the log-spectral magnitude MMSE estimators generally suppress more than their spectral magnitude counterparts it is expected that the increase in segmental SNR is due to more noise suppression and will therefore lead to somewhat more speech suppression as well.

To verify that the increase in segmental SNR is due to more noise suppression we measure performance in terms of $\mathrm{SSNR}_{\mathrm{seg}}$ versus $\mathrm{NSNR}_{\mathrm{seg}}$ as a function of the ν-parameter and input SNR. These results are shown in Fig. 3. As expected, the log-spectral magnitude MMSE estimator leads to a much higher performance in terms of $\mathrm{NSNR}_{\mathrm{seg}}$, i.e., noise suppression performance, but a somewhat lower speech quality $\mathrm{SSNR}_{\mathrm{seg}}$. The $\times$ in Fig. 3 indicates the estimators for $\nu = 0.1$. The ν-parameter increases along the curves. We see that as the ν-parameter increases, noise suppression performance decreases as well as eventually speech quality. The $*$ in Fig. 3 indicates the special case for $\nu = 1$, which was proposed in [7]. Compared to this log-spectral magnitude MMSE estimator under Gaussian distributional assumptions, the proposed estimators lead for certain choices of ν to a much better noise suppression while maintaining similar speech quality.

4.2. Subjective Evaluation

To evaluate the subjective quality of the proposed estimator, an OAB listening test was performed with six participants, the authors not included. Here, O is the original noisy signal and A and B are two different enhanced signals. The participants had to judge which of the two signals, A or B, had best quality. Two different listening tests were performed. In experiment 1, method A is the proposed log-spectral magnitude estimator with ν in Eq. (3) set to $\nu = 0.3$ and method B is the log-spectral magnitude estimator under the assumption that both the speech and the noise DFT coefficients are complex-Gaussian distributed, i.e., the estimator presented in [7]. In experiment 2 method A is again the proposed log-spectral magnitude MMSE estimator with ν in Eq. (3) set to $\nu = 0.3$ and method B is the spectral magnitude MMSE estimator under exactly the same super-Gaussian distributional assumptions, i.e., the estimator presented in [5]. Further, in all experiments the maximum suppression was limited to 0.05, for perceptual reasons. The listener was presented first the noisy signal followed by two different enhanced versions A and B. Each series was repeated 4 times with the enhanced versions played in random order. Four different speech signals were used, two male and two female speakers, all degraded with white noise at an SNR of 5 and 15 dB. The average preference for method A, the proposed method, is given in Table 1. In experiment 1 the average preference for method A over method B was 80 % and 76 % for input SNRs of 5 and 15 dB, respectively. In experiment 2 the average preference for method A over method B was 95 % and 83 % for input SNRs of 5 and 15 dB, respectively. A Wilcoxon statistical significance test [13] is used to determine whether the outcome of the listening test is significant. In Table 1, the P-values of this Wilcoxon test are given. These are the probabilities of observing the given result, or more extreme, by chance if the null hypothesis (the two tested methods have equal quality) is true. We can conclude that method A is in all experiments significant better than method B for significance levels larger than $1.4 \cdot 10^{-3}$.

Table 1: Average preference for method A and P-values of Wilcoxon test.

input	experiment 1		experiment 2	
SNR	pref. for A	P-value	pref. for A	P-value
5 dB	80 %	$0.5 \cdot 10^{-3}$	95 %	$0.5 \cdot 10^{-5}$
15 dB	76 %	$1.4 \cdot 10^{-3}$	83 %	$0.7 \cdot 10^{-6}$

5. Conclusions

In this paper magnitude MMSE estimators are presented that minimize the mean-squared error of the log-spectra under assumption that speech DFT coefficients are super-Gaussian distributed. By combining this perceptual more meaningful distortion measure with distributional assumptions that match with measured histograms of speech DFT coefficients, an estimator is obtained that improves both in terms of objective as well in subjective experiments over state-of-the-art reference noise reduction methods.

A MATLAB toolbox containing implementations of the presented estimators can be downloaded from the website http://ict.ewi.tudelft.nl/%7Erichard.

6. References

[1] H. L. van Trees, *Detection, Estimation and Modulation Theory*, vol. 1, John Wiley and Sons, 1968.

[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] 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, Sept. 2005.

[4] 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.

[5] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio Speech and Language Processing*, vol. 15, no. 6, pp. 1741 – 1752, August 2007.

[6] P. Loizou, *Speech enhancement theory and practice*, CRC Press, 2007.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[8] D. R. Brillinger, *Time Series: Data Analysis and Theory*, SIAM, Philadelphia, 2001.

[9] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series and Products*, New York: Academic, 6th ed. edition, 2000.

[10] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New-York, ninth dover printing, tenth gpo printing edition, 1964.

[11] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 3, pp. 541–553, March 2008.

[12] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, NJ, 2000.

[13] D. J. Sheskin, *Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 3rd edition edition, 2004.