# ON LINEAR VERSUS NON-LINEAR MAGNITUDE-DFT ESTIMATORS AND THE INFLUENCE OF SUPER-GAUSSIAN SPEECH PRIORS

*Richard C. Hendriks* and Richard Heusdens*

Delft University of Technology, The Netherlands
{R.C.Hendriks, R.Heusdens}@TUDelft.nl

## ABSTRACT

Although the linear mean-squared error (MSE) complex-DFT estimator, i.e., the Wiener filter, is well-known, its magnitude-DFT (MDFT) counterpart has never been considered in the context of speech enhancement. Therefore, certain theoretical questions regarding MDFT estimators remained unanswered. For example, it is unknown to which extend the performance of existing MSE MDFT estimators depends on the chosen speech prior, or on the non-linearity of the estimators.

In this paper we present linear MSE MDFT estimators for speech enhancement. In contrast to the linear complex-DFT estimator, the presented linear MSE MDFT estimators do depend on the assumed distribution of the speech DFT coefficients. Based on objective and subjective experiments, it can be concluded that the chosen speech prior, i.e., Gaussian versus super-Gaussian has a significant effect on the performance of MDFT estimators, while the linearity as compared to non-linearity has only a minor influence.

***Index Terms***— speech enhancement, magnitude-DFT estimator

## 1. INTRODUCTION

A common strategy to increase robustness of speech processing applications is to equip them with a noise reduction algorithm. These algorithms are often applied on a frame-by-frame basis in the Fourier domain, e.g., by using the discrete Fourier transform (DFT) as in [1]. One of the most well-known estimators for DFT-based speech enhancement is the Wiener filter [2], which is the optimal linear mean-squared error (MSE) estimator of the complex speech DFT coefficients as well as the optimal non-linear Bayesian MSE estimator when the distribution of speech and noise DFT coefficients is Gaussian. Based on the argumentation that the phase of speech DFT coefficients is less important than its magnitude, it was proposed in [1] to estimate the magnitude-DFT (MDFT) instead of complex-DFT coefficients. The resulting estimator is non-linear and was derived by minimizing the Bayesian MSE of the MDFT coefficients while assuming that speech and noise DFT coefficients are Gaussian distributed. More recently, based on histograms of speech DFT coefficients, Bayesian MSE estimators have been proposed under super-Gaussian instead of Gaussian distributions for both MDFT and complex-DFT estimators, e.g., [3][4][5].

Considering the development outlined above, it is remarkable that the linear MSE MDFT estimator is missing in this line of development. Clearly, non-linear MSE estimators lead to an equal or lower MSE than linear estimators. However, due to the lacking existence of the linear MSE MDFT estimator, some theoretical questions remain unanswered. More specifically, it is unknown to which extend the performance of existing MSE MDFT estimators depends on the chosen speech prior, or on the non-linearity of the estimators.

Besides theoretical interest, linear MDFT estimators might also be of practical interest. More specifically, the fact that linear estimators are often of a simpler form than their non-linear counterparts, might lead to some practical or implementational advantages of linear over non-linear MDFT estimators.

In this paper we study and derive linear estimators for clean speech MDFT coefficients. We will show that, in contrast to linear complex-DFT estimators, linear MSE MDFT estimators do depend on the assumed distribution of the speech DFT coefficients. Therefore, for a certain distribution of the speech DFT coefficients, a linear as well as a non-linear MSE MDFT estimator can be obtained.

Further, by means of objective and subjective evaluations we will analyse the linear and non-linear MSE MDFT estimators, leading to more insight on the difference between these estimators and on the influence of the assumed underlying prior distribution.

## 2. NOTATION AND BASIC ASSUMPTIONS

We assume that the noisy microphone signal consists of clean speech degraded by additive noise. Further, we assume that the speech and noise process are independent. Let $Y(k,i)$ be a noisy DFT coefficient for a frequency-bin $k$ and time-frame $i$. Due to linearity of the Fourier transform and the assumed additive noise model it holds that

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

where $X$ and $N$ are the clean speech and noise DFT coefficient, respectively. The DFT coefficients $Y$, $X$ and $N$ are assumed to be complex zero-mean random variables statistically independent across time and frequency. We will use uppercase letters to denote random variables and the corresponding lowercase letters for their realizations. Since 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 mathematical convenience we use a polar domain notation for the random variables $Y$ and $X$, i.e., $Y = Re^{j\Theta}$ and $X = Ae^{j\Phi}$, where $j = \sqrt{-1}$.

Based on histograms of speech DFT coefficients published in [5] we assume that the speech DFT-phase $\Phi$ is uniformly distributed and independent from the DFT-magnitude $A$. To be in line with the fact that speech DFT coefficients are super-Gaussian distributed, see e.g., [4], we model the speech MDFT coefficients with the generalized-Gamma distribution given by

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

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

where we set $\gamma = 2$ from which it follows that $\beta = \nu/\sigma_X^2$, with $\sigma_X^2 = E\left[|X|^2\right]$ [5]. Whether the distribution of speech DFT coefficients is super-Gaussian depends on the parameter $\nu$ in Eq. (2). For $0 < \nu < 1$, the distribution of $X$ is super-Gaussian and for $\nu = 1$ the distribution of $X$ is complex Gaussian.

The noise DFT coefficients are assumed to have a complex zero-mean Gaussian distribution. This is based on the fact that the time-span of dependency [6] is relatively low for many noise sources, see e.g., [4]. Based on this distributional assumption, together with the assumed independence of speech and noise DFT coefficients, it follows that the distribution $f_{R|A}(r|a)$ is given by

$$f_{R|A}(r|a) = \frac{2r}{\sigma_N^2} \exp\left(-\frac{r^2 + a^2}{\sigma_N^2}\right) \mathcal{I}_0\left(\frac{2ar}{\sigma_N^2}\right), \tag{3}$$

where $\sigma_N^2 = E\left[|N|^2\right]$ and $\mathcal{I}_0(\cdot)$ is the modified Bessel function of the first kind and order zero.

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

## 3. LINEAR MAGNITUDE-DFT ESTIMATORS

The linear estimator $\hat{A}$ of $A$ with a possibly non-zero mean is given by [7, Ch. 2]

$$\hat{A} = GR + m, \tag{4}$$

with $G \in \mathbb{R}$ and $m \in \mathbb{R}$. Notice that strictly speaking this estimator is not linear, but affine, due to the shift $m$. However, the estimator can always be mapped to a linear estimator using a translation. Let the MSE between $A$ and $\hat{A}$ be defined as $h(G, m) = E[|A - \hat{A}(G, m)|^2]$. Since magnitudes are non-negative by definition, the estimator needs to satisfy the constraint $\hat{A} \geq 0$. This leads to the following optimization problem

$$\min_{G,m} h(G, m) \text{ subject to } \hat{A} \geq 0. \tag{5}$$

Using the Karush-Kuhn-Tucker conditions for constrained optimization [8] the following solution is obtained,

$$\hat{A} = \max\left(\frac{E\left[AR\right] - E\left[R\right]E\left[A\right]}{E\left[R^2\right] - E\left[R\right]^2}\left(R - E\left[R\right]\right) + E\left[A\right], 0\right). \tag{6}$$

The estimator in Eq. (6) is dependent on the the first and second moments of $A$ and $R$, and the cross-correlation $E\left[AR\right]$. Similar as for the Wiener filter, the estimator in Eq. (6) can be written as a function of the second order moments of the speech and noise DFT coefficients only. However, to do so, it is not sufficient to make use of the assumptions that speech and noise are additive and independent, as is sufficient for the Wiener filter. In addition it is necessary to make assumptions on the actual distributions of the speech and noise DFT coefficients. In the following subsections, expressions for $E\left[R\right]$, $E\left[A\right]$ and $E\left[AR\right]$ will be derived.

### 3.1. Computing $E\left[R\right]$, $E\left[A\right]$ and $E\left[AR\right]$

The first order moment of $R$ can be computed as

$$E\left[R\right] = \int_0^\infty r f_R(r) dr. \tag{7}$$

The distribution $f_R(r)$ is given by

$$f_R(r) = \int_A f_A(a) f_{R|A}(r|a) da. \tag{8}$$

Let $\eta$ and $\rho$ be defined as $\eta = \nu(\nu + \xi)^{-1}$ and $\rho = \xi(\nu + \xi)^{-1}$, respectively. Using [9, Eqs. 6.643.2 and 9.220.2], we obtain

$$f_R(r) = 2re^{-\frac{r^2}{\sigma_N^2}} \frac{\eta^\nu}{\sigma_N^2} \mathcal{M}\left(\nu, 1, \zeta\rho\right), \tag{9}$$

with $\mathcal{M}$ the confluent hypergeometric function [9]. Substitution of Eq. (9) into Eq. (7), followed by using [9, Eq. 7.621.1] leads to

$$E\left[R\right] = \Gamma(3/2)\eta^\nu \sigma_N \mathcal{F}\left(3/2, \nu, 1, \rho\right), \tag{10}$$

where $\mathcal{F}(a, b, c, d)$ and $\Gamma\left(\cdot\right)$ denote the hypergeometric and Gamma function [9], respectively.

The first order moment of $A$ can be computed as

$$E\left[A\right] = \int_0^\infty a f_A(a) da. \tag{11}$$

Substitution of Eq. (2) into Eq. (11) followed by using [9, Eq. 3.381.4] leads to

$$E\left[A\right] = \Gamma(\nu + 1/2)\sigma_X / (\Gamma(\nu)\sqrt{\nu}). \tag{12}$$

The cross-correlation between $A$ and $R$ is given by

$$E\left[AR\right] = \int_0^\infty \int_0^\infty a r f_{R|A}(r|a) f_A(a) dr da. \tag{13}$$

Substitution of Eqs. (2) and (3) into Eq. (13), followed by using [9, Eqs. 6.643.2 and 7.621.1] leads to

$$E\left[AR\right] = \frac{\Gamma(3/2)\Gamma(\nu + \frac{1}{2})\sigma_N^2}{\Gamma(\nu)\sqrt{\nu/\xi}} \eta^{\nu + \frac{1}{2}} \mathcal{F}\left(\nu + \frac{1}{2}, \frac{3}{2}, 1, \rho\right). \tag{14}$$

### 3.2. Linear Magnitude-DFT Estimator

The linear MDFT estimator is given by substitution of Eqs. (10), (12) and (14) into Eq. (6), that is

$$\hat{A} = \max\left(GR + m, 0\right), \tag{15}$$

with $G$ and $m$ given by

$$G = \frac{\eta^\nu \Gamma(\frac{3}{2})\Gamma(\nu + \frac{1}{2})}{\Gamma(\nu)\sqrt{\nu/\xi}} \frac{\left(\sqrt{\eta}\mathcal{F}(\nu + \frac{1}{2}, \frac{3}{2}, 1, \rho) - \mathcal{F}(\frac{3}{2}, \nu, 1, \rho)\right)}{\xi + 1 - \left(\Gamma(\frac{3}{2})\eta^\nu \mathcal{F}(\frac{3}{2}, \nu, 1, \rho)\right)^2}, \tag{16}$$

and

$$m = \frac{\Gamma(\nu + 1/2)\sigma_X}{\Gamma(\nu)\sqrt{\nu}} - G\Gamma(3/2)\eta^\nu \sigma_N \mathcal{F}(3/2, \nu, 1, \rho). \tag{17}$$

From Eqs. (15)-(17) we see that the final estimator is a function of the *a priori* SNR $\xi$, and the $\nu$-parameter, which specifies the assumed distribution for the magnitude-DFT coefficients. Notice that in contrast to the linear MSE complex-DFT estimator, the linear MSE MDFT estimator is dependent on the assumed distribution for the speech DFT coefficients via the parameter $\nu$.

To compute the linear estimator, several special functions have to be evaluated (the hypergeometric function and the Gamma function). However, this is not necessarily an issue, since in practical systems the gain function is often tabulated. Notice that for a given $\nu$-value, this estimator only needs tabulation of the function $G$ in Eq. (16) and the last term in Eq. (17), i.e., $\mathcal{F}(3/2, \nu, 1, \rho)$. Both are a function of only one parameter, namely the *a priori* SNR $\xi$. Therefore, this estimator requires significantly less memory as compared to the gain functions of the estimator in e.g., [1][5], which are a function of both *a priori* SNR and *a posteriori* SNR.

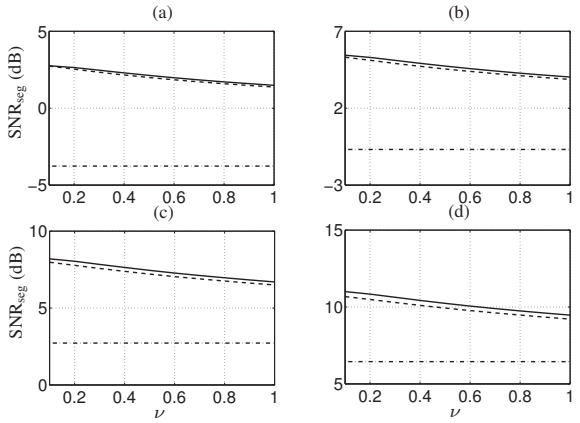


**Fig. 1**. Segmental SNR for non-linear estimator (solid line), linear estimator (dashed line), and noisy data (dash-dotted line) for speech degraded by white Gaussian noise at an input SNR of (a) 0 dB (b) 5 dB (c) 10 dB (d) and 15 dB.

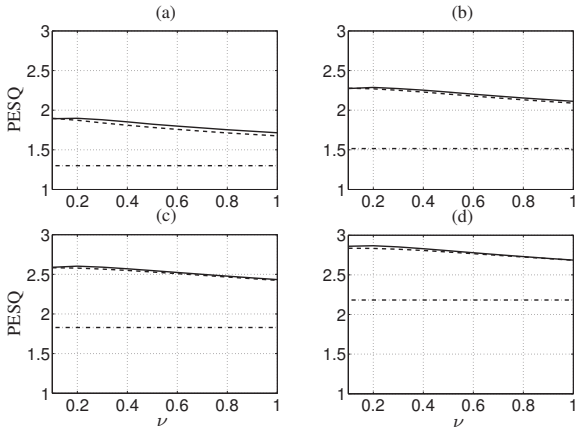


**Fig. 2**. PESQ scores for non-linear estimator (solid line), linear estimator (dashed line), and the noisy data (dash-dotted line) for speech degraded by white Gaussian noise at an input SNR of (a) 0 dB (b) 5 dB (c) 10 dB (d) and 15 dB.

## 4. OBJECTIVE EVALUATION

To analyse to which extend the performance of MSE MDFT estimators depends on the chosen speech prior or on the non-linearity of the estimator, we first present objective simulation experiments and compare the presented linear MSE MDFT estimator with the non-linear MSE MDFT estimator presented in [5].

For evaluation we use more than 7 minutes of Danish speech spoken by 9 female and 8 male speakers. The speech signals were degraded by train noise and computer generated white noise at input SNRs of 0, 5, 10 and 15 dB. All signals are sampled at a frequency of $f_s = 8$ kHz and start with a noise-only period of 0.5 seconds. All algorithms use the first 0.1 seconds for initialization, which is therefore excluded in performance measurements. The time-frames have a length of $K = 256$ samples with 50% overlap, and are windowed using a square-root-Hann window. Estimation of the speech PSD is performed using the decision-directed approach [1]. For noise PSD estimation we use the method presented in [10].

For evaluation we use PESQ [11] and segmental SNR defined as

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{I}\sum_{i=0}^{I-1} \mathcal{T}\left\{10\log_{10}\frac{\|x_t(i)\|^2}{\|x_t(i)-\hat{x}_t(i)\|^2}\right\} \quad [\mathrm{dB}],$$

where $x_t(i)$ and $\hat{x}_t(i)$ denote time-frame $i$ of the clean speech signal $x_t$ and the enhanced speech signal $\hat{x}_t$, respectively, and $\mathcal{T}(x) = \min\{\max(x,-10),35\}$.

In Figs. 1 and 2 the linear and non-linear estimator are compared in terms of segmental SNR and PESQ, respectively. The performance loss in terms of segmental SNR and PESQ is slightly dependent on the input SNR, but generally less than 0.2 dB segmental SNR and less than 0.03 points in terms of PESQ.

For comparison: the performance loss that is obtained when assuming speech DFT coefficients to be Gaussian distributed instead of super-Gaussian distributed with $\nu \approx 0.1$ is around 1.4 dB segmental SNR and 0.2 points in terms of PESQ for both the linear and the non-linear estimators. Therefore, it is to be expected that the difference between the linear and non-linear estimator is perceptually insignificant in comparison to the difference that is obtained by using estimators based on super-Gaussian instead of Gaussian priors.

Fig. 3 shows a similar comparisons in terms of segmental SNR and PESQ for speech degraded by passing train noise. Similar as for white Gaussian noise, the performance loss of using a linear instead of non-linear estimator for this non-stationary noise source is less than 0.2 dB segmental SNR and 0.03 points in terms of PESQ.

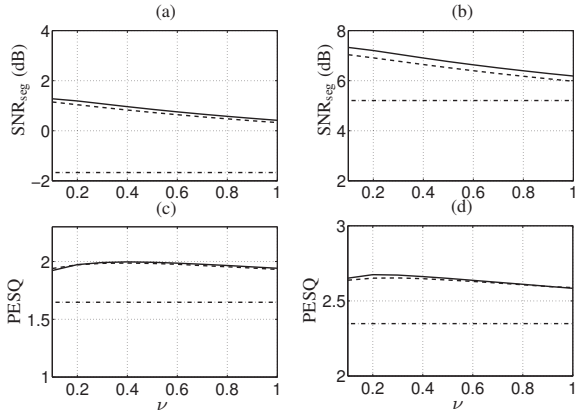


**Fig. 3**. (a)-(b) segmental SNR and (c)-(d) PESQ scores, for non-linear estimator (solid line), linear estimator (dashed line), and noisy data (dash-dotted line) for speech signals degraded by train noise at an input SNR of (a) 0 dB (b) 10 dB (c) 0 dB (d) 10 dB.

## 5. SUBJECTIVE EVALUATION

From the objective evaluation in Sec. 4 the following two hypotheses can be conducted. At first, the performance loss due to using the linear instead of the non-linear estimator, is negligible. Secondly, the performance of the linear and the non-linear MSE MDFT estimator when derived under the assumption that speech DFT coefficients are super-Gaussian distributed have a better performance than when these estimators are derived under the assumption that speech DFT coefficients are Gaussian distributed. To verify these two hypotheses, a Mushra-like [12] listening test is performed with two male and two female sentences, degraded by white noise at input SNRs of 5 dB and 10 dB. The number of participants in the listening test was six, the authors not included. The listeners were presented the original clean signal as a reference signal, and six blind signals. Among these six signals were the clean signal, the noisy signal as an an-

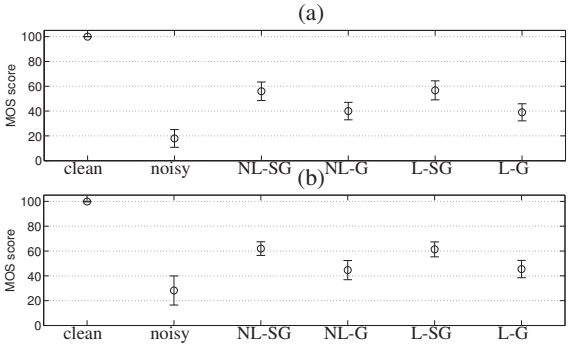


**Fig. 4**. Listening test results for input SNR of (a) 5 dB (b) 10 dB.

chor, the non-linear MSE MDFT estimator derived under a super-Gaussian distribution with $\nu = 0.2$ abbreviated as NL-SG, the non-linear MSE MDFT estimator derived under a Gaussian distribution abbreviated as NL-G, the linear MSE MDFT estimator derived under a super-Gaussian distribution with $\nu = 0.2$ abbreviated as L-SG and the linear MSE MDFT estimator derived under a Gaussian distribution abbreviated as L-G. The participants were asked to compare the blind signals with the clean reference signal in terms of quality and score them between 0 (poor) and 100 (excellent) quality. The order in which the participants were presented the speech signals, as well as the order of the blind signals per subject, were randomized.

The results of this listening test, averaged over all participants and signals, in combination with the 95 % confidence intervals, are shown in Fig. 4. With respect to the non-linear versus the linear MSE estimators it follows that the NL-SG and the L-SG processed signals, as well as the NL-G and the L-G processed signals were given approximately the same score, i.e., less than 1 point difference on a scale of 100. At the same time, both super-Gaussian based estimators, i.e, NL-SG and L-SG, were ranked approximately 16 points higher than the Gaussian based estimators, i.e., NL-G and L-G. These results acknowledge the aforementioned two hypotheses.

To verify the statistical significance of the listening test, a paired t-test for two dependent samples [13] is applied. From this statistical analysis it follows that at both input SNRs and for both the linear and the non-linear MSE MDFT estimator, the super-Gaussian based estimators have a significantly better performance than the Gaussian based estimators at a significance level of $\alpha = 5{\cdot}10^{-5}$. This is an acknowledgement of earlier findings on the improved performance of super-Gaussian based estimators over Gaussian based estimators for speech enhancement. Further, it follows that the linear and the non-linear MSE estimators are indeed not significantly different. This holds for both the situation that speech DFT coefficients are assumed to be Gaussian distributed and when speech DFT coefficients are assumed to be super-Gaussian distributed.

## 6. CONCLUDING REMARKS

In this paper linear magnitude-DFT (MDFT) MSE estimators for speech enhancement are presented. In contrast to linear complex-DFT MSE estimators, these linear estimators do depend on the assumed distribution of the speech DFT coefficients. Therefore, for a certain distribution of the speech DFT coefficients, a linear as well as a non-linear MSE MDFT estimator can be obtained.

Based on objective experiments it can be concluded that the loss that is obtained by using the linear MSE MDFT estimators instead of non-linear MSE MDFT estimators is negligible. This is further investigated with a listening test, from which it indeed follows that there is no statistical significant difference between the non-linear and linear MSE magnitude DFT estimators. On the other hand, both the linear and the non-linear MSE MDFT estimators show a statistical significant improvement when using super-Gaussian instead of Gaussian prior distributions for the speech DFT coefficients. It can therefore be concluded that the performance of magnitude-DFT MSE estimators mainly depends on the chosen speech prior and not so much on the linearity or non-linearity of the estimator.

Besides the similar performance of non-linear linear MSE MDFT estimators, the linear MSE MDFT estimators have some potential practical advantages. More specifically, when tabulating the estimators for use in a practical system, the presented linear estimator requires significantly less memory as compared to the non-linear MSE MDFT estimators.

## 7. REFERENCES

[1] 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.

[2] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*, MIT Press, principles of electrical engineering series edition, 1949.

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

[7] W. Utschick, H. Boche, and R. Mathar (Eds.), *Linear Estimation and Detection in Krylov Subspaces*, vol. 1, Springer, 2007.

[8] W. Karush, "Minima of functions of several variables with inequalities as side constraints," M.S. thesis, Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

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

[10] R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "Fast noise psd estimation with low complexity," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2009, pp. 3881–3884.

[11] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.

[12] ITU-R BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems - general requirements," Tech. Rep., 2001.

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